Determining NBER Recession Points Using Machine Learning

Luca Lavarini

Dissertation written under the supervision of Dr. Jonathan Tepper

Dissertation submitted in partial fulfilment of requirements for the MSc in International Finance at Universidade Católica Portuguesa and for the MSc in Business Analytics at Aston University, September 2024.

Declaration

I declare that I have personally prepared this report and that it has not in whole or in part been submitted for any other degree or qualification. Nor has it appeared in whole or in part in any textbook, journal or any other document previously published or produced for any purpose. The work described here is my/our own, carried out personally unless otherwise stated. All sources of information, including quotations, are acknowledged by means of reference, both in the final reference section and at the point where they occur in the text.

Acknowledgements

I would like to express my deepest gratitude to my supervisor, Johnatan Tepper, for his invaluable guidance, support, and patience throughout this project. His insights and encouragement were crucial to the completion of this work.

I am also profoundly grateful to my family, whose love, optimism and support have been my foundation throughout this journey. Thank you for always believing in me.

Finally, a huge thanks to all my friends who contributed to this journey, whether through stimulating discussions, much-needed laughs, or just being there when I needed it most. You all played a part in this achievement, and I'm truly grateful for your support.

Birmingham, UK, September 11th 2024 Luca Lavarini

Abstract

Title: “Determining NBER Recession Points Using Machine Learning”

Author: Luca Lavarini

Key words: Machine Learning, NBER recession, Business Cycle Forecast

The financial crises cause significant challenges due to their profound impact on the economy and the inherent difficulty in predicting such events. Successfully forecasting a financial crisis could offer remarkable advantages, enabling preemptive measures to mitigate its adverse effects. Previous research has highlighted the importance of various indicators in predicting economic downturns, including the inverted term spread, real GDP, and unemployment rates. Additionally, machine learning methods have shown potential in identifying non-linear patterns among these variables, making them valuable in forecasting NBER recessions.

In this study, we evaluated several machine learning classification and non-linear regression algorithms such as Support Vector Machine, K-Nearest Neighbours, Decision Tree, Extreme Gradient Boosting, Adaptive Boosting, Random Forest, Extra Trees, and Categorical Boosting other than traditional time series models like ARIMA and AR. The best forecast of the NBER recession points from 0 to 12 months ahead was obtained by inputting the best machine learning models’ prediction as one of the exogenous variables of an *ARIMA(1,0,1)*. The forecasts obtained were especially effective between $t + 0$ and $t + 4$, with real GDP being the most relevant macroeconomic feature. Additionally, one version of the forecast was better suited to predict market troughs than official NBER recessions.

Future research could extend this work by exploring the impact of different types of recessions, developing models tailored to emerging markets, or training models on specific big debt crises, such as using data from the 2008 financial crisis to forecast recessions similar to Japan’s 1990 economic downturn.

Resumo

Título: “Determinação dos Pontos de Recessão do NBER Usando Machine Learning”

Autor: Luca Lavarini

Palavras-chave: Machine Learning, Recessão do NBER, Previsão de Ciclo Econômico

As crises financeiras representam grandes desafios devido ao seu profundo impacto econômico e à dificuldade de prever esses eventos com antecedência. Prever uma crise financeira com sucesso pode oferecer vantagens consideráveis, permitindo a adoção de medidas preventivas para mitigar seus efeitos. Pesquisas anteriores destacaram a importância de indicadores como a inversão da curva de juros, o PIB real e as taxas de desemprego na previsão de recessões. Além disso, métodos de têm mostrado potencial na identificação de padrões não lineares entre essas variáveis, tornando-os valiosos para prever recessões do NBER.

Neste estudo, avaliamos diversos algoritmos de como Support Vector Machine, K-Nearest Neighbours, Decision Tree, Extreme Gradient Boosting, Adaptive Boosting, Random Forest, Extra Trees e Categorical Boosting, além de modelos tradicionais como ARIMA e AR. A melhor previsão dos pontos de recessão do NBER de 0 a 12 meses foi obtida ao inserir as previsões dos melhores modelos de como variáveis exógenas em um ARIMA(1,0,1). As previsões mais eficazes entre t + 0 e t + 4, o PIB real o indicador mais relevante. Além disso, uma versão do modelo mostrou-se mais adequada para prever os mínimos de mercado do que as recessões oficiais do NBER.

Pesquisas futuras podem explorar o impacto de diferentes recessões, desenvolver modelos para mercados emergentes ou usar dados de crises, como a crise de 2008, para prever recessões semelhantes à recessão do Japão de 1990.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

Autoregressive-Moving-Average	(ARIMA)
Autoregressive	(AR)
Autoregressive term of order one	(AR(1))
Area Under the Receiver Operating Characteristic Curve	(AUROC)
Capital Asset Pricing Model	(CAPM)
Consumer price index	(CPI)
Center for Research in Security Prices	(CRSP)
Change of direction	(COD)
Emerging markets	(EMEs)
Federal Reserve Economic Database	(FRED)
Gross domestic product	(GDP)
Great financial crisis	(GFC)
Gilchrist and Zakrajšek	(GZ)
Hyperparameter	(HP)
In-Sample	(IS)
Linear Vector Quantization	(LVQ)
Moving average term of order one	(MA1))
Machine Learning	(ML)
Multi-recurrent Neural Network	(MRN)
Mean squared error	(MSE)
National Bureau of Economic Research	(NBER)
Neural network	(NN)
Out-Of-Sample	(OOS)
Root mean squared error	(RMSE)
Spport Vector Machine	(SVM)
Vector autoregressive	(VAR)

1. INTRODUCTION

1.1 Recession Forecasting And Its Importance

Over the past century, the United States have experienced 16 economic recessions (St. Louis Fed, 2024). In some cases, these downturns unleashed catastrophic waves of unemployment, with the rate surpassing 25% during the Great Depression (Wheelock, 2020), reflecting the nation's economic collapse. This unprecedented surge in unemployment was mirrored by a catastrophic 26% decline in real Gross Domestic Product (GDP) between 1929 and 1933 (Dalio, 2022). Similarly, the Covid-19 crisis saw unemployment nearly hit 15%, and during the first quarter of 2020, the GDP contracted at an annual rate of 5%, plummeting by a staggering 32.9% annualised decline in the second quarter (Wheelock, 2020).

Given the immense strain recessions place on the economy, their accurate forecast becomes crucial. Predicting economic downturns can be highly beneficial to markets, businesses, and the broader economy, allowing people to prepare in advance and alleviate some of the problems. Moreover, such a forecast is critical for policymakers, as an early warning can enable them to take preemptive actions to mitigate the damages and attenuate the impacts of the economic fall, such as avoiding bubbles and managing economic and financial resources.

Although the National Bureau of Economic Research (NBER) is responsible for determining recession periods, this analysis is conducted retrospectively, only identifying recessions after they have occurred. Many academic studies have investigated the dynamics of recessions, assessing different variables as possible predictors and implementing different techniques; Estrella and Mishkin (1998) were among the first to do so, followed by Ponka (2016), Bruneau, Christensen and Meh (2018), and more recently Borio, Drehmann, and Xia (2019), and Sham (2019). Additionally, Qi (2001) was one of the pioneers in employing Machine Learning (ML) to analyse the NBER recession, followed, for example, by Giusto and Piger (2013), Chakraborty and Joseph (2017) and more recently, Orojo et al. (2023). Nevertheless, to the best of our knowledge, no method has already been able to forecast a recession in advance.

1.2 Aims And Objectives

This study addresses the following aims:

1. To develop a statistical ML model capable of predicting the onset and conclusion of NBER recession points in advance.
2. To Determine the optimum forecast horizon for robust prediction of NBER recession points.
3. To evaluate the impact of key macroeconomic variables on the predictive model's performance.
4. Evaluate the financial benefits derived from the forecasts.

These aims will be achieved by the following objectives:

1. Reviewing the latest research in forecasting the NBER recession points using statistical and ML techniques.
2. Acquiring preprocessing and transforming appropriate NBER and macroeconomic time series data to make them suitable for the forecasting task.
3. Evaluating various forecasting methods, such as binary and multi-nomial classification ML techniques and non-linear regression algorithms.
4. Examining linear time series forecasting approaches, including Autoregressive (AR) and Autoregressive-Moving-Average (ARIMA) models.
5. Evaluating the effectiveness of these models and assessing the best approach across different forecast horizons, ranging from $t + 0$ to $t + 12$ months.
6. Analysing the p-values and coefficients of key macroeconomic exogenous variables in the models employed across different forecasting time horizons to assess their significance, magnitude, and impact on the model's performance.
7. Developing a trading strategy based on the forecasts to capitalise on potential market opportunities.
8. Assessing the current state of the economy using the most recent data, to gauge the likelihood of a 2024 recession

The remainder of this paper is organised as follows. Section 2 describes the previous literature on the topic. Section 3 details the variables implemented and how they were constructed. Section 4 elucidates the different models implemented. Section 5 presents the results of the different approaches and the model through which final results were achieved. Section 6 illustrates a trading strategy based on the forecasts obtained and attempts to gauge the likelihood of a 2024 recession. Section 7 concludes by discussing how the results met the

research objectives and summarising the main limitations and the direction of possible further studies.

2. LITERATURE REVIEW

The literature on forecasting the NBER recession encompasses a broad spectrum of variables. Over time, the forecasting power of many different variables has been investigated, and different combinations of predictors have been explored, providing valuable insights and enhancing our understanding of the effectiveness of different variables and the key indicators of economic downturns.

2.1 Traditional Statistical Models

One of the most prevalent methodologies in this domain has been the use of probit models. These models are well suited for binary outcomes, such as predicting the occurrence of a recession. Estrella and Mishkin (1998) were among the first to systematically test various economic indicators by implementing a series of probit equations to evaluate the forecasting power of the variables to predict an NBER recession. Their research focused on interest rates and spreads, stock price main indexes, such as the S&P 500, Dow Jones Industrials and NYSE composite, monetary aggregates, individual macroeconomic indicators, like real GDP, and inflation, among other variables. While some variables are considered at time 0, others are included with a lag of one to three months. They concluded that stock price indexes were useful for the prediction within one to three-quarters horizons, while the yield curve slope was revealed to be a robust predictor beyond one-quarter, especially if considered individually.

Liu and Moench (2016) analysed a range of variables, including the term spread, the annual returns on the S&P 500 index, interest rates, various components of the Confidence Board's Leading Economic Index, and financial indicators as the manufacturers' new orders of capital goods and balances in Broker-Dealer margin accounts. They also considered a six-month lag of the term spread. Their research considered forecast horizons of 3, 6, 12, 18, and 24 months. They used univariate and multivariate probit models to assess the likelihood of future recession based on the predictors gathered. The best prediction was achieved at the three-month forecast horizon, employing a model that accounted for the annual return of the S&P 500, the term spread, and its six-month lag, which resulted in an area under the receiver

operating characteristic curve (AUROC) of 0.95. Similarly, the model attained an AUROC of 0.91 at a six-month horizon.

In addition, Bruneau, Christensen and Meh (2018) investigated the relationship between the NBER recession and the real estate market. They discovered that the large negative co-movements in house prices across US cities tend to cluster over time, being closely related to recession periods. Interestingly, they showed how cities that contribute more to the GDP or have more population were more influential in signalling a recession.

Ponka (2016) further expanded the literature by considering a broad range of credit variables, such as corporate bond spreads (Gilchrist and Zakrajšek (GZ) Credit Spread, introduced by Gilchrist and Zakrajšek (2012)), excess bond spread, default spread, which is considered as the difference between Baa and Aaa rated corporate bonds yields, total consumer credit, and real estate loans. In addition, he included some of the 'classic' recession predictors as the term spread, stock returns, computed as the logarithmic first difference of the S&P 500, federal funds rate, and the University of Michigan Consumer Sentiment Index. The research applied different lag structures for each independent variable with a maximum lag length of 12 months. The specific lags were chosen based on the Bayesian Information Criterion. The forecast horizons included 1, 3, 6, and 12 months ahead. The study used probit models to predict the US recession. The best Out-of-Sample (OOS) result was obtained using a model that included GZ credit spread and the excess bond premium as predictors, particularly for the 1-month ahead forecast; such a model had an AUROC of 0.915, underlying the forecasting power of the predictors involved. Additionally, the AUROC remained high also for the 3-month ahead forecast for the models including the excess bond premium and the GZ credit spread.

Borio, Drehmann, and Xia (2019) also considered the term spread and financial cycle proxies such as credit-to-GDP ratio, real property prices, and the debt service ratio, which measures interest payments plus amortisations divided by GDP. They applied panel probit models to forecast the NBER recession 12, 24 and 36 months ahead, comparing the predictive power of the term spread versus financial cycle measures. They showed that the debt service ratio outperformed the term spread with an OOS AUROC of 0.98 for a one-year forecast horizon and 0.82 for a two-year horizon.

2.2 Non-Linear Machine Learning Models

Many studies that cover the investigation of NBER recession predictors were implemented using probit or other linear models. On the one hand, this kind of model is straightforward, especially suited for binary outcomes, and its results are immediate to interpret. Moreover, it allows for easy inclusion of multiple predictors. On the other hand, they assume a linear relationship between the independent variables and the log odds of the outcome, which may risk reducing the complexity of the association between predictors. Additionally, probit models struggle to capture non-linear interactions and dependencies between predictors that, especially during a recession, may act upon complex and hidden non-linear interactions. Moreover, probit models are sensitive to the choice of the predictors and their lag structure.

Given the limitations of probit and other classic econometric models and that economic cycles are asymmetric and cannot be represented by linear constant parameter single-index models (Qi, 2001), new methods have emerged in handling non-linear relationships and considering the complexity of predictors' interactions. Qi (2001) was the first to use ML to predict the NBER recession. Specifically, he applied a three-layer feedforward neural network (NN) model over the same dataset that Estrella and Mishkin (1998) employed, with the same lag structure. The study examined forecasting horizons from 1 to 8 quarters. He concluded that the term spread is the single OOS best indicator for the prediction between 2 and 6 quarters ahead, based on the OOS mean squared forecast error. Additionally, they showed that the prediction could benefit when the S&P 500 (considered as monthly average) was added as a predictor. Finally, it is shown that NN can generate very clear binary signals of recession without implementing a cutoff value to gauge the presence of the recession.

Giusto and Piger (2013) focused on non-farm payroll, industrial production, real income, and real manufacturing and trade sales with their one-month lag as independent variables to forecast the NBER recession at time $t + 0$, meaning they aimed to forecast the recession in real-time. To do so, they employed a linear vector quantization (LVQ) model, a classification algorithm that is a form of neural network. The study demonstrated that LVQ *nowcasted* all five recessions between December 1976 and August 2013, the sample period under analysis, with an average delay of 126 days for peaks and 192 days for troughs in real-time.

Berge (2015) analysed a broad range of interest rates and interest spreads, such as the different kinds of term spread BAA and AAA corporate bond spreads and the TED spread, financial variables such as the S&P 500 index and the real money growth, calculated as M2 deflated by the consumer price index (CPI), in addition to macroeconomic indicators like the industrial production and real personal income. Based on these variables, the study focused on forecasting the NBER 0, 6, 12, 18, and 24 months ahead by implementing Bayesian Model Averaging, linear and non-linear boosted algorithms. The research showed that non-linear boosting provided the best OOS prediction with an AUROC of 0.966 for the nowcast and 0.943 for the 12-month prediction.

Vrontos, Galakis and Vrontos (2021) considered 56 macroeconomic and financial market-related variables, including the term spread, unemployment, productivity measures, monetary aggregates, corporate bond spreads, real GDP and other industrial production metrics, S&P 500 monthly returns and change in consumer sentiment, among others. They attempted to forecast the NBER recession with different ML algorithms 1, 3, 6, and 12 months ahead. Different lag structures were implemented depending on the time horizon prediction, but a combination of 3, 6, and 12 months past lag was generally considered. The study tested various ML models, including penalised logit regression models, k-nearest neighbours, discriminant analysis, tree-based models, ensemble methods, Naïve Bayes, and Bayesian generalised linear models. They demonstrated that ML techniques, particularly penalised logit regression models and k-nearest neighbours, provide relevant improvements in the NBER forecasting over econometric models such as probit and logit. Additionally, they confirm that the term spread was confirmed to be an important feature in forecasting the NBER, especially for the 12-month horizon.

Following Giusto and Piger (2013), Orojo et al. (2023) implemented a Multi-recurrent Neural Network (MRN) to forecast the NBER recession at $t + 1$. They used the four indicators mentioned above; however, they introduced their change of direction (COD)[1] for a total of eight

[1] The COD variables recorded of the difference in magnitude of any give variable at time t and $t - 1$; their values were assigned as follows: +1 for a positive change, -1 for a negative change, and 0 for no change.

predictors. Finally, they showed how MRN performed better than LVQ and Support Vector Machine (SVM).

2.3 Discussion

The studies concerning the NBER recession and its prediction have been characterised by the implementation of econometric models, particularly probit models. These models have long been used for their simplicity and ability to handle binary outcomes like the occurrence of a recession. The work of Estrella and Mishkin (1998) set the fundamentals in this domain, testing a broad range of indicators and demonstrating the robustness of the term spread slope in anticipating the NBER recession. From there, many studies have been conducted, expanding the knowledge concerning the performance of macroeconomic and financial variables. For example, Liu and Moench (2016) demonstrated that including the annual returns of the S&P 500 alongside the term spread and its six-month lag achieved high predictive accuracy. These many studies highlight the effectiveness of traditional econometric models in forecasting the NBER. Yet, they also present some limitations. In particular, they are unsuitable for capturing non-linear relationships and complex interactions among predictors.

In fact, one of the critical limitations of traditional models lies in their assumption of linearity and potential oversimplification of more complex relationships. However powerful, probit models struggle to capture non-linear interactions that, as crises often happen unexpectedly, characterize economic downturns.

The emergence of ML has been revealed to be a powerful tool to deepen the research, offering a high degree of flexibility in handling complex, high-dimensional datasets. Qi (2001) was a pioneer in applying ML to predict the NBER recession, showing the potential of such an approach. Further studies, such as those by Berge (2015) and Vrontos, Galakis and Vrontos (2021), have employed a broad range of ML algorithms, consistently demonstrating the high forecasting power of ML models across different forecast horizons.

Despite its many advantages, ML is not without challenges. As Chakraborty and Joseph (2017) observed, ML models function as “black boxes”, making it difficult to interpret the relationships between variables and understand how the predictions are drawn. Such a lack of transparency can be problematic as the interpretability of models is crucial for market players’ and policymakers’ decisions. Consequently, despite ML’s potential, its application should be cautiously approached.

In light of the existing literature, our study aims to further explore the forecasting power of different ML models in predicting the NBER recession. By considering variables previously demonstrated to be connected to the business cycle, not only we seek to forecast economic downturns from 0 to 12 months ahead, but, on the work of Gómez-Cram (2021), who showed the practical benefits of recession forecasting for investment purposes, we also test a trading strategy that switches from stock to treasury bills during predicted recessions, extending the application of this study to the real world of financial decision making.

3. VARIABLES AND DATASET

3.1 NBER Data

The dependent variable is the NBER recession period, which is a binary variable determined ex-post by the NBER's Business Cycle Dating Committee and defined as "the period between a peak of economic activity and its subsequent trough, or lowest point" (National Bureau of Economic Research, 2022). A value of one indicates a recession, while a zero points out an economic expansion. This variable was retrieved from the Federal Reserve Economic Database (FRED) of the Federal Reserve Bank of St. Louis.

This study considered monthly data from June 1st 1948, to December 1st 2021, for a total of 124 recession months, 14% of the total sample, and 12 distinct recession periods. The longest expansion period lasted 128 months, from the Great Financial Crisis (GFC) to Covid-19, while the shortest expansion period was 12 months, occurring during the middle of the oil shock of the early 1980s. Over these more than 70 years considered, the longest period of economic contraction was 18 months during the GFC, and the shortest was 2 months during Covid-19. On average, economic expansion phases lasted 67 months, with a standard deviation of 40 months, while recessions averaged 10 months in duration, with a standard deviation of 5 months.

3.2 Macroeconomic And Financial Features

All the independent variables were analysed with a monthly frequency and considered at the beginning of the month. When they were available at a different recurrence, an interpolation was applied.

Following Giusto and Piger (2013) from FRED, we retrieved the total non-farm payroll, industrial production, and real personal income. Considering Estrella and Mishkin (1998), we gathered the 3-month Treasury bill rate and the 10-year US Treasury bond market yield (the data before March 1st, 1953, were gathered from Amit Goyal's website), the Real Gross Domestic Product, and the CPI by computing the percentage change over 12 months. In light of Sham (2019) we included the unemployment rate. Furthermore, we included the balance on the current account, the spot West Texas Intermediate price. From MacroTrends.com, we retrieved the nominal price of gold and silver.

In accordance with Estrella and Mishkin (1998), Liu and Moench (2016) from Wharton Research Data Services, specifically the Center for Research in Security Prices (CRSP), the value of the S&P 500 index. Finally, from Goyal and Welch (2008) dataset, we retrieved the Dow Jones Industrial Average book-to-market ratio, the 12-month moving sum of dividends paid on the S&P 500, the 12-month moving sum of earnings paid on the S&P 500, the 12-month moving sum of net issues by NYSE listed stocks divided by the total end-of-year market capitalisation of NYSE stocks; S&P 500 variance, computed as the sum of squared daily returns on the S&P 500; and following Berge (2015) and Ponka (2016) we also considered both the yields on AAA and BAA corporate bonds. Table 1 summarises the macroeconomic and financial variables.

Table 1: Macroeconomic and financial variables

Series	Macroeconomic and fiancial variables (FRED ticker in parenthesis) (monthly frequency)	Dataset
PAYROLL	All Employees, Total non-farm (PAYMES)	FRED
IND_PRO	Industrial Production: Total Index (INDPRO)	FRED
R_INCOME	Real personal income excluding current transfer receipts (PIECTR)	FRED
BILL	3-Month Treasury Bill Secondary Market Rate, Discount Basis (TB3MS)	FRED
BOND	Market Yeld on US Treasury Securities at 10-year Constant Maturity, Quoted on an Investment Basis (GS10)	FRED
R_GDP	Real Gross Domestic Product (GDPC1)	FRED
INFL	Consumer Price Index for All Urban Consumers: All Items in US. City Average (CPIAUCLS)	FRED
UNP	Unemployment Rate (UNRATE)	FRED
CRAC	Balance on Current Account, NIPA's (NETFI)	FRED
WTI	Spot Crude Oil Price: West Texas Intermediate (WTISPLC)	FRED
GOLD	Spot gold price	MacroTrends.com
SILVER	Spot silver price	MacroTrends.com
SP500	S&P 500 Index	CRSP
DIVIDENDS	12-month moving sum of dividends paid on the S&P 500	Amit Goyal's website
EARNINGS	12-month moving sum of earnings paid on the S&P 500	Amit Goyal's website
DJIA_BTM	Dow Jones Industrial Average book-to-market ratio	Amit Goyal's website
COR_ISS_ACTIVITY	12-month moving sum of net issues by NYSE listed stocks divided by the total end-of-year market capitalisation of NYSE stocks	Amit Goyal's website
SP500_SIGMA	S&P 500 variance	Amit Goyal's website
CORP_BOND_AAA_YIELD	AAA corporate bond rated yield	Amit Goyal's website
CORP_BOND_BAA_YIELD	BAA corporate bond rated yield	Amit Goyal's website
SMB	Small Minus Big factor	Kenneth R. French Data Library
HML	High Minus Low factor	Kenneth R. French Data Library

3.3 Bespoke Features

In accordance with Estrella and Mishkin (1998), the term spread was built by the 3-month bill rate to the 10-year US bond yield; additionally, we calculated the Short-Long ratio by dividing the 3-month bill rate by the 10-year US bond yield, and following Liu and Moench

(2016) we included the monthly returns on the S&P 500. Furthermore, we built the real income over real GDP and gold-to-silver ratios. In fact, gold is considered a safe-haven asset, while silver tends to be favoured during expansions due to its greater industrial applications. It follows that such a ratio is supposed to increase when investors favour safe-haven assets (such as in economic downturns) and decrease during economic expansions.

Finally, recognising that some industries can influence the prediction of the NBER recession, such as real estate (Bruneau, Christensen and Meh, 2018), and the credit and banking industry (Ponka, 2016), while others may be directly impacted by a recession, such as automobiles, oil, manufacturing industries, especially during periods of deleveraging (Dalio, 2022), we calculated the cumulative returns for the following industries using data from the Kenneth R. French Data Library: Agriculture, Entertainment, Construction Materials, Constructions, Steel Works Etc, Machinery, Automobiles and Trucks, Petroleum and Natural Gas, Utilities, Communication, Business Services, Retail, Banking, and Real Estate. We also added the Small Minus Big and High Minus Low factors to our dataset.

All variables retrieved from the Kenneth R. French Data Library, Amit Goyal's website, and CRSP are recorded as monthly performance data at the end of each month. For consistency, we considered them at the beginning of the following month. For instance, the return of the agriculture industry for January 2000, known at January 31st, was recorded on February 1st. Table 2 summarises the bespoke variables.

Table 2: Bespoke variables

Series	Bespoke vairables (monthly frequency)
TERM_SPREAD	Term spread (BOND - BILL)
SL_RATIO	Short-Long ratio (BILL / BOND)
R_INCOME_R_GDP	Real personal income to real GDP ratio (R_ICNOME / R_GDP)
GOLD_SILVER	Gold to silver ratio (GOLD / SILVER)
SP500_RET	S&P 500 monthly returns
CR_AGRIC	Cumulative returns of agriculture industry
CR_FUN	Cumulative returns of entertainment industry
CR_BLDMT	Cumulative returns of construction materials industry
CR_CNSTR	Cumulative returns of constructions industry
CR_STEEL	Cumulative returns of steel works etc. industry
CR_MACH	Cumulative returns of machinery industry
CR_AUTOS	Cumulative returns of automobiles and trucks industry
CR_OIL	Cumulative returns of petroleum and natural gas industry
CR_UTIL	Cumulative returns of utilities industry
CR_TELCM	Cumulative returns of communications industry
CR_BUSSV	Cumulative returns of business services industry
CR_RTAIL	Cumulative returns of retail industry
CR_BANKS	Cumulative returns of banking industry
CR_RLEST	Cumulative returns of real estate industry

Subsequently, based on Orojo et al. (2023), we built the COD for the 41 predictors so far retrieved. At this stage, the dataset included 82 independent variables plus the target, covering the period from January 1st 1948, the earliest date for which all the variables were available, to January 1st, 2024.

3.4 In-Sample, Validation, and Out-Of-Sample Split

The In-Sample (IS) set ranged from July 1st 1948 to June 1st 2006, and comprised 696 monthly observations, while, at $t + 0$, the OOS set went from July 1st 2006 to December 1st 2021, for a total of 186 observations. in this way an 80:20 ratio between the IS and OOS datasets was maintained.

The IS set was further divided into training and validation sets, maintaining an 80:20 training-to-validation ratio for all the forecast time points. This division ensured that the models were trained on the data covering the 9 NBER recessions between July 1st 1948 and July 1st 1994, validated between August 1st 1994 and June 1st 2006, on the 2001 recession, and finally tested OOS on the 2008 and 2019 crises.

3.5 Data Processing

During the training-validation phase, the training set was initially standardized, and the same standardisation parameters were subsequently applied to the validation set. During the IS training and OOS testing phase, the standardisation was first applied to the IS set, and then the same standardisation parameters were applied to the OOS set. Following Orojo et al. (2023), in both cases, the COD variables were not standardised.

4. MODEL FITTING AND SELECTION METHODOLOGY

Throughout the project, various methodologies were explored and implemented. As the study progressed, less effective approaches were discarded, and other procedures were employed until the most suitable solution for accurate forecasting was identified on the validation set. This section explains how the different approaches were implemented and different models were built.

4.1 Lag Structure And Forecast Horizons

The final dataset comprised the 41 predictors, their COD, and the predictors lagged of $t - 6$ for a total of 123 variables. The lag structure was defined considering that the larger the

dataset was, the less reliable the models' validation output was, although a short-time lag structure always benefitted the forecast. Various machine learning algorithms were then applied to predict NBER recessions at the following monthly time horizons: t + 0 (nowcast), *t + 1*, *t + 2*, up to *t + 12* months ahead.

4.2 Binary Classifiers

Initially, the target was treated as a binary variable, with values of 1 in case of a recession and 0 otherwise, and the forecast was approached as a binary classification problem. As such, the *LogisticRegression* and *LinearDiscriminantAnalysis* linear ML classifier were implemented, as well as the following non-linear ones: *KNeighborsClassifier*, *DecisionTreeClassifier*, *AdaBoostClassifier*, *XGBClassifier*, *RandomForestClassifier*, *ExtraTreesClassifier*, and *CatBoostCalssifier*. *SVC*, was considered both in its linear and non-linear version by allowing the HyperParameter (HP) tuning process to choose between linear and non-linear kernels. Specifically, the implementation of *RandomForestClassifier*, *AdaBoostClassifier*, *KNeighborsClassifier*, and *LinearDiscriminantAnalysis* was inspired by Vrontos, Galakis, and Vrontos (2021), while the choice of non-linear boosted methods by Berge (2015). Such a broad range of algorithms allowed for a more thorough consideration of various data patterns and relationships.

Subsequently, the HP grid was defined for each algorithm, and the HPs were selected within an appropriate range. Afterwards, we applied the BayesSearchCV method on the training set and employed the F1 scoring method to approximate the best HPs to tune each algorithm. The number of iterations of the *BayesSearchCV* was set to 50 except for *LogisticRegression DecisionTreeClassifier* and *CatBoostClassifier*, which were reduced due to computational and time limitations. Table 3 summarises the HP grid employed.

4.3 Multi-nomial Classifiers

Subsequently, as the study explored the multi-nomial classification, the dependent variable was assigned values of 1 during an NBER recession, 0.8 one month before, 0.6 two months prior, 0.4 three months in advance, and 0 for all the other non-recession months these five values were transformed into classes. As the classes were heavily imbalanced (at $t = 0$, in the training set, *Class 0: 433; Class 1: 96; Classes 0.4; 0.6; and 0.8: 9 values*; while, in the validation set, *Class 0: 128; Class 1: 9; Classes 0.4; 0.6; and 0.8: 1 value*), the *smote* technique was implemented to balance the training set, and obtaining better validation predictions. After

trying different proportions of the five categories, we selected the following: *Category 0: 400 observations; Category 1: 150 observations; Categories 0.4, 0.6 and 0.8: 30 observations*. We did not further increase the categories between 0 and 1 because of overfitting concerns.

After dividing the target variable into five different classes, as for the binary classification, the same ten algorithms were tuned and trained (see Table 3)

4.4 Linear Regression Models

Two datasets were employed for the implementation of the linear regression models. Firstly, a reduced version of the full dataset was used. Similarly to Orojo et al. (2023), it comprised the total non-farm payroll, industrial production, and real personal income, all with their COD for a total of six variables. No lags were included.

The reduced dataset's predictors were first input into ARIMA and AR models as exogenous variables. The models *ARIMA(p,d,q)* and *AR(p)* were tried with p and q ranging from 0 to 12, while the parameter d ranged from 0 to 2. After that, the full dataset was considered, and its predictors were input into the linear models as exogenous variables.

4.5 Non-linear Regression Model

With this approach, the target variable was treated as a numerical one. As for the multinomial approach, the dependent variable was assigned with the values of 1 during an NBER recession, 0.8 one month before, 0.6 two months prior, and 0.4 three months in advance. Moreover, in this case, the non-recession values were substituted with a random value of mean 0.025 and standard deviation 0.005 to simulate a semi-linear pattern rather than five different classes.

The following non-linear regression algorithms were then employed: *KNeighborsRegressor, DecisionTreeRegressor, CatBoostRegressor, SVR, AdaBoostRegressor, XGBRegressor, RandomForestRegressor,* and *ExtraTreesRegressor*. All of them were tuned as previously described in the classification approach, except that they were selected by considering their negative mean squared error. Their HP grid is summarised in Table 4.

4.6 Additional Methodological Considerations

Lastly, we combined the regression and the ARIMA(1,0,1) methods one after the other. The target variable was first processed in the same way as in the regression approach. Subsequently, as previously, the eight algorithms were tuned and trained on the training set and

assessed on the validation set (obviously, at this stage, the same forecasts were obtained as the dataset and algorithms were the same). Thereafter, the best training and validation algorithms' forecasts were stored in a variable we called *pred*. To build this variable, we considered the validation root mean squared error (RMSE), the graphical shape of the *ARIMA* output against the target and the number of significant variables in the summary of the ARIMA model prediction.

In some cases, *pred* was obtained by summing up to three different algorithms' predictions. We also created a variable called *stress*, which recorded the variable *pred* shifted back of three months to capture the stress in the economic cycle before the NBER recession prediction. Then, the two new variables were added to the set of predictors. Finally, the 125 predictors were input as exogenous variables into an *ARIMA(1,0,1)* for all the forecast time points.

This approach led to two outcomes: either the ARIMA assigned a coefficient of *1* to the *pred* variables and *0* to all the other predictors, replicating the output from the best regression algorithms, or as previously, all predictors attributed with a coefficient different from 0 but always insignificant, failing to identify the recession and stress periods.

As it was evident that a smaller set of predictors had to be input in the *ARIMA* as exogenous variables, we selected the *n* predictors that most positively and negatively correlated with the target at each point in time of the forecast, with *n* set to *3*, *5*, *10*, and *15*. At this point, two versions of this approach were constructed.

In the *first version* of this additional approach, the *n* predictors were selected after the regression algorithms were run and the variables *pred* and *stress* built. In this case, the vairbales input in the *ARIMA* models were *2n*. With this approach, as *n* changed, the algorithms did not change their validation prediction, nor did *pred*. This is because all the algorithms were always tuned and trained on the same dataset. However, the *ARIMA*'s result was influenced when *n* changed, as this affected the number of exogenous variables input into the model, which had repercussions on the validation RMSE, the significance of the predictors input, and the value of the monthly forecast over the validation set. For all the forecast time points, *pred* and *stress* were always included in the *ARIMA* model, as they were always the most highly correlated variables with the target.

In the *second version*, the variables were selected before running the regression algorithms, whose results were then affected by being trained on a different set of predictors. As previously, the variables *pred* and *stress* were built and then added to the set of exogenous variables for a total of *2n + 2* features.

Table 3: HP grid for binary and multi-nomial classification algorithms

LogisticRegression	***LinearDiscriminantAnalysis***	***KNeighborsClassifier***	***CatBoostClassifier***
C: Real(0.1, 10)	Solver: Categorical('lsqr')	n_neighbors: Integer(3,25)	learning_rate: Real(0.01, 1.0, prior='log-uniform')
Solver: Categorical(['lbfgs', 'newton-cg', 'sag', 'saga'])	Shrinkage: Real(0,1)	metric: Categorical(['euclidean', 'manhattan', 'minkowski'])	depth: Integer(4, 15)
		p: Integer(1,2)	l2_leaf_reg: Real(1e-2, 10, prior='log-uniform')

DecisionTreeClassifier	***SVC***	***AdaBoostClassifier***
criterion: Categorical(['gini', 'entropy'])	C: Real(0.1, 10)	base_estimator__criterion: Categorical(['gini', 'entropy'])
max_depth: Integer(5,30)	kernel: Categorical(['linear', 'poly', 'rbf', 'sigmoid'])	base_estimator__max_depth: Integer(1, 40)
min_samples_split: Integer(2, 25)	degree: Integer(2, 5)	base_estimator__min_samples_leaf: Integer(1, 8)
min_samples_leaf: Integer(1, 50)	gamma: Categorical(['scale', 'auto'])	n_estimators: Integer(10, 30)
max_features: Categorical(['auto', 'sqrt', 'log2'])	coef0: Real(0.0, 1.0)	learning_rate: Real(0.01, 3, prior='log-uniform')
max_leaf_nodes: Integer(10, 50)		algorithm: Categorical(['SAMME', 'SAMME.R'])

XGBClassifier	***RandomForestClassifier***	***ExtraTreeClassifier***
max_depth: Integer(3, 30)	n_estimators: Integer(10, 40)	n_estimators: Integer(10, 30)
min_child_weight: Integer(1, 6)	max_depth: Integer(1, 30)	max_depth: Integer(1, 30)
n_estimators: Integer(10, 30)	min_samples_split: Integer(2, 15)	min_samples_split: Integer(2, 10)
learning_rate: Real(0.01, 1, prior='log-uniform')	min_samples_leaf': Integer(1, 8),	min_samples_leaf: Integer(1, 10)
subsample: Real(0.5, 1.0, prior='uniform')	max_features: Categorical(['auto', 'sqrt', 'log2'])	max_features: Categorical(['auto', 'sqrt', 'log2'])
colsample_bytree: Real(0.5, 1.0, prior='uniform')	bootstrap: Categorical([True, False])	bootstrap: Categorical([True, False])
gamma: Real(0, 5, prior='uniform')		

Table 4: HP grid for non-linear regression algorithms

DecisionTreeRegressor	*SVR*	*AdaBoostRegressor*
max_depth: Integer(5,30)	C: Real(0.1, 10)	base_estimator__criterion: Categorical(['mse', 'friedman_mse', 'mae'])
min_samples_split: Integer(2, 25)	kernel: Categorical(['poly', 'rbf', 'sigmoid'])	base_estimator__max_depth: Integer(1, 40)
min_samples_leaf: Integer(1, 50)	degree: Integer(2, 3)	base_estimator__min_samples_leaf: Integer(1, 8)
max_features: Categorical(['auto', 'sqrt', 'log2'])	gamma: Categorical(['scale', 'auto'])	n_estimators: Integer(10, 30)
max_leaf_nodes: Integer(10, 50)	coef0: Real(0.0, 1.0)	learning_rate: Real(0.01, 3, prior='log-uniform')

XGBRegressor	*RandomForestRegressor*	*ExtraTreeRegressor*
max_depth: Integer(3, 30)	n_estimators: Integer(10, 40)	n_estimators: Integer(10, 30)
min_child_weight: Integer(1, 6)	max_depth: Integer(1, 30)	max_depth: Integer(1, 30)
n_estimators: Integer(10, 30)	min_samples_split: Integer(2, 15)	min_samples_split: Integer(2, 10)
learning_rate: Real(0.01, 1, prior='log-uniform')	min_samples_leaf': Integer(1, 8),	min_samples_leaf: Integer(1, 10)
subsample: Real(0.5, 1.0, prior='uniform')	max_features: Categorical(['auto', 'sqrt', 'log2'])	max_features: Categorical(['auto', 'sqrt', 'log2'])
colsample_bytree: Real(0.5, 1.0, prior='uniform')	bootstrap: Categorical([True, False])	bootstrap: Categorical([True, False])
gamma: Real(0, 5, prior='uniform')		

KNeighborsRegressor	*CatBoostRegressor*
n_neighbors: [3]	learning_rate: Real(0.01, 1.0, prior='log-uniform')
metric: Categorical(['euclidean', 'manhattan', 'minkowski'])	depth: Integer(4, 15)
p: Integer(1, 2)	l2_leaf_reg: Real(1e-2, 10, prior='log-uniform')

5. MODEL IMPLEMENTATION AND OOS RESULTS

5.1 Binary Classifiers

After the tuning and training phase, the best HPs, the validation F1 and the accuracy scores of each algorithm were retrieved. Nevertheless, the highest F1 and accuracy scores were not always as informative as the graph of the forecasted variable against the target, through which the algorithms were primarily evaluated. In fact, the graphical representation allowed us to see better the entity of the mistakes and how far it was from the 2001 recession.

Although the validation results were accurate in identifying the recession, particularly within the short time horizon, between $t + 0$ and $t + 2$, with some minor errors during the crisis, by further increasing the forecast time horizon, the predictive power dropped, although in many cases a period of high volatility in the forecast output could be identified in correspondence of the recession. This could be because the dataset was highly imbalanced, and consequently, the algorithms could not distinguish between business cycle stress periods and real recession times.

5.2 Multi-nomial Classifiers

Based on the confusion matrix, the F1 and accuracy scores, and the graphical output of the forecasted prediction against the target, the algorithms performed well within $t + 0$ and $t + 3$, although with some minor mistakes. However, the best algorithms encountered difficulties distinguishing the five categories across all forecast time points, though they could differentiate between recession and non-recession outcomes. As the forecast time horizons increased, the prediction became less precise, identifying only some of the 2001 crisis points while missing others. In such cases, the algorithms yielded a more volatile forecast in correspondence with the recession. Appendix 1 presents the validation output.

Although the 2001 recession could be recognised by correctly classifying some points or detecting a more volatile period, these two approaches did not perfectly classify all the recession points, nor were they informative concerning the state of the business cycle throughout the validation period (see Section 5.6).

5.3 Linear Regression Models

In this case, the results became less reliable as the forecast time horizon was increased, especially after $t + 3$. Moreover, considering the validation RMSE at the same forecast time point, there was no substantial difference between when the parameters p and q were equal to 1 and when they were higher, for example, 12. While the parameter d diverged from 0, the forecast became inaccurate in most cases. Consequently, we concluded that *ARIMA(1,0,1)* and *AR(1)* were the best models.

When the full dataset was applied, even at $t + 0$, the *ARIMA(1,0,1)* and *AR(1)* models could not yield reliable forecasts, as all the variables were attributed with coefficients different from 0, but always statistically nonsignificant at 95% confidence level, and the models could not identify the most relevant predictors. Therefore, with the same models at $t + 0$, further trials were attempted by choosing a random subset of the full dataset. The number of chosen

predictors was set to *5*, *10*, and *15*. The variable selection was repeated 1000 times per each set of variables, and the best combination of predictors was retrieved based on the lower RMSE on the validation set of the models. Nevertheless, even at t + 0, the results were not always accurate and reliable, and this method was highly time-consuming.

5.4 Non-linear Regression Model

After the best HPs and the training and validation RMSE of each algorithm were retrieved, their prediction was also assessed by graphing the forecasted prediction and the target variable. As before, the graphical representation gave more insights concerning the reliability of the forecast during the recession period. Figure 1 and Figure 2 exemplify this concept. They represent the validation output of *ExtraTreesRegressor* and *DecisionTreeRegressor*, respectively, at *t + 3*. They had an RMSE of 0.159 and 0.143 respectively. Even if the second RMSE is lower, signalling a better forecast, it fails to distinguish between the recession period and a stress point throughout the business cycle, while the same did not happen with *ExtraTreesRegressor,* although it had a higher RMSE. Consequently, the graphical output was crucial for assessing the different algorithms.

With this approach, the predictions demonstrated strong performance across all forecasting time points, consistently correctly identifying the validation's recession points. Among the non-linear algorithms employed, *CatBoostRegressor* was revealed to be the best algorithm, yielding the best forecast for many forecast time points. As before, as the forecast time horizon increased, the prediction became less reliable, and other non-recession periods were increasingly highlighted along with the 2001 points, especially after *t + 6*.

Figure 1: ExtraTreesRegressor best model's validation output at t + 3

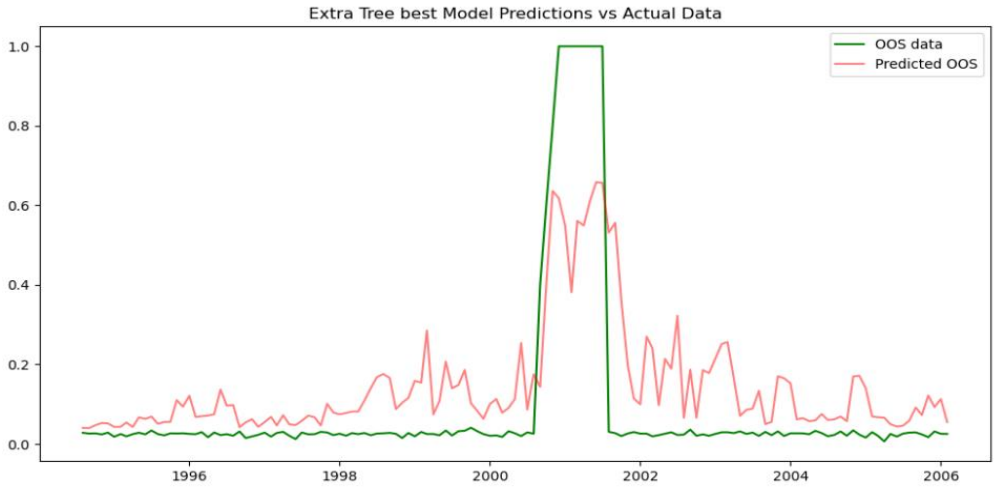
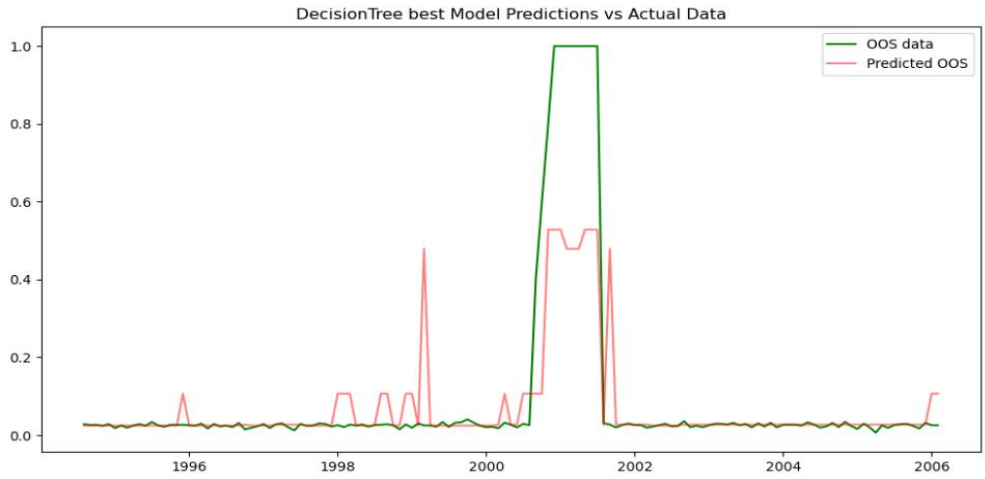


Figure 2: DecisionTreeRegressor best model's validation output at t + 3



5.5 Additional Methodological Considerations

Finally, we retrieved the validation results of both approaches for all the validation forecast time points and the values of *n*. In both cases, the larger the number *n* of predictors input in the *ARIMA(1,0,1)*, the worse the final predictions; in fact, the most reliable results were obtained with $n = 3$ or $n = 5$.

The output obtained could identify the 2001 recession throughout the whole forecasting window. However, as *t* increased, the predictions became more volatile. Nevertheless, the *ARIMA* output could almost always identify periods of strain in the economic business cycle, though it could not always distinguish between periods of simple economic stress and the dot-com recession, especially from $t + 8$ onwards.

Overall, the validation results of the two variations were not remarkably different, the same was true for the *ARIMA*'s RMSE score, although there were changes in the significance level of the exogenous variables input in the model, with the first approach revealing to be more reliable in this regard.

5.6 Methodology Selection

Although the multi-nomial classification output forecasted recession points in the same months where economic turmoil was detected with the non-linear regression and the combination of non-linear regression and *ARIMA* models, it still was too close to a binary classification, as it could not distinguish between the target's different categories introduced. Conversely, in the case of non-linear regression algorithms, the prediction's spikes could be measured, differentiating business cycle slowdown and actual recession. Moreover, the multi-

nomial classification difficulty spotted the entire recession period but only some months of it, which prevented a clear understanding of when a possible recession could have started and finished. Figure 3 and Figure 4 show the graphical output of *CatBoostRegressor* and *CatBoostClassfier* at $t + 3$, respectively. Although in Figure 3, the prediction did not achieve the value of 1 during the recession, as the target, its output was more insightful and precise compared to the multi-nomial classification's output in Figure 4.

On the other hand, both the non-linear regression and the combination of non-linear regression and *ARIMA* models yielded more insightful results. Nevertheless, the latter approach was judged as more complete, as it could shed light on how the variables influenced the prediction while accounting for the non-linear regression models' best prediction through the variables *pred* and *stress*. Moreover, the fact that *pred* was input into the *ARIMA* model made the prediction more flexible, allowing the *ARIMA* model to attribute a higher coefficient in case of a reliable prediction and a lower coefficient otherwise.

This final approach was, in fact, the most informative with respect to the NBER recession and the importance and magnitude of the variables, which is given by the predictors' coefficients and significance level. As such, the methodology selected for the OOS test was the combination of the non-linear regression and ARIMA approaches, and since there was little difference between the two versions implemented, both were tested.

Figure 3: CatBoostRegressor best model's validation output at t + 3 (best non-linear regression algorithm at t + 3)

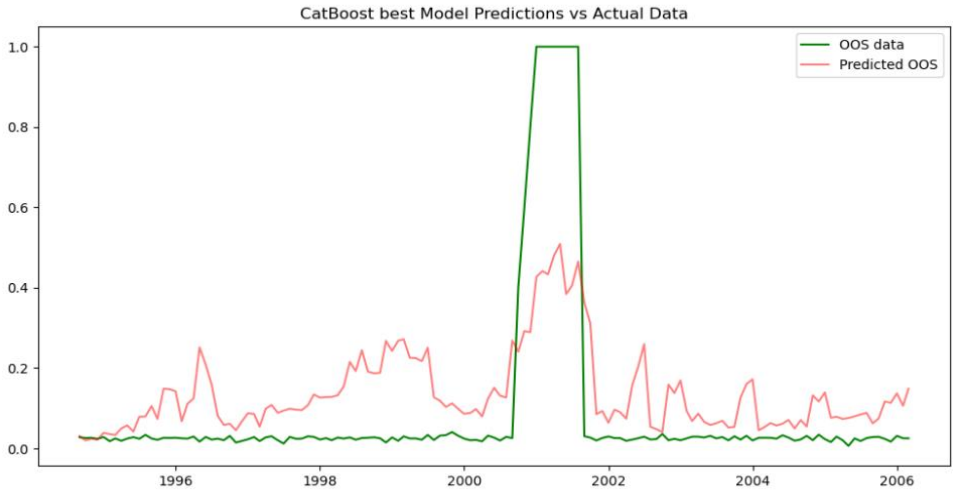
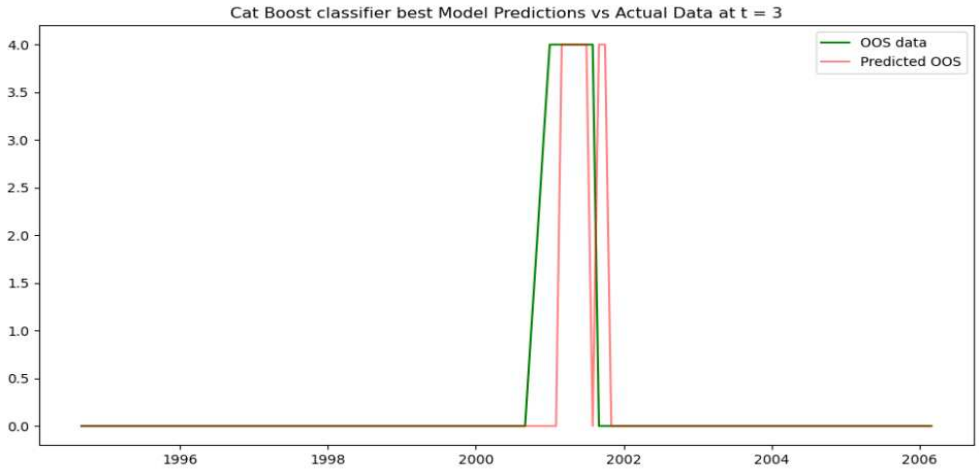


Figure 4: ExtraTreesClassifier best model's validation output at t + 3 (best multi-nomial algorithm at t + 3)



5.7 OOS Model Implementation

After the target variable was modified as in the non-linear regression approach in both the IS and OOS sets, per each point in time of the forecast, the four variations of the validation set results (one per each value of *n*) were retrieved and compared. The best validation prediction was chosen based on its RMSE, the number of significant variables, their significance level, and their graphical shape compared to the realised dot-com recession (not the validation target, as it was modified to account for stress periods).

Subsequently, a cutoff value was established to distinguish between recession and expansion periods. Initially, this was attempted by building a variable whose values were 1 if the forecast was above the cutoff and 0 otherwise. The cutoff value was adjusted between 0 and 1, with the initial focus on finding the value that minimized the mean squared error (MSE) between the actual recession points and the predicted ones. However, this approach became inefficient as the number of forecast time points increased. In some cases, the MSE also captured spikes from other stress periods, misclassifying them as recessions, which inflated the MSE score. As a result, a higher-than-optimal cutoff value was often chosen, excluding many true recession points.

Consequently, at each forecast time point, the cutoff was the result of the following formula:

$$Cutoff\ value(t) = \mu[validation\ predictions(t)] + \sigma[validation\ prediction(t)] \times k$$

The parameter *k* was manually adjusted to set the cutoff at a level that fully captured the dot-com recession, minimized its value, and reduced the impact of spikes from stress periods, avoiding false positive signals and maintaining a certain degree of flexibility.

Next, we applied the tuned non-linear regression algorithms to the OOS set, using the models trained on the training set at time *t* and validated on the corresponding validation set, with the previously selected *n*, to construct the variables pred and stress. These algorithms were then trained on the IS data and tested on the OOS set. The forecasts from both IS and OOS were stored, and the pred and stress variables were generated. Finally, the reduced set of exogenous variables was input into the ARIMA(1,0,1) model, and its output was compared to the predetermined cutoff to obtain the recession prediction.

Such a modelling structure was employed for both versions of the combination of the non-linear regression algorithms and the *ARIMA* models, the results of which are discussed below.

5.8 Version 1 OOS Results

The OOS results of the first version of the approach that combined non-linear regression and *ARIMA* models for all the forecast time points were obtained with $n = 3$, which means the *ARIMA(1,0,1)* employed six exogenous variables.

The nowcast ($t + 0$) could identify the two recessions very efficiently; overall, it correctly recognised Covid-19 and the whole GFC except for the first month at the beginning, with a recall score of 95%, though it included 13 false positive cases. As the forecast increased its time horizon, the model became less reliable. At $t + 5$, its recall dropped from 85% (at $t + 4$) to 40%, and at $t + 6$, it decreased to 25%, while the F1 score went from 0.567 at $t + 4$, to 0.327 and 0.263 at $t + 5$ and $t + 6$, respectively. At $t + 7$, the forecast completely missed the GFC, and at $t + 11$, Covid-19. Table 5 summarises the key performance metrics.

Concerning the false positive points, between $t + 2$ and $t + 5$, in addition to the points around the two crises, the forecast detected some points at the end of 2010, the beginning of 2011, and mid-2011. This aligns with the anticipation of the US debt ceiling ‘stress’ period in July 2011. Additionally, a short period of false recession was signalled at the end of 2013 and the beginning of 2014, anticipating a remarkable drop in the oil price in the latter half of the year, a condition that can coincide with a potential weakening of the global economy and consequently, a possible downturn.

Moreover, the analysis of the *ARIMA* output at each point in time of the forecast revealed that the most important variables were the *pred* and the autoregressive term of order one (AR(1)) based on their significance level, with a confidence level of 99% for both variables at any point in time of the forecast, and their coefficients, which were always the highest among the predictors. Moreover, the third most important predictor was the COD of the real GDP, which was statistically significant up to $t + 6$, with a confidence level of 99%, although its coefficient was close to zero except between $t + 2$ and $t + 4$, making it a relevant variable for the short-run forecast. Additionally, up to $t + 3$, the COD of payroll and real income was statistically significant with a 99% confidence level and had somewhat influential forecasting power. Throughout the 13 forecast points, the first-order moving average term (MA(1)) had a p-value of zero in eight cases. Despite its sign fluctuating frequently, it generally had a negative coefficient, which, in most cases, was the lowest among the model's variables, (see Table 6).

Considering the 13 forecast time points, the average of the coefficients statistically significant with a confidence level of 99% was 0.296, and the mean of the coefficient with a significance level higher than 10% was 0.003, meaning that the models' results were essentially attributed to the statistically significant variables as the others did not have a remarkable influence. Table 6 summarises the results of the first version of the approach that combines the non-linear regression and *ARIMA* models.

The fact that the confidence level of the variable *pred* and AR(1) was 99% for all the forecast time points means that there is strong evidence that the coefficients of these variables are different from zero in 99% of the cases. This does not imply that all the models were accurate or useful in forecasting recessions, which was determined by the performance metrics in Table 5 and the graphical performance of the prediction against the target. Therefore, even though the term spread was employed between $t + 6$ and $t + 12$, and was always significant with a confidence level of 95%, except for $t + 7$, it was not possible to assess its role in the forecast since for those forecast time points, the prediction was very poorly, and the recessions were almost always missed (see Table 5, and the OOS output below). The graphical output of the forecasts obtained with this method follows.

Figure 5: OOS version 1 predictions at t + 0

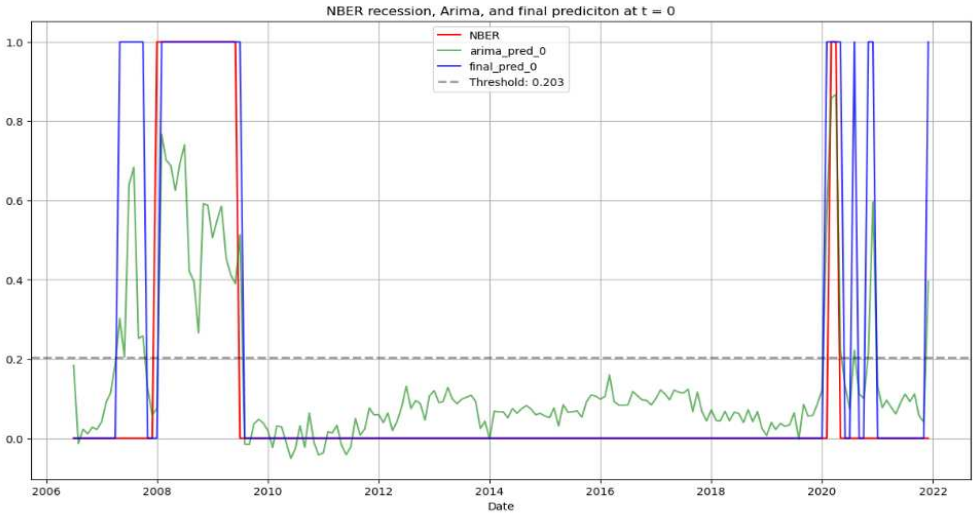


Figure 6: OOS version 1 predictions at t + 1

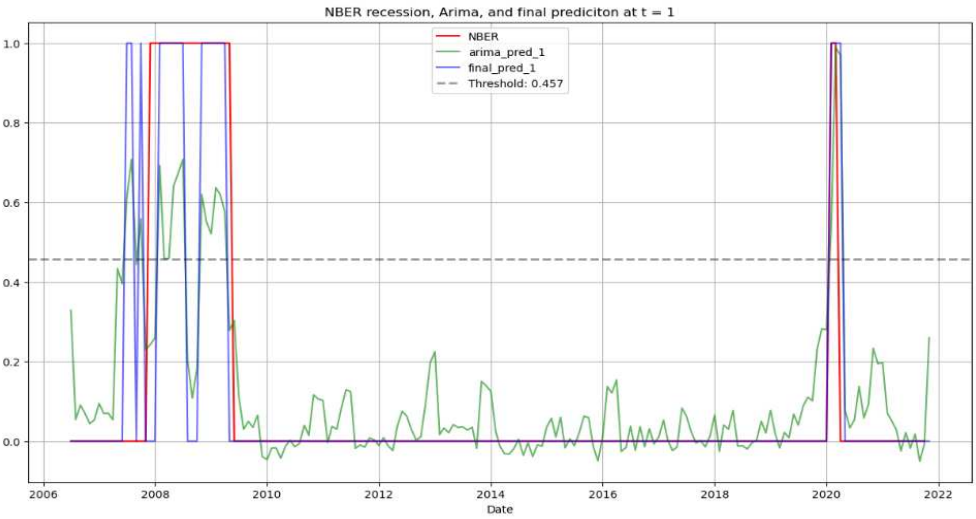


Figure 7: OOS version 1 predictions at t + 2

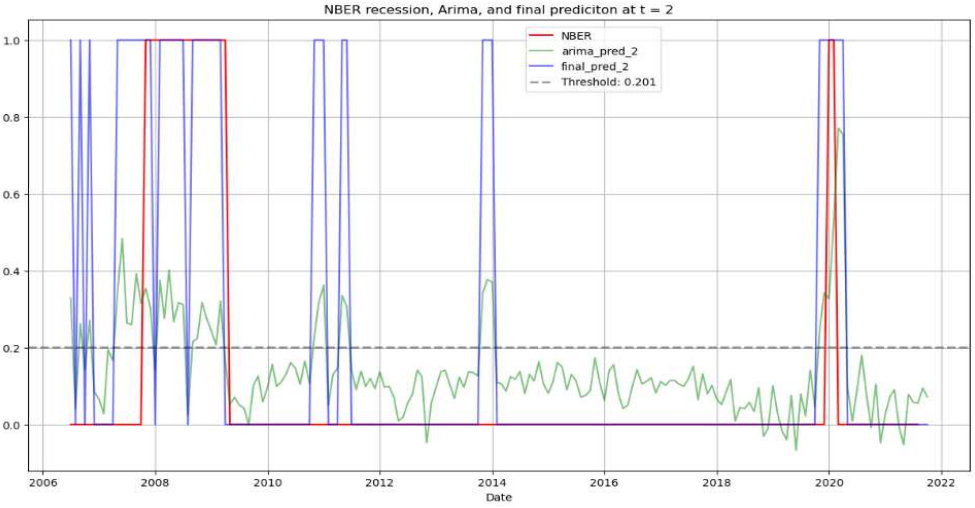


Figure 8: OOS version 1 predictions at t + 3

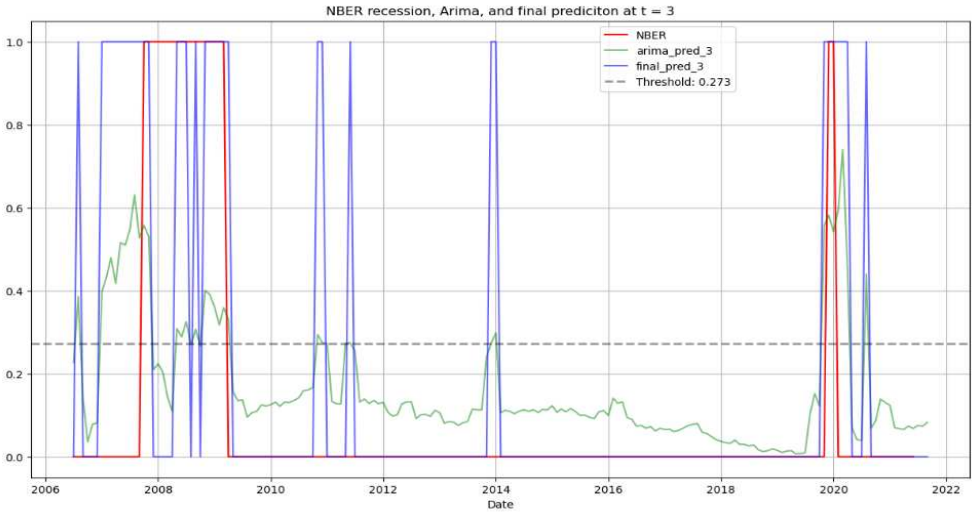


Figure 9: OOS version 1 predictions at t + 4

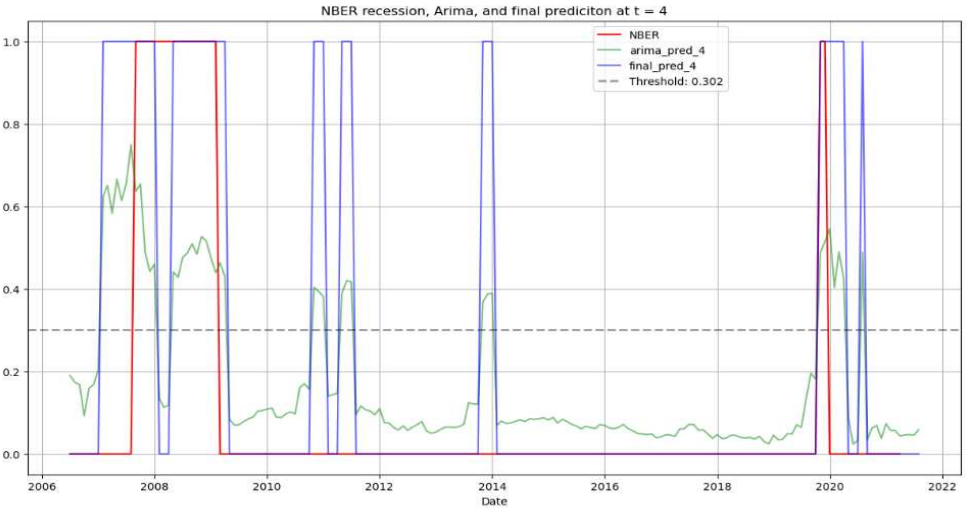


Figure 10: OOS version 1 predictions at t + 5

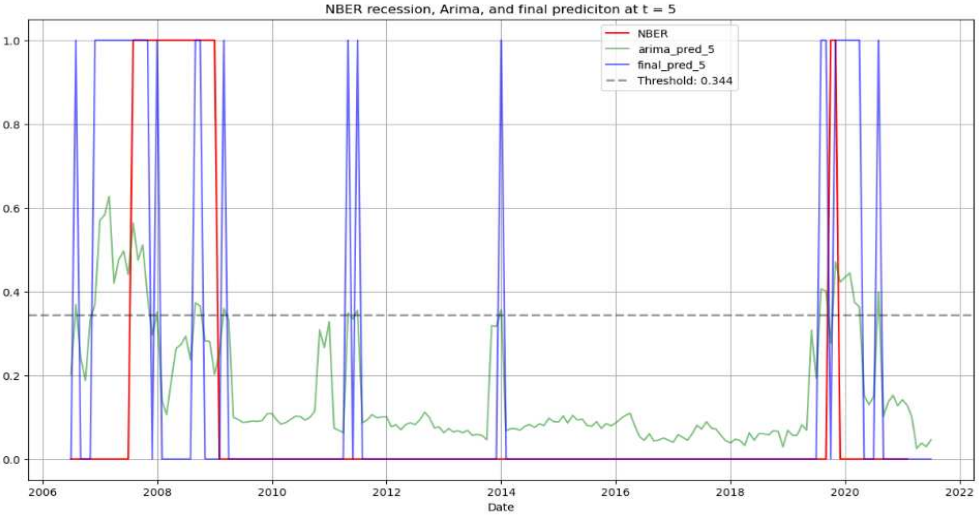


Figure 11: OOS version 1 predictions at t + 6

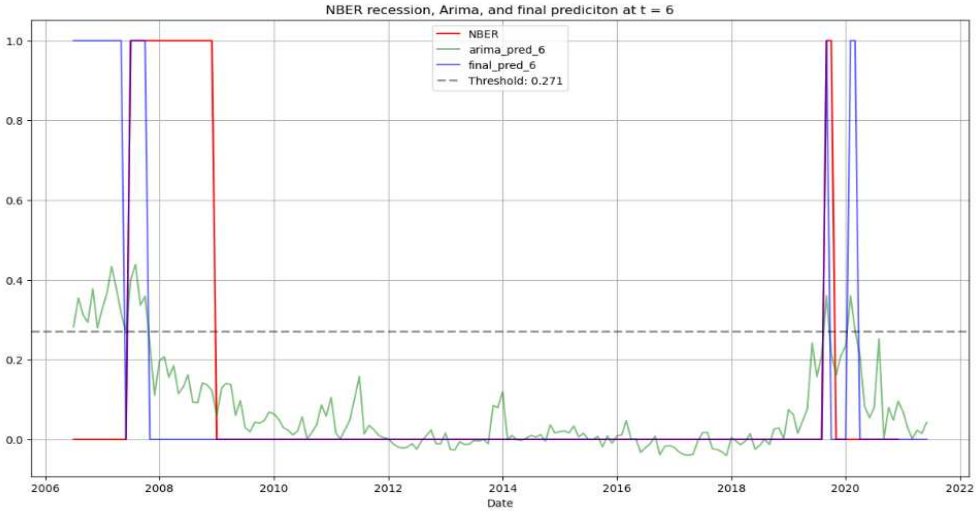


Figure 12: OOS version 1 predictions at t + 7

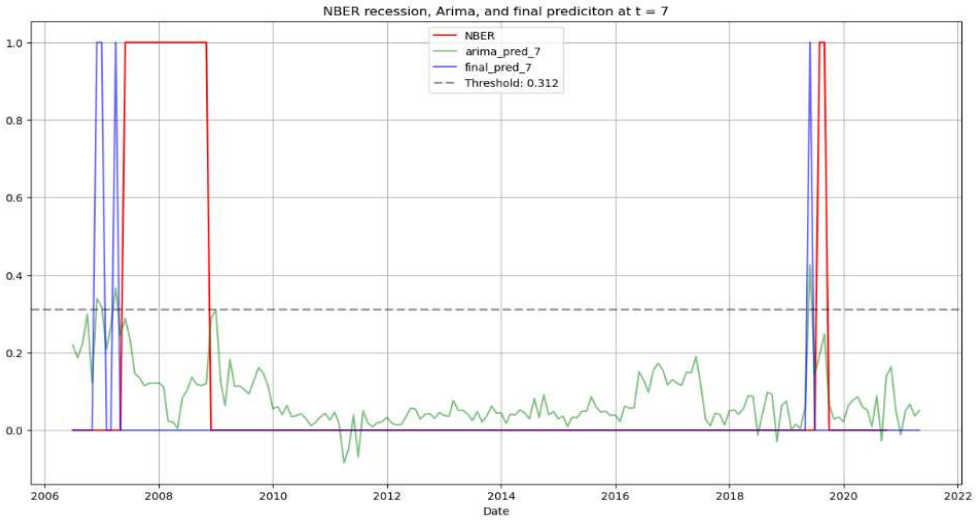


Figure 13: OOS version 1 predictions at t + 8

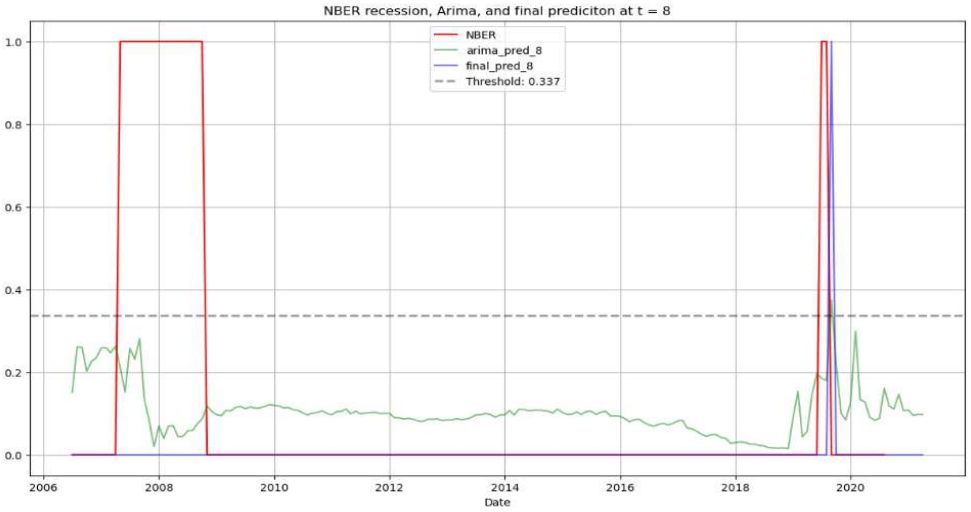


Figure 14: OOS version 1 predictions at t + 9

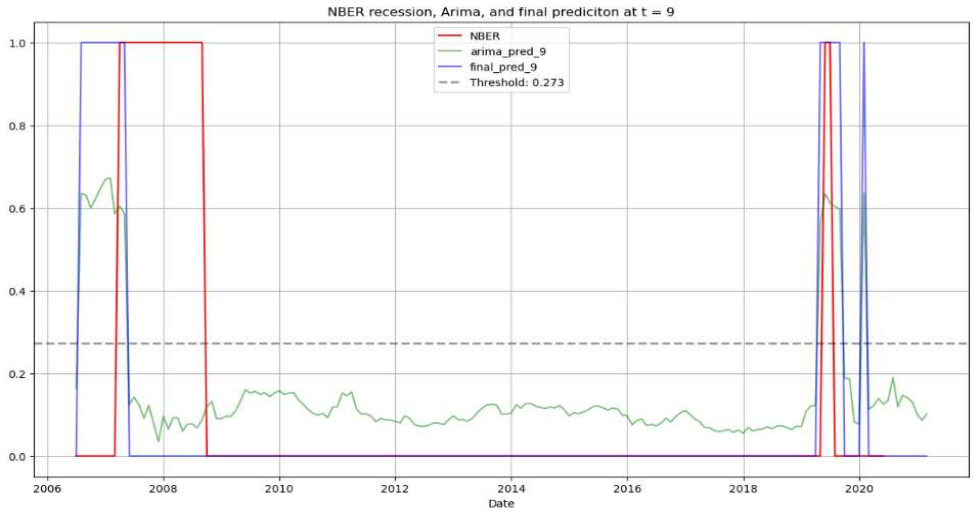


Figure 15: OOS version 1 predictions at t + 10

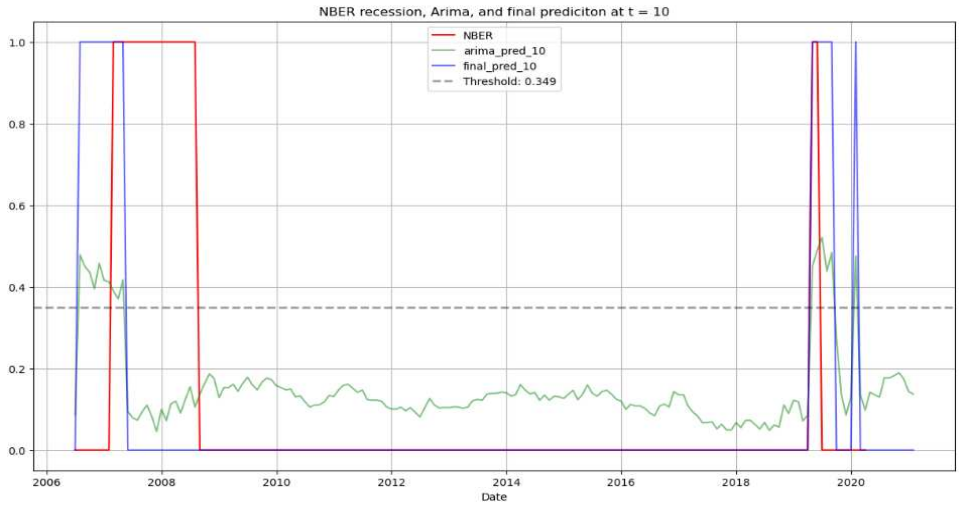


Figure 16: OOS version 1 predictions at t + 11

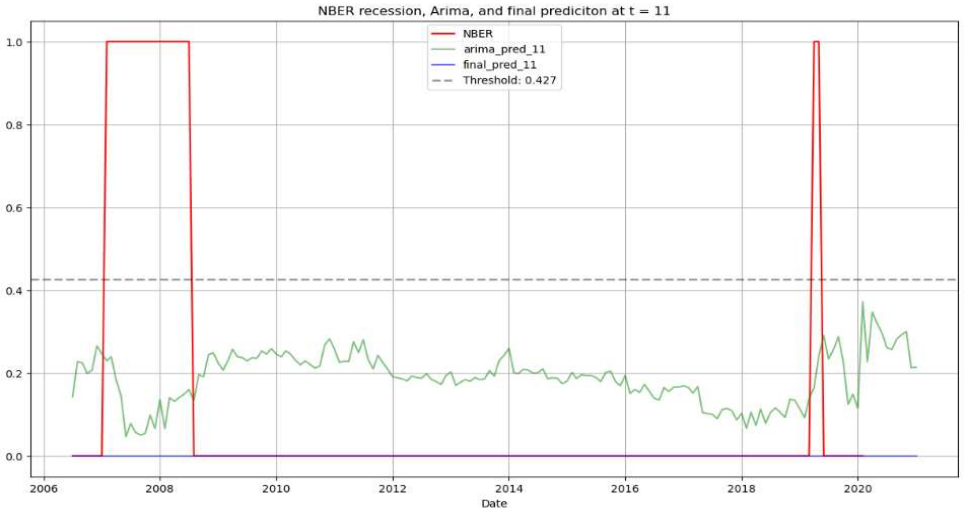
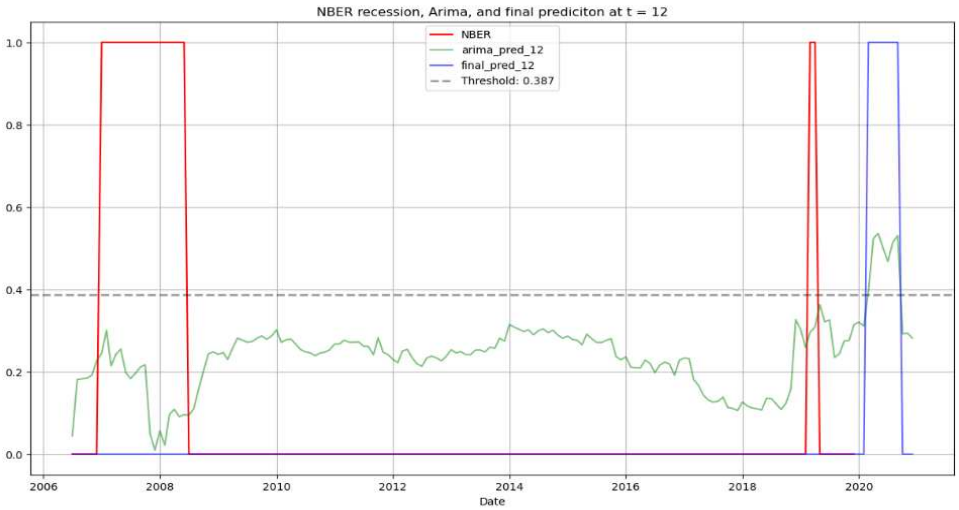


Figure 17: OOS version 1 predictions at t + 12



5.9 Version 2 OOS Results

The OOS results of the second version of the combination of the non-linear regression and *ARIMA* models were obtained with $n = 5$ for all the forecast time points except for $t + 5$, $t + 7$, $t + 9$, and $t + 10$, where n was set to 3. Hence, the results were based on 14 predictors except in the four cases mentioned before, where the predictors were 10.

In this case, the nowcast prediction was not as precise as in the previous model, as it missed 11 out of the 20 months of the GFC recession with only 2 false positives. As in the previous version, the forecasting power decreased after $t + 4$: at $t + 5$, the F1 dropped from 0.556 to 0.386, although the recall remained high, from 75% to 85%; however, the false positives at $t + 5$ were 51 monthly observations. Even though the GFC was completely missed at $t + 8$, interestingly, the Covid-19 recession was never entirely missed with this approach. Table 7 summarises the most relevant performance metrics of this approach.

Similar to the previous version, in this case, some false recessions were identified between late 2010 and early 2011, as well as between late 2013 and early 2014, between the intervals from $t + 2$ to $t + 5$.

Again, the analysis of the *ARIMA* predictors at each forecast time point revealed that the most relevant variables were *pred* and AR(1), although they were not statistically significant with a confidence level of 95% at $t + 0$ and $t + 1$ for the AR(1) and at $t + 8$ and $t + 12$ for the *pred* variable. Also in this case, the COD of the real GDP played a relevant role, with a statistically significant coefficient with a confidence level of 99% up to $t + 7$, and significant with a 95% confidence level at $t + 9$ and $t + 11$. Moreover, also the COD of

payroll, real income, and industrial production were statistically significant, with a confidence level of 99% up to *t + 4* (*t + 3* for the COD of payroll).

Overall, other major variables and their lags, included in the *ARIMA* model, such as unemployment, inflation, term spread, short-long ratio, and 3-month treasury spread, in many cases were not statistically significant at 95% confidence level or were employed after *t + 4*, hence, not contributing to a reliable forecasting model.

Finally, the average of the coefficients with a p-value of lower than 0.001 was 0.282, and the mean of the coefficient with a significance level higher than 10% was – 0.004, meaning that the models' results were essentially attributed to the statistically significant variables. Table 8 summarises the coefficient of the *ARIMA* model and their significance level for each point in time of the forecast. The graphical output of the forecasts obtained with this method follows.

Figure 18: OOS version 2 predictions at t + 0

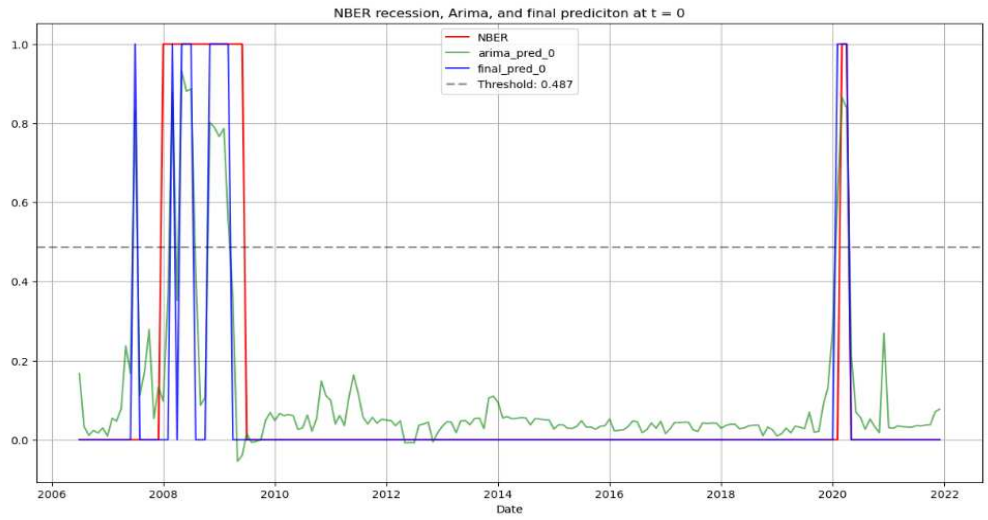


Figure 19: OOS version 2 predictions at t + 1

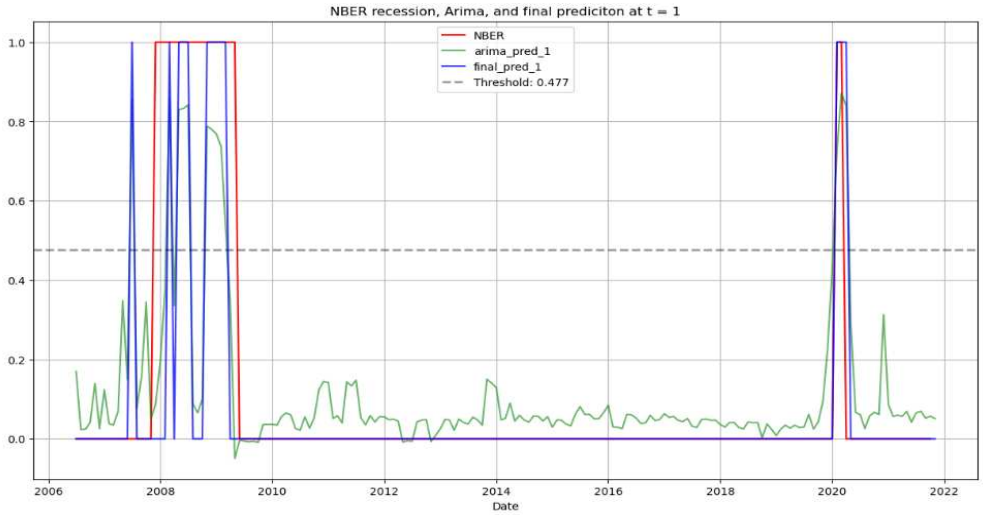


Figure 20: OOS version 2 predictions at t + 2

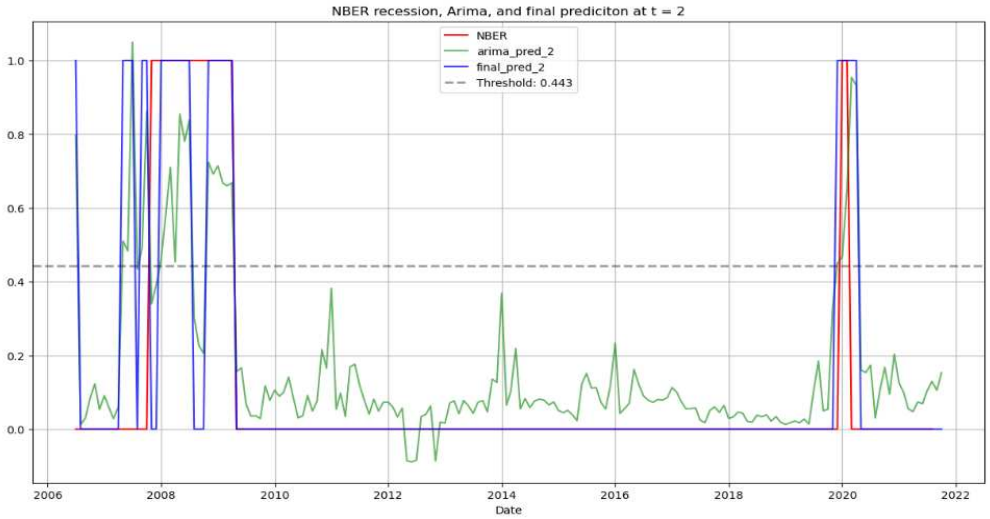


Figure 21: OOS version 2 predictions at t + 3

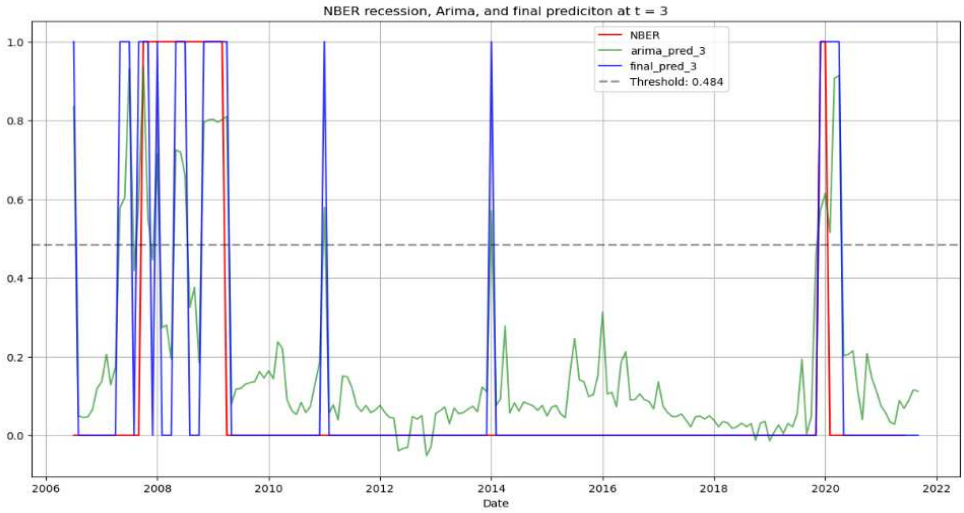


Figure 22: OOS version 2 predictions at t + 4

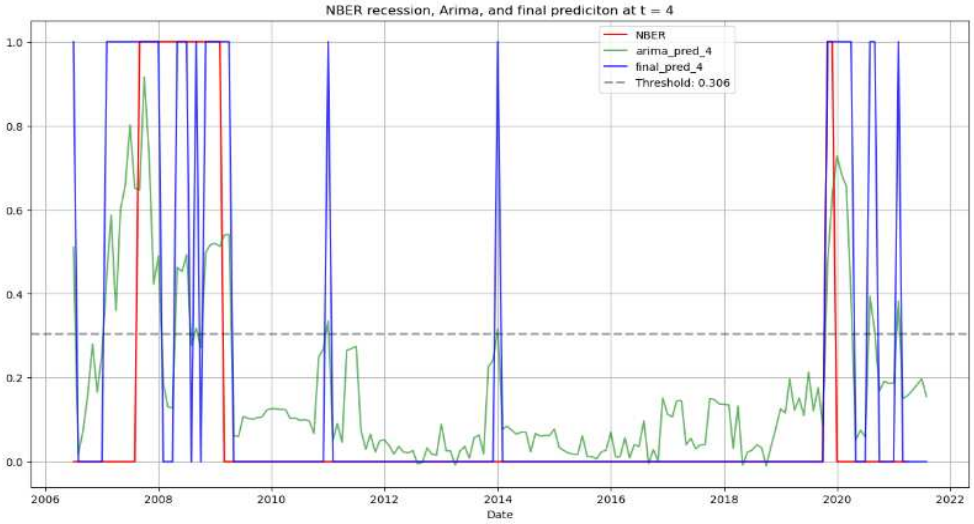


Figure 23: OOS version 2 predictions at t + 5

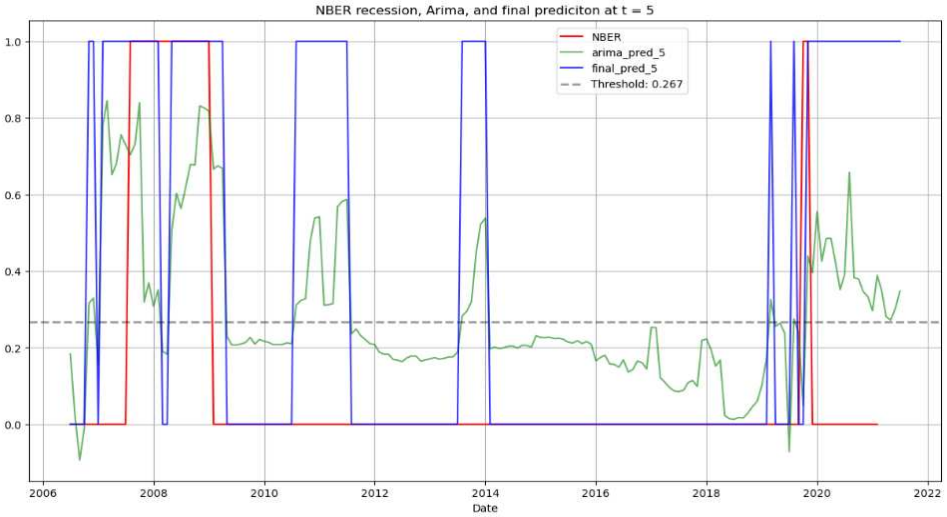


Figure 24: OOS version 2 predictions at t + 6

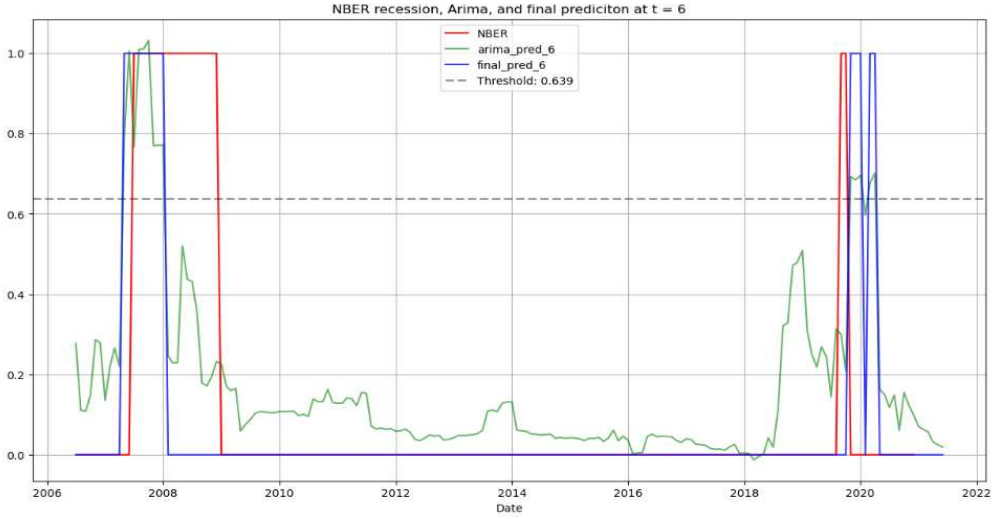


Figure 25: OOS version 2 predictions at t + 7

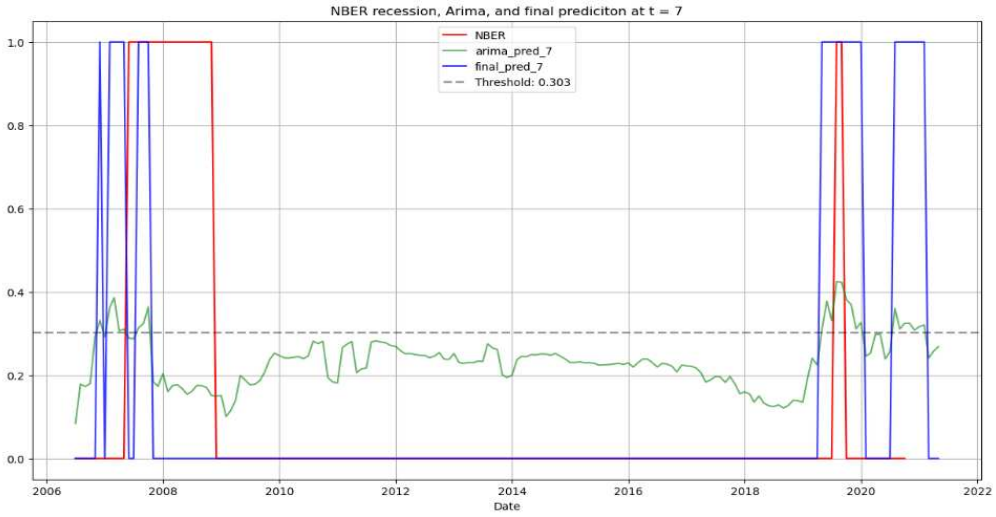


Figure 26: OOS version 2 predictions at t + 8

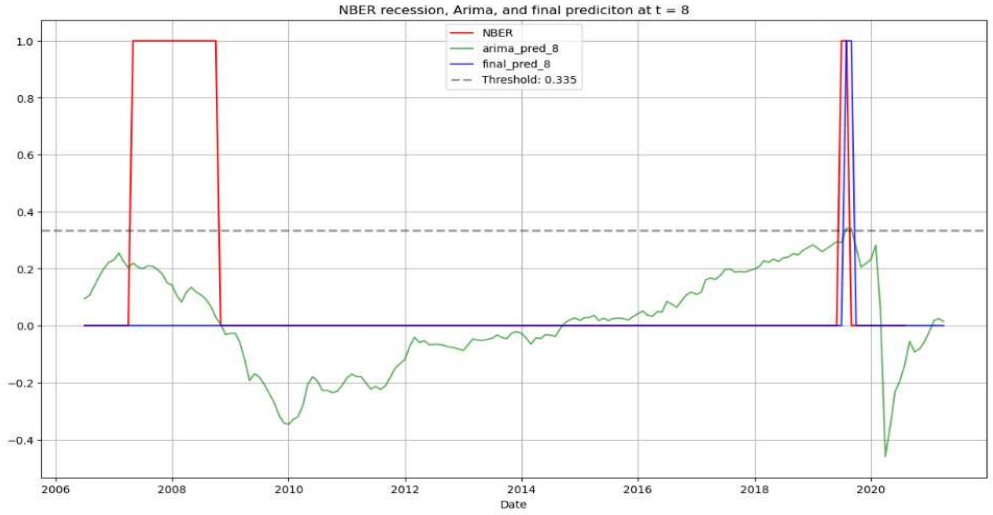


Figure 27: OOS version 2 predictions at t + 9

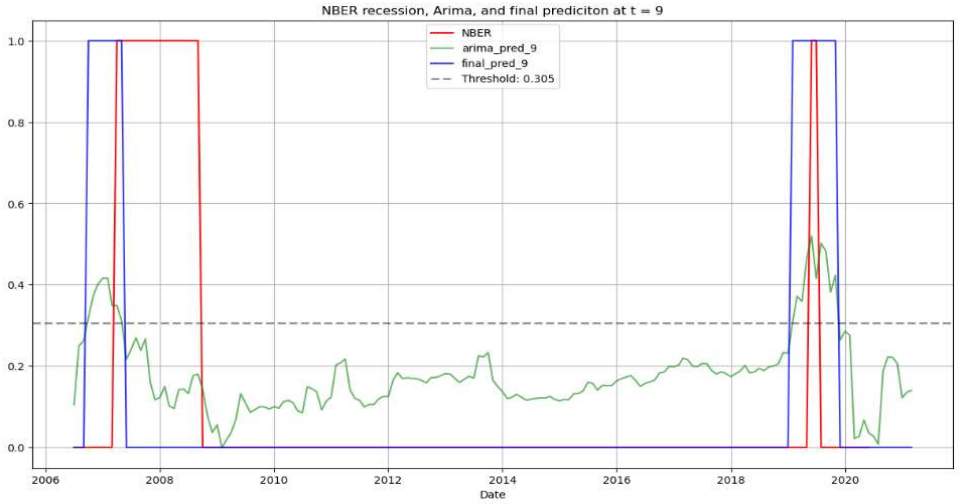


Figure 28: OOS version 2 predictions at t + 10

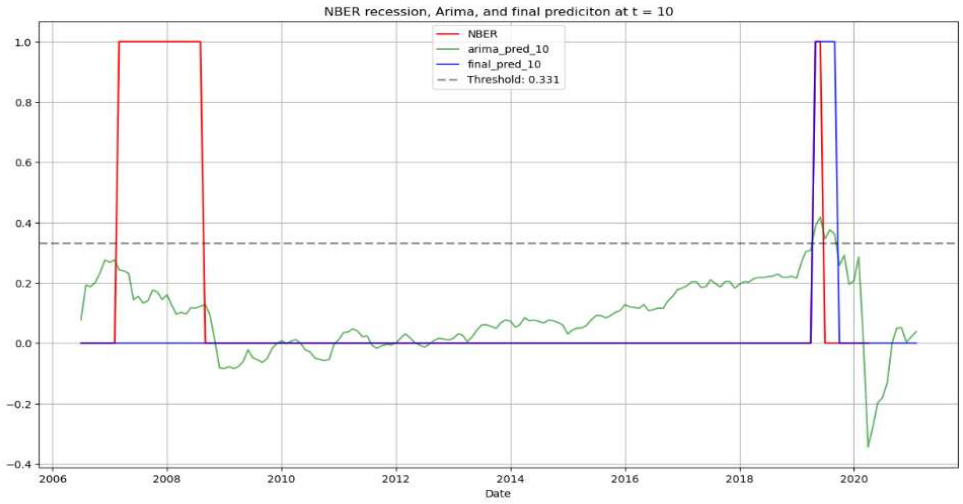


Figure 29: OOS version 2 predictions at t + 11

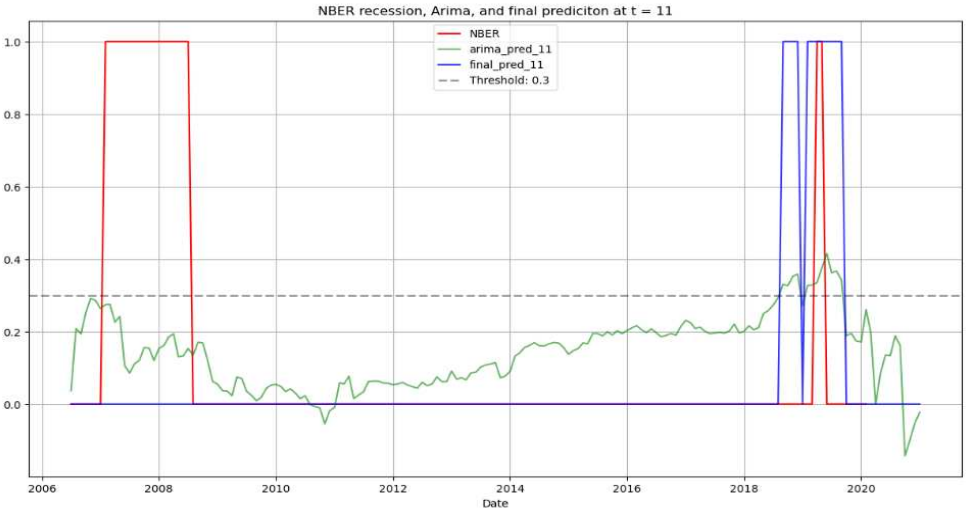
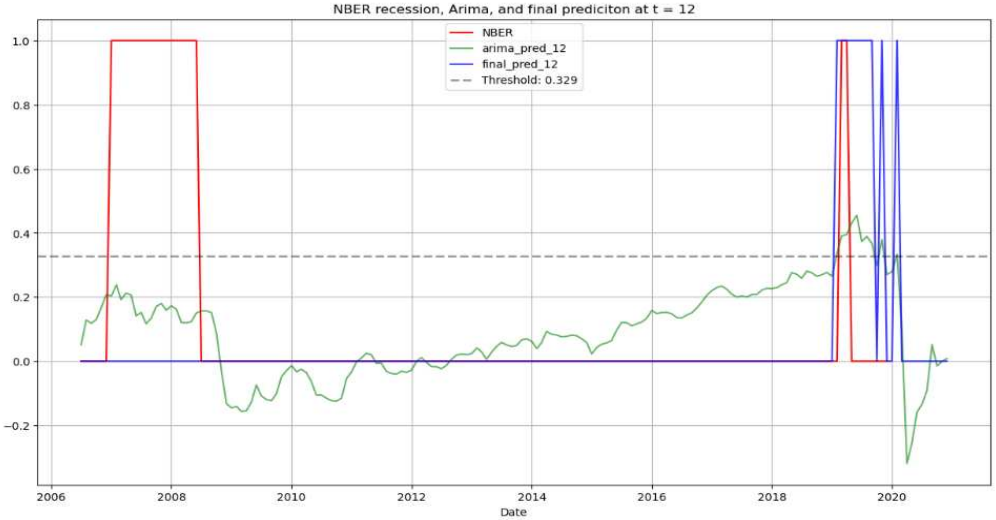


Figure 30: OOS version 2 predictions at t + 12



5.10 Result Discussion

By looking at the prediction, the most remarkable difference between the two methods is that the first version has a higher recall score up to $t + 4$, after which the opposite is true. This does not mean the second version is better for the long-run predictions. In fact, the F1 score of both versions decreases drastically after $t + 4$.

Interestingly, although the average AR(1) coefficients were quite similar, 0.623 in the first case compared to 0.66 in the second, the *pred* variable showed substantial differences in both its average and standard deviation. Specifically, with the first approach, the average and standard deviation were 0.751 and 0.159, respectively, while in the second case, they were 0.462 and 0.256. This suggested that the *ARIMA* model found the forecast of the regression algorithms more reliable when they were trained on the original dataset of 123 predictors. Nevertheless, the *pred* variable of the second approach could have been penalised by the larger number of exogenous variables input into the *ARIMA*.

Table 5: Version 1 OOS performance metrics at different forecast time points (recession points in the period: 20)

Metrics	t+0	t+1	t+2	t+3	t+4	t+5	t+6	t+7	t+8	t+9	t+10	t+11	t+12
Accuracy	0.925	0.946	0.870	0.847	0.857	0.818	0.844	0.866	0.882	0.842	0.852	0.886	0.845
Precision	0.594	0.778	0.447	0.382	0.425	0.276	0.278	0.000	0.000	0.250	0.313	0.000	0.000
Recall	0.950	0.700	0.850	0.650	0.850	0.400	0.250	0.000	0.000	0.200	0.250	0.000	0.000
F1 score	0.731	0.737	0.586	0.482	0.567	0.327	0.263	0.000	0.000	0.222	0.278	0.000	0.000
false positive	13	4	21	21	23	21	13	4	1	12	11	0	7
false negative	1	6	3	7	3	12	15	20	20	16	15	20	20

Table 6: Version 1 OOS forecast models variables' coefficients

Variables	t+0	t+1	t+2	t+3	t+4	t+5	t+6	t+7	t+8	t+9	t+10	t+11	t+12
pred	0.775	1.003	0.851	0.783	0.662	0.880	0.817	0.507	0.470	0.510	0.894	0.790	0.826
stress	0.042	-0.003	0.001	0.036	0.076	0.004	0.009	0.003	-0.001	0.059	0.003	-0.029	0.036
cod_unp	-0.003	0.000	-	-	-	-	-	-	-	-	-	-	
cod_r_gdp	0.044	0.004	0.113	0.212	0.138	0.001	0.053	0.003	0.003	0.010	-	-	
cod_payroll	0.052	0.013	0.045	-	-	-	-	-	-	-	-	-	
cod_r_income	0.042	0.301	0.139	-0.006	0.007	0.132	-	-	-	-	-	-	
sl_ratio_0_-6	-	-	-0.016	-0.015	-0.009	-0.010	0.000	-	-	-	-	-	
term_spread_0_-6	-	-	-	0.028	0.022	0.023	0.020	0.001	0.005	-0.008	0.001	0.011	0.005
sl_ratio_0	-	-	-	-	-	-	-	-0.002	-0.020	-0.011	-0.028	-0.042	-0.059
term_spread_0	-	-	-	-	-	-	0.0227	0.001	0.014	0.045	0.030	0.016	0.021
unp	-	-	-	-	-	-	-	-	-	-	-0.008	0.0201	0.0137
ar. L1	0.487	0.387	0.470	0.801	0.620	0.771	0.913	-0.527	0.849	0.823	0.819	0.903	0.787
ma. L1	-0.258	-0.043	0.040	-0.177	0.108	-0.238	-0.534	0.673	-0.292	-0.018	-0.188	-0.169	0.003

Heatmap legend
p-val ≤ 0.000
0.000 < p-val < 0.050
0.050 ≤ p-val < 0.100
p-val ≥ 0.100

Table 7: Version 2 OOS performance metrics at different forecast time points (recession points in the period: 20)

Metrics	t+0	t+1	t+2	t+3	t+4	t+5	t+6	t+7	t+8	t+9	t+10	t+11	t+12
Accuracy	0.941	0.941	0.924	0.902	0.868	0.702	0.889	0.810	0.888	0.831	0.881	0.840	0.851
Precision	0.846	0.846	0.625	0.542	0.441	0.250	0.500	0.208	0.500	0.222	0.400	0.167	0.200
Recall	0.550	0.550	0.750	0.650	0.750	0.850	0.350	0.250	0.050	0.200	0.100	0.100	0.100
F1 score	0.667	0.667	0.682	0.591	0.556	0.386	0.412	0.227	0.091	0.211	0.160	0.125	0.133
false positive	2	2	9	11	19	51	7	19	1	14	3	10	8
false negative	9	9	5	7	5	3	13	15	19	16	18	18	18

Table 8: Version 2 OOS forecast models variables' coefficients

Variables	t+0	t+1	t+2	t+3	t+4	t+5	t+6	t+7	t+8	t+9	t+10	t+11	t+12
cod_unp	-0.002	-0.005	-0.006	-0.006	-	-	-	-	-	-	-	-	-
cod_r_gdp	0.016	0.028	0.109	0.116	0.074	0.124	0.062	0.064	0.008	0.031	-	0.024	-
cod_payroll	0.020	0.019	0.041	0.032	-	-	-	-	-	-	-	-	-
cod_r_income	0.026	0.029	0.091	0.075	0.057	0.005	0.011	-	-	-	-	-	-
sl_ratio_0_-6	-0.002	-0.003	-0.008	-0.008	-0.022	-0.002	0.019	-0.025	-0.035	-0.074	-0.039	-0.013	-0.041
term_spread_0_-6	0.006	0.0061	0.014	0.020	0.025	0.027	0.010	0.014	-0.058	-0.054	-0.022	-0.001	-0.044
sl_ratio_0	-	-	-	-	0.021	-0.044	-0.024	-0.032	0.050	0.044	0.043	-0.023	0.049
term_spread_0	-	-	-	-	0.030	-	0.015	-0.008	0.034	0.015	0.053	0.032	0.033
unp	-	-	-	-	-	-	-	-	-0.074	-	-0.046	-0.023	-0.054
DJIA_btm	-0.002	-0.004	-0.007	-0.011	-0.015	-	-	-	-	-	-	-	-
r_income_r_gdp	-0.003	0.000	-	-	-	-	-	-	-	-	-	-	-
infl_-6	-0.002	0.000	-0.007	-0.011	-0.008	-	-0.007	-	0.098	-	-	-	-0.032
infl	-	-	-0.006	-0.001	-0.008	0.002	-0.017	0.035	-0.021	0.025	0.051	0.049	0.087
cod_ind_pro	0.009	0.007	0.017	0.018	0.015	-	-	-	-	-	-	-	-
bill	-	-	-	-	-	-	-0.005	-	0.024	-	-	0.008	-0.024
unp_-6	-	-	-	-	-	-	0.018	-	-0.010	-	-	-0.048	-0.010
cod_corp_bond_Baa_yield	-	-	-	-	-	-	-	-	-	-	-	-0.004	-
cod_cr_Autos	-	-	-	-	-	-	-	-	-	-	-	-	0.000
pred	0.557	0.571	0.812	0.841	0.809	0.419	0.607	0.292	0.058	0.298	0.219	0.436	0.084
stress	0.025	0.026	0.021	0.009	0.018	0.082	0.143	0.135	-0.052	0.310	0.141	0.079	0.165
ar. L1	0.274	0.267	0.333	0.539	0.821	0.780	0.272	0.890	0.894	0.875	0.888	0.862	0.891
ma. L1	-0.109	-0.095	0.095	-0.057	-0.369	-0.120	0.154	0.088	0.145	0.151	0.102	0.074	0.134

Heatmap legend
p-val ≤ 0.000
0.000 < p-val < 0.050
0.050 ≤ p-val < 0.100
p-val ≥ 0.100

6. RECESSION FORECAST

6.1 OOS Trading Strategy

Following Gómez-Cram (2021), the nowcast prediction was used to build a trading strategy that invested in the S&P 500 during economic expansions and shifted to 3-month treasury bills during recessions. The trading strategy was applied to the OOS set and employed in both versions of the non-linear regression and *ARIMA* models, and then the trading strategy was compared to buying and holding the S&P 500 during the period (in my calculations, we included transaction fees of 0.2% and assumed the risk-free rate as the average of the 3-month treasury bill throughout the OOS set, which annualised was of 0.93%).

The cumulated return of the buy-and-hold strategy was 259.60% against the 433.67% and 671.03% of the cumulated return of the first and second versions of the trading strategy, respectively. In fact, even though the first version better aligned with the NBER, the second version tracked more closely the drawdowns of the S&P 500, enabling the strategy to capitalize on market dips during recessions. The annualised Capital Asset Pricing Model (CAPM) alpha was 6.07% and 7.14% for the first and second strategy, respectively, confirming that the second strategy provided a higher return.

Consequently, to assess the two trading results with respect to market risk, the Sharpe and Sortino ratios were calculated. The first method yielded Sharpe and Sortino ratios of 1.011

and 1.346, respectively, while the second approach produced ratios of 1.110 and 1.625, proving the superiority of the second methodology for trading purposes.

Additionally, the strategy was implemented using both methodologies up to *t + 4*. The best results were obtained in both cases at *t + 1* in terms of annualised mean return, Sharpe and Sortino ratios, and annualised CAPM alpha. Furthermore, the second approach outperformed the first up to t + 3, as evidenced by its higher Sharpe and Sortino ratios (see Table 9 and Table 10)

Interestingly, while the first approach better predicted NBER recessions, the second approach was more effective in forecasting market sell-offs, resulting in improved positioning for trading performance. Even more notable is that, in both strategies, S&P 500-related variables were excluded from the final ARIMA models (see Table 6 and Table 8). Additionally, in the second approach, stock market-related predictors were not even used as features in the non-linear regression algorithms to construct the *pred* and *stress* variables.

The graphs of the cumulative returns and drawdowns at *t + 0* and the summary tables of the performance metrics for the trading strategy implemented by the two approaches follow.

Figure 31: Version 1 cumulative returns of the trading strategy

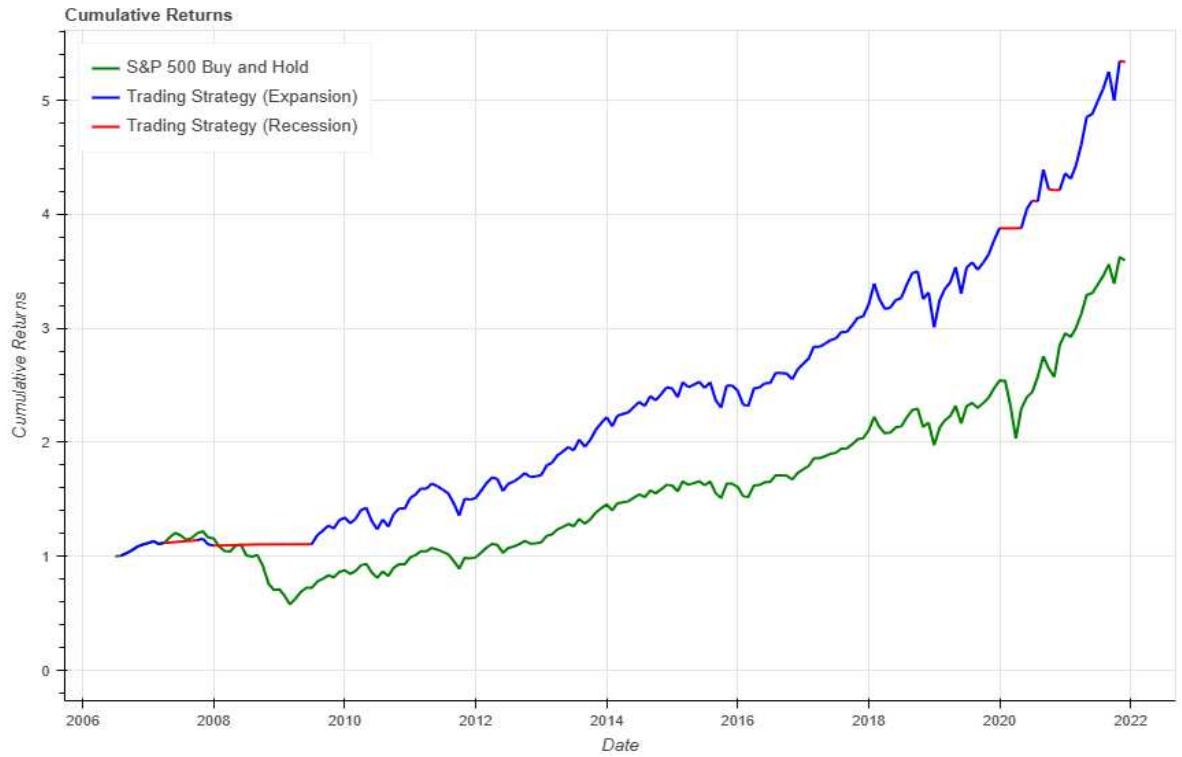


Figure 32: Version 2 cumulative returns of the trading strategy

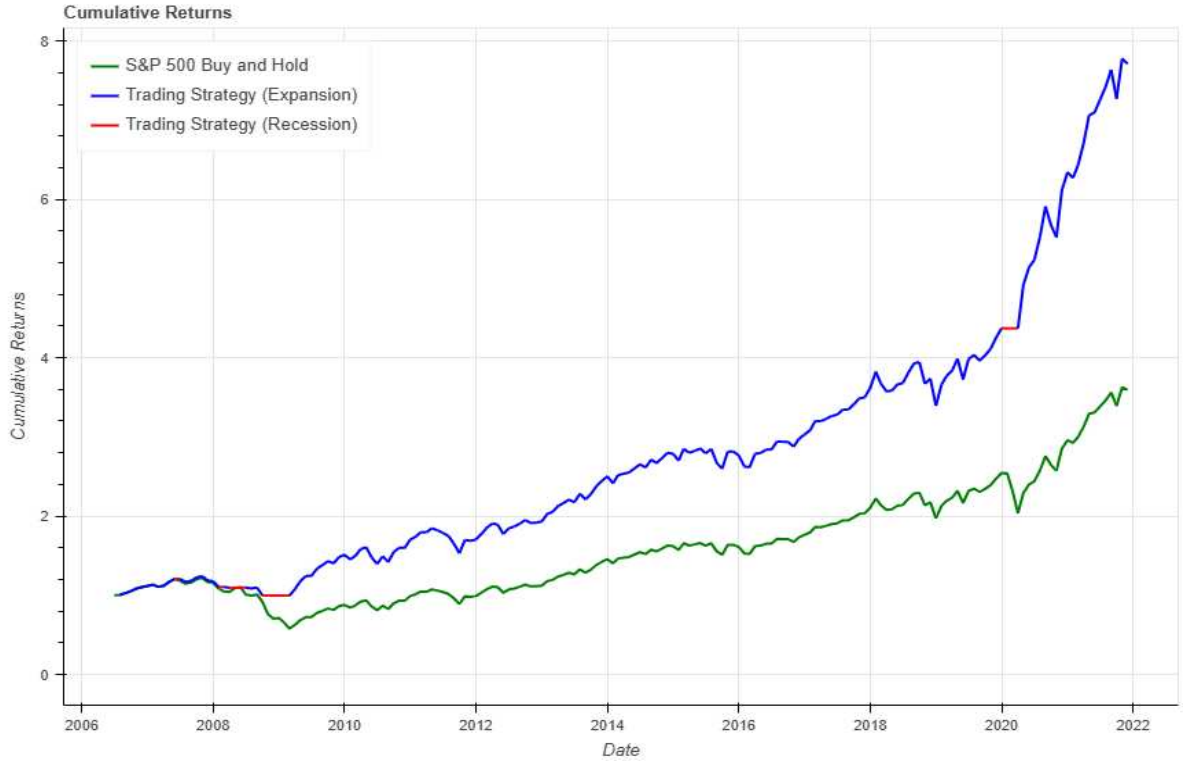


Figure 33: Version 1 drawdowns of the trading strategy

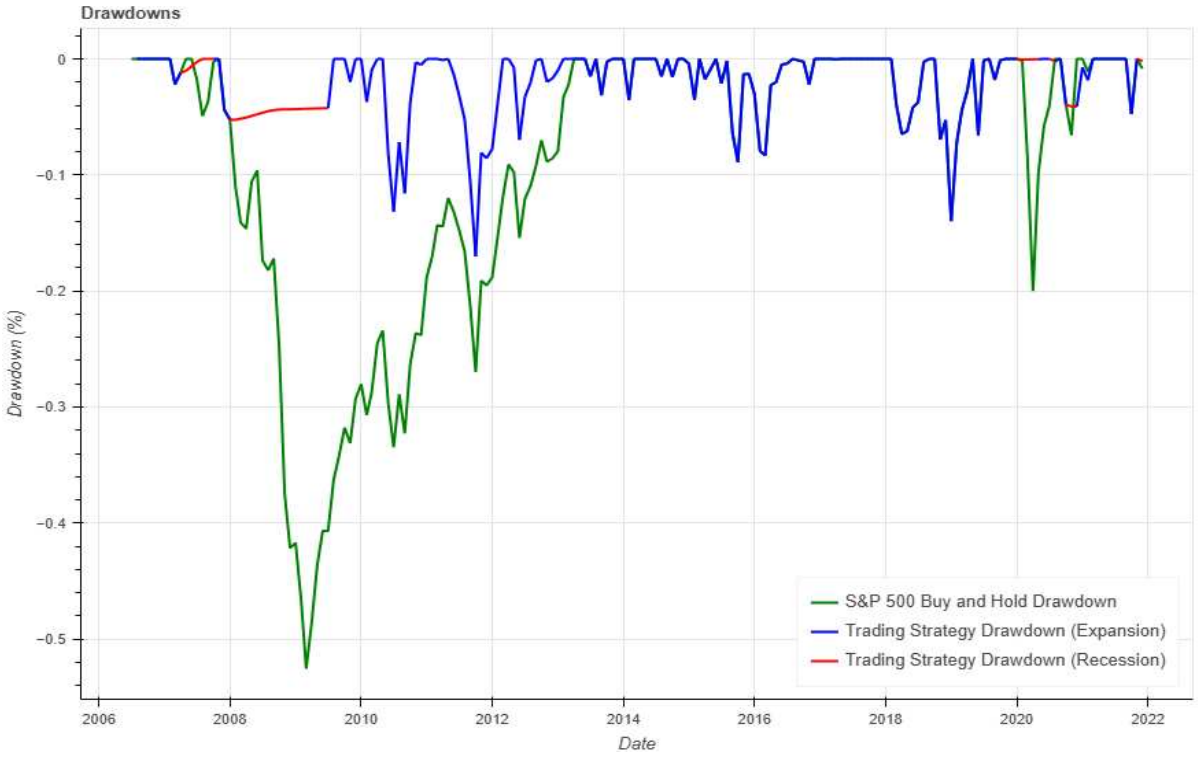


Figure 34: Version 2 drawdowns of the trading strategy

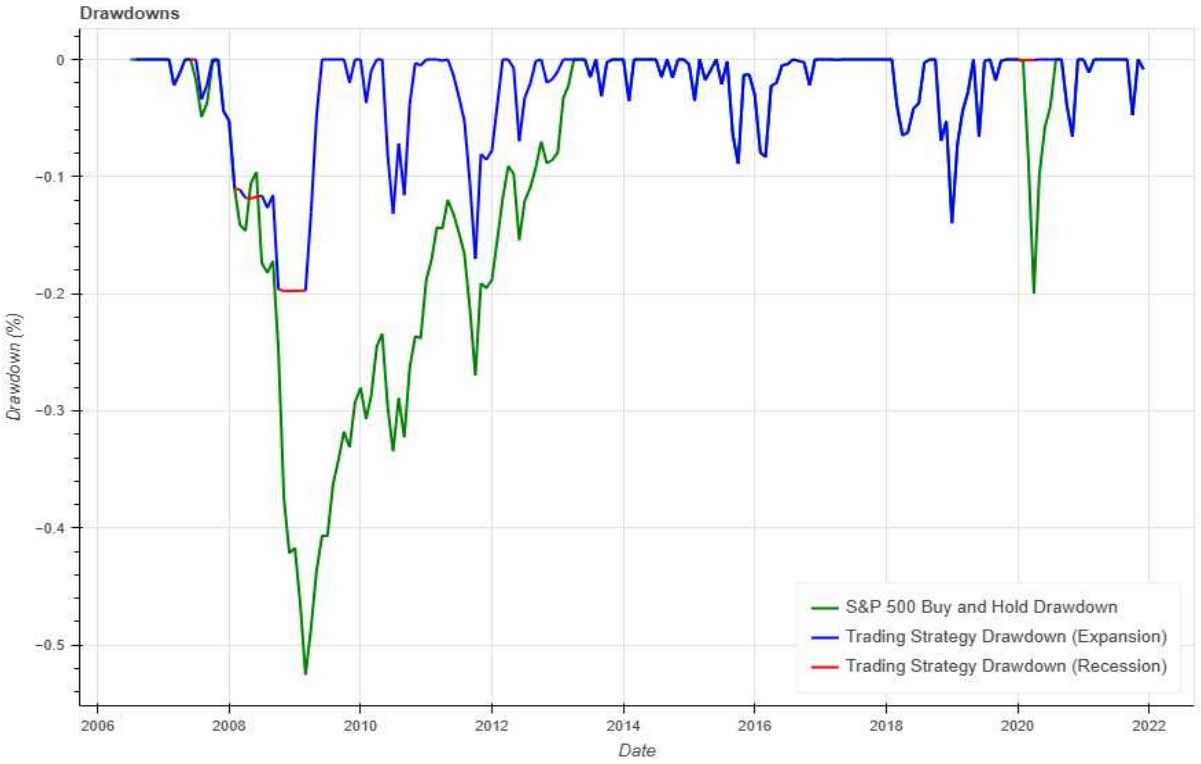


Table 9: Version 1 performance metrics for the trading tratégy

	S&P 500	*t + 0*	*t + 1*	*t + 2*	*t + 3*	*t + 4*
Annualised mean return	0.099	0.122	0.150	0.130	0.119	0.123
Annualised volatility	0.151	0.111	0.124	0.120	0.124	0.119
Sharpe ratio	0.593	1.011	1.143	1.010	0.890	0.953
Sortino ratio	0.745	1.346	1.618	1.457	1.242	1.326
Maximum drawdown	-0.526	-0.170	-0.170	-0.163	-0.237	-0.148
Annualised CAPM alpha	0	0.061	0.075	0.063	0.047	0.055
CAPM beta	1	0.548	0.695	0.639	0.677	0.622

Table 10: Version 2 performance metrics for the trading strategy

	S&P 500	*t + 0*	*t + 1*	*t + 2*	*t + 3*	*t + 4*
Annualised mean return	0.099	0.151	0.152	0.135	0.123	0.118
Annualised volatility	0.151	0.127	0.128	0.124	0.124	0.122
Sharpe ratio	0.593	1.110	1.119	1.017	0.915	0.890
Sortino ratio	0.745	1.625	1.634	1.434	1.300	1.213
Maximum drawdown	-0.526	-0.198	-0.198	-0.170	-0.243	-0.195
Annualised CAPM alpha	0	0.071	0.072	0.063	0.049	0.048
CAPM beta	1	0.773	0.733	0.685	0.687	0.652

6.2 Actual State Of The Economy

Finally, both versions of the combination of non-linear regression and *ARIMA* models were used to assess the current state of the economy for the following 12 months, up to January 1st 2025. The training set consisted of 906 monthly observations, spanning from July 1st, 1948,

to December 1st, 2023, while the test set included only one observation: January 1st, 2024. This was the last date for which all 123 predictors were available at time $t + 0$. Forecasts for the next 12 months were generated by running the models to predict all future time points without changing the training or test sets. Due to the lagging of the target variable to predict future values, the forecast made on January 1st was based on the model's prediction for $t + lag\ period$, where t is the last available date at which all data are available, and the lag corresponds to the number of months by which the dependent variable was shifted. For example, if the target were lagged by 1 month, the prediction at $t + 1$ would be for 1 month ahead.

The very same methodology used for the OOS prediction was employed with this 'larger dataset, and both versions of the combination of non-linear regression and *ARIMA* models were applied. The two methods did not forecast recession points, as the final *ARIMA* prediction was always below the cutoff value.

7. CONCLUSION

7.1 Discussion

This study aimed to address the critical problem of accurately forecasting economic recessions, which, if anticipated correctly, can provide significant benefits for policymakers, markets, and businesses. Reliable predictions can help mitigate declines in GDP, asset prices, wealth, and unemployment spikes, which are usually typical of economic downturns. Timely forecasts could allow policymakers to take preemptive actions, such as adjusting monetary policy and implementing fiscal stimulus, to better manage recessions' negative impacts.

To achieve the objectives of the research, we developed two distinct forecasting models by combining ML non-linear regression algorithms with *ARIMA* models. We assessed NBER recession forecasts over a time window from $t + 0$ to $t + 12$. Among the non-linear regression models, boosting algorithms, such as *CatBoost*, *AdaBoost*, and *XGBoost*, performed best up to approximately $t + 8$, after which bagging algorithms, notably *ExtraTrees* and *RandomForest*, yielded the best results. However, predictions remained reliable only up to $t + 4$, after which the forecasting power significantly diminished. Notably, these predictions proved sensitive to economic stress periods, signalling recessions during the 2011 debt ceiling turmoil and the 2014 oil price slowdown, though these instances were ultimately classified as false recessions, between $t + 2$ and $t + 5$. Nevertheless, based on the accuracy scores of both

forecasting versions, our models slightly underperformed the forecasts of Berge (2015) and Vrontos, Galakis, and Vrontos (2021).

The most relevant variables throughout the different time points were the initial forecast made by the non-linear regression algorithms, the AR(1), the COD of real GDP, real income, payroll, and industrial production. Nevertheless, assessing the impact of macroeconomic variables on model performance proved to be more complex. In fact, however, the final *ARIMA* model provided valuable information concerning the sign, magnitude, and statistical significance of the exogenous variables, the variable selection process was relatively simplistic as the variables were selected based on their correlation with the target at each forecasting time point. Therefore, while this approach offered a straightforward understanding of the relevance of the chosen exogenous variables, it did not allow for a robust ranking of their importance across the 123 initial features considered.

Lastly, the financial benefits of predicting the NBER recession were assessed by using the two forecasting methods to implement a trading strategy, which consisted of investing in the S&P 500 during expansions and switching to t-bills during recessions. The comparisons of these two versions showed that one could nowcast the NBER at $t + 0$, though such a prediction was not aligned with stock market dynamics, while the second version allowed one to identify the S&P 500 trough and invest immediately after that point, even though still in an NBER recession, hence, yielding the best financial performance.

Finally, the fact that the prediction performs well in the short run implies that it has to be applied to real-time data. However, such an absence limited our ability to assess the present state of the economy as the models' forecasts for $t + 7$ approximately coincided with the current state of the economy at $t + 0$. Even though this problem makes the study of little value for many businesses and common investors, the research can still be useful for policymakers and practitioners who work with real-time measurements.

7.2 Limitations And Future Work

One of the primary limitations was the insufficient availability of historical data, which was the main reason why some variables were discarded, even if considered by the previous literature relevant to the NBER forecast, such as the real manufacturing and trade industries sales, money aggregates M1, M2, and M3, public debt, public debt as a percentage of GDP,

bank credit and bank credit as a percentage of GDP. In addition, due to a lack of computational power, NN approaches were too time-consuming, so they had to be disregarded.

Moreover, because of the lack of real-time data, the forecast is of no value for all those investors who do not have access to real-time measurements as the data retrieved from the databases mentioned in section 3 are published with more than four months of delay, going beyond the four months within which the forecast proved to be reliable. However, future studies can consider a reduced version of the original dataset, where all predictors not issued within four months are excluded.

Additionally, even if the study assessed the most important predictors, this analysis was only possible on the *ARIMA* model's output. In fact, as explained by Chakraborty and Joseph (2017), ML behaves like a black box, and it is impossible to dissect the web of relationships between all the 123 predictors used as independent variables for the classification and regression algorithms.

Furthermore, in the study, all recessions are treated equally, and 2008 GFC constitutes 90% of OOS's recession points. But as explained by Dalio (2022), the 2008 crisis was a deleveraging, not technically just a recession. Opposite to a recession, which is caused by central banks tightening monetary policy, usually to fight inflation, a deleveraging is a downward cause-effect spiral where assets lose their value, credit contracts, the economic output falls, cash flows decrease, and market players are forced to sell their assets to meet their obligations creating a negative self-reinforcing cycle. Such deleveragings happen every 75 years, give or take 25; examples are Japan in the 1990s and the US in the 1930s (Dalio, 2022). Consequently, as the IS set covers nine recessions, it is plausible that the methods elaborated in this study perform more poorly during deleveragings and forecast better normal recessions (in fact, the Covid-19 recession was better forecasted than the GFC). Eventually, further improvements could be made by training the algorithms on the US 2008 GFC and testing it on Japan's 1990s deleveraging.

Likewise, while developed countries experience domestic currency debt crises, EMEs tend to have balance of payment/currency problems, which may lead to hyperinflation dynamics. Hence, further studies may explore EMEs recessions and investigate the most important variables in such cases; in fact, while in our research, inflation was not revealed to

be very important (see Table 8), it is legitimate to expect it will have a more relevant impact in EMEs.

References

Andrea Giusto & Jeremy Piger, 2013. "Nowcasting US Business Cycle Turning Points with Vector Quantization," Working Papers, Dalhousie University, Department of Economics.

Berge, T.J., 2015. Predicting recessions with leading indicators: Model averaging and selection over the business cycle. *Journal of Forecasting*, 34(6), pp.455–471. doi:10.1002/for.2345

Borio, C.E.V., Drehmann, M. and Xia, F.D. (2019). *Predicting Recessions: Financial Cycle versus Term Spread*. [online] papers.ssrn.com. Available at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3468418.

Bruneau, G., Christensen, I. and Meh, C. (2018). Housing market dynamics and macroprudential policies. *Canadian Journal of Economics/Revue canadienne d'économique*, [online] 51(3), pp.864–900. doi:https://doi.org/10.1111/caje.12346.

Chakraborty, C. and Joseph, A. (2017). Machine Learning at Central Banks. *SSRN Electronic Journal*. doi:https://doi.org/10.2139/ssrn.3031796.

Dalio, R. (2022). *Principles for Navigating Big Debt Crises*. Simon and Schuster.

Estrella, A. and Mishkin, F.S. (1998). Predicting US Recessions: Financial Variables as Leading Indicators. *Review of Economics and Statistics*, 80(1), pp.45–61. doi:https://doi.org/10.1162/003465398557320.

Gilchrist, S. and Zakrajšek, E. (2012). Credit Spreads and Business Cycle Fluctuations. *American Economic Review*, [online] 102(4), pp.1692–1720. doi:https://doi.org/10.1257/aer.102.4.1692.

GÓMEZ-CRAM, R. (2021). Late to Recessions: Stocks and the Business Cycle. *The Journal of Finance*, 77(2), pp.923–966. doi:https://doi.org/10.1111/jofi.13100

Liu, W. and Moench, E. (2016). What predicts US recessions? *International Journal of Forecasting*, 32(4), pp.1138–1150. doi:https://doi.org/10.1016/j.ijforecast.2016.02.007.

NAES, R., SKJELTORP, J.A. and ØDEGAARD, B.A. (2011). Stock Market Liquidity and the Business Cycle. *The Journal of Finance*, 66(1), pp.139–176. doi:https://doi.org/10.1111/j.1540-6261.2010.01628.x.

National Bureau of Economic Research (2022). Business Cycle Dating. [online] National Bureau of Economic Research. Available at: https://www.nber.org/research/business-cycle-dating.

Oluwatamilore Orojo, Tepper, J., McGinnity, T.M. and Mahmud, M. (2023). The Multi-Recurrent Neural Network for State-Of-The-Art Time-Series Processing. Procedia Computer Science, 222, pp.488–498. doi:https://doi.org/10.1016/j.procs.2023.08.187.

Ponka, H. (2016). The Role of Credit in Predicting US Recessions. *Journal of Forecasting*, 36(5), pp.469–482. doi:https://doi.org/10.1002/for.2448.

Qi, M. (2001). Predicting US recessions with leading indicators via neural network models. *International Journal of Forecasting*, 17(3), pp.383–401. doi:https://doi.org/10.1016/s0169-2070(01)00092-9.

Sahm, C., 2019. Direct stimulus payments to individuals. *Recession ready: Fiscal policies to stabilize the American economy*, pp.67-92.

Stlouisfed.org. (2024). *NBER based Recession Indicators for the United States from the Peak through the Trough*. [online] Available at: https://fred.stlouisfed.org/series/USRECDM [Accessed August 9th 2024].

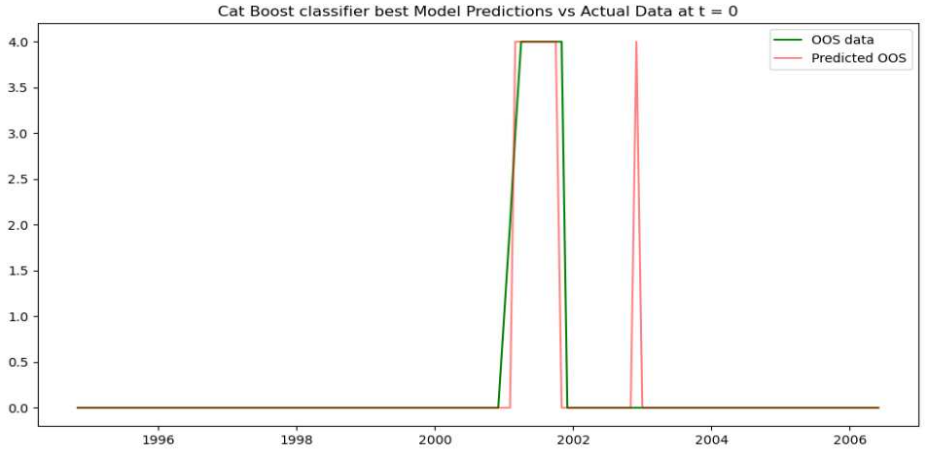
Vrontos, S.D., Galakis, J. and Vrontos, I.D., 2021. Modeling and predicting US recessions using machine learning techniques. *International Journal of Forecasting*, 37(2), pp.647–671. doi:10.1016/j.ijforecast.2020.08.005.

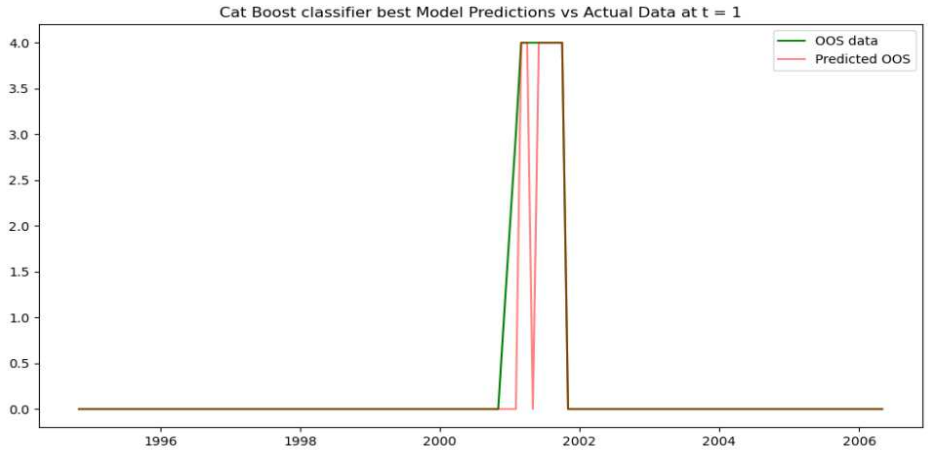
Welch, I. and Goyal, A. (2008). A Comprehensive Look at The Empirical Performance of Equity Premium Prediction. Review of Financial Studies, 21(4), pp.1455–1508. doi:https://doi.org/10.1093/rfs/hhm014.

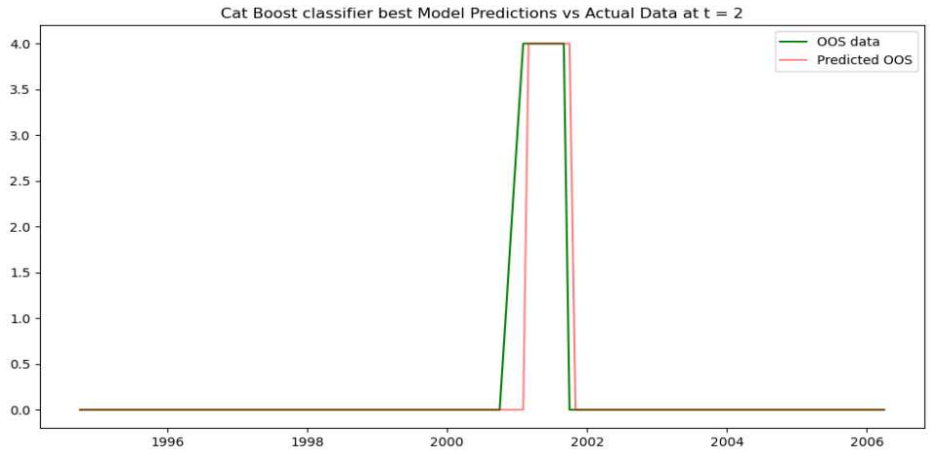
Wheelock, D.C. (2020). Comparing the COVID-19 Recession with the Great Depression. *Economic Synopses*, 2020(39). doi:https://doi.org/10.20955/es.2020.39.

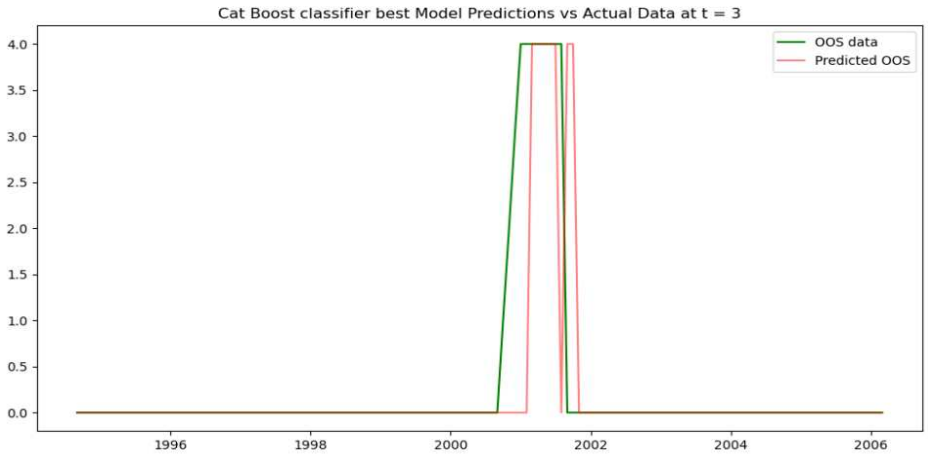
APPENDIX 1

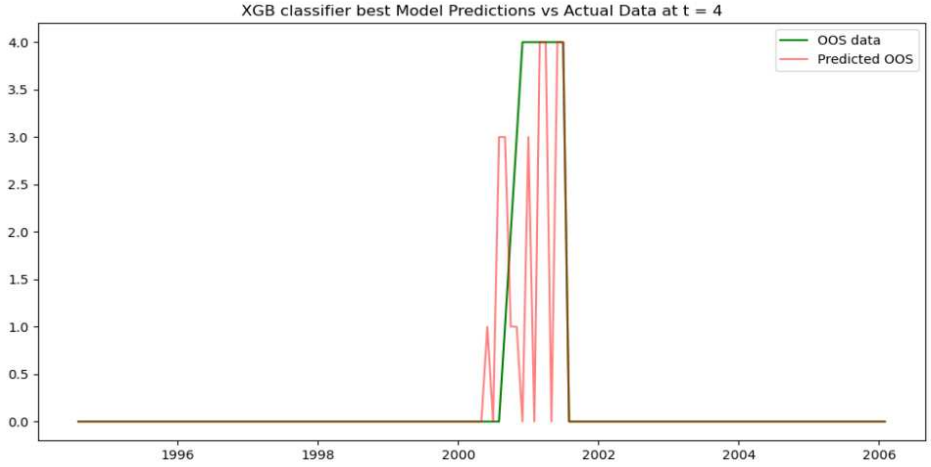
Appendix 1 summarises the output of the multi-nomial classification from $t + 0$ to $t + 12$ on the validation set.

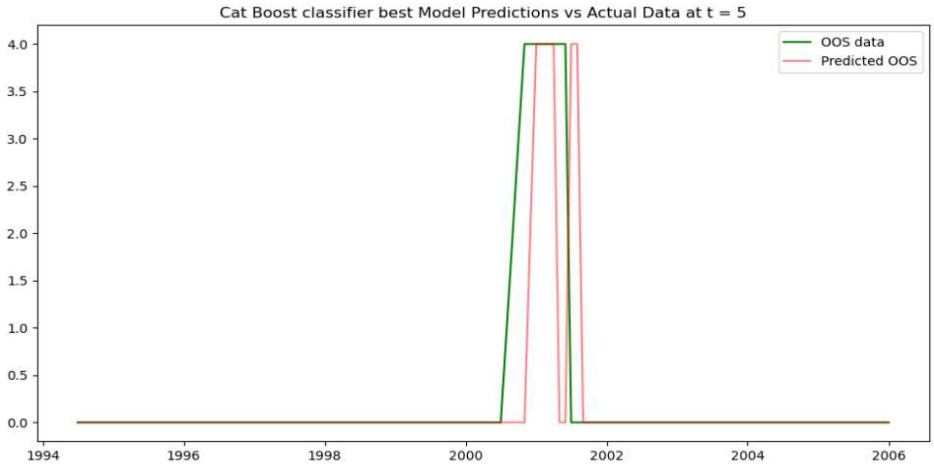


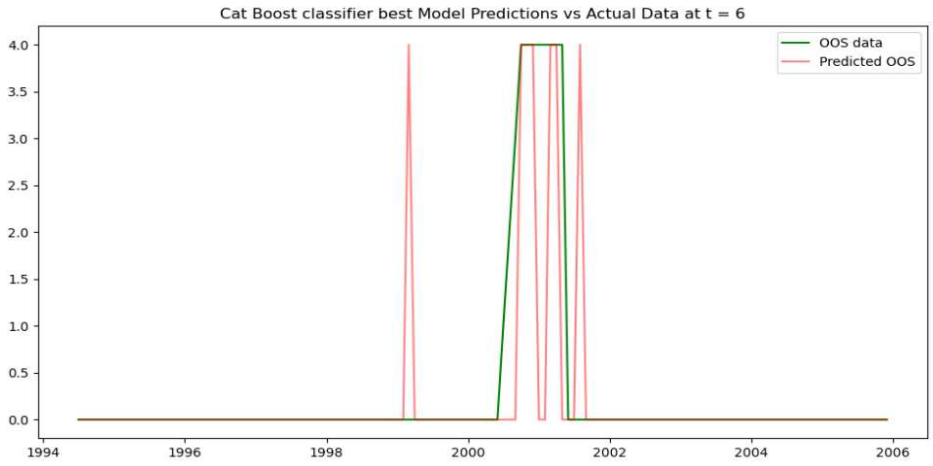


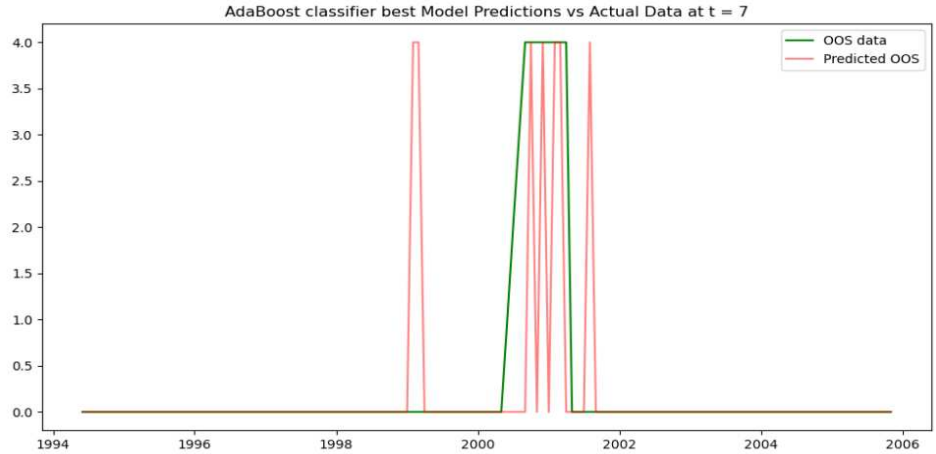


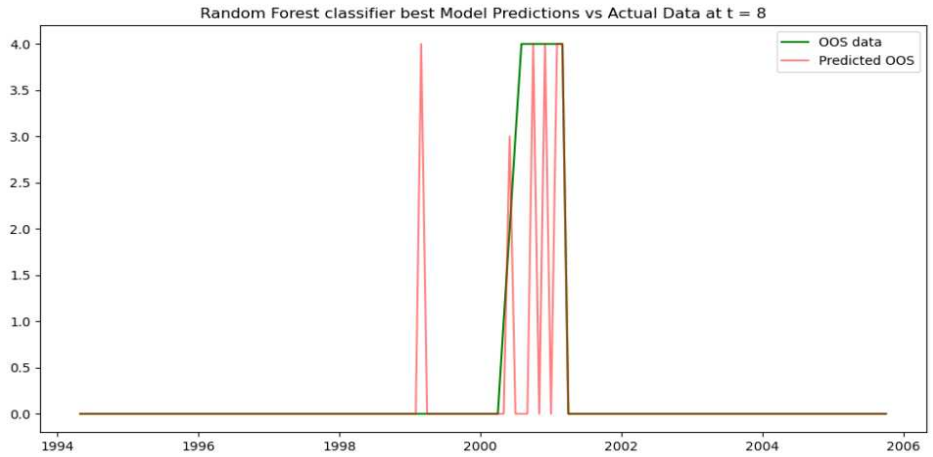

Cat Boost classifier best Model Predictions vs Actual Data at t = 3
OOS data
Predicted OOS
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
1996
1998
2000
2002
2004
2006

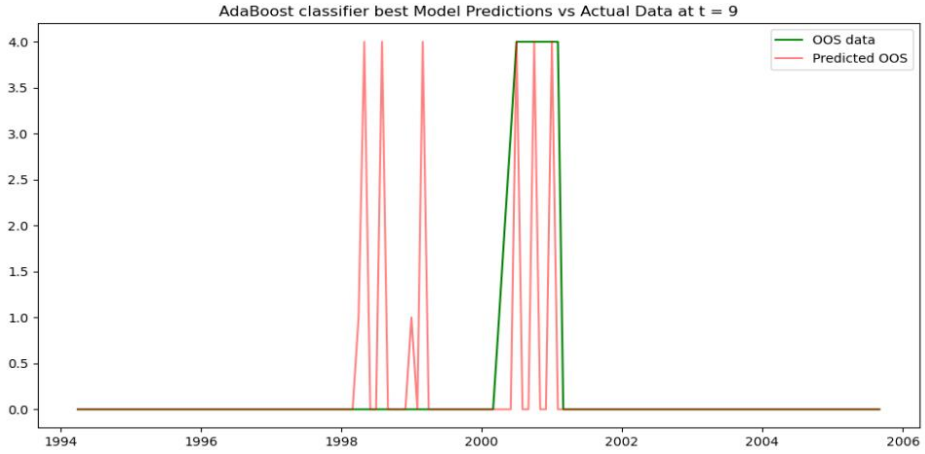

XGB classifier best Model Predictions vs Actual Data at t = 4
OOS data
Predicted OOS
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
1996
1998
2000
2002
2004
2006

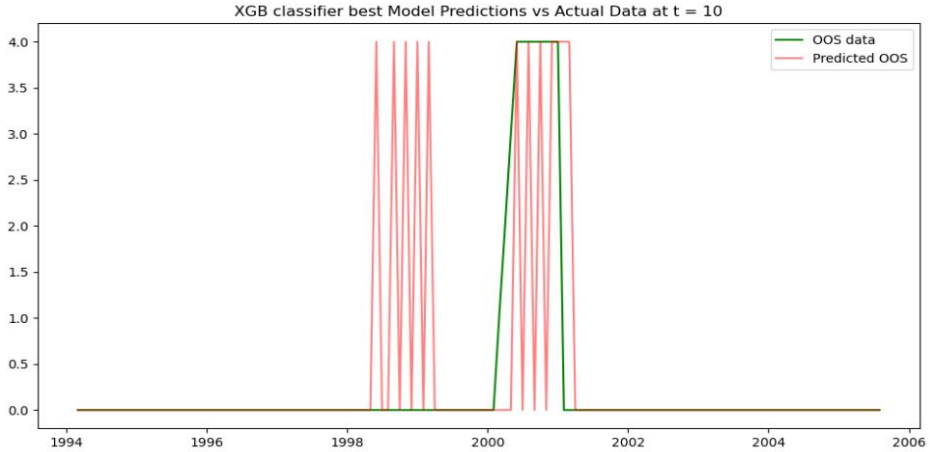

Cat Boost classifier best Model Predictions vs Actual Data at t = 5
OOS data
Predicted OOS
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
1994
1996
1998
2000
2002
2004
2006

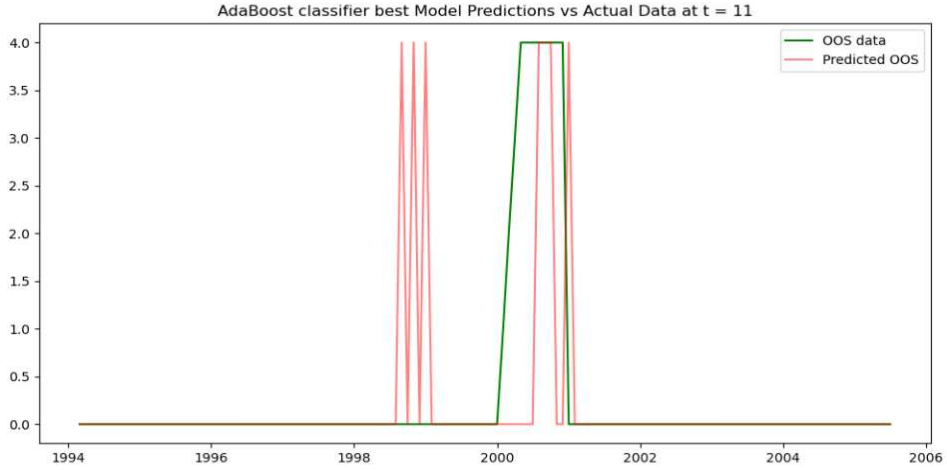

Cat Boost classifier best Model Predictions vs Actual Data at t = 6
OOS data
Predicted OOS
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
1994
1996
1998
2000
2002
2004
2006

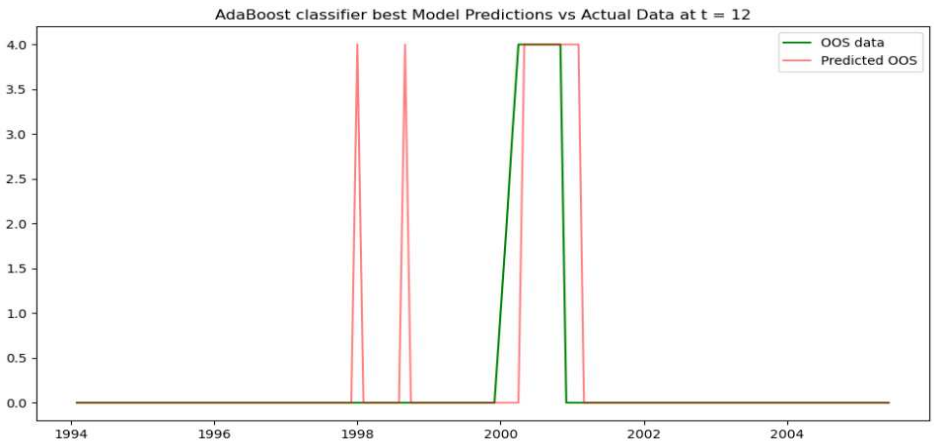

AdaBoost classifier best Model Predictions vs Actual Data at t = 7
OOS data
Predicted OOS
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
1994
1996
1998
2000
2002
2004
2006


Random Forest classifier best Model Predictions vs Actual Data at t = 8
OOS data
Predicted OOS
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
1994
1996
1998
2000
2002
2004
2006


AdaBoost classifier best Model Predictions vs Actual Data at t = 9
OOS data
Predicted OOS
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
1994
1996
1998
2000
2002
2004
2006


XGB classifier best Model Predictions vs Actual Data at t = 10
OOS data
Predicted OOS
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
1994
1996
1998
2000
2002
2004
2006


AdaBoost classifier best Model Predictions vs Actual Data at t = 11
OOS data
Predicted OOS
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
1994
1996
1998
2000
2002
2004
2006


AdaBoost classifier best Model Predictions vs Actual Data at t = 12
OOS data
Predicted OOS
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
1994
1996
1998
2000
2002
2004