Modeling and Predicting U.S. Recessions Using Machine Learning Techniques

Spyridon D. Vrontos[a], John Galakis[b] and Ioannis D. Vrontos[c] *

[a]Department of Mathematical Sciences, University of Essex, United Kingdom
[b]Iniohos Advisory Services, Geneva, Switzerland
[c]Department of Statistics, Athens University of Economics and Business, Greece

Abstract

The most represenative machine learning techniques are implemented for modeling and forecasting U.S. economic activity and recessions in particular. An elaborate, comprehensive and comparative framework is employed in order to estimate U.S. recession probabilities. The empirical analysis explores the predictive content of numerous well-followed macroeconomic and financial indicators, but also introduces a set of less studied predictors. The predictive ability of the underlying models is evaluated using a plethora of statistical evaluation metrics. The results strongly support the application of machine learning over more standard econometric techniques in the area of recession prediction. More specifically, the analysis indicates that penalized Logit regression models, k-nn method, and Bayesian generalized linear models largely outperform 'original' Logit/Probit models in the prediction of U.S. recessions, as they achieve higher predictive accuracy across long, medium and short-term forecast horizons.

Keywords: Forecasting; Recession; Binary Probit/Logit; Classification and Regression Trees, Penalized Likelihood Models

*Corresponding author: Ioannis Vrontos, Department of Statistics, Athens University of Economics and Business, 76 Patission Str., 10434, Athens, Greece, Email: vrontos@aueb.gr

1 Introduction

Over time, a great deal of attention has focused on measuring business cycles and identifying their underlying turning points in particular, as changes in aggregate economic activity have repercussions, albeit asymmetric, on households, companies and governments. The development of an effective 'early warning' framework, however, has not been a straightforward task, as business cycle expansions and contractions share a set of 'typical' or 'common' patterns, but, at the same time, exhibit their own unique characteristics. The nature and causes of recessions are not the same; that is, recessions are not triggered by the same shocks. The most recent U.S. recession (2007-2009), one of the most severe downturns since the Second World War, was triggered by a crisis in the housing and financial markets. By contrast, the 2001 recession was primarily caused by the collapse of the so-called tech bubble and the 9/11 attacks. This is the primary reason that recession forecasting is rather challenging.

It is thus not accidental that central banks, government related agencies, and numerous research organizations have been active in producing modeling frameworks and sets of leading indicators to effectively track and predict shifts in aggregate economic activity. Similarly, several financial institutions have developed related forecasting models to assist them in asset allocation decisions, while numerous academic studies have put forward various recession forecasting models.

The earlier research focused on the application of binary dependent variable models, predominantly Probit models, to accurately identify and predict turning points in the underlying economic environment, as they are considered more precise and stable relative to those modeling continuous measures of economic activity. The specific studies employ a variety of covariates to model and forecast U.S. economic conditions that can be defined or categorized as lagging, coincident and leading indicators, respectively. The slope of the U.S. yield curve, the spread between the 10-year Treasury yield and the 3-month Treasury bill or variants, however, has received most attention, as it is regarded as the single best out-of-sample predictor of U.S. recessions.

Since the early 90's static Probit models have been augmented to capture dynamic elements, such as the persistence in the business cycle by including lagged autoregressive parameters of the dependent-recession variable and allowing for time variation in the structure of the model. At the same time, alternative versions of Markov switching models have been developed and extensively used in the literature for determining and forecasting changes in economic regimes.

This paper focuses on the application of a wide variety of machine learning techniques for U.S. recession forecasting. More specifically, the machine learning methods employed include regularization techniques, such as Ridge, Least Absolute Shrinkage and Selection Operator (LASSO), and Elastic Net, Discriminant Analysis classifiers, Bayesian classifiers, and classification and regression trees (CART), such as Bagging, Random Forest, and Boosting. Moreover, Probit and Logit benchmark models that employ the yield curve as a recession predictor, are included in the analysis for comparison reasons with the existing literature.

Prior research has documented that various economic and financial variables contain predictive information regarding U.S. recessions. The pioneering work of Estrella and Mishkin (1998) in particular has put forward numerous predictive variables apart from the yield curve that can be considered as leading recession indicators, including equity prices, money supply measures, and leading indicator indices. Since then, various interest rate, credit market, and different employment market related variables have been considered as potential predictors. Recently, more niche leading indicators have been employed; a primary example is the Leading Credit Index (LCI) that reflects potential structural changes in financial markets (Liu and Moench, 2016), and certain subcomponents like balances in broker-dealer margin accounts that proxy for financial leverage.

Following the trend of new variable detection, a number of less studied predictors are considered in the underlying analysis; more specifically, variables such as the change in the ratio of residential investment to GDP, the change in the ratio of short-term household liabilities to disposable personal income, heavy truck sales and financial conditions indices are included in the analysis.

The predictability of recession is assessed for the 1979-2019 period. To take into account the serial correlation structure in the data, the sample is divided into three disjoint time periods; an in-sample period which is used to estimate the different modeling approaches for specific tuning parameter values, a validation period that is used to tune the hyperparameters and select their optimal values and an out-of-sample period that spans 2000-2019 where forecasts are generated and the predictive performance of the modeling approaches is evaluated. The main findings can be summarized as follows. The predictive power of the yield curve over a 12-month ahead horizon is reconfirmed. Moreover, the analysis reveals that the yield curve remains one of the most robust and important predictors of upcoming recessions, as it appears to be a significant predictor for all different forecast horizons employed in the analysis. Having said that, the inclusion of additional variables alongside the term spread can enhance the underlying predictive accuracy. More importantly, however, the implementation of various machine learning techniques increases accuracy even further. Penalized likelihood binary Logit models (LASSO and Elastic Net) and the k-nn method seem to be the most consistent across short, medium, and long-term forecast horizons, followed by the regularized discriminant analysis and Bayesian techniques, as well as Random Forest, especially for shorter-term forecasts. Overall, there is strong evidence supporting the implementation of machine learning models for economic activity prediction. To our knowledge, this is the first comprehensive, comparative study of machine learning techniques in the area of economic recession forecasting.

The remainder of the paper is organized as follows. Section 2 discusses the relevant literature; Section 3 describes the research methodology, i.e. introduces the different model specifications and the machine learning techniques under consideration; Section 4 presents the data and Section 5 the empirical design and findings; Section 6 offers a practical guide for recession forecasting and Section 7 concludes.

2 Literature Review

The idea of a potential set of early warning indicators that could spot changes in the business cycle and correctly predict expansions and contractions is not new. As early as 1946, Burns and Mitchell concentrated on determining whether a time series has specific cycles, and if so, to specify the beginning and end dates of each cycle. They provide an elaborate statistical description of the cyclical aspects of numerous time series, and classify macroeconomic-related variables as lagging, coincident and leading. Based on their seminal work, the U.S. Department of Commerce has been generating economic indicators that are widely followed and used to predict business cycle turning points.

Closely related is Stock and Watson's (1989, 1991, and 1993) modeling framework, where business cycles are measured by co-movements in various components of economic activity to obtain an alternative index to the Department of Commerce indicators. Recessions (expansions) are generated by negative (positive) symmetric shocks to a linear and dynamically stable time series system.

As the most critical shift in economic activity is that from a state of expansion to that of recession and vice versa, one of the primary aims of the research is to focus on the predictability of recession over different time horizons. In the existing literature there are two broad approaches in forecasting business cycles; the first is through the use of continuous models that forecast aggregate macroeconomic variables, such as economic growth, usually through the application of Hidden Markov models, while the second concentrates on the prediction of different macroeconomic regimes or states typically through the use of binary Probit/Logit models.

Early studies using Logit/Probit models explore the nature of financial crises in the U.S. in the pre-Great Depression period (Canova, 1994). Estrella and Mishkin (1998) examine the performance of numerous financial variables as predictors of subsequent U.S. recessions. More specifically, using a Probit model they examine the ability of financial variables to predict whether the U.S. economy will be in a recession between one and eight quarters ahead. They find that one of the most successful models, is that containing the yield curve slope or term spread (10-year Treasury yield and the 3-month Treasury bill spread) as the sole explanatory power for forecasting a recession four quarters ahead. The estimated probability of recession from the specific model is 10% when the spread averages 0.76% over the quarter, and rises to 90% when the spread averages -2.40%. Apart from the yield curve slope, Estrella and Mishkin (1998) include interest rates, equity market indices, monetary aggregates, macroeconomic and leading indicator variables as inputs, but conclude that the yield curve slope is the single-most powerful predictor of U.S. recessions in the medium-term.

The combination of financial and macroeconomic variables as explanatory variables is not accidental, as there is substantial evidence that financial variables contain ample information regarding the future path of economic activity, as financial markets are expectations driven, descent aggregators of information due to the broad participation, and participant's efforts to price assets as close as possible to their underlying fundamental

values. The specific type of models is static in the sense that they do not capture the autocorrelation in the dependent variable time series, i.e. in the binary recession indicator. There have been numerous extensions of the static models that take into account the autocorrelation in the binary recession indicator for prediction; Dueker (1997, 2005), Chauvet and Potter (2005), Kauppi and Saikkonen (2008), and Nyberg (2010) are primary examples. Dueker (1997), for example, includes a lagged recession parameter, i.e. a lagged dependent variable, in the specification of the Probit model and finds that the inclusion of the lagged dependent variable appears to complement the explanatory power of the slope of the yield curve. Moreover, he extends the dynamic Probit model by allowing for some time variation in the structure of the model. More specifically, the coefficients are allowed to change values based on a latent binary state variable that follows a Markov process; in this case, the coefficients take either of the two values, depending on the value of the state variable, altering the shock necessary to induce a recession. Dueker (1997) finds that the Markov process assists the forecast, as it is able to predict the length of the recession, once it is ongoing. He concludes that it is important to incorporate dynamic serial correlation in the Probit model, while time variation in the coefficients is not particularly significant for short-term horizons.

Chauvet and Potter (2005) also extend the Probit model by estimating a specification with business cycle specific variance and an uncorrelated latent variable, a Probit specification with constant variance and an autoregressive latent process, as well as a Probit model with business cycle specific variance and an autoregressive latent process.

More recently, Kauppi and Saikonnen (2008) extend the dynamic Probit model by adding a lagged value of the underlying probability, while a further extension of this model can be obtained by adding an interaction term between the recession indicator (dependent variable) and a set of explanatory variables that could be either the same or different than those in the original set of predictors. Nyberg (2010) estimates various dynamic Probit models for the U.S. and German economy and concludes that allowing for dynamic specifications adds significant additional predictive power and reaffirms that the term spread is a useful recession predictor.

Apart from Probit/Logit models, Markov switching models have been used extensively in the literature to determine and forecast economic regime changes. Hamilton (1989) introduces a hidden Markov model to describe changes as a latent first-order Markov process through realized macroeconomic data. The specific model has been widely used for time series that the underlying autoregressive parameters switch between different regimes. Empirical studies that implement hidden Markov models include among others Hamilton (1990), Hamilton and Perez-Quiros (1996), Marsh (2000), Banachewicz, Lucas, and Vaart (2008), Pinson and Madsen (2012), Nguyen (2018).

Besides the above-mentioned studies that implement hidden Markov models, other Markov switching models have also showed success in forecasting financial and economic time series. Chauvet and Hamilton (2006) introduce a Markov switching model in which the binary recession indicator intertwines with real GDP

rates to form a Markov chain. More recently, Barsoum and Stankiewicz (2015) analyze business cycle patterns in macroeconomic time series with Markov switching mixed data sampling (MIDAS) models. Nyberg (2018) introduces a regime switching vector autoregressive model with time-varying regime probabilities, where the switching dynamics are generated by an observable binary response variable simultaneously predicted with the variables subject to regime changes. The model reveals a strong bidirectional link between U.S. interest rates and the business cycle as defined by the National Bureau of Economic Research. Last but not least, Tian and Shen (2019) provide extensions to the implementation of Markov switching models in economic prediction. More specifically, they examine the out-of-sample predictive performance of a set of macroeconomic and financial variables in predicting U.S. recessions one to twelve months ahead.

Lately, there has been quite some interest in applying machine learning algorithms in economic activity and recession forecasting. Gogas et al. (2015) create a forecasting model using the information content of the U.S. yield curve to predict future economic activity (above or below trend real GDP growth) using a Support Vector Machine (SVM) classifier. More specifically, the empirical study focuses on forecasting the cyclical component of U.S. real GDP one, two and three quarters ahead for the 1967-2011 period. Their results show that the model exhibits significant ability to forecast upcoming recessions. Moreover, in a follow-up study using both a Probit model and an SVM approach, Gogas, Papadimitriou and Chrysanthidou (2015) reaffirm their previous research that the yield curve is a useful tool for assessing future economic activity.

Ng (2014) explores the effectiveness of boosting as a classification tool in providing recession warning signals three, six and twelve months ahead. The analysis covers a broad range of macroeconomic and financial variables for the 1961-2011 period and reveals that only a limited set of variables (around 10) exhibits systematic predictive power. The main conclusion is that there is considerable variation through time in the number and composition of the pertinent set of predictors; in general, spread-related variables seem to have information content regarding recessions, even though there are distinct differences in business cycle evolution.

Berge (2015) investigates U.S. business cycle turning point forecasting by way of four competing model selection methods, equally-weighted forecasts, Bayesian Model Averaging (BMA), and linear and non-linear machine learning boosting algorithms. He concludes that forecasts generated by BMA and boosting models outperform equally-weighted forecasts both in and out-of-sample. In addition, non-linear boosting models seem to produce superior forecasts versus their linear counterparts. Once again, the yield curve emerges as a robust and consistent predictor of future economic activity in the medium-term, while spreads related to corporate bonds are better suited for longer-term prediction. Last but not least, the analysis reconfirms that various macroeconomic and financial indicators contain information that could be useful in identifying and predicting business cycle turning points, but at different forecast horizons.

Dopke et al. (2017) use boosted regression trees to explore the importance of several leading indicators in predicting German economic recessions. More specifically, they employ financial market, real economy, and

composite leading indicators, as well as survey based and price related data, 35 variables in total, for business cycle forecasting. Their results reveal that short rates and the term spread are crucial leading indicators for recession prediction during the 1974-2014 period. In addition, they document that the relative importance of the term spread in particular and that of equity market returns has increased over time, while that of the short-term interest rate has decreased. Most importantly, they show that the predictive accuracy of boosted regression trees is significantly better than that of standard probit models.

Pierdzioch and Gupta (2019) estimate boosted regression trees for a long U.S. data set that goes back to 1889 to investigate the information content of disaggregated news-based uncertainty indices regarding U.S. recessions alongside a set of control variables. More specifically, they use the measures of uncertainty developed by Manela and Moriera (2017) that create monthly text-based uncertainty metrics using front-page articles in the Wall Street Journal, News Implied Volatility (NVIX), related to government policy, financial intermediation, natural disasters, financial markets, wars and those that are uncategorized. They find that based on predictor relative performance the war-related NVIX measure is among the best predictors of recession across different forecast horizons, while the importance of the NVIX metric related to financial markets has risen during the second half of the sample period. Similarly, the government-related uncertainty seems to be relevant at a 12-month forecast horizon.

Davig and Smalter Hall (2019) demonstrate that the use of a Naive Bayes model consistently outperforms competing models and the Survey of Professional Forecasters in real-time recession forecasting up to twelve months ahead. Using the FRED-MD data set as a starting point (135 variables), they settle on four 'core' variables in most of their models; the ISM Manufacturing Production Index, Total non-farm payroll growth, the term spread and the S&P 500 Index. Wherever applicable they use both revised and real-time versions of the underlying data and show that in both cases their Naive Bayesian framework outperforms Logit regression models.

3 Model Specifications and Machine Learning Techniques

This section outlines the models used to forecast U.S. recessions. More specifically, standard Probit and Logit regression models, as well as a plethora of machine learning techniques are briefly discussed, with primary goal to create a comprehensive framework for modeling and predicting U.S. recessions.

3.1 Binary Probit and Logit Regression Models

Probit and Logit regression models have been extensively employed to model and predict the probability of recession $Prob(Y_t)$, where Y_t is a binary-valued dependent variable that takes values $Y_t = 1$, if the economy is in recession in month t, and $Y_t = 0$, otherwise, based on a set of predictor variables X_t. Assuming that each Y_t is the

outcome of independent Bernoulli random variables with probability function $p_t^{y_t}(1-p_t)^{1-y_t}$, and probability of recession p_t, the aim of the analysis is to accurately model/predict the conditional probability of recession $Prob(Y_t = 1|\Re_{t-h})$ given the information set of covariates at lag h, $\Re_{t-h}$, and a function p_t parameterized by a parameter vector β.

In the case of the Logit regression model, $\ln\left(\frac{p_t}{1-p_t}\right)$ is a linear function of the predictor variables and the unknown parameters β, and can be written in the following form:

$$\ln\left(\frac{p_t}{1-p_t}\right) = X'_{t-h}\beta = \beta_0 + \sum_{i=1}^{N}\beta_i x_{i,t-h}, \tag{1}$$

where $p_t = Prob(Y_t = 1|\Re_{t-h})$ is the conditional probability of recession at time t given the information set of covariates at time $t-h$, $\Re_{t-h}$, $\ln\left(\frac{p_t}{1-p_t}\right)$ is the Logit transformation (log-odds), X_{t-h} is a vector, augmented by one for the constant, that contains the predictor variables with values $x_{i,t-h}$, $i = 1,\ldots,N$, $t = 1,\ldots,T$, and $\beta = (\beta_0,\beta_1,\ldots,\beta_N)$ is the parameter vector to be estimated. Equation (1) suggests that the conditional probability of recession is given by

$$p_t = \frac{\exp\left(\beta_0 + \sum\limits_{i=1}^{N}\beta_i x_{i,t-h}\right)}{1+\exp\left(\beta_0 + \sum\limits_{i=1}^{N}\beta_i x_{i,t-h}\right)}.$$

Logit regression models can be estimated using Maximum Likelihood (ML) methods. Let T denote the sample size, $(y_t, x_{1,t},\ldots,x_{N,t})$ denote the values of $(Y_t, X_{1,t},\ldots,X_{N,t})$ for the observation at time t, $t = 1,\ldots,T$. Assuming that each Y_t follows a Bernoulli distribution with probability of recession p_t, the log-likelihood function for a sample of T observations can be written as:

$$\begin{aligned}
\ln f(Y|X,\beta) &= \sum_{t=1}^{T}\left[y_t \ln p_t + (1-y_t)\ln(1-p_t)\right] = \\
&= \sum_{t=1}^{T}\ln\left[1-p_t\right] + \sum_{t=1}^{T} y_t \ln\left[\frac{p_t}{1-p_t}\right] = \\
&= -\sum_{t=1}^{T}\ln\left[1+\exp\left(\beta_0 + \sum_{i=1}^{N}\beta_i x_{i,t-h}\right)\right] + \sum_{t=1}^{T} y_t\left(\beta_0 + \sum_{i=1}^{N}\beta_i x_{i,t-h}\right).
\end{aligned}$$

An alternative class of models used to estimate the probability of recession is the well-known Probit regression model. Based on the Probit specification, the conditional probability of a recession $Prob(Y_t = 1|\Re_{t-h})$ is modeled as follows:

$$Prob(Y_t = 1|\Re_{t-h}) = \Phi(\Psi_t)$$

where $\Phi(.)$ is the cumulative normal distribution function, h denotes the lags of the predictive variables used in the analysis, and

$$\Psi_t(\beta) = X'_{t-h}\beta$$

where X is the vector of the predictor values (augmented by ones), and β is the vector of the unknown coefficients. The parameters of the binary Probit models defined above are estimated by maximum likelihood, as follows:

$$\hat{\beta}_{ML} = argmax_{\beta} \left\{ \sum_{\{Y_t=1\}} \ln \Phi\Big(\Psi_t(\beta)\Big) + \sum_{\{Y_t=0\}} \ln\Big(1 - \Phi\big(\Psi_t(\beta)\big)\Big) \right\}. \tag{2}$$

For more details, please refer to Estrella and Mishkin (1998).

3.2 Penalized Likelihood Binary Logit Regression Models

A number of penalized likelihood binary Logit models for recession probability estimation are outlined in this section. More specifically, the Ridge, the Least Absolute Shrinkage and Selection Operator (LASSO) and the Elastic Net regularization techniques are presented, as they are implemented in the empirical analysis. The specific techniques are penalized least squares methods that impose shrinkage in the regression coefficients and allow for automatic variable selection. The inclusion of a parameterization penalty in the likelihood function is strongly recommended in order to avoid problems with overfitting, especially in cases of a large predictor set, and to improve out-of-sample predictive performance. For more details, refer to Hoerl and Kennard (1970), Tibshirani (1996) and Zou and Hastie (2005).

3.2.1 The Ridge Logit Model

The Ridge Logit model (based on the Ridge regression model introduced by Hoerl and Kennard, 1970) is a modeling approach that can result in better prediction accuracy, by regularizing or shrinking the estimated coefficients towards zero. The specific process has a number of benefits, as it leads to lower variance, decreases the complexity of the underlying model, but does not reduce the number of predictors, it just dampens their effect. Given the log-likelihood function of the ordinary Logit model, the Ridge log-likelihood Logit function introduces a shrinkage penalty that employs the ℓ_2-norm of β, $\|\beta\|_2 = \sqrt{\sum_{j=1}^{N} \beta_j^2}$ and a tuning parameter λ, $\lambda > 0$, that controls the degree of regularization. Increasing the value of λ tends to reduce the magnitude of the estimated coefficients, but does not result in the exclusion of any of the predictors. The maximum likelihood estimator of the Ridge Logit model is given by:

$$\hat{\beta}_R^{Logit} = argmax_{\beta} \left\{ \ln f\,(Y|X,\beta) - \lambda \sum_{j=1}^{N} \beta_j^2 \right\}.$$

3.2.2 The LASSO Logit Model

An alternative approach is the Least Absolute Shrinkage and Selection Operator (LASSO) model, introduced by Tibshirani (1996). By imposing a different type of shrinkage the LASSO Logit model yield better interpretability and can also perform variable selection, in order to identify predictors that are strongly associated

with recessions. Estimation of the LASSO Logit coefficients is achieved by maximizing their corresponding log-likelihood functions, while imposing a shrinkage penalty based on the ℓ_1-norm of β. The vector $\hat{\beta}_L^{Logit}$ is obtained by

$$\hat{\beta}_L^{Logit} = argmax_{\beta} \left\{ \ln f(Y|X,\beta) - \lambda \sum_{j=1}^{N} |\beta_j| \right\}$$

where $\|\beta\|_1 = \sum_{j=1}^{N} |\beta_j|$ is the ℓ_1-vector norm of β and λ, $\lambda > 0$, is the LASSO-related tuning parameter that controls the degree of shrinkage of β. The ℓ_1 penalty function shrinks the coefficients towards zero, forcing some of the coefficient estimates to be exactly equal to zero, when the tuning parameter λ is sufficiently large. In this way, the LASSO performs variable selection, while the resulting models are generally much easier to interpret relative to those produced by Ridge regression. This modeling approach eliminates entirely a number of predictors and, thus, could provide a set of predictive variables that are the most important to accurately forecast recession probabilities; the specific approach, therefore, reduces model complexity.

3.2.3 The Elastic Net Binary Logit Model

The Elastic Net technique is also used in the analysis; the Elastic Net is an alternative modeling approach, proposed by Zou and Hastie (2005) that has been extended to the Logit regression framework. The Elastic Net was introduced as an improved technique (to LASSO and Ridge), that is able to handle high correlated variables in the predictor set and account for the collinearity problem that is inherent in the analysis of high dimensional data. It combines the penalty terms of the Ridge and LASSO regression by using a convex combination scheme of the two techniques that is supposed to preserve their primary advantages and overcome their drawbacks. In practice, this can be achieved as the ℓ_1-norm of the Elastic Net performs automatic variable selection, while the ℓ_2-norm stabilizes the solution, and thus increases the out-of-sample predictive performance in the case of highly correlated predictors. The Elastic Net Logit coefficient estimates are obtained by maximizing the log-likelihood functions that penalize the size of the model coefficients based on both the ℓ_1-vector norm and ℓ_2-vector norm of β. Thus, the parameter estimates of the Elastic Net Logit regression model are obtained by:

$$\hat{\beta}_{EN}^{Logit} = argmax_{\beta} \left\{ \ln f(Y|X,\beta) - \lambda \Big((1-\alpha) \sum_{j=1}^{N} \frac{\beta_j^2}{2} + \alpha \sum_{j=1}^{N} |\beta_j| \Big) \right\}$$

respectively, where λ and α are the Elastic Net tuning parameters. In the empirical analysis we employ $\alpha = 0.5$.

3.3 Discriminant Analysis

Discriminant Analysis methods are used to classify if the economy is in recession at a specific period of time. More specifically, Linear and Regularized Discriminant analysis is applied in the analysis. These techniques are alternative classification tools that overcome certain limitations linked to traditional Probit/Logit type

regression models, i.e. when the recession binary-valued dependent variable is well separated or when there are just a few observations in a certain state (e.g. recessions), increasing the probability that the Logit regression estimates become unstable. In addition, the specific methods can be extended to multi-class classification problems.

3.3.1 Linear Discriminant Analysis

Linear Discriminant Analysis (LDA) is a multivariate statistical technique that can be used to classify and make predictions with respect to a binary dependent variable based on a set of predictors. The basic idea is to derive a linear combination of the predictors that 'best' differentiates the events related to a specific state (recession or expansion). There are several ways to develop discriminant criteria. One method is to use a Bayesian approach for classification and derivation of recession probabilities. Based on the Bayes' theorem, the conditional probability of Y_t at time t, given a set of predictors X_{t-h} is estimated by:

$$Pr(Y_t = k|X_{t-h} = x_{t-h}) = \frac{\pi_k f_k(x_{t-h})}{\sum_{i=0}^{1} \pi_i f_i(x_{t-h})}, \quad for \quad k = 0,1$$

where $k = 0$ and $k = 1$ denote the two possible outcomes of the dependent variable, i.e. the economy is in recession ($k = 1$) or not ($k = 0$), π_k denotes the prior probability, $f_k(x_{t-h}) = Pr(X_{t-h} = x_{t-h}|Y = k)$ denotes the conditional density of X_{t-h} given $Y = k$, while $Pr(Y_t = k|X_{t-h} = x_{t-h})$ denotes the posterior distribution of recession or expansion, given the set of predictors X_{t-h}. Assuming that the conditional distribution of the set of predictors X_{t-h} given $Y = k$ has a multivariate normal density with mean vector μ_k and a common covariance matrix $S_k = S$ for $k = 0,1$, i.e. $f_k(x_{t-h}) \sim N(\mu_k, S)$, every observation $X_{t-h} = x_{t-h}$ is assigned to each regime/state based on the largest value of

$$x_{t-h}^{'} S^{-1} \mu_k - \frac{1}{2} \mu_k^{'} S^{-1} \mu_k + log\pi_k$$

for $k = 0,1$. For more details, please refer to James et al. (2013).

3.3.2 Regularized Discriminant Analysis

Linear Discriminant Analysis is a popular technique for tackling classification problems. It becomes rather impractical, however, in high dimensional empirical problems with small sample size datasets and correlated predictors. In such cases, the estimated covariance matrix may be highly variable (unstable), suffers from singularity problems and may be non-identifiable. The use of the Regularized Discriminant Analysis (RDA) technique has been proposed as one of the likely solutions in the literature. The specific technique replaces the class specific covariance estimates by their average, pooled covariance matrix and utilizes regularization

to improve performance. The regularized covariance matrices have the following form:

$$S_k(\lambda) = (1-\lambda)S_k + \lambda S$$

where S is the pooled covariance matrix used in LDA, S_k, $k = 0, 1$, is the class specific covariance matrix used in Quadratic Discriminant Analysis, and $\lambda \in [0, 1]$ is a regularization parameter. In practice, a convex combination that allows S_k to be shrunk towards a scaled identity matrix, using the shrinkage parameter γ can be adopted as follows:

$$S_k(\lambda, \gamma) = (1-\gamma)S_k(\lambda) + \gamma \frac{1}{d} \mathrm{Tr}[S_k(\lambda)]I$$

where $\frac{1}{d}\mathrm{Tr}[S_k(\lambda)]$ is the mean of the diagonal elements of $S_k(\lambda)$, representing the mean variance of the class predictors. The RDA classifier is obtained by using the following equation:

$$\Psi_k^{RDA}(x_{t-h}) = \left\{ (x_{t-h} - \bar{x}_{k,t-h})' S_k^{-1}(\lambda, \gamma)(x_{t-h} - \bar{x}_{k,t-h}) + log|S_k(\lambda, \gamma)| \right\}$$

where λ is a tuning parameter that controls the degree of shrinkage of the individual class covariance matrix estimates towards the pooled estimates and γ is an additional regularization parameter that controls shrinkage toward a multiple of the identity matrix for a given value of λ. Please refer to Friedman (1989) for more details.

3.4 Classification and Regression Trees

In this section, the classification problem of the recession binary dependent variable is addressed by using tree-based classification methods. Classification and regression tree (CART) models, introduced by Breiman et al. (1984), provide a flexible method for analyzing non-linear relationships between the dependent variable and the set of predictors. In tree-based models, the space of the predictor variables is partitioned into a number of simple regions created recursively using a sequence of binary decisions. The terminal nodes of the tree correspond to distinct and non-overlapping regions of the partition, and the partition is determined by splitting rules associated with the set of predictors at each of the internal (splitting) nodes.

The estimation procedure of the tree structure is usually based on a top-down, greedy algorithm to grow a tree together with a prunning algorithm to avoid overfitting. Starting from a single node (all observations belong to a single region) at the top of the tree, and then successively splitting the predictor space, the tree is growing by sequentially choosing the new partitions, i.e. two new branches further down on the tree, based at each step on a model fitting criterion, such as the residual sum of squares (for regression trees), and the Gini index or the entropy (for classification trees). This sequential procedure generates a maximal binary tree that is corrected by pruning through a model selection criterion, such as cost complexity pruning, cross-validation, or an information criterion like AIC or BIC. Based on the estimated tree structure, prediction for

a given observation at time t can be constructed by using the mean (for regression trees) and the mode (for classification trees) of the training observations in the partition to which it belongs.

Tree-structured models can be displayed graphically, are simple and easy to interpret, they suffer, however, from reduced predictive accuracy/performance and non-robustness (small changes in the data can lead to different tree-structures and therefore to different predictions). Ensemble methods, which combine different tree models, have attracted a lot of attention in the literature, as a plethora of theoretical and empirical analysis has shown that the predictive performance can be substantially improved by aggregating many decision trees; see, for example, Breiman (1998), Bauer and Kohavi (1999), Buhlmann and Yu (2002), Biau, Devroye and Lugosi (2008), Dietterich (2000). The ensemble methods include Bagging (Breiman, 1996), Boosting (Freund and Schapire, 1997, Friedman, 2001), and Random Forest (Breiman, 2001) among several other extensions that use trees as building blocks to construct more powerful predictive models. Next, we present the classification models applied in this study, which include Bagging, Random Forest, Adaptive Boosting and Gradient Boosting.

3.4.1 CART Recursive Partitioning Algorithm

Recursive Partitioning Algorithms (rpart) are used to construct classification and regression models; that is, the resulting models can be represented as binary trees. The aim of the analysis is to predict/classify if the economy is in recession or not based on a set of predictor variables. The tree structured model is constructed using a two stage procedure; in the first step, a single predictor variable is found which best splits the data into two groups. There are several measures that can be used to find the best split; possible choices can be the information index and the Gini index. The dataset is partitioned using this splitting variable into two child nodes, and this process is applied separately to each sub-group (child node). This is done recursively until the subgroups or the terminal nodes either reach a minimum size or until no improvement can be made. This tree structure may be too complicated. The second step of the procedure consists of pruning the full tree. A complexity hyperparameter cp, which measures the cost of adding another variable, can be used. More details can be found at Therneau and Atkinson (2019).

3.4.2 Bagging

Bootstrap aggregating or Bagging (Breiman, 1996) is a widely applied ensemble method for classification. It improves the predictive accuracy of a classifier by generating multiple versions of the classifier based on bootstrap replicates of the training data set, and then combines these classifiers in order to construct a single one. The specific aggregated procedure produces a stable forecast/classifier with smaller variance and, thus, succeeds in achieving substantial gains in terms of accuracy.

To briefly outline the bagging algorithm, consider $(Y_t, X_{t-h}), t = 1, \ldots, T$ to be the training sample, where Y_t is

the recession binary dependent variable at time t, and X_{t-h} is the vector of predictors at time $t-h$. In the analysis (recession classification problem), we are interested in classifying if the economy is in recession or not, or stated in a different way, we are interested in estimating the $Prob(Y_t = 1|\Re_{t-h})$, or $Prob(Y_t = 0|\Re_{t-h})$, i.e. the conditional probability of recession or not at time t, given the information set of predictors at time $t-h$, $\Re_{t-h}$. The core idea of the algorithm is to construct B bootstrap samples $S_1^*, S_2^*, \ldots, S_B^*$. From each bootstrap sample S_i^*, $i = 1, \ldots, B$, the quantity of interest, i.e. the classifier, say $\hat{C}_i^*$, is estimated based on the same learning procedure. Then, the bagged estimator/classifier, $\hat{C}_{bag}$, can be obtained by aggregating the different bootstrap classifiers. Breiman (1996) proposes the use of the majority vote, that is, the overall classifier is the most occurring class among the B bootstrap classifiers. An alternative approach is to estimate the conditional probability of recession from each bootstrap sample and then average the bootstrapped probabilities.

Yet another approach based on a similar bootstrapped procedure is the following: Consider a parameterized training sample of the form $(\tilde{Y}_t, X_{t-h})$, $t = 1, \ldots, T$, where X_{t-h} is the vector of predictors at time $t-h$, and $\tilde{Y}_t \in \{-1, 1\}$; $\tilde{Y}_t = 1$ or $\tilde{Y}_t = -1$ indicates that the economy is in recession or not, respectively. The classification into one of the two groups is based on the following expression:

$$\hat{\Psi}(x_{t-h}) = sign\Big(\hat{\delta}(x_{t-h}) - \tau\Big),$$

where $\hat{\delta}(x_{t-h})$ is the decision tree base classifier obtained by $\hat{\delta}(x_{t-h}) = 2\hat{p}(x_{t-h}) - 1$, $\hat{p}(x_{t-h})$ is the probability of recession estimated by the tree, and τ is a cut-off value (usually zero). If $\hat{\delta}(x_{t-h}) > \tau$, then $\hat{\Psi}(x_{t-h}) = 1$, indicating that the economy is in recession, while $\hat{\delta}(x_{t-h}) < \tau$ implies $\hat{\Psi}(x_{t-h}) = -1$ and the economy not in recession. Based on the bagging algorithm, for each bootstrap sample S_i^* a classifier can be estimated employing the score functions $\hat{\delta}_i(x_{t-h})$, $i = 1, \ldots, B$. Afterwards, these functions are aggregated into a bagged score based on:

$$\hat{\delta}_{bag}(x_{t-h}) = \frac{1}{B}\sum_{i=1}^{B}\hat{\delta}_i(x_{t-h}),$$

and the final classification is obtained as follows:

$$\hat{\Psi}_{bag}(x_{t-h}) = sign\Big(\hat{\delta}_{bag}(x_{t-h}) - \tau\Big).$$

3.4.3 Random Forest

The Random Forest (Breiman, 2001) constitutes an alternative machine learning technique that generates improved predictive accuracy/performance compared to the standard classification and regression tree models, as well as bagged trees. Random Forest can reduce the classifier variance by aggregating a number of low correlated trees obtained using different bootstrap training samples. This can be achieved by constructing multiple decision trees, where the splitting variables of the internal nodes are based on a random process, i.e.

by introducing randomness in the tree-growing procedure.

To provide more clarity, the estimation procedure of a classifier based on the Random Forest approach is presented in the context of a recession classification problem. Consider the training sample (Y_t, X_{t-h}), $t = 1,\ldots,T$. As in the case of bagging, B bootstrap samples $S_1^*, S_2^*, \ldots, S_B^*$ are generated from the training data set. For each bootstrap sample, a Random Forest classification tree is constructed, T_i^{rf}, and the classifier $\hat{C}_i^{rf}$, $i = 1,\ldots,B$, estimated based on the same learning procedure. Then, the Random Forest classifier, $\hat{C}_{rf}$, can be obtained by aggregating the different bootstrap classifiers using the majority vote rule, i.e. $\hat{C}_{rf} =$ majority vote $\{\hat{C}_i^{rf}\}_1^B$.

The Random Forest classifier found by aggregating different bootstrapped classifiers is similar to that obtained by the Bagging technique. However, there is a substantial modification in the way that Random Forest classification trees are constructed. Below, the growing process of a Random Forest tree and its main advantages are outlined. The process employs a top-down, greedy algorithm that creates a maximal in-depth classification tree using the bootstrap sample. Starting from a single node at the top of the tree, the tree grows by recursively splitting the temporary terminal node into two new child nodes. The splitting variable is chosen to be the 'best' variable among a random subset of m candidate variables taken from the full set of the N predictive variables. A new random sample of m candidate splitting variables is taken at each splitting node of the tree. Thus, the Random Forest selects at each node of each tree, a random subset of variables, and only these variables are used as candidates to find the best split for the node. The use of different bootstrap samples, i.e. construct different classification trees, together with the introduction of this randomness at each splitting node of each tree, provides a number of uncorrelated trees and as a consequence, the Random Forest forecast/classifier achieves reduced variance.

3.4.4 Adaptive Boosting

Boosting is an ensemble technique originally introduced to improve the performance of classification algorithms. It combines several weak classifiers (e.g. a classifier that predicts marginally better than random) to produce a more accurate and powerful classifier. The basic idea is to combine classifiers that are iteratively constructed through the resampling of the training data by assigning increased weight to observations that were misclassified, to produce a new classifier that could boost the performance in previous problematic cases. This process is repeated, generating several classifiers, which are then combined into a final classifier by applying weighted majority vote.

The Adaptive Boosting (Adaboost) algorithm proposed by Freund and Schapire (1997) is employed in the analysis, to classify if the economy is in recession or not, given the set of predictive variables. Consider a parameterized training sample of the form $(\tilde{Y}_t, X_{t-h})$, $t = 1,\ldots,T$, where X_{t-h} is the vector of predictors at time $t-h$, and $\tilde{Y}_t \in \{-1,1\}$; $\tilde{Y}_t = 1$ or $\tilde{Y}_t = -1$ indicates that the economy is in recession or not. The classification

into one of the two events, using the AdaBoost algorithm (see, also, Hastie, Tibshirani and Friedman, 2009) is outlined in the following steps:

1. Initialize the observation weights $\omega_t = \frac{1}{T}, t = 1, \ldots, T$, i.e. each observation shares the same weight.
2. For $l = 1, \ldots, L$
   (a) Fit a weak classifier $C_l(x_{t-h})$ to the training data using observation weights ω_t.
   (b) Compute the weighted misclassification error for C_l: $err_l = \frac{\sum_{t=1}^{T} \omega_t I[\tilde{Y}_t \neq C_l(x_{t-h})]}{\sum_{t=1}^{T} \omega_t}$
   (c) Compute $\alpha_l = ln[\frac{1-err_l}{err_l}]$
   (d) Update the observation weights giving more weight to misclassified observations and less weight to correctly classified observations: $\omega_t \leftarrow \omega_t exp(\alpha_l I[\tilde{Y}_t \neq C_l(x_{t-h})]), t = 1, \ldots, T$, and normalize them.
3. Compute the boosted classifier $\hat{C}_{boost}(x_{t-h}) = sign\left[\sum_{l=1}^{L} \alpha_l C_l(x_{t-h})\right]$. If this sum is positive, then classify the observation as $+1$, i.e. recession, otherwise assign -1, i.e. not recession.

A very interesting feature of the Adaptive Boosting algorithm was presented by Friedman, Hastie and Tibshirani (2000) who showed that Adaptive Boosting can be interpreted as a forward stagewise additive model that minimizes exponential loss. The specific important finding allowed the extension and generalization of the technique.

3.4.5 Gradient Boosting

Friedman (2001, 2002) developed a framework for the creation of a new generation of boosting techniques, based on boosting's link to the statistical principles of additive modeling and maximum likelihood (Friedman, Hastie and Tibshirani, 2000). More specifically, Friedman (2001) presented a general method for estimating/approximating a function $F(x_{t-h})$, mapping the lagged predictor variables X_{t-h} to the dependent variable, based on numerical optimization in the function space. The function $F(x_{t-h})$ can be expressed in terms of forward stagewise additive modeling and the application of steepest-descent minimization. A generic gradient descent 'boosting' algorithm is developed (Gradient Boosting) for additive expansions based on different fitting criteria. The principle idea of the Gradient Boosting algorithm is to construct additive models by sequentially fitting a base (weak) learner to current pseudo-residuals at each iteration. Thus, the learning process fits new models consecutively, to provide a more accurate estimate of the dependent variable.

The main focus of the underlying analysis is to model and predict the probability of recession Y_t given a set of predictor variables X_{t-h}. Assuming that each Y_t ($Y_t = 1$ denotes recession at time t) is the outcome of independent Bernoulli random variables with probability of recession p_t, the quantity of interest is the log-odds, denoted by $\lambda(x_{t-h})$, where $\lambda(x_{t-h}) = \ln\left(\frac{p_t}{1-p_t}\right)$ that is a function of the set of predictors X_{t-h}. Under

the framework of Gradient Boosting, $\lambda(x_{t-h})$ is the function $F(x_{t-h})$ that can be expressed based on additive expansions of the form:

$$\lambda(x_{t-h}) = F(x_{t-h}) = \sum_{l=1}^{L} g_l(x_{t-h};\gamma_l)$$

where $g_l(x_{t-h};\gamma_l)$ is the base learner (e.g. a tree), and γ_l is the parameter vector to be estimated (for trees, it contains identification of splitting variables, their splitting values, and the constants in the terminal nodes). The steps of the Gradient Boosting algorithm (see, for example, Friedman, 2001, Efron and Hastie, 2016) are given by:

1. Start with $\hat{F}_0(x_{t-h}) = 0$, and set the shrinkage parameter $\varepsilon > 0$ that controls the rate at which boosting fits-overfits the data.

2. For $l = 1,\ldots,L$

   (a) Compute the pseudo-residuals, i..e. the negative gradient of the loss function being minimized at the current fit at each training data point:
   $R_t = -\frac{\partial L(Y_t,F(x_{t-h}))}{\partial F(x_{t-h})}|_{F(x_{t-h})=\hat{F}_{l-1}(x_{t-h})}$

   (b) Approximate the negative gradient by a tree of depth-d by solving the following optimization:
   $minimize_\gamma \sum_{t=1}^{T}(R_t - g(x_{t-h};\gamma))^2$.

   (c) Update $\hat{F}_l(x_{t-h}) = \hat{F}_{l-1}(x_{t-h}) + \hat{g}_l(x_{t-h})$, where $\hat{g}_l(x_{t-h}) = \varepsilon g(x_{t-h};\hat{\gamma}_l)$.

3. Returns the sequence $\hat{F}_l(x_{t-h})$, $l = 1,\ldots,L$.

3.5 Other Machine Learning Techniques

3.5.1 Naive Bayes

An alternative approach that can be used in the recession classification problem is the Naive Bayes learning scheme (see, for example, Bauer and Kohavi, 1999). More specifically, the scheme can be used to classify and make predictions with respect to the recession binary dependent variable based on a set of predictive variables. By applying the Bayes' theorem and assuming conditional independence of the predictors given the classifier, the Naive Bayes classifier can be found by applying:

$$\Psi^{NB}(X_{t-h}) = argmax_k \left\{ \pi_k \prod_{i=1}^{N} Pr(X_{i,t-h} = x_{i,t-h}|Y = k) \right\}$$

where $\pi_k = P(Y = k)$ denotes the prior distribution of the recession dependent variable, which is estimated by the sample proportions, $Pr(X_{i,t-h} = x_{i,t-h}|Y = k)$ denotes the conditional density of X given $Y = k$, for $k = 1$ (recession) and $k = 0$ (expansion). It is evident that the method assigns an observation $X_{t-h} = x_{t-h}$ to the class with the largest posterior probability, known as the maximum posterior decision rule.

The Naive Bayes method relies on the assumption that the predictive variables are conditionally independent given a specific class. Even though this assumption is not valid in empirical applications, the method is often more accurate than more sophisticated techniques; see, for example Domingos and Pazzani (1997) and the references therein. The method can be applied in classification problems in particular, i.e. when the dependent variable is qualitative, as the independence assumption is less restrictive than might be expected, and usually the method assigns maximum probability to the correct class. Its good performance can be attributed to the use of the zero-one loss function.

3.5.2 Bayesian Generalized Linear Models

Generalized linear models (GLM) are used to model the dependence of the binary recession variable Y_t on a set of predictor variables X_{t-h}. GLM models assume that the mean μ of an observation of Y is related to the predictor variables through a link function and a linear predictor model, i.e. $g(\mu) = \eta = X'_{t-h}\beta$, where $g(\mu)$ is the link function, and η is the linear predictor. Regarding the Logit regression model, the link function $g(\mu)$ is the Logit transformation, that is, $g(\mu_t) = \ln\left(\frac{p_t}{1-p_t}\right)$, where p_t is the probability of recession.

The Bayesian approach to inference in order to obtain stable Logit regression coefficients is briefly outlined. In Bayesian analysis, inference about the model parameters is based on the posterior distribution that plays a fundamental role, and is computed by using the Bayes theorem:

$$f(\beta_m|D) = \frac{f(D|\beta_m)f(\beta_m)}{f(D)} = \frac{f(D|\beta_m)f(\beta_m)}{\int f(D|\beta_m)f(\beta_m)d\beta_m} \propto f(D|\beta_m)f(\beta_m)$$

where $f(\beta_m|D)$ is the posterior distribution of a specific model m parameters, β_m, given the data D, $f(D|\beta_m)$ is the likelihood function, and $f(\beta_m)$ is the prior density of model m parameters. The posterior distribution summarizes all that is known about the model parameters after (i.e. posterior to) seeing the data. To reflect the ex ante opinion of the uncertainty regarding the model parameters, a prior distribution is assigned. Conditional on having observed the data, the prior opinion can then be updated to a posterior opinion on model parameters using the Bayes theorem.

To implement the Bayesian generalized linear model in the problem of interest (recession forecasting), the approach proposed by Gelman et al. (2009) is applied, using the 'bayesglm' function, that finds the approximate posterior mode and variance using extensions of the classical generalized linear model computations. Gelman et al. (2009) propose the use of a new prior distribution, the Student-t family, focusing on the Cauchy sequence as a conservative choice, for the Logit regression model parameters by first scaling the predictor variables. The proposed prior distribution setting enables the production of stable, regularized estimates, even when there is separation in Logit regression, and allows for an automated procedure in applied data analysis.

3.5.3 k-Nearest Neighbor

k-Nearest Neighbor (k-nn) is a non-parametric technique used for classification. Given the training data set $(Y_t, X_{t-h}), t = 1, \ldots, T$, the aim of the analysis is to classify if the economy is in recession or not in a specific point in time given the predictive variables. The target is to classify Y_t given the set of predictors X_{t-h}. It is a distance based classifier, that takes into account the similarity of the lagged predictive variables at a specific time period with the observations of the training set. Then, the new observation is classified based on the majority vote of its neighbors, i.e. the new observation is assigned to the class that is most common among its k nearest neighbors. Implementation of this technique, requires the specification of a distance metric, that is needed to calculate the distance between the new observation and that of the training sample, and a positive integer k, which determines the number of neighbors used in the voting process, as well as standardization of the predictor variables. A Euclidean distance metric can be used for computing distances in a multi-dimensional predictor space with quantitative variables, as is the case in the analysis. The choice of tuning parameter k is very important, as it affects the classifier obtained by the technique. A solution is to use a holdout validation period to select the value of k; details are presented in the empirical application section.

This classification method does not require any model to be fitted, and does not impose any strong assumption regarding the underlying data. Thus, it can be applied easily to empirical classification problems. The basic idea behind the algorithm is that is reasonable to consider that observations that are close together/similar (based on an appropriate metric) will have the same classification.

3.6 Statistical Performance Evaluation

The majority of the performance evaluation metrics reported are based on the so-called confusion matrix. The confusion matrix is a concise way to summarize the prediction results regarding a classification problem. The number of correct and incorrect predictions are compared to the underlying count values and separated by each class. It provides useful insight into the errors made by the classifier, but, more importantly, on the types of errors being made. The confusion matrix reports different combinations of predicted and actual values. More specifically, the matrix consists of the following elements: True Positive (TP): The actual value is positive and its prediction positive, False Positive (FP): The actual value is negative, but its prediction positive, False Negative (FN): The actual value is positive, but its prediction negative, and True Negative (TN): The actual value is negative and its prediction negative. The confusion matrix representing the classification problem of interest is outlined below;

The confusion matrix		
	Positive (Predicted)	Negative (Predicted)
Positive (Actual)	TP (Recession)	FN
Negative (Actual)	FP	TN (No Recession)

while, the following metrics are used to evaluate the accuracy and correctness of the estimated classification models:

1. Area Under the Curve

The Area under the Curve (AUC) is a summary statistic for the goodness of a predictor in a binary classification problem. It is perceived as an accurate classifier performance measure. Let pp_t be a real-valued scalar which denotes a probability prediction about Y_t, a linear indicator regarding the underlying economic regime. pp_t along with a threshold value tv designate a recession whenever $pp_t \geq tv$, and expansion whenever $pp_t \leq tv$. Given the specific variables, the following conditional probabilities can be defined as:

$$TP(tv) = Pr[pp_t \geq tv | Y_t = 1], FP(tv) = Pr[pp_t \geq tv | Y_t = 0]$$

where $TP(tv)$ is the true positive rate and $FP(tv)$ is the false positive rate. The Receiver Operating Characteristic (ROC) curve plots the complete set of possible combinations of $TP(tv)$ and $FP(tv)$ for $tv \in (-\infty, \infty)$. As $tv \to \infty$, $TP(tv) = FP(tv) = 0$, and conversely as $tv \to -\infty$, $TP(tv)$ =FP(tv)= 1. As such, the ROC curve is an increasing function in $[0,1] \times [0,1]$ space. The AUC is a measure of overall classification ability, a summary of all the trade-offs in the ROC curve, and is defined as follows:

$$AUC = \int_0^1 ROC(r)dr$$

where the ROC curve is represented with the Cartesian convention $\{ROC(r), r\}_{r=0}^{1}$, with $ROC(r) = TP(tv)$ and $r = FP(tv)$, and $AUC \in [0.5, 1]$. An alternative specification of AUC is as follows:

$$AUC = \int_0^1 \frac{TP}{TP+FN} d\frac{FP}{FP+TN}$$

The AUC takes values between 0.5 and 1.0. A value of 0.5 implies that the predictions are not better than random, while a value of 1.0 implies perfect predictions. For more details please refer to Berge and Jorda (2011).

2. **Accuracy**

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN}$$

Accuracy, also known as the Correct Prediction Ratio (CPR), is the total number of events that were classified correctly (both recessions and no recessions), out of the total sample size. Misclassification Error (MCE), $MCE = 1 - Accuracy$, indicates how often the classifier is incorrect overall.

3. Precision

$$Precision = \frac{TP}{TP+FP}$$

Precision shows the percentage of months that were correctly classified as recessions out of the total number of months classified as recessions, i.e. when a model predicts recession, precision measures how often it is correct.

4. Sensitivity

$$Sensitivity = \frac{TP}{TP+FN}$$

Sensitivity, also known as Recall or True Positive Rate, is the percentage of the months classified as recessions over the actual number of recessions.

5. Specificity

$$Specificity = \frac{TN}{TN+FP}$$

Likewise, specificity or True Negative Rate is the proportion of the months classified as non-recessions over the actual number of no recessions.

6. Balanced Accuracy A metric that combines the Sensitivity and Specificity measures is Balanced Accuracy.

$$BalancedAccuracy = \frac{Sensitivity + Specificity}{2}$$

7. F-Score As it is difficult to compare two models with high Precision and low Sensitivity (Recall) or vice versa, the F-Score is used to make them comparable.

$$F_1Score = \frac{2}{\frac{1}{Recall} + \frac{1}{Precision}} = \frac{2TP}{2TP+FP+FN}.$$

The F_1Score is a measure that combines the Precision and Recall metrics into a single measure, as an accurate model should exhibit high values in both metrics.

8. Kappa Statistic

The Kappa coefficient is a statistic used to measure intra-rater reliability for categorical problems. It is similar to the Accuracy metric, but takes into account the accuracy that would have been achieved anyway through random forecasts. The Kappa statistic, denoted by κ, is estimated as follows:

$$\kappa = \frac{p_0 - p_e}{1 - p_e} = 1 - \frac{1 - p_0}{1 - p_e}$$

where p_0 is the relative observed agreement among the classes, and p_e is the proportion of agreement that is expected by chance.

9. Directional Predictability Test

The test was introduced by Pesaran and Timmermann (1992) and employed by Granger and Pesaran (2000) for evaluating directional forecasting or predictability performance and market timing. In the analysis it is used for business cycle and recession predictability in particular. We examine the null hypothesis, H_0, of "No statistically significant recession predictability" against the alternative hypothesis, H_A, of "There is statistically significant recession predictability". See, also, Nyberg (2011), and Bergmeir, Costantini and Benitez (2014).

4 Data

As one of the primary conclusions of the existing business cycle turning point related research is that there is considerable variation in the macroeconomic and financial variables that contain useful information in identifying and predicting economic activity turning points, and that, to a large extent, their effectiveness is time dependent, a relatively large number of potential predictive variables, with leading, coincident and lagging properties, is included in the analysis.

More specifically, the set of predictive variables contains 56 macroeconomic and financial market related indicators, the majority of which are widely-followed by both policy makers and practitioners, and have been used in the existing literature for predictability of recession. Apart from the 'usual suspects', i.e. predictive variables such as the yield curve, the default spread, financial market and economic activity related indicators, a number of less studied indicators, such as the change in the ratio of residential investment to GDP, the change in the ratio of short-term household liabilities to disposable personal income, heavy truck sales and financial conditions indices are included in the analysis.

In general, the forecasting variables are representative of categories related to output and income, the labor market, the housing market, orders and inventories, money and credit, interest rates, prices, and the financial markets. A comprehensive list is presented in Table A1 in the Appendix. The data is obtained from the Federal Reserve Bank of St. Louis' FRED database and cover the period between January 1979 to June 2019 (486 monthly observations, or 40.5 years). The full sample is divided into three disjoint subsamples; an in-sample (training), a holdout (validation) and an out-of-sample (testing) period; the initial in-sample period ranges from January 1979 to December 1996 (216 monthly observations, or 18 years) and follows an expanding window scheme, the validation period is a 36-month rolling window scheme, while an out-of-sample period ranging from January 2000 to June 2019 (234 monthly observations, or 19.5 years), has been employed in order to evaluate forecasting performance.

It is recognized that several of the macroeconomic-related time series used are subject to multiple revisions.

As the primary focus of this paper is neither related to real-time identification of business cycle turning points, nor an evaluation of and/or comparison to the NBER's Business Cycle Dating Committee methodology and eventual dating (Chauvet and Piger, 2008), only revised data is included in the analysis; numerous studies have followed a similar path (Ng, 2014, Dopke et al., 2017, Pierdzioch and Gupta, 2019, among others).

The recession indicator is observable and derived from the National Bureau of Economic Research's Business Cycle Dating Committee chronology of the U.S. business cycle. The chronology consists of alternating dates of peaks and troughs in economic activity, representing expansions and contractions. The recession indicator is a binary variable that takes the value of 1, if the economy is in recession in month t and 0 otherwise.

As there is considerable deviation in the effectiveness of the variables to predict economic activity over different horizons, the analysis centers around four distinct windows, an imminent, a short, a medium and a long-term one. More specifically, the forecast horizon is equal to $1, 3, 6$ and 12 months for each of the forecasting models, using the corresponding h-step ahead forecasts.

The four windows based on the different forecast horizons are the following:

1. Long-Term Setup: compute 12-step ahead forecasts using the predictive variables, $X_1 - X_{56}$, lagged by 12-months.

2. Medium-Term Setup: compute 6-step ahead forecasts using the predictive variables, $X_1 - X_{56}$, lagged by 6 and 12-months.

3. Short-Term Setup: compute 3-step ahead forecasts using the predictive variables, $X_1 - X_{56}$, lagged by 3, 6 and 12-months.

4. One-month ahead Setup: compute 1-step ahead forecasts using the predictive variables, $X_1 - X_{56}$, based on several lagged predictive variables, i.e. lagged by 12-months only, by 6 and 12-months, and by 3, 6 and 12-months, respectively.

As part of the 'original' dataset has quarterly frequency, it needs to be transformed into monthly equivalents. Variables such as real Gross Domestic Product (GDP), real Gross Domestic Income (GDI), and corporate profits, are transformed into monthly frequencies using natural cubic spline interpolation.

5 Empirical Design and Analysis

5.1 Empirical Design: Sample splitting and tuning via validation

In this section, the empirical design of the underlying analysis is outlined. The main objective is to assess the out-of-sample performance of the different modeling approaches under consideration. The performance of the

machine learning models, however, depends on the choice of hyperparameters (or tuning parameters). The approach of Gu, Kelly and Xiu (2020) is employed to tune the hyperparameters of the relevant machine learning techniques in order to optimize their out-of-sample performance. Based on the specific approach, the sample is divided into three disjoint time periods (subsamples) that maintain the temporal ordering of the underlying data, and, thus, take into account the serial correlation structure in the data. The first subsample period is the 'training set', in which each model is estimated subject to a specific set of hyperparameter values. The second subsample is the 'validation set', during which the hyperparameter values are tuned. This can be done by estimating the alternative models, for each value of the set of tuning parameters adopted, using the data from the training subsample to construct forecasts for each data point in the validation set. Then, based on the forecasts in the validation set, an objective function is estimated (maximized), which in our case is the accuracy metric, and (for all tuning parameter values) the optimal value(s) of the tuning parameter(s) that maximize the accuracy metric is sought iteratively. As pointed out by Gu, Kelly and Xiu (2020), hyperparameter tuning amounts for a degree of model complexity that tends to produce reliable out-of-sample performance, since, based on the validation set, this procedure is used to evaluate the predictive performance of modeling approaches for different tuning parameter values, and selects the optimal one. Finally, the third subsample is the 'testing set', which is used to examine/evaluate the out-of-sample performance of the different modeling approaches. Out-of-sample forecasts for each data point in the testing set are obtained for every machine learning model based on the optimal tuning parameter value(s) found in the validation set.

The analyzed data cover the period between January 1979 to June 2019 (486 monthly observations, or 40.5 years). The full sample is divided into three disjoint time periods:

a training, a validation and a testing or out-of-sample period; the initial training period ranges from January 1979 to December 1996 (216 monthly observations, or 18 years), the respective initial validation period ranges from January 1997 to December 1999 (36 monthly observations, or 3 years), while the testing or out-of-sample period from January 2000 to June 2019 (234 monthly observations, or 19.5 years) has been employed in order to evaluate forecasting performance. To obtain out-of-sample forecasts for the testing period, a 'hybrid' scheme is adopted, that uses a 'recursive' in-sample scheme that gradually includes more recent observations in the training period and a 'fixed' 3-year validation scheme. To clarify the performance evaluation scheme, the procedure followed is described in detail for constructing one step-ahead forecasts. In the first iteration, that is, to construct forecasts for the first out-of-sample period (i.e. January 2000), the training set consists of 216 monthly observations (January 1979 to December 1996) which is used to estimate the machine learning models for a specific set of hyperparameter values. The corresponding validation set consists of 36 monthly observations (January 1997 to December 1999) that is used to tune the hyperparameters of the different modeling approaches. Based on the selected optimal tuning parameter value(s) in the validation set, one step-ahead forecasts are constructed for January 2000. Next, for the second iteration, i.e. to construct forecasts for Febru-

ary 2000, the training set is recursively increased by one observation, thus, it consists of 217 observations (from January 1979 to January 1997) and all the models/techniques are re-estimated for specific tuning parameters; the respective validation set consists of a fixed window scheme of 36 observations (February 1997 to January 2000) that is used to select the optimal tuning parameters for each model, and based on the optimal model hyperparameter values one step-ahead forecasts are computed for February 2000. This procedure is repeated recursively using an expanding/recursive window for the training set, and a fixed size rolling sample for the validation set, in order to produce out-of-sample forecasts for the testing period. The advantage of this procedure is that the recursive scheme always retains the entire history in the training sample, thus its window size gradually increases by one recent observation. Following Gu, Kelly and Xiu (2020), there is no cross-validation, in order to maintain the temporal ordering of the data for prediction, and, therefore, to respect the serial dependence of the time series data under consideration. A similar procedure is adopted in order to construct short (3-step ahead), medium (6-step ahead) and long (12-step ahead) forecasts.

5.2 Tuning Hyperparameters Range and Values

The values, or the corresponding range, of the tuning hyperparameters adopted in the different machine learning models are outlined below. Several machine learning techniques used in the analysis rely on the choice of tuning parameters that control model complexity and, thus, optimize their underlying predictive performance. The hyperparameters are different for each implemented technique; in the case of penalized likelihood models (Lasso, Ridge and Elastic Net), for example, the tuning parameters are the penalization parameters λ and α. Similarly, hyperparameters are the regularization parameters in Regularized Discriminant Analysis, the distribution type, the Laplace correction and the bandwidth adjustment in the Naive Bayes method, the randomly selected predictors in Random Forest, the number of boosting iterations, the depth of the trees, the learning rate in Adaptive and Gradient Stochastic Boosting, the number of neighbors in the k-nn method and the complexity parameter in the recursive partitioning method. Table A2 of the Appendix provides implementation details with respect to the tuning parameter values applied in the empirical analysis, as well as the corresponding R functions/methods that were used for model estimation. For example, for the estimation of the penalized binary Logit models, the 'glmnet' method/function and different values for the regularization parameters were used. For the LASSO, Ridge and Elastic Net models fixed α parameter values (i.e. $\alpha = 1$, $\alpha = 0$ and $\alpha = 0.5$, respectively), and two different sets of values for the penalized parameter, λ, i.e. λ ranges in $(10^{-5}, 10^{-1})$ with step 10^{-4}, as well as $\log(\lambda)$ ranges in $(-11, -1)$ with step 0.01, were employed. In Regularized Discriminant Analysis, the regularized parameters γ and λ range in $(0.1, 1)$ and $(0, 1)$ with 0.1 step, respectively. Moving on to the Random Forest, the 'rf' method was implemented, where the randomly selected predictors, $mtry$, belong to the set, $mtry \in \{2, 3, 5, 10, 20, 30\}$ with 300 trees, while for the Adaptive Boosting technique the method 'ada' was used for estimation, with three tuning parameters i.e. $iter$ (number of boosting itera-

tions), *maxdepth* (maximum tree depth),and *nu* (learning rate) assuming the following values: $iter \in \{400, 500\}$, $maxdepth \in \{1, 2, 3\}$ and $nu \in \{0.01, 0.1\}$. A more detailed description for all the estimated models is provided in Table A2 of the Appendix.

5.3 Empirical Findings

The empirical results are presented in this section. More specifically, Tables 1, 3, 5 and 7 show the out-of-sample performance of the underlying models for each forecasting horizon (long, medium, short-term and one-month ahead), based on a variety of statistical evaluation metrics, defined in Section 3.

As the sample is rather unbalanced, in the sense that only a very small part falls under the recession regime, the choice of a 50% threshold for classification and evaluation purposes might not be optimal. Numerous suggestions have been put forward in the literature. Chauvet and Hamilton (2006) propose that recession be declared when the smoothed probability estimate of the latent state exceeds 65% in a Markov switching model context. In a similar framework Chauvet and Piger (2008) require the smoothed probability to exceed 80% and stay above the specific level for three consecutive months. Berge and Jorda (2011) look for the optimal threshold between the 30% and 60% range, such that the Chauvet and Piger (2008) estimated probabilities would best fit the NBER Dating Committee's dates. Ng (2014) points out that applying a relatively high threshold, 65% for instance, would underestimate the risk of recession; she focuses on the evolution of the relative probability and estimates the threshold as the mean plus 1.65 times the standard deviation of the fitted probabilities over the sample, yielding thresholds of 30% and 43.5% for the 12 and 3-month forecasting window, respectively. Moreover, given that based on history the risk of recession increases considerably when the fitted probability of the standard benchmark model exceeds 30%, the chosen threshold for classification and evaluation purposes is fixed at 33% in our analysis.

The starting and focal point of the analysis is the ability of the U.S. yield curve (10-year Treasury yield minus 3-month Treasury bill rate) to predict shifts in economic activity, as it is considered by many as the single best out-of-sample predictor of U.S. recessions. The slope of the curve has long been regarded as one the most reliable signals of an upcoming economic recession, as the curve inverted ahead of each of the last seven recessions, and had a single false signal in the past 50 years. Since the late 60's, every U.S. recession has been led by a period during which the term spread was negative. It is thus not accidental that economists and financial analysts have monitored the position of the curve in order to assess the probability of an economic downturn. This is the primary reason that the Logit and Probit models including the yield curve as the sole predictive variable explicitly serve as benchmark or reference models for out-of-sample performance evaluation.

Please insert Table 1 around here

The predictive accuracy of the static binary Logit model using the term spread as the only covariate is relatively high in the long-term forecast horizon case (Table 1); based on metrics such as Area Under the ROC

curve (AUC) (83.8%) or Accuracy (89.7%), the yield curve does a rather good job in providing consistent early warning signals regarding upcoming recessions during the 2000-2019 period. In general, the results are quite similar with regard to the static Probit model. The overall results are consistent and confirm recent studies, such as Liu and Moench (2016). The benchmark model's out-of-sample forecasts are depicted in Figure 1, while the corresponding Receiver Operating Characteristic (ROC) curve is plotted in Figure 2.

Please insert Figure 1 around here

Please insert Figure 2 around here

Moving on to alternative models, both a kitchen sink approach and stepwise regression techniques using the Akaike Information Criterion (AIC) were implemented as part of the analysis, and confirmed that especially model/variable selection based on such approaches is rather inadvisable, due to the high dimensionality (large number of regressors) and the inherent multicollinearity issues in the underlying data that results in estimated parameter instability in Logit/Probit type models. By contrast, certain machine learning techniques, like LASSO and Elastic Net, are able to account for the above mentioned drawbacks.

As a consequence, the expectation is that, at least a number of machine learning models should exhibit higher predictive ability. It is indeed evident from Table 1 that the application of various machine learning techniques produce superior results relative to the yield curve (benchmark) model. Penalized likelihood binary Logit models seem to produce the most consistent out-of-sample recession forecasts, with the LASSO and Elastic Net techniques achieving the highest marks with respect to accuracy. More specifically, both the LASSO and the Elastic Net models exhibit AUC values in excess of 92%, while Accuracy levels are also above the specific mark. Both models have among the highest Pesaran-Timmermann statistics, pointing to statistically significant recession predictability. Additional models with notable performance are the Bayes GLM, the Regularized Discriminant Analysis (RDA), and the k-nn ones, as they rank among the highest performers on numerous evaluation metrics. Interestingly, tree-based techniques, such as Bagging and Random Forest, as well as Adaptive and Gradient Boosting, do not exhibit high predictive accuracy within the specific forecast window. The predictive performance of several representative models is shown in Figures 1 and 2 (Panel (a)).

Please insert Table 2 around here

To provide more insight with regard to predictive variable selection and inclusion, the LASSO and the Elastic Net models will be used as case studies. Table 2 presents the predictive variables that are included at each out-of-sample iteration in the LASSO model, ranked on their underlying frequency of inclusion. More specifically, the table illustrates the variables selected by the LASSO model that achieve at least a 75% rate of inclusion. It is rather interesting that apart from the usual suspects, the yield curve, real money supply growth (M1), the change in the ISM Manufacturing Index, and real GDP growth, the list contains less highlighted predictive variables, such as the Chicago Fed National Financial Conditions Index, the Chicago Fed National Financial Conditions Leverage Sub Index, heavy truck sales, the change in the ratio of residential investment to

GDP. In addition, on the top ranked list are also indicators that are used by the NBER's Cycle Dating Committee to determine periods of expansion and contraction, like real personal income excluding transfers, as well as employment related indicators (initial claims and average hourly earnings). Last but not least, certain producer price indices exhibit consistent predictive ability.

Moving on to the Elastic Net model, there is a great degree of consistency regarding the predictive variables included in the out-of-sample period. The majority of the variables included in the specific model are common, with the exception of the Consumer Sentiment Index, the change in the real money supply M1, and the 3-month Treasury bill rate.

Please insert Table 3 around here

Table 3 presents the out-of-sample performance of the models regarding the medium-term forecast horizon (6-months). It is apparent that, once again, the k-nn and the Regularized Discriminant Analysis models are among the models with the most consistent out-of-sample performance, as they enjoy among the highest rates in terms of forecast accuracy, with misclassification error (MCE) rates at 8.1% and 9.4% respectively, even though the RDA model exhibits lower AUC (74.5%). Moreover, both models exhibit high Pesaran-Timmermann statistics, 8.695 and 7.026, respectively. However, the technique that ranks highest in terms of Accuracy (92.7%), Precision (100%) and has the highest PT-statistic (8.653) is Naive Bayes. It is notable that the Random Forest and the Bayes GLM techniques are the best performing models based on AUC, 92.0% and 91.7%, respectively. Penalized likelihood Logit models continue to demonstrate robust performance, as both Ridge and LASSO models have MCE rates below 10.0%. The Ridge model, however, exhibits the best overall performance. Once again, tree-based techniques are not among the best performing models, with the exception of Random Forest that is within the better performers in the specific forecast window. (Figures 1-2, Panel b).

As expected, the predictive accuracy of the benchmark model is worse compared to that in the long-term forecasting window; both the static binary Logit and Probit models have AUC values close to 76% versus the 84% values in the case of the long-term forecast horizon. The specific finding is consistent with and confirms previous studies, such as Estrella and Mishkin (1998) that show that the yield curve exhibits higher predictive ability at longer forecast horizons and, more recently, Liu and Moench (2016) that reconfirm its superior forecasting accuracy at the 12-month ahead horizon. It is important to note that the number of machine learning techniques that outperform the benchmark model has increased in the medium versus the long-term forecast window.

Please insert Table 4 around here

For comparability reasons, the LASSO and the Elastic Net models are used, once again, as case studies for variable selection. To a large extent, the included variables are similar (Table 4); the main difference between the medium and the long-term horizon models is the inclusion of the default spread, the return of the S&P

500 Index and the change in corporate profits at the expense of labor related variables, such as initial claims, and leading economic indicators (ISM Manufacturing Index), as well as the fact that in some cases variables are included with a shorter lag length (6-months). A primary example are average hourly earnings and heavy truck sales that enter the model with a 6-month lag, instead of a 12-month one. Another difference is that some variables, such as the change in Chicago Fed National Financial Conditions Index, are included at different lag lengths (6 & 12-months). The picture is rather similar in the case of the Elastic Net model; newly included top ranked variables are the 10-year minus 2-year term spread, the 5-year minus 3-month term spread, and the 3-month Treasury T-bill rate.

Please insert Table 5 around here

The main conclusions with respect to model-related predictive accuracy are quite consistent also in the case of the short-term forecast horizon (Table 5). The k-nn model continues to maintain its high predictive accuracy. The Accuracy of the specific model is the highest and close to the 95% mark. In addition, penalized likelihood binary Logit models, once again, display high predictive performance, as Elastic Net and LASSO share low MCE rates (below 7%) and high PT statistics (above 9). The most notable difference compared to the long and medium-term forecast window is the enhanced performance of tree-related techniques. The Random Forest technique clearly ranks among the top performing models with an AUC value over 95.0%, Accuracy close to 94.0% and relatively high sensitivity and specificity rates. Moreover, Bagging exhibits improved performance on several evaluation metrics; the same is true regarding the performance of single-tree models, RPART. In the same spirit, boosting-related models show considerably enhanced performance with high AUC and Accuracy rates. It seems that the predictive ability of tree-based techniques is a function of the underlying forecast horizon, as the shorter the forecast window the higher their predictive accuracy. (Figures 1-2, Panel c). It is clear that all the machine learning techniques display superior forecasting ability versus the benchmark model, indicating that it might be worthwhile to implement a machine learning framework for recession forecasting.

In terms of variable selection the short-term LASSO picks primarily financial market related indicators, such as the default spread, the return of the S&P 500 Index, the yield curve, the change in Chicago Fed National Financial Conditions Index, the change Chicago Fed National Financial Conditions Leverage Sub Index, the change in corporate profits, and the change in the ratio of short-term household liabilities to disposable personal income, as well as producer price related variables (Table 6). In addition, heavy truck sales continue to play a crucial role for recession forecasting as in the cases of the long and medium-term forecasting horizons.

Please insert Table 6 around here

As expected, there are more predictive variables in the Elastic Net model compared to the LASSO model. To a large extent, the included variables are similar to those in its medium-term counterpart; the main differences center on the inclusion of variables like the change in housing permits and the change in the ratio of average GDP plus GDI to potential GDP.

As a final robustness check, the Long, Medium and Short-term Set up is implemented for a very short forecasting window. More specifically, 1-month ahead forecasts are generated for all underlying models and the corresponding performance evaluation metrics are estimated (Table 7, Panels A, B, C). The results largely confirm the superior predictive ability of machine learning techniques over the benchmark model across the different setups. In addition, the consistency of certain techniques, such as k-nn, Bayes GLM, RDA, and penalized models is reaffirmed. Given the shorter forecasting window, tree-related techniques, once again, exhibit high and consistent predictive performance.

The main takeaways of the analysis are twofold: (a) the application of machine learning techniques for recession prediction are a significant step forward compared to more established econometric techniques, irrespective of the forecasting horizon, and (b) there are notable differences in the prediction accuracy of the underlying machine learning models, and especially for tree-based models as they seem to be a function of the underlying forecasting horizon.

6 A Practical Guide for Recession Forecasting

It is evident from the preceding section that the implementation of numerous machine learning techniques in U.S. recession forecasting substantially improves the underlying predictive ability and, thus accuracy, relative to more standard and widely-used econometric models (Logit, Probit models). In addition, it is a fact that there is a certain degree of divergence in the forecasting ability of the underlying machine learning techniques. Having said that, it would be rather cumbersome in practice for both economists and financial analysts to apply numerous machine learning techniques relative to more established econometric techniques, as they are computationally intensive and time consuming.

A practical way to address these issues would be to use a machine learning technique that can perform effective variable selection and use the selected predictive variables as inputs in a more standard econometric framework. As a practical example for the design of an early warning recession framework, the LASSO model is used for out-of-sample variable selection over the long-term forecast horizon (12-months).

More specifically, the predictive variables that are included in, at least, around 90% of the out-of-sample iterations over the 2000-2019 period, form the basic set of 'candidate' predictors that will be employed in the Probit/Logit econometric framework. The final pool of candidate predictors are the following: the yield curve, real money supply growth, real GDP growth, the Chicago Fed National Financial Conditions Leverage Sub Index, average hourly earnings, producer price inflation for intermediate demand, and producer price inflation for metals and metal products, heavy truck sales and the change in the ratio of residential investment to GDP. A kitchen sink Logit and Probit regression is then performed to generate out-of-sample recession probabilities. The results are presented in Table 8.

Please insert Table 8 around here

Both the Logit and the Probit kitchen sink models exhibit superior forecasting ability compared to the yield curve reference model, with AUC values that are over 96.5% (Table 8). A sizeable gain versus the AUC values of the simple yield curve model and an additional confirmation of the LASSO model's effectiveness in automatic variable selection. The specific framework represents an acceptable compromise for handy recession forecasting that could be used by practitioners.

7 Conclusion

The main findings of this study can be summarized as follows; first, consistent with the existing literature, the ability of the yield curve to act as an early warning system for predicting U.S. recessions is reconfirmed. More specifically, the yield curve remains a consistent and reliable recession predictor at the 12-month forecast horizon. It is thus a rational and pragmatic choice to treat forecasts generated by the specific predictor as benchmark or reference models. Having said that the addition of alternative macroeconomic and financial market related predictors enhances recession predictability compared to forecasts based solely on the term spread.

Second and most important, the implementation of machine learning techniques seems to be a clear step forward for U.S. recession prediction compared to mainstream econometric techniques. Generating forecasts from an elaborate set of machine learning techniques improves significantly the predictive ability and accuracy at all forecast horizons in the vast majority of the cases relative to both univariate and multiple Logit and Probit models.

Third, as anticipated, there is a certain degree of divergence in the predictive accuracy of the underlying machine learning models. Model accuracy is not consistent across forecast horizons, some techniques, however, exhibit more consistent relative performance. Penalized likelihood binary models (LASSO, Elastic Net, Ridge), the k-nn method, and the Bayes GLM technique, seem to generate the most reliable forecasts across different forecast horizons, while the Random Forest model is strong in medium and short-term horizons.

Fourth, the analysis employs an elaborate set of predictive variables containing 56 macroeconomic and financial market related indicators, the majority of which have been widely-followed and used in the existing literature for predictability of recession, but introduces a number of less studied indicators, such as the change in the ratio of residential investment to GDP, the change in the ratio of short-term household liabilities to disposable personal income, heavy truck sales and financial conditions indices. These novel predictors appear to contain useful information for upcoming recessions.

Last but not least, as the implementation of machine learning techniques can be computationally intensive and time consuming, a forecasting framework is outlined that represents an intermediate step between

standard econometric and machine learning techniques, in practice.

References

Banachewicz, K., Lucas, A., and Vaart, A.V.D. (2008). Modeling portfolio defaults using hidden Markov models with covariates. Econometrics Journal, 11, 1, 155-171.

Barsoum, F., and Stankiewicz, S. (2015). Forecasting GDP growth using mixed-frequency models with switching regimes. International Journal of Forecasting, 31, 33-50.

Bauer, E., and Kohavi, R. (1999). An Empirical Comparison of Voting Classification Algorithms: Bagging, Boosting, and Variants. Machine Learning, 36, 105-139.

Berge, T.J. (2015). Predicting Recessions with Leading Indicators: Model Averaging and Selection over the Business Cycle. Journal of Forecasting, 34, 455-471.

Berge, T.J., and Jorda, O. (2011). Evaluating the Classification of Economic Activity into Recessions and Expansions. American Economic Journal: Macroeconomics,3,2, 246-277.

Bergmeir, C., Costantini, M., and Benítez, J.M. (2014). On the usefulness of cross-validation for directional forecast evaluation. Computational Statistics and Data Analysis, 76, 132-143.

Biau, G., Devroye, L., and Lugosi, G. (2008). Consistency of random forests and other averaging classifiers. Journal of Machine Learning Research, 9, 2015-2033.

Breiman L. (1996). Bagging predictors. Machine Learning, 24, 2, 123-140.

Breiman L. (1998). Arcing classifiers. The Annals of Statistics, 26, 3, 801-849.

Breiman, L. (2001). Random Forests. Machine Learning, 45, 5-32.

Breiman, L., Friedman, J., Stone, C.J., and Olshen, R.A. (1984). Classification and Regression Trees. CRC press.

Buhlmann, P., and Yu, B. (2002). Analyzing bagging. The Annals of Statistics, 30, 4, 927-961.

Burns, A.F., and Mitchell, W.C. (1946). Measuring Business Cycles. New York: National Bureau of Economic Research.

Canova, F. (1994). Were financial crises predictable? Journal of Money, Credit and Banking, 26, 1, 102-124.

Chauvet, M., and Hamilton, J.D. (2006). Dating business cycle turning points. In Milas, C., Rothman, P., van Dijk, D., and Wildasin, D.E. (Eds.), Nonlinear time series analysis of business cycles: Contributions to economic analysis (pp. 1–54), Vol. 276. Bingley, UK: Emerald Group.

Chauvet, M. and Piger, J. (2008). A Comparison of the Real-Time Performance of Business Cycle Dating Methods. Journal of Business & Economic Statistics, 26, 1, 42-49.

Chauvet, M., and Potter, S. (2005). Forecasting recession using the yield curve. Journal of Forecasting, 24, 77-103.

Davig T., and Smalter Hall, A. (2019). Recession forecasting using Bayesian classification. International Journal of Forecasting, 35, 3, 848-867.

Dietterich, T.G. (2000). An experimental comparison of three methods for constructing ensembles of deci-

sion trees: Bagging, boosting, and randomization. Machine Learning, 40, 2, 139-157.

Domingos, P., and Pazzani, M. (1997). Beyond independence: Conditions for the optimality of the simple Bayesian classifier. Machine Learning, 29, 103-130.

Dopke, J., Fritsche, U., and Pierdzioch, C. (2017). Predicting recessions with boosted regression trees. International Journal of Forecasting, 33, 4, 745-759.

Dueker, M.J. (1997). Strengthening the case for the yield curve as a predictor of U.S. recessions. Federal Reserve Bank of St. Louis Review, 79, 41-51.

Dueker, M.J. (2005). Dynamic forecasts of qualitative variables: A Qual VAR model of U.S. recessions. Journal of Business and Economic Statistics, 23, 96-104.

Efron, B., and Hastie, T. (2016). Computer Age Statistical Inference, Algorithms, Evidence, and Data Science, Cambridge University Press.

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

Freund, Y., and Schapire, R.E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. Journal of Computer and System Sciences, 55, 1, 119-139.

Friedman, J.H. (1989). Regularized discriminant analysis. Journal of the American Statistical Association, 84, 405, 165-175.

Friedman, J.H. (2001). Greedy function approximation: A gradient boosting machine. The Annals of Statistics, 29, 5, 1189-1232.

Friedman, J.H. (2002). Stochastic Gradient Boosting. Computational Statistics and Data Analysis, 38, 4, 367-378.

Friedman J.H., Hastie, T., and Tibshirani R. (2000). Additive logistic regression: a statistical view of boosting. The Annals of Statistics, 28, 2, 337-407.

Gelman, A.Jakulin, A., Pittau, M.G., and Su, Y. (2009). A Weekly Informative Default Prior Distribution for Logistic and other Regression Models. The Annals of Applied Statistics, 2, 4, 1360-1383.

Gogas, P., Papadimitriou, T., Matthaiou, M., and Chrysanthidou, E. (2015). Yield curve and recession forecasting in a machine learning framework. Computational Economics, 45, 4, 635-645.

Gogas, P., Papadimitriou, T., and Chrysanthidou, E. (2015). Yield curve point triplets in recession forecasting. International Finance, 18, 2, 207-226.

Granger, C.W.J., and Pesaran, M.H. (2000). Economic and statistical measures of forecast accuracy. Journal of Forecasting, 19, 7, 537-560.

Gu, S., B., Kelly, and Xiu, D. (2020). Empirical Asset Pricing via Machine Learning. The Review of Financial Studies, 33, 2223-2273.

Hamilton, J.D. (1989). A New Approach to the Economic Analysis of Nonstationary Time Series and the

Business Cycle. Econometrica, 57, 2, 357-384.

Hamilton, J.D. (1990). Analysis of time series subject to changes in regime. Journal of Econometrics, 45, 39-70.

Hamilton, J.D., and Perez-Quiros, G. (1996). What do leading indicators lead? Journal of Business, 69, 1, 27-49.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). The Elements of Statistical Learning, Data Mining, Inference, and Prediction. Springer Series in Statistics, Springer, Berlin.

Hoerl, A.E., and Kennard, R.W. (1970). Ridge regression: Applications to nonorthogonal problems. Technometrics, 12, 1, 69-82.

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013). An introduction to statistical learning, Springer.

Kauppi, H., and Saikkonen, P. (2008). Predicting U.S. recessions with dynamic binary response models. Review of Economics and Statistics, 90, 4, 777-791.

Liu, W., and Moench, E. (2016). What predicts US recessions? International Journal of Forecasting, 32, 1138-1150.

Manela, A., and Moreira, A. (2017). News Implied Volatility and Disaster Concerns. Journal of Financial Economics, 123, 137-162.

Marsh, I.W. (2000). High frequency Markov switching models in the foreign exchange market. Journal of Forecasting, 19, 123-134.

Ng, S. (2014). Boosting Recessions. Canadian Journal of Economics, 47, 1, 1-34.

Nguyen, N. (2018). Hidden Markov Model for Stock Trading. International Journal of Financial Studies, 6, 2, 1-17.

Nyberg, H. (2010). Dynamic probit models and financial variables in recession forecasting. Journal of Forecasting, 29, 215-230.

Nyberg, H. (2011). Forecasting the direction of the US stock market with dynamic binary probit models. International Journal of Forecasting, 27, 561-578.

Nyberg, H. (2018). Forecasting US interest rates and business cycle with a nonlinear regime switching VAR model. Journal of Forecasting, 37, 1-15.

Pesaran, M.H., and Timmermann, A. (1992). A Simple Nonparametric Test of Predictive Performance. Journal of Business and Economic Statistics, 10, 4, 461-465.

Pierdzioch, C., and Gupta, R. (2019). Uncertainty and Forecasts of U.S. Recessions. Studies in Nonlinear Dynamics & Econometrics, in press.

Pinson, P., and Madsen, H. (2012). Adaptive modeling and forecasting of offshore wind power fluctuations with Markov-switching autoregressive models. Journal of Forecasting, 31, 281-313.

Stock, J.H., and Watson, M.W. (1989). New indexes of coincident and leading economic indicators. In

Blanchard, O., and Fischer, S. (Eds.), NBER Macroeconomics Annual, Cambridge: MIT Press, 351-409.

Stock, J.H., and Watson, M.W. (1991). A probability model of the coincident economic indicators. In Lahiri K., and Moore, G.H. (Eds.), Leading economic indicators: New Approaches and forecasting records, Cambridge: MIT Press.

Stock, J.H., and Watson, M.W. (1993). A procedure for predicting recessions with leading indicators: Econometric issues and recent experience. In Stock, J.H., and Watson, M.W. (Eds.), Business cycles, indicators and forecasting, Chicago: University of Chicago Press for NBER, 255-284.

Therneau, T.M., and Atkinson, E.J. (2019). An Introduction to Recursive Partitioning Using the RPART Routines, Mayo Foundation.

Tibshirani, R. (1996). Regression shrinkage and selection via the LASSO. Journal of the Royal Statistical Society, Series B, 267-288.

Tian R., and Shen G. (2019). Predictive power of Markovian models: Evidence from US recession forecasting. Journal of Forecasting, 1–27.

Zou, H., and Hastie, T. (2005). Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society, Series B, 67, 2, 301-320.

Table (1) Performance Evaluation Measures – Long-Term (12-month ahead forecasts)

Method	AUC	MCE	Accuracy	Kappa	Sensitivity	Specificity	Precision	$F_1 Score$	Balanced Accuracy	Pesaran - Timmermann
Panel A: Logit and Probit Models										
Logit-YC	0.838	0.103	0.897	0.289	0.231	0.981	0.600	0.333	0.606	5.028
Probit-YC	0.838	0.107	0.893	0.242	0.192	0.981	0.556	0.286	0.587	4.327
Panel B: Penalized Logit Models - $\lambda \in (10^{-5}, 10^{-1}, step: 10^{-4})$										
LASSO	0.928	0.060	0.940	0.687	0.692	0.971	0.750	0.720	0.832	10.513
Ridge	0.920	0.094	0.906	0.428	0.385	0.971	0.625	0.476	0.678	6.777
EN	0.921	0.073	0.927	0.626	0.654	0.962	0.680	0.667	0.808	9.577
Panel C: Penalized Logit Models - $\log(\lambda) \in (-11, -1, step: 0.01)$										
LASSO	0.953	0.051	0.949	0.740	0.769	0.971	0.769	0.769	0.870	11.326
Ridge	0.921	0.094	0.906	0.428	0.385	0.971	0.625	0.476	0.678	6.777
EN	0.930	0.077	0.923	0.597	0.615	0.962	0.667	0.640	0.788	9.142
Panel D: Discriminant Analysis Models										
Linear DA	0.784	0.222	0.778	0.299	0.692	0.788	0.290	0.409	0.740	5.237
Regular. DA	0.825	0.077	0.923	0.512	0.423	0.986	0.786	0.550	0.704	8.283
Panel E: Bayesian Models										
Bayes GLM	0.966	0.077	0.923	0.550	0.500	0.976	0.722	0.591	0.738	8.587
Naive Bayes	0.767	0.115	0.885	-0.008	0.000	0.995	0.000	0.000	0.498	-0.354
Panel F: Trees and Forests										
rpart	0.703	0.111	0.889	0.193	0.154	0.981	0.500	0.235	0.567	3.561
Bagging	0.665	0.274	0.726	-0.011	0.192	0.793	0.104	0.135	0.493	-0.172
Random Forest	0.834	0.128	0.872	0.069	0.077	0.971	0.250	0.118	0.524	1.272
Panel G: Boosting and k-nn										
AdaBoost	0.473	0.124	0.876	-0.024	0.000	0.986	0.000	0.000	0.493	-0.616
gbm	0.765	0.120	0.880	-0.016	0.000	0.990	0.000	0.000	0.495	-0.502
k-nn	0.806	0.068	0.932	0.629	0.615	0.971	0.727	0.667	0.793	9.661

The table reports the performance evaluation measures of the long-term horizon (12-month step ahead) forecasts obtained by standard Logit/Probit models and several machine learning modeling approaches for the out-of-sample period (2000:1-2019:6).

Table (2) LASSO Model – Long-Term Performance (12-month ahead forecasts)

Variable	Lag	Inclusion frequency (# of times)	Inclusion frequency (%)
Yield Curve (10y-3m) (%)	12	234	100.0%
Average Hourly Earnings Total Private yoy (%)	12	234	100.0%
Producer Price Index by Commodity for Intermediate Demand by Commodity Type: Unprocessed Goods for Intermediate Demand (yoy %)	12	234	100.0%
Producer Price Index by Commodity Metals and metal products: Primary nonferrous metals (yoy %)	12	224	95.7%
Chicago Fed National Financial Conditions Leverage SubIndex	12	217	92.7%
Change in ratio of residential investment to GDP (yoy, %)	12	216	92.3%
Heavy truck sales (D, #)	12	212	90.6%
Real Money Supply M2 (%)	12	209	89.3%
RGDP Growth (yoy %)	12	209	89.3%
Real Personal Income excluding Transfers (yoy %)	12	207	88.5%
Initial Claims (#, Month Av)	12	200	85.5%
Chicago Fed National Financial Conditions Index	12	193	82.5%
Real Money Supply M1 (%)	12	188	80.3%
Change in Chicago Fed National Financial Conditions Index	12	183	78.2%
ISM Manufacturing Index - monthly difference (%)	12	179	76.5%

The table reports the most relevant predictive variables and their underlying frequency of inclusion obtained by the LASSO model regarding the long-term horizon (12-month ahead) forecasts for the out-of-sample period (2000:1-2019:6).

Table (3) Performance Evaluation Measures - Medium-Term (6-month ahead forecasts)

Method	AUC	MCE	Accuracy	Kappa	Sensitivity	Specificity	Precision	$F_1 Score$	Balanced Accuracy	Pesaran - Timmermann
Panel A: Logit and Probit Models										
Logit-YC	0.759	0.124	0.876	0.230	0.231	0.957	0.400	0.293	0.594	3.680
Probit-YC	0.758	0.137	0.863	0.053	0.077	0.962	0.200	0.111	0.519	0.914
Panel B: Penalized Logit Models - $\lambda \in (10^{-5}, 10^{-1}, step: 10^{-4})$										
LASSO	0.743	0.098	0.902	0.334	0.269	0.981	0.636	0.378	0.625	5.678
Ridge	0.830	0.098	0.902	0.436	0.423	0.962	0.579	0.489	0.692	6.770
EN	0.718	0.111	0.889	0.324	0.308	0.962	0.500	0.381	0.635	5.128
Panel C: Penalized Logit Models - $\log(\lambda) \in (-11, -1, step: 0.01)$										
LASSO	0.755	0.098	0.902	0.334	0.269	0.981	0.636	0.378	0.625	5.678
Ridge	0.758	0.090	0.910	0.465	0.423	0.971	0.647	0.512	0.697	7.302
EN	0.722	0.103	0.897	0.349	0.308	0.971	0.571	0.400	0.639	5.652
Panel D: Discriminant Analysis Models										
Linear DA	0.640	0.188	0.812	0.305	0.577	0.841	0.313	0.405	0.709	4.980
Regular. DA	0.745	0.094	0.906	0.450	0.423	0.966	0.611	0.500	0.695	7.026
Panel E: Bayesian Models										
Bayes GLM	0.917	0.111	0.889	0.324	0.308	0.962	0.500	0.381	0.635	5.128
Naive Bayes	0.866	0.073	0.927	0.485	0.346	1.000	1.000	0.514	0.673	8.653
Panel F: Trees and Forests										
rpart	0.730	0.128	0.872	-0.031	0.000	0.981	0.000	0.000	0.490	-0.713
Bagging	0.820	0.184	0.816	0.192	0.346	0.875	0.257	0.295	0.611	2.981
Random Forest	0.920	0.103	0.897	0.422	0.423	0.957	0.550	0.478	0.690	6.531
Panel G: Boosting, k-nn										
AdaBoost	0.636	0.115	0.885	0.252	0.231	0.966	0.462	0.308	0.599	4.137
gbm	0.825	0.128	0.872	0.278	0.308	0.942	0.400	0.348	0.625	4.299
k-nn	0.815	0.081	0.919	0.567	0.577	0.962	0.652	0.612	0.769	8.695

The table reports the performance evaluation measures of the medium-term horizon (6-month step ahead) forecasts obtained by standard Logit/Probit models and several machine learning modeling approaches for the out-of-sample period (2000:1-2019:6).

Table (4) LASSO Model - Medium-Term Performance (6-month ahead forecasts)

Variable	Lag	Inclusion frequency (# of times)	Inclusion frequency (%)
Yield Curve (10y-3m) (%)	12	234	100.0%
Average Hourly Earnings Total Private yoy (%)	6	214	91.5%
Producer Price Index by Commodity for Intermediate Demand by Commodity Type: Unprocessed Goods for Intermediate Demand (yoy %)	12	209	89.3%
Producer Price Index by Commodity Metals and metal products: Primary nonferrous metals (yoy %)	12	196	83.8%
Real Money Supply M1 (%)	12	188	80.3%
RGDP Growth (yoy %)	12	188	80.3%
Corporate Profits After Tax (yoy %)	6	188	80.3%
Change in Chicago Fed National Financial Conditions Index	6	188	80.3%
Chicago Fed National Financial Conditions Leverage SubIndex	12	188	80.3%
Heavy truck sales (D, #)	6	186	79.5%
S&P 500 Index monthly return (%)	6	185	79.1%
Moodys BAA Yield over 10-year Treasury Yield (Credit Spread)	6	184	78.6%
Change in Chicago Fed National Financial Conditions Index	12	180	76.9%
Change in ratio of residential investment to GDP (yoy, %)	12	176	75.2%

The table reports the most relevant predictive variables and their underlying frequency of inclusion obtained by the LASSO model regarding the medium-term horizon (6-month ahead) forecasts for the out-of-sample period (2000:1-2019:6).

Table (5) Performance Evaluation Measures – Short-Term (3-month ahead forecasts)

Method	AUC	MCE	Accuracy	Kappa	Sensitivity	Specificity	Precision	$F_1 Score$	Balanced Accuracy	Pesaran - Timmermann
Panel A: Logit and Probit Models										
Logit-YC	0.805	0.124	0.876	0.230	0.231	0.957	0.400	0.293	0.594	3.680
Probit-YC	0.801	0.128	0.872	0.069	0.077	0.971	0.250	0.118	0.524	1.272
Panel B: Penalized Logit Models - $\lambda \in (10^{-5}, 10^{-1}, step: 10^{-4})$										
LASSO	0.907	0.068	0.932	0.629	0.615	0.971	0.727	0.667	0.793	9.661
Ridge	0.920	0.081	0.919	0.534	0.500	0.971	0.684	0.578	0.736	8.293
EN	0.925	0.064	0.936	0.646	0.615	0.976	0.762	0.681	0.796	9.946
Panel C: Penalized Logit Models - $\log(\lambda) \in (-11, -1, step: 0.01)$										
LASSO	0.912	0.064	0.936	0.646	0.615	0.976	0.762	0.681	0.796	9.946
Ridge	0.922	0.081	0.919	0.534	0.500	0.971	0.684	0.578	0.736	8.293
EN	0.923	0.060	0.940	0.663	0.615	0.981	0.800	0.696	0.798	10.251
Panel D: Discriminant Analysis Models										
Linear DA	0.754	0.141	0.859	0.414	0.615	0.889	0.410	0.492	0.752	6.512
Regular. DA	0.846	0.068	0.932	0.642	0.654	0.966	0.708	0.680	0.810	9.827
Panel E: Bayesian Models										
Bayes GLM	0.928	0.085	0.915	0.519	0.500	0.966	0.650	0.565	0.733	8.019
Naive Bayes	0.888	0.077	0.923	0.597	0.615	0.962	0.667	0.640	0.788	9.142
Panel F: Trees and Forests										
rpart	0.676	0.094	0.906	0.428	0.385	0.971	0.625	0.476	0.678	6.777
Bagging	0.928	0.094	0.906	0.593	0.769	0.923	0.556	0.645	0.846	9.224
Random Forest	0.954	0.064	0.936	0.681	0.731	0.962	0.704	0.717	0.846	10.417
Panel G: Boosting, k-nn										
AdaBoost	0.910	0.068	0.932	0.629	0.615	0.971	0.727	0.667	0.793	9.661
gbm	0.947	0.077	0.923	0.567	0.538	0.971	0.700	0.609	0.755	8.763
k-nn	0.889	0.051	0.949	0.740	0.769	0.971	0.769	0.769	0.870	11.326

The table reports the performance evaluation measures of the short-term horizon (3-month ahead) forecasts obtained by standard Logit/Probit models and several machine learning modeling approaches for the out-of-sample period (2000:1-2019:6).

Table (6) LASSO Model – Short-Term Performance (3-month ahead forecasts)

Variable	Lag	Inclusion frequency (# of times)	Inclusion frequency (%)
Yield Curve (10y-3m) (%)	12	232	99.1%
S&P 500 Index monthly return (%)	6	232	99.1%
Producer Price Index by Commodity Metals and metal products: Primary nonferrous metals (yoy %)	6	232	99.1%
Producer Price Index by Commodity for Intermediate Demand by Commodity Type: Unprocessed Goods for Intermediate Demand (yoy %)	3	229	97.9%
Yield Curve (10y-3m) (%)	6	215	91.9%
Change in Chicago Fed National Financial Conditions Index	3	215	91.9%
Producer Price Index by Commodity for Intermediate Demand by Commodity Type: Unprocessed Goods for Intermediate Demand (yoy %)	12	215	91.9%
Heavy truck sales (D, #)	6	211	90.2%
Change in ratio of average (GDP+GDI) to potential GDP	3	210	89.7%
Chicago Fed National Financial Conditions Leverage SubIndex	12	206	88.0%
Corporate Profits After Tax (yoy %)	6	194	82.9%
Producer Price Index by Commodity Metals and metal products: Primary nonferrous metals (yoy %)	12	194	82.9%
RGDP Growth (mom %)	3	192	82.1%
Real Money Supply M2 (%)	12	189	80.8%
RGDP Growth (yoy %)	12	189	80.8%
Moodys BAA Yield over 10-year Treasury Yield (Credit Spread)	3	188	80.3%
Change in ratio of short term household liabilities to Disposable Personal Income (yoy %)	3	187	79.9%
Real Money Supply M2 (%)	3	186	79.5%
Heavy truck sales (D, #)	3	185	79.1%

The table reports the most relevant predictive variables and their underlying frequency of inclusion obtained by the LASSO model regarding the short-term horizon (3-month ahead) forecasts for the out-of-sample period (2000:1-2019:6).

Table (7a) Performance Evaluation Measures (1-month ahead forecasts - Long- Term Setup – Lag 12

Method	AUC	MCE	Accuracy	Kappa	Sensitivity	Specificity	Precision	$F_1 Score$	Balanced Accuracy	Pesaran - Timmermann
Panel A: Logit and Probit Models										
Logit-YC	0.878	0.094	0.906	0.349	0.269	0.986	0.700	0.389	0.627	6.056
Probit-YC	0.876	0.103	0.897	0.255	0.192	0.986	0.625	0.294	0.589	4.706
Panel B: Penalized Logit Models - $\lambda \in (10^{-5}, 10^{-1}, step: 10^{-4})$										
LASSO	0.988	0.034	0.966	0.833	0.885	0.976	0.821	0.852	0.930	12.747
Ridge	0.982	0.051	0.949	0.749	0.808	0.966	0.750	0.778	0.887	11.465
EN	0.977	0.034	0.966	0.833	0.885	0.976	0.821	0.852	0.930	12.747
Panel C: Penalized Logit Models - $\log(\lambda) \in (-11, -1, step: 0.01)$										
LASSO	0.986	0.026	0.974	0.870	0.885	0.986	0.885	0.885	0.935	13.311
Ridge	0.982	0.047	0.953	0.773	0.846	0.966	0.759	0.800	0.906	11.854
EN	0.977	0.034	0.966	0.833	0.885	0.976	0.821	0.852	0.930	12.747
Panel D: Discriminant Analysis Models										
Linear DA	0.951	0.051	0.949	0.749	0.808	0.966	0.750	0.778	0.887	11.465
Regular. DA	0.965	0.034	0.966	0.833	0.885	0.976	0.821	0.852	0.930	12.747
Panel E: Bayesian Models										
Bayes GLM	0.987	0.034	0.966	0.833	0.885	0.976	0.821	0.852	0.930	12.747
Naive Bayes	0.945	0.051	0.949	0.740	0.769	0.971	0.769	0.769	0.870	11.326
Panel F: Trees and Forests										
rpart	0.829	0.081	0.919	0.551	0.538	0.966	0.667	0.596	0.752	8.491
Bagging	0.969	0.051	0.949	0.757	0.846	0.962	0.733	0.786	0.904	11.614
Random Forest	0.964	0.043	0.957	0.791	0.846	0.971	0.786	0.815	0.909	12.106
Panel G: Boosting, k-nn										
AdaBoost	0.961	0.043	0.957	0.776	0.769	0.981	0.833	0.800	0.875	11.884
gbm	0.974	0.051	0.949	0.740	0.769	0.971	0.769	0.769	0.870	11.326
k-nn	0.915	0.034	0.966	0.827	0.846	0.981	0.846	0.846	0.913	12.649

The table reports the performance evaluation measures of the 1-month ahead forecasts obtained by standard Logit/Probit models and several machine learning modeling approaches, using only 12-month lagged values of the predictive variables, for the out-of-sample period (2000:1-2019:6).

Table (7b) Performance Evaluation Measures (1-month ahead forecasts - Medium-Term Setup – Lag 6, 12)

Method	AUC	MCE	Accuracy	Kappa	Sensitivity	Specificity	Precision	$F_1 Score$	Balanced Accuracy	Pesaran - Timmermann
Panel A: Logit and Probit Models										
Logit-YC	0.821	0.115	0.885	0.312	0.308	0.957	0.471	0.372	0.632	4.897
Probit-YC	0.819	0.115	0.885	0.219	0.192	0.971	0.455	0.270	0.582	3.713
Panel B: Penalized Logit Models - $\lambda \in (10^{-5}, 10^{-1}, step: 10^{-4})$										
LASSO	0.960	0.038	0.962	0.802	0.808	0.981	0.840	0.824	0.894	12.271
Ridge	0.975	0.043	0.957	0.784	0.808	0.976	0.808	0.808	0.892	11.988
EN	0.979	0.038	0.962	0.802	0.808	0.981	0.840	0.824	0.894	12.271
Panel C: Penalized Logit Models - $\log(\lambda) \in (-11, -1, step: 0.01)$										
LASSO	0.958	0.038	0.962	0.802	0.808	0.981	0.840	0.824	0.894	12.271
Ridge	0.975	0.038	0.962	0.802	0.808	0.981	0.840	0.824	0.894	12.271
EN	0.977	0.038	0.962	0.802	0.808	0.981	0.840	0.824	0.894	12.271
Panel D: Discriminant Analysis Models										
Linear DA	0.937	0.034	0.966	0.827	0.846	0.981	0.846	0.846	0.913	12.649
Regular. DA	0.959	0.030	0.970	0.851	0.885	0.981	0.852	0.868	0.933	13.022
Panel E: Bayesian Models										
Bayes GLM	0.980	0.030	0.970	0.851	0.885	0.981	0.852	0.868	0.933	13.022
Naive Bayes	0.943	0.047	0.953	0.766	0.808	0.971	0.778	0.792	0.889	11.720
Panel F: Trees and Forests										
rpart	0.896	0.111	0.889	0.295	0.269	0.966	0.500	0.350	0.618	4.775
Bagging	0.979	0.034	0.966	0.838	0.923	0.971	0.800	0.857	0.947	12.859
Random Forest	0.975	0.038	0.962	0.809	0.846	0.976	0.815	0.830	0.911	12.371
Panel G: Boosting, k-nn										
AdaBoost	0.965	0.038	0.962	0.795	0.769	0.986	0.870	0.816	0.877	12.189
gbm	0.962	0.038	0.962	0.809	0.846	0.976	0.815	0.830	0.911	12.371
k-nn	0.935	0.034	0.966	0.833	0.885	0.976	0.821	0.852	0.930	12.747

The table reports the performance evaluation measures of the 1-month ahead forecasts obtained by standard Logit/Probit models and several machine learning modeling approaches, using 6 and 12-month lagged values of the predictive variables, for the out-of-sample period (2000:1-2019:6).

Table (7c) Performance Evaluation Measures (1-month ahead forecasts - Short-Term Setup – Lag 3, 6, 12)

Method	AUC	MCE	Accuracy	Kappa	Sensitivity	Specificity	Precision	$F_1 Score$	Balanced Accuracy	Pesaran - Timmermann
Panel A: Logit and Probit Models										
Logit-YC	0.837	0.103	0.897	0.349	0.308	0.971	0.571	0.400	0.639	5.652
Probit-YC	0.832	0.111	0.889	0.264	0.231	0.971	0.500	0.316	0.601	4.401
Panel B: Penalized Logit Models - $\lambda \in (10^{-5}, 10^{-1}, step: 10^{-4})$										
LASSO	0.960	0.043	0.957	0.776	0.769	0.981	0.833	0.800	0.875	11.884
Ridge	0.975	0.038	0.962	0.802	0.808	0.981	0.840	0.824	0.894	12.271
EN	0.971	0.051	0.949	0.731	0.731	0.976	0.792	0.760	0.853	11.199
Panel C: Penalized Logit Models - $\log(\lambda) \in (-11, -1, step: 0.01)$										
LASSO	0.958	0.043	0.957	0.776	0.769	0.981	0.833	0.800	0.875	11.884
Ridge	0.978	0.030	0.970	0.846	0.846	0.986	0.880	0.863	0.916	12.944
EN	0.967	0.043	0.957	0.776	0.769	0.981	0.833	0.800	0.875	11.884
Panel D: Discriminant Analysis Models										
Linear DA	0.934	0.030	0.970	0.856	0.923	0.976	0.828	0.873	0.950	13.117
Regular. DA	0.951	0.026	0.974	0.870	0.885	0.986	0.885	0.885	0.935	13.311
Panel E: Bayesian Models										
Bayes GLM	0.976	0.034	0.966	0.833	0.885	0.976	0.821	0.852	0.930	12.747
Naive Bayes	0.951	0.047	0.953	0.766	0.808	0.971	0.778	0.792	0.889	11.720
Panel F: Trees and Forests										
rpart	0.826	0.073	0.927	0.598	0.577	0.971	0.714	0.638	0.774	9.219
Bagging	0.973	0.034	0.966	0.838	0.923	0.971	0.800	0.857	0.947	12.859
Random Forest	0.977	0.038	0.962	0.815	0.885	0.971	0.793	0.836	0.928	12.486
Panel G: Boosting, k-nn										
AdaBoost	0.981	0.038	0.962	0.795	0.769	0.986	0.870	0.816	0.877	12.189
gbm	0.979	0.034	0.966	0.821	0.808	0.986	0.875	0.840	0.897	12.570
k-nn	0.954	0.021	0.979	0.894	0.923	0.986	0.889	0.906	0.954	13.673

The table reports the performance evaluation measures of the 1-month ahead forecasts obtained by standard Logit/Probit models and several machine learning modeling approaches, using 3, 6 and 12-month lagged values of the predictive variables, for the out-of-sample period (2000:1-2019:6).

Table (8) Practical Guide for Recession Forecasting: Long-Term Performance (12-month ahead forecasts)

Method	AUC	MCE	Accuracy	Kappa	Sensitivity	Specificity	Precision	$F_1 Score$	Balanced Accuracy	Pesaran - Timmermann
Panel A: Logit Models										
Yield Curve	0.838	0.103	0.897	0.289	0.231	0.981	0.600	0.333	0.606	5.028
Nine selected predictors	0.967	0.064	0.936	0.726	0.923	0.938	0.649	0.762	0.930	11.339
Panel B: Probit Models										
Yield Curve	0.838	0.107	0.893	0.242	0.192	0.981	0.556	0.286	0.587	4.327
Nine selected predictors	0.966	0.060	0.940	0.748	0.962	0.938	0.658	0.781	0.950	11.719

The table reports the performance evaluation measures of the long-term horizon (12-month ahead) forecasts obtained by standard Logit/Probit models and by the kitchen sink Logit/Probit regression model for the out-of-sample period (2000:1-2019:6).

Figure 1: Out-of-sample Recession Probabilities for the January 2000 to June 2019 period: Panel (a) displays the predicted recession probabilities based on the Yield Curve Logit, the LASSO, the Bayes GLM, and the k-nn model related to the long-term (12-month ahead) forecast horizon, Panel (b) displays the predicted recession probabilities based on the Yield Curve Logit, the Ridge, the Naive Bayes, the k-nn, and the Random Forest model related to the medium-term (6-month ahead) forecast horizon, Panel (c) displays the predicted recession probabilities based on the Yield Curve Logit, the Elastic Net, the k-nn, and the Random Forest model related to the short-term (3-month ahead) forecast horizon, and Panel (d) displays the predicted recession probabilities based on the Yield Curve Logit, the LASSO, the k-nn, and the Bayes GLM related to the 1-month ahead forecast horizon using only 12-month lagged values of the predictive variables.

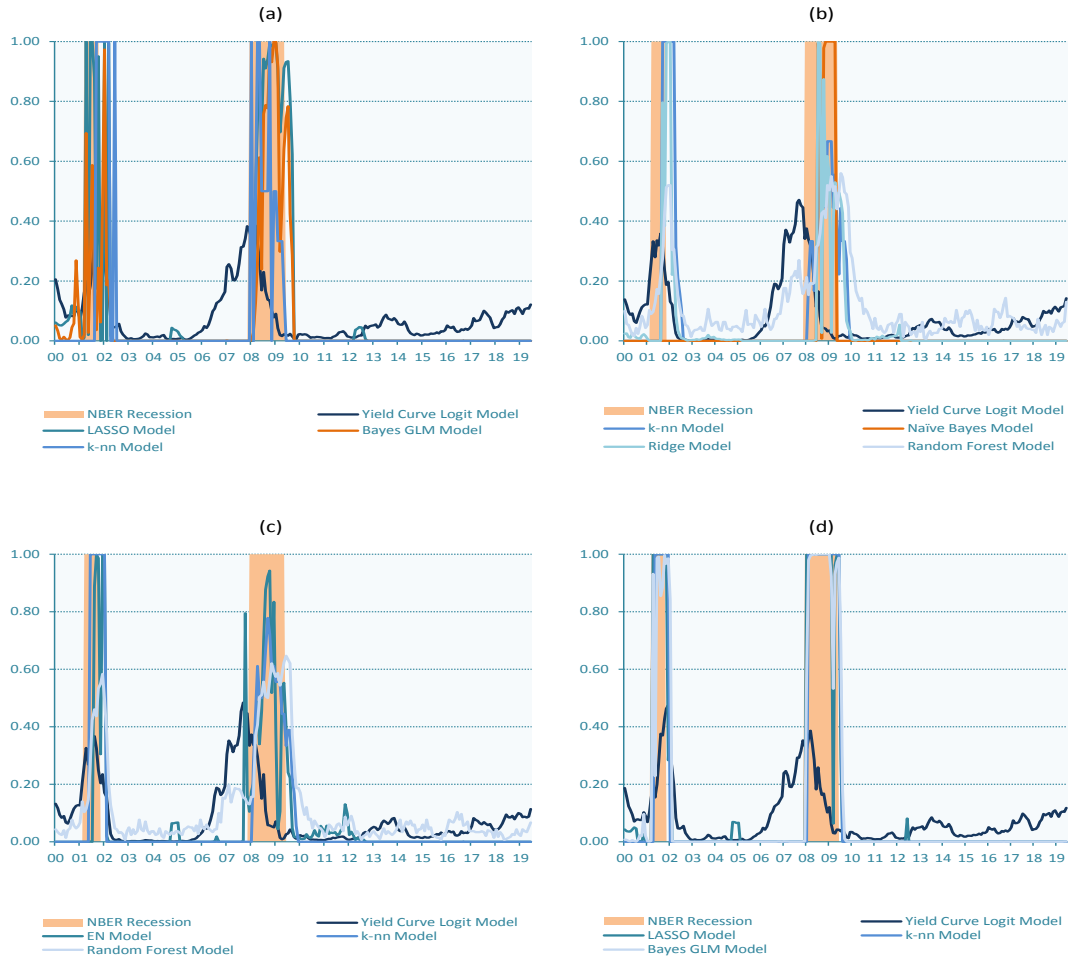
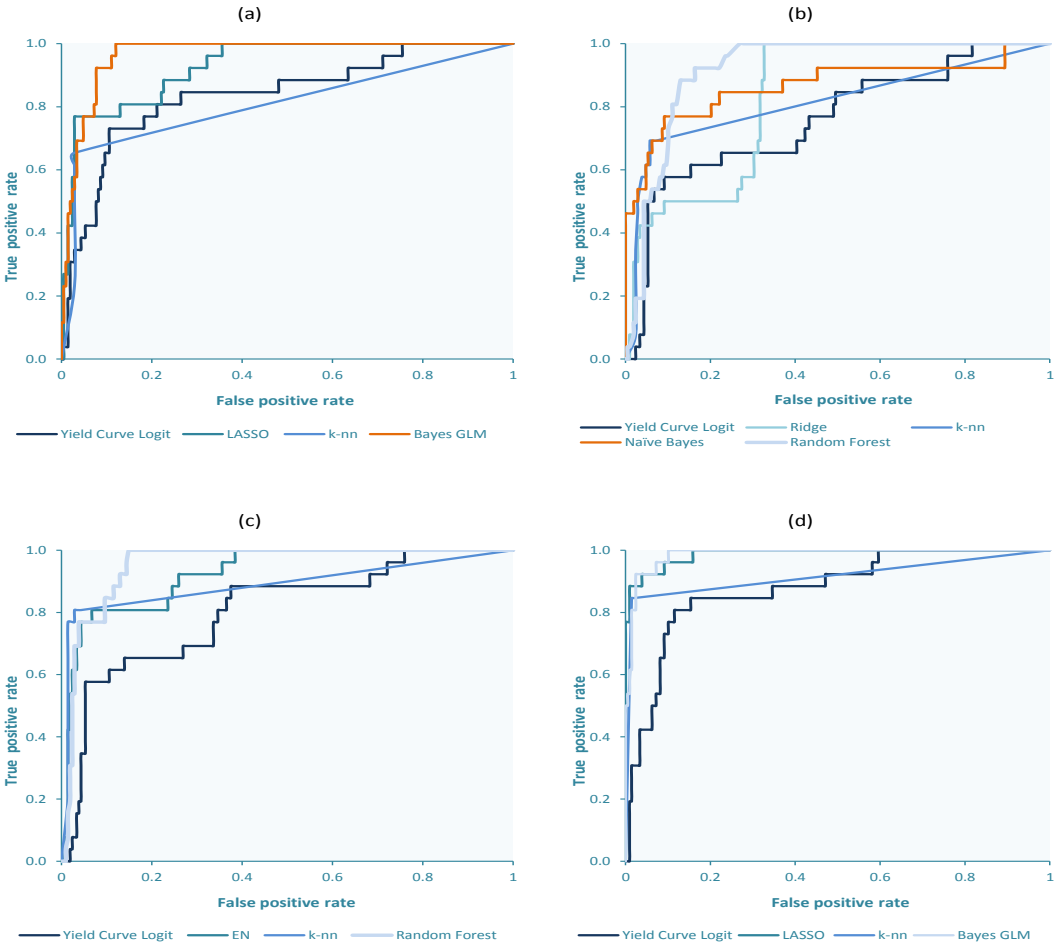


Figure 2: Receiver Operating Characteristic (ROC) curves for the out-of-sample period, January 2000 to June 2019: Panel (a) displays the ROC curves based on the Yield Curve Logit, the LASSO, the Bayes GLM, and the k-nn model related to the long-term (12-month ahead) forecast horizon, Panel (b) displays the ROC curves based on the Yield Curve Logit, the Ridge, the Naive Bayes, the k-nn, and the Random Forest model related to the medium-term (6-month ahead) forecast horizon, Panel (c) displays the ROC curves based on the Yield Curve Logit, the Elastic Net, the k-nn, and the Random Forest model related to the short-term (3-month ahead) forecast horizon, and Panel (d) displays the ROC curves based on the Yield Curve Logit, the LASSO, the k-nn, and the Bayes GLM model related to the 1-month ahead forecast horizon using only 12-month lagged values of the predictive variables.



Appendix

Table (A1) Set of Predictors - Overview

Code	Predictive Variable	Unit	Transformation
1	Term spread - 10-year Treasury yield -3-month Tbill rate	(%)	-
2	Term spread - 10-year Treasury yield - 2-year Treasury yield	(%)	-
3	Term spread - 5-year Treasury yield -3-month Tbill rate	(%)	-
4	10-year Treasury yield	(%)	-
5	3-month Treasury Bill rate	(%)	-
6	Initial Claims	#	month average
7	Initial Claims	#	4-week moving average
8	Rate of unemployment	(%)	-
9	Long-term Non Accelerating Inflation Rate of Unemployment	(%)	-
10	Unemployment Gap - Unemployment rate - Long-term NAIRU	(%)	
11	Average Weekly Hours Total Private	(%)	year-on-year % change
12	Average Hourly Earnings Total Private	(%)	year-on-year % change
13	Non-Farm Payrolls - Total Private	#	-
14	Non-Farm Payrolls - Total Private	(%)	year-on-year % change
15	Real Money Supply M1	(%)	month-on-month % change
16	Real Money Supply M1	(%)	year-on-year % change
17	Real Money Supply M2	(%)	month-on-month % change
18	Real Money Supply M2	(%)	year-on-year % change
19	Moodys BAA Yield	(%)	-
20	Moodys BAA Yield	(%)	monthly difference
21	Credit Spread - Moodys BAA Yield over 10-year Treasury Yield	(%)	-
22	Effective Federal Funds rate	(%)	-
23	S&P 500 Index monthly return	(%)	month-on-month % change
24	Heavy truck sales	#	-
25	Heavy truck sales	#	monthly difference
26	Motor Vehicle Retail Sales: Domestic Autos	#	-
27	Motor Vehicle Retail Sales: Domestic Autos	#	monthly difference

The Table reports detailed information about the set of predictive variables used in the analysis.

Table (A1) Set of Predictors - Overview (Continued)

Code	Predictive Variable	Unit	Transformation
28	Industrial Production Growth	(%)	year-on-year % change
29	Real GDP Growth *	(%)	month-on-month % change
30	Real GDP Growth *	(%)	year-on-year % change
31	Real GDI Growth *	(%)	month-on-month % change
32	Real GDI Growth *	(%)	year-on-year % change
33	Output Gap (Real GDP - Real Potential GDP) *	(%)	-
34	Housing starts	(%)	year-on-year % change
35	Housing permits	(%)	year-on-year % change
36	Ratio of residential investment to GDP *	(X)	-
37	Change in ratio of residential investment to GDP (yoy, %) *	(%)	year-on-year % change
38	Change in ratio of short term household liabilities to Disposable Personal Income (yoy, %)	(%)	year-on-year % change
39	Change in ratio of average (GDP+GDI) to Potential GDP *	(%)	-
40	Corporate Profits	(%)	year-on-year % change
41	Corporate Profits After Tax	(%)	year-on-year % change
42	Chicago Fed National Financial Conditions Index	Z-score	normalized
43	Change in Chicago Fed National Financial Conditions Index	Z-score	monthly difference
44	Chicago Fed National Financial Conditions Leverage SubIndex	Z-score	normalized
45	Change in University of Michigan consumer sentiment	(%)	yearly change
46	Real Personal Income excluding Transfers	(%)	year-on-year % change
47	Real Manufacturing and Trade Industries Sales	(%)	year-on-year % change
48	Capacity Utilization Rate	(%)	-
49	Producer Price Index by Commodity for Final Demand: Finished Goods	(%)	year-on-year % change
50	Producer Price Index by Commodity for Intermediate Demand by Commodity Type: Unprocessed Goods for Intermediate Demand	(%)	year-on-year % change
51	Producer Price Index by Commodity Metals and metal products: Primary nonferrous metals	(%)	year-on-year % change
52	Personal Consumption Expenditures	(%)	year-on-year % change
53	ISM Manufacturing Index	(%)	-
54	ISM Manufacturing Index	(%)	monthly difference
55	ISM Manufacturing Index New Orders	(%)	-
56	ISM Manufacturing Index New Orders	(%)	monthly difference

Notes. * Monthly frequency generated by applying natural cubic spline.
The table reports detailed information about the set of predictive variables used in the analysis.

Table (A2) Model Estimation Details - Hyperparameter Tuning

Models	Method	Tuning parameters	Values of the tuning parameters
Panel A: Logit and Probit Models			
Logit	"glm"	No tuning parameters	—
Probit	"glm"	No tuning parameters	—
Panel B: Penalized Logit Models			
LASSO	"glmnet"	One tuning parameter: λ (α =1.0)	$\lambda \in (10^{-5}, 10^{-1}, step: 10^{-4})$, $\log(\lambda) \in (-11, -1, step: 0.01)$
Ridge	"glmnet"	One tuning parameter: λ (α =0)	$\lambda \in (10^{-5}, 10^{-1}, step: 10^{-4})$, $\log(\lambda) \in (-11, -1, step: 0.01)$
EN	"glmnet"	One tuning parameter: λ (α =0.5)	$\lambda \in (10^{-5}, 10^{-1}, step: 10^{-4})$, $\log(\lambda) \in (-11, -1, step: 0.01)$
Panel C: Discriminant Analysis Models			
Linear DA	"lda"	No tuning parameters	—
Regular. DA	"rda"	Two tuning parameters: γ, λ	$\gamma \in (0.1, 1, step: 0.1)$ $\lambda \in (0, 1, step: 0.1)$
Panel D: Bayesian Models			
Bayes GLM	"bayesglm"	No tuning parameters	—
Naive Bayes	"nb"	Three tuning papameters: fL, $useKernel$, $adjust$	$fL \in \{0.5, 1.0, 1.5, 2.0\}$ $useKernel \in \{TRUE, FALSE\}$ $adjust \in \{0.5, 1.0\}$
Panel E: Trees and Forests			
rpart	"rpart"	One tuning parameter: cp	$cp \in (0, 0.5, step: 0.02)$
Bagging	"treebag"	No tuning parameters	$nbagg = 300$
Random Forest	"rf"	One tuning parameter: $mtry$	$ntree = 300$, $nodesize = 5$, $mtry \in \{2, 3, 5, 10, 20, 30\}$
Panel F: Boosting, k-nn and Recursive Partitioning			
AdaBoost	"ada"	Three tuning papameters: $iter$, $maxdepth$, nu	$iter \in \{400, 500\}$ $maxdepth \in \{1, 2, 3\}$ $nu \in \{0.01, 0.1\}$
gbm	"gbm"	Four tuning papameters: $n.trees$, $interaction.depth$, $shrinkage$, $n.minobsinnode$	$n.trees \in \{400, 500\}$ $interaction.depth \in \{1, 2, 3\}$ $shrinkage \in \{0.01, 0.1\}$ $n.minobsinnode \in \{5, 10\}$
k-nn	"knn"	One tuning parameter: k	$k \in \{1, 2, \ldots, 20\}$

The table outlines the set of hyperparameters and their 'potential' values used for tuning each machine learning model.