Viewpoint: Boosting Recessions

Serena Ng *Department of Economics, Columbia University*

Abstract. This paper explores the effectiveness of boosting, often regarded as the state of the art classification tool, in giving warning signals of recessions 3, 6, and 12 months ahead. Boosting is used to screen as many as 1,500 potentially relevant predictors consisting of 132 real and financial time series and their lags. Estimation over the full sample 1961:1–2011:12 finds that there are fewer than 10 important predictors and the identity of these variables changes with the forecast horizon. There is a distinct difference in the size and composition of the relevant predictor set before and after mid-1980. Rolling window estimation reveals that the importance of the term and default spreads are recession specific. The Aaa spread is the most robust predictor of recessions three and 6 months ahead, while the risky bond and 5-year spreads are important for 12 months ahead predictions. Certain employment variables have predictive power for the two most recent recessions when the interest rate spreads were uninformative. Warning signals for the post-1990 recessions have been sporadic and easy to miss. The results underscore the challenge that changing characteristics of business cycles pose for predicting recessions. JEL classification: C5, C6, C25, C35

Prévoir les récessions. Ce texte explore l'efficacité de la méthode dite du 'boosting', qu'on considère souvent comme un instrument de classification qui est à la fine pointe de l'art de prévoir les récessions 3, 6 et 12 mois à l'avance. Cette méthode est utilisée pour passer au crible quelques 1500 prédicteurs potentiellement pertinents construits à partir de 132 séries chronologiques de variables réelles et financières plus ou moins décalées. Des estimations de l'échantillon complet pour la période du début de 1961 à la fin de 2011 révèlent qu'aussi peu que dix prédicteurs sont importants, et que l'identité de ces variables change selon l'horizon de prévision considéré. Il y a aussi une différence marquée dans la taille et la composition de cet ensemble de prédicteurs avant et après le milieu des années 1980. Il appert que l'importance des écarts de crédit (écarts des taux d'intérêt et des risques de défaut de paiement) est spécifique à la récession particulière. L'écart Aaa est le prédicteur le plus robuste des récessions dans trois et six mois, alors que la débenture risquée et l'écart des taux d'intérêt pour la fenêtre de 5 ans sont les prédicteurs importants pour un horizon temporel de 12 mois. Certaines variables reliées à l'emploi ont eu un certain pouvoir de prédiction pour les deux dernières récessions quand les écarts de taux d'intérêt n'ont pas été éclairants. Les signaux des clignotants pour les récessions d'après 1990 ont

I would like to thank the organizers for the opportunity to present this paper as a State of the Art Lecture at the 2013 Canadian Economics Association Meeting in Montreal, Quebec. Financial support from the National Science Foundation (SES-0962431) is gratefully acknowledged. E-mail: serena.ng@columbia.edu

Canadian Journal of Economics/Revue canadienne d'Economique, Vol. 47, No. 1
February / février 2014. Printed in Canada / Imprimé au Canada

0008-4085 / 14 / 1–34 /

été sporadiques et faciles à manquer. Les résultats soulignent le défi que posent à ceux qui font des prévisions de récessions les caractéristiques changeantes des cycles économiques.

1. Introduction

The central theme of business cycle analysis is to study the reasons why the economy goes through periods of contractions and expansions. In order to do so, we need to document features in the historical data during these two phases of economic activity. This involves determining the dates when recessions began and ended or, in other words, establishing the business cycle chronology. In the United States, this task falls to the NBER Business Cycle Dating Committee (Committee 2008), while the Center for Economic Policy Research (CEPR) has taken up this responsibility for the euro area since 2002. In Canada, the Business Cycle Council of the C.D. Howe Institute not only dates but also grades the severity of each recession. Cross and Bergevin (2012) find that at least for Canada, the 1929 recession was the only one in a century of data deemed to be a category five.

Recessions are understood at a general level to be periods of significant, persistent, and pervasive declines of economic activity, while expansions are periods of prosperity. However, there is no objective measure of economic activity, nor are the notions of pervasiveness and persistence universally defined. Wikipedia cites two consecutive quarters of negative GDP growth, or a 1.5% rise in unemployment within 12 months, as possible definitions of a recession. The U.K. Treasury simply calls a recession when there are two or more consecutive quarters of contraction in GDP. A U.S. recession is officially defined to be the period between a peak and a trough, and an expansion is the period between a trough and a peak. The turning points are then determined by considering monthly industrial production, employment, real income, as well as manufacturing and wholesale-retail sales. The rationale for not focusing on GDP data is that the monthly estimates tend to be noisy, and the quarterly data can be subject to large revisions. Indeed, when the NBER announced that the U.S. was in a recession at the start of 2008, quarterly GDP growth was still positive.

These recession announcements are important signals about state of the economy and tend to receive a good deal of public attention. However, the committees do not have explicit models or formulas for how they arrived at the dates, and furthermore, the announcements were made retroactively. For example, the NBER announced that economic activity peaked in December 2008 and bottomed out in September 2010, more than a full year after activity actually peaked (i.e., in July 2007) and bottomed (i.e., in June 2009.) This has spawned a good deal of interest in providing a formal analysis of the business cycle chronology in the hope that a better understanding of the past would enable better predictions of future recessions, even though such events cannot be totally avoided. But three features make the exercise challenging. First, the true duration and turning

points of business cycles remain unknown even after the fact. A model could be seen to give a wrong classification relative to the announced dates, but such false positives could be valuable signals of what lies ahead. As such, there is no unique criterion to validate the model predictions. This issue is especially relevant for the U.S., since its reference cycle is not based on any observed variable per se. Second, recessions are time dependent, not single-period events, and there are far fewer recessions than non-recession periods. In the U.S., only 15% of the observations between 1961 and 2012 are deemed to be recession months, which can affect our ability to identify the recessions from the data. Third, while the committees officially look at a small number of variables, it is almost surely the case that many other series are unofficially monitored. A researcher typically pre-selects a few predictors for analysis. Omitting relevant information is a distinct possibility.

But what does an econometrician with lots of data at his disposal have to offer to policy makers on the issue of which variables to monitor? The question is useful even if the answer is "not much," because we would then know that information has been used exhaustively. With this in mind, this paper considers the usefulness of a methodology known as *boosting* in giving warning signals of recessions and, in so doing, identify the predictors of recessions in the U.S. over the sample 1961:1 to 2011:12. Boosting is an ensemble scheme that combines models that do not perform particularly well individually into one with much improved properties. It was originally developed as a classification rule to determine, for example, if a message is spam or if a tumour is cancerous given gene expression data. Subsequent analysis shows that boosting algorithms are useful beyond precise classification. The two features of boosting algorithms that drew my attention are their abilities to perform estimation and variable selection simultaneously, and to entertain a large number of predictors. If N is the number of potential predictors and T is the number of time series observations, boosting allows N to be larger than T.

Boosting is applied in this paper to the problem of predicting recessions. In line with the ensemble nature of boosting, the recession probability estimates are based on a collection of logit models. In my application, each model has only one predictor. This is unlike standard logit models that put all predictors into a single model. The application to recession dates creates two interesting problems, both relating to the dependent nature of the data. The first arises from the fact that some variables lead, some lag, while others move concurrently with the reference cycle. A predictor may be useful at one lag and not at another. The second problem is that parameter instability is a generic feature of economic time series, and the relevant predictor set may well evolve over time. I adapt existing boosting algorithms to accommodate these problems.

The analysis aims to shed light on three problems. The first is to identify which variables and at which lags are informative about recessions. The second is to understand if predictors are recession and horizon specific. The third is to

learn more about the characteristics of recent recessions. I find that a handful of variables are systematically important predictors over the 50-year period, but their relative importance has changed over time. While the model provides warning signals for the post-1990 recessions, the signals especially of the 2008 recession are sporadic and easy to miss.

The rest of the paper proceeds as follows. Section 2 begins with a review of existing work on business cycle dating. Section 3 then turns to Adaboost – the algorithm that initiated a huge literature in machine learning – before turning to recent boosting algorithms that can be justified on statistical grounds. The empirical analysis is presented in Section 4. Boosting is far from perfect for analyzing recessions. The paper concludes with suggestions for future work.

2. Related Work

Academic research on business cycle chronology takes one of two routes: fit existing models using better predictors, or find better models taking a small number of predictors as given. This section gives a brief overview of this work. A more complete survey can be found in Marcellino (2006), Hamilton (2011), and Stock and Watson (2010b).

Let Y_t^* be the latent state of the economy. We observe $Y_t = 1$ (as determined by the NBER, for example) only if period t is in a recession and zero otherwise. That is,

$$Y_t = 1 \quad \text{if } Y_t^* > c_*,$$

where c_* is an unknown threshold. As Y_t^* is not observed, it seems natural to replace it by observed indicators x_t, allowing for the relation between Y^* and x to be phase shifted by h periods. A model for recession occurence would then be

$$Y_t = 1 \quad \text{if } x_{t-h} > c_x.$$

Once x is chosen, a binomial likelihood can be maximized and the estimated probability for $Y_t = 1$ can be used for classification, given some user-specified threshold c_x. The simplest approach is to take x_t to be scalar. Popular choices of x_t are GDP and labour market data such as unemployment. These variables are also officially monitored by various dating committees. Lahiri and Yang (2013) provide a review of the literature on forecasting binary outcomes.

An increase in the short rate is indicative of economic slowdown due to monetary policy tightening. Because some recessions in the past are of monetary policy origin, interest rate spreads have been a popular recession predictor. Indeed, recessions tend to be preceded by an inverted yield curve, with short-term rates

higher than long-term rates.[1] The difference between the 10-year and a short-term rate on Treasury bills was used in work by Estrella and Mishkin (1998), Chauvet and Hamilton (2006), Wright (2006), Rudebusch and Williams (2009), among others. Also popular are spreads between corporate bonds of different grades (such as Baa and Aaa bonds) and a risk-free rate. These are considered to be measures of liquidity risk (of selling in thin markets) and credit risk (of default). They are countercyclical, as defaults and bankruptcies are more prevalent during economic downturns. A shortcoming of credit spreads is that the definition of the credit ratings varies over time. On the other hand, credit spreads could better reflect the changing developments in financial markets.

Data provided by the Institute for Supply Management (ISM) are also widely watched indicators of business activity, as documented in Klein and Moore (1991), Dasgupta and Lahiri (1993). The ISM surveys purchasing managers of 250 companies in 21 industries about new orders, production, employment, deliveries, and inventory. A weighted average of the five components (in decreasing order of importance) is used to construct a purchasing manager index (PMI), which is interpreted as a measure of excess demand. The ISM also produces a NAPM price index measuring the fraction of respondents reporting higher material prices. The index is seen as an inflation indicator. The NAPM data have two distinct advantages: the data are released on the first business day after the end of the month for which they are indicating, and they are not subject to revisions.

The exercise of predicting recessions would be easy if we had perfect indicators of economic activity, but this, of course, is not the case. As noted earlier, GDP growth was positive when the 2008 financial crisis was in full swing. The NAPM data are limited to manufacturing business activity, which is narrow in scope, especially when manufacturing has been a declining fraction of overall economic activity. The risky spread between commercial paper and Treasury bills worked well prior to 1990 but failed to predict the 1990–91 recession. The yield curve was inverted in August 2006, but as Hamilton (2011) pointed out, this recession signal is at odds with the fact that the level of the three-month rate was at a historical low. Stock and Watson (2001) reviewed evidence of various asset prices and found their predictive power not to be robust. Gilchrist and Zakrajsek (2012) documented that spreads between yields of bonds traded in the secondary market and a risk-free rate are better predictors than standard default spreads, especially in the 2008 recession. The finding is consistent with the view that business cycles are not alike.

An increasing number of studies have gone beyond using a single proxy to incorporate more information by way of diffusion indexes. These are scalar variables constructed as weighted averages of many indicators of economic activity. Examples include CFNAI (Chicago Fed National Activity Index) comprising 85 monthly indicators; the ADS index of Aruoba, Diebold, and Scotti (2009),

1 The importance of the term spread in recession analysis is documented in http://www.newyorkfed.org/research/capital_markets/ycfaq.html.

which tracks movements of stocks and flows data at high frequencies. These diffusion indexes are typically based on a dynamic factor model in which N data series collected into a vector x_t are driven by a common cyclical variable F_t and idiosyncratic shocks. The estimation exercise consists of extracting F_t from x_t. Stock and Watson (1989) used data for $N = 4$ series to estimate parameters of the model by maximum likelihood. A recession is declared if F_t follows a pattern, such as if $(F_{t-1}, F_{t-2}, \ldots, F_{t-8})$ is in a set designed to mimic the NBER recession dates. This work was re-examined in Stock and Watson (2010a) using different ways to estimate the coincident indexes.

A popular parametric framework for analyzing the two phases of business cycles is the Markov switching model of Hamilton (1989). Chauvet (1998) extends the four-variable dynamic factor model of Stock and Watson (1989) to allow for Markov switching properties. Monthly updates of the smoothed recession probabilities are published on the author's website and in the FRED database maintained by the Federal Reserve Bank of St. Louis. Chauvet and Hamilton (2006) takes as starting point that the mean growth rate of GDP is state dependent, being lower during recession than non-recession months. The sole source of serial dependence in GDP growth is attributed to the persistence in the recession indicator, but the regime-specific parameters are constant over time. Like a dynamic factor model, inference about the latent state can be obtained without directly using the NBER recession dates. Chauvet and Hamilton (2006) suggest that a recession be declared when the smoothed probability exceeds 0.65, and that the recession ends when the probability falls below 0.35. Turning points are then the dates when the probabilities cross a threshold.

Of the non-parametric methods, the algorithm by Bry and Boschan (1971) developed some 40 years ago remains to be widely used. The algorithm first identifies peaks as observations in the 12-month moving-average of a series that are lower over a two-sided window of five months. Analogously, troughs are points associated with observations in the five-month window that are higher. The algorithm then applies censoring rules to narrow the turning points of the reference cycle. In particular, the duration must be no less than 15 months, while the phase (peak to trough or trough to peak) must be no less than five months. The Bry-Boschan algorithm treats expansions and recessions symmetrically. Moench and Uhlig (2005) modified the algorithm to allow for asymmetries in recessions and expansions. They find that the identified number of recessions is sensitive to the censoring rules on phase length. It is noteworthy that the censoring rules have not changed since 1971, even though the economy has changed in many dimensions.

Various papers have compared the accuracy of different methods proposed. Chauvet and Piger (2008) used a real time data set with x_t consisting of employment, industrial production, manufacturing and trade sales, and real personal, income, as in Stock and Watson (1989). They find that both parametric and non-parametric methods produce a few false positives and can identify NBER troughs almost six to eight months before the NBER announcements, but these

methods cannot improve upon the NBER timing in terms of calling the peaks of expansions. Berge and Jorda (2011) compared the adequacy of the components of diffusion indices vis-à-vis the indices themselves. They find that the turning points implied by the diffusion indices are well aligned with the NBER dates. However, the diffusion indices do not predict future turning points well. Within a year, their predictions are no better than a coin toss. However, some components of the Conference Board's leading indicator, notably term spreads and new orders for consumer goods, seem to have some information about future recessions.

Most studies in this literature make use of a small number of highly aggregated predictors. But calling recessions using aggregate data is different from calling recessions by aggregating information from disaggregate series, as originally suggested by Burns and Mitchell (1946). This is likely due to the computational difficulty in parametrically modelling a large number of series, and also that the Bry-Boschan algorithm is designed for univariate analysis. Harding and Pagan (2006) suggest to remedy the second problem by identifying the turning points of the reference cycle from individual turning points, but their analysis remains confined to four variables. As for the first problem, Stock and Watson (2010a, 2010b) assumed knowledge that a recession has occurred in an interval to focus on the task of dating the peaks and troughs. They considered a 'date and aggregate' method that estimates the mode (or mean/median) from the individual turning point dates. From an analysis of 270 monthly series, they find that with the exception of four episodes (defined to be the NBER turning point plus or minus twelve months), the results are similar to an 'aggregate and date' approach that looks at turning points from an aggregate time series constructed from the sub-aggregates.

The present analysis centers on identifying the relevant predictor set. In practice, this means screening lots of potential predictors and selecting only those that are actually relevant. My 'lots of data' focus is close in spirit to Stock and Watson (2010b), but their task is dating the peaks and troughs conditional on knowing that a recession has occurred in an interval. My interest is in narrowing down the predictors to only those that are relevant. Since I do not estimate the turning points, determining when economic activity peaked and bottomed is outside the scope of my analysis. I consider out of sample predictions without modelling the latent state. In this regard, my analysis is close in spirit to Berge and Jorda (2011). However, I screen a much larger set of potential predictors and allow the predictor set to change over time.

Business cycles have changing characteristics. Compared with the features documented in Zarnowitz (1985), Ng and Wright (2013) find that the business cycle facts in the last two decades have changed again. An important consideration in my boosting analysis is that predictors useful in one recession may not be useful in other recessions. Before turning to the main analysis, the next section presents boosting first as a machine-learning algorithm and then as a non-parametric model-fitting device.

3. Boosting

For $t = 1, \ldots, T$, let $Y_t = \{1, 0\}$ be a binary variable. In the application to follow, "1" indicates month t was in a recession according to the NBER dating committee. It will also be useful to define $y_t = 2Y_t - 1 = \{1, -1\}$. The objective of the exercise is to fit Y_t with a model given N potentially relevant economic variables and eventually use the model for prediction. For now, I simply denote the predictor set at time t by $x_t = (x_{1,t}, \ldots, x_{N,t})'$, dropping subscripts that indicate the predictors are lagged values.

Consider the population problem of classifying Y given predictors x using a rule $F(x) = \{-1, 1\}$ to minimize a loss function, say, $\mathcal{J}(y, F(x))$. If y was continuous, the squared loss function $\mathcal{J} = E[y - F(x)]^2$ would be the obvious choice. But since both y and F are binary indicators, a more appropriate criterion is the classification *margin*, $yF(x)$ which is negative when a wrong prediction is made. It plays the role of residuals in regressions with continuous data. An algorithm that makes important use of the classification margin is AdaBoost due to Freund (1995) and Schapire (1990).

3.1. Discrete Adaboost

1. Let $w_t^{(1)} = \frac{1}{T}$ for $t = 1, \ldots, T$ and $F_0(x) = 0$.
2. For $m = 1, \ldots M$:
   a. Find $f_m(x)$ from the set of candidate models to minimize the weighted error

$$\epsilon_m = \sum_{t=1}^{T} w_t^{(m)} 1(y_t \neq f_m(x_t)).$$

   b. If $\epsilon_m < 0.5$, update $F_m(x_t) = F_{m-1}(x_t) + \alpha_m f_m(x_t; \theta)$ and

$$w_t^{(m+1)} = \frac{w_t^{(m)}}{Z_m} \exp\left(- \alpha_m y_t f_m(x_t; \theta) \right),$$

   where $Z_m = 2\sqrt{\epsilon_m(1 - \epsilon_m)}$, $\alpha_m = \frac{1}{2} \log\left(\frac{1-\epsilon_m}{\epsilon_m} \right)$.
3. Return classifier $\text{sign}(F_m(x))$.

An example that uses three predictors to classify twelve recession dates is provided in the Appendix.

Adaboost is an algorithm that has its roots in PAC (probably approximately correct) learning theory. Given covariates x and outcome y, a problem is said to be strongly PAC learnable if there exists a classifier (or learner) $f(x)$ such that the error rate ERROR $= E[1(f(x) \neq y]$ is arbitrarily small. That is, P(ERROR

$< \varepsilon) \geq 1 - \delta$ for all $\delta > 0$ and all $\varepsilon > 0$. Now a random guess has a classification error of $\varepsilon = 1/2$. An algorithm is weakly learnable if there exists $\gamma > 0$ such that $P(\text{ERROR} \leq \frac{1}{2} - \gamma) \geq 1 - \delta$. Weak learnability thus only requires $f(x)$ to perform slightly better than random guessing. Obviously, strong learnability implies weak learnability. The question is whether weak learnability implies strong learnability. Schapire (1990) showed that the strong and weak learnable class are identical. This fundamentally important result implies that a weak classifier $f(x)$ that performs only slightly better than random guessing can be boosted to be a strong learner. Adaboost is the first of such algorithms.

In the boosting algorithm, the classifier chosen at stage m is the weak learner while a strong learner is the one that emerges at the final step M. These are denoted $f_m(x)$ and $F_M(x)$ respectively. A weak learner is a function parameterized by θ that maps the features of x into class labels $\{-1, 1\}$. The weaker learner could first obtain least squares estimates of θ and assign $f_m(x) = \text{sign}(x\theta)$. It can also be a decision stump that assigns the label of 1 if the condition $(x \geq \theta)$ holds. While each $f_m(x)$ provides a classification, the final class label $F_M(x)$ is determined by the sign of a weighted sum of $f_m(x)$. Hence it is a weighted majority vote. The classification margin $yF_M(x)$ is a measure of the confidence of the model. The closer it is to 1, the more confidence there is that the final $F_M(x)$ is correct, and the closer it is to -1, the more confidence that $F_M(x)$ is incorrect. A margin that is close to zero indicates little confidence. The parameter M is a stopping rule to prevent overfitting. By suitable choice of M, we can identify which n of the N predictors are useful.

Dettling (2004) and Breiman (1996, 1998) noted that Adaboost is the best off-the-shelf classifier in the world. The crucial aspect of Adaboost is that it adjusts the weight on each observation so that the mis-classified observations are weighted more heavily in the next iteration. This can be seen by noting that

$$w_t^{(m+1)} = \frac{w_t^{(m)}}{Z_m} \begin{cases} \exp(-\alpha_m) & y_t = f_m(x_t; \theta) \\ \exp(\alpha_m) & y_t \neq f_m(x_t; \theta) \end{cases}.$$

Thus, the weight on y_t is scaled by $\exp(\alpha_m)$ in iteration $m + 1$ if it is misclassified in iteration m. Correspondingly, observations that are correctly classified previously receive smaller weights. The algorithm effectively forces the classifier to focus on training the misclassified observations. One can interpret ϵ_m (when divided by T) as the sample analog of the expected misclassification rate $\epsilon_m = E_w[1(y \neq f^{(m)}(x))]$ with w_t as weights. The normalizing factor Z_m is optimally chosen so that $w_t^{(m+1)}$ sums (over t) to one.

Adaboost is presented above as a classification algorithm, but is it associated with a loss function and what are its optimality properties? Freund and Schapire (1996) showed that boosting can be interpreted as a two-player game in which a learner has to form a random choice of models to make a prediction in each of a sequence of trials, and the goal is to minimize mistakes. The Adaboost

solution emerges upon applying the weighted majority algorithm of Littlestone and Warmuth (1994) to the dual of the game. For our purpose, the interesting angle is that Adaboost turns out to minimize a monotone transformation of the zero-one loss function $\mathcal{J}_{0-1}$, defined as

$$\mathcal{J}_{0-1} = E[1(yF(x) < 0] = P(yF(x) < 0).$$

As $yF(x)$ is negative only when the sign of y does not agree with the classifier $F(x)$, minimizing $\mathcal{J}_{0-1}$ is the same as minimizing the misclassification rate. The zero-one loss function is neither smooth nor convex:[2] Consider the exponential transformation

$$\begin{aligned}\mathcal{J}_{EXP} &= E[\exp(-yF(x)|x)] \\ &= P(y = 1|x)\exp(-F(x)) + P(y = -1|x)\exp(F(x)).\end{aligned}$$

Notably, if $\mathcal{J}_{EXP}$ is zero, the zero-one loss will also be zero. Because $\mathcal{J}_{EXP} \geq \mathcal{J}_{0-1}$, $\mathcal{J}_{EXP}$ is an upper bound for $\mathcal{J}_{0-1}$. Minimizing $\mathcal{J}_{EXP}$ with respect to $F(x)$ gives

$$F^*(x) = \frac{1}{2}\log\frac{P(y = 1|x)}{P(y = -1|x)}. \tag{1}$$

The classifier defined by

$$\text{sign}(F^*(x;\theta)) = \text{argmax}_y P(y|x)$$

coincides with Bayes classification based on the highest posterior probability. Equivalently, y is labelled 1 if the posterior probability exceeds $\frac{1}{2}$.

3.2. The Statistical View

This subsection first presents the statistical underpinnings of Adaboost and then considers generic features of boosting. The key link to statistical analysis is an additive logistic model. Recall that a parametric logit model maps the log-odds ratio to the predictors x via a finite dimensional parameter vector β. With $Y = \{0, 1\}$ and class probability defined as

$$P_t = P(Y_t = 1|x_t) = \frac{\exp(f(x_t;\beta))}{1 + \exp(f(x_t;\beta))}, \tag{2}$$

the log-odds ratio is modelled as

$$\log\frac{P(Y_t = 1|x_t)}{P(Y_t = 0|x_t)} = f(x_t;\beta) = x_t\beta. \tag{3}$$

2 A different approximation to the zero-one loss is given in Buhlmann and Yu (2003).

Given T observations, the sample binomial likelihood is

$$\log L(x;\beta) = \sum_{t=1}^{T} Y_t \log P_t + (1 - Y_t)\log(1 - P_t)$$
$$= \sum_{t} Y_t x_t'\beta - \log(1 + \exp(x_t'\beta)).$$

As is well known, β can estimated by a gradient descent (Newton-Ralphson) procedure that iteratively updates $\beta^{(1)} = \beta^{(0)} - (L''(\beta))^{-1}L'(\beta)$ till convergence, where $L'(\beta)$ and $L''(\beta)$ are the first and second derivatives of $L(\beta)$ with respect to β. For the logit model, $L'(\beta) = X'(Y - P)$ and $L''(\beta) = -X'WX$. Let W be a $T \times T$ diagonal matrix with the tth entry being the weight $P_t(1 - P_t)$. The updating rule can be written as

$$\beta^{(1)} = \beta^{(0)} + (X'WX)^{-1}X'(Y - P).$$

Upon defining $Z = W^{-1}(Y - p)$ as the adjusted response variable, we also have

$$\beta^{(1)} = \beta^{(0)} + (X'WX)^{-1}X'WZ.$$

The parameters can be conveniently updated by running a weighted regression of Z on X.

With the parametric model as the backdrop, consider now a non-parametric analysis that replaces $x_t'\beta$ by $F(x)$. Define

$$p_t = P(Y_t = 1|x_t) = \frac{1}{1 + \exp(-2F(x_t))} = \frac{\exp(F(x_t))}{\exp(F(x_t)) + \exp(-F(x_t))}. \quad (4)$$

With $y_t = \{1, -1\}$, the sample binomial likelihood

$$\log L(y, p) = -\sum_{t=1}^{T} \log(1 + \exp(-2y_t F(x_t)))$$

is maximized at the true value of $p_t = P(y_t = 1|x_t)$, or equivalently at

$$F(x) = \frac{1}{2}\log\frac{P(y = 1|x)}{P(y = -1|x)}.$$

This solution evidently differs from the standard logit one given in (3) by a factor of a two. But observe that this is precisely the Adaboost solution given in (1). This interesting result is not immediately obvious because $\mathcal{J}_{EXP}$ is itself not a proper likelihood, but merely an approximation to the zero-one loss. Nonetheless, the

two objective functions are second order equivalent, as

$$\ln(1+\exp(-2z))+1-\ln 2 \approx 1-z+\frac{z^2}{2},+\ldots,$$

while

$$\exp(-z) \approx 1-z+\frac{z^2}{2}+\ldots.$$

In general, $\log(1+\exp(-2z)) \leq \exp(-z)$. Adaboost imposes a larger penalty for mistakes because $\log(1+\exp(-2z))$ grows linearly as $z \to -\infty$, while $\exp(-z)$ grows exponentially.

Having seen that the Adaboost solution also maximizes the binomial likelihood with p_t defined as in (4), we will now use the insight of Breiman (1999) and Friedman (2001) to see Adaboost from the perspective of fitting an additive model. To maximize the expected binomial log likelihood defined for $Y_t = \{1, 0\}$ with p_t defined as in (4), the method of gradient descent suggests to update from the current fit $F_m(x)$ to $F_{m+1}(x)$ according to

$$F_{m+1}(x) = F_m(x) - \frac{L'(F_m+f)}{L''(F_m+f)}|_{f=0} \equiv F_m(x)+f_m(x),$$

where $L'(\cdot)$ and $L''(\cdot)$ are the first and second derivatives of the expected log likelihood with respect to f, evaluated at $f=0$. But under the assumptions of analysis, $L'(x) = 2E[Y-p|x]$ and $L''(x) = -4E[(1-p)|x]$. The update

$$f_m(x) = \frac{1}{2}\frac{E[Y-p|x]}{E[p(1-p)|x]} = \frac{1}{2}E_w\left[\frac{Y-p}{p(1-p)}\Big|x\right]$$

is designed to use the weighted expectation of the residual $Y-p$ to improve the fit. In practice, this amounts to taking $f_m(x)$ to be the fit from a weighted regression of the adjusted response $z_t = \frac{Y_t-p_t}{p_t(1-p_t)}$ on x_t with $p_t(1-p_t)$ as weights. The important difference compared with the parametric logit analysis is that now the function $f_m(x_t)$ at each t, not the parameters, is being estimated. For this reason, the procedure is known as functional gradient descent. After M updates, the log odds ratio is represented by

$$F_M(x) = \sum_{m=1}^{M} f_m(x^m),$$

which is an ensemble of M component functions $f_m(\cdot)$. The functional gradient descent algorithm essentially fits a stagewise regression, meaning that variables are included sequentially in a stepwise regression, and no change is made to the coefficients of the variables already included. The size of the ensemble is

determined by M. This parameter also controls which predictors are dropped, as variables not chosen in steps one to M will automatically have a weight of zero.

The ensemble feature of boosting is preserved even when the functional gradient algorithm is applied to other objective functions. The generic boosting algorithm is as follows:

Gradient Boosting. For minimizing $\mathcal{J}(x) = EJ(x)$:

1. For $t = 1, \ldots, T$, initialize w_t and $F_0(x_t)$.
2. For $m = 1, \ldots M$:
   a. Compute the negative gradient $J'(x_t) = -\frac{\partial J(y_t, F_t)}{\partial f}|_{F_{m-1}(x_t)}$.
   b. Let $f_m(x_t; \theta_m)$ be the best fit of $J'(x_t)$ using predictor x_t.
   c. Update $F_m(x_t) = F_{m-1}(x_t) + c(f_m(x_t; \theta))$.
3. Return the fit $F_M(x)$ or the classifier $C(F_M(x))$.

Step (2a) computes the adjusted response and step (2b) obtains the best model at stage m. Step (2c) uses the negative gradient to update the fit. For quadratic loss $J(x) = (y_t - F(x_t))^2$, $J'(x_t)$ is the residual $y_t - F_{m-1}(x_t)$. Then gradient boosting amounts to repeatedly finding a predictor to fit the residuals not explained in the previous step. By introducing a parameter ν to slow the effect of $f_m(\cdot)$ on $F(\cdot)$, step (2c) can be modified to control the rate at which gradient descent takes place:

$$F_m(x_t) = F_{m-1}(x_t) + \nu\ c(f_m(x_t; \theta)).$$

The regularization parameter $\nu \in [0, 1]$ is also known as the learning rate. Obviously, the parameters M and ν are related, as a low learning rate would necessitate a larger M.

Seen from the perspective of gradient boosting, Adaboost uses $J(x) = \exp(-yF(x))$, while the Logitboost of Friedman, Hastie, and Tibshirani (2000) specifies $J(x) = \ln(1 + \exp(-2yF(x)))$. The two are second-order equivalent, as discussed earlier.[3] Some software packages use the terms interchangeably. There are, however, important differences between a boosting based logit model and the classical logit model even though both minimize the negative binomial likelihood. The predictors in a logit model are selected prior to estimation, and the fit is based on a model that considers multiple predictors jointly. In contrast, gradient boosting performs variable selection and estimation simultaneously, and the final model is built up from an ensemble of models.

We have thus seen that Adaboost is gradient boosting applied to a specific loss function. Many variations to this basic theme have been developed. Instead of fitting a base learner to the observed data, a variation known as stochastic boosting randomly samples a fraction of the observations at each step m (Friedman

3 Logitboost uses initialization $w_t = \frac{1}{T}$, $F_0(x_t) = 0$, $p(y_t|x_t, \theta) = \frac{1}{2}$.

2002). If all observations are randomly sampled, the bagging algorithm of Breiman (1996) obtains. Bagging tends to yield estimates with lower variances. By letting the subsampling rate to be between zero and one, stochastic boosting becomes a hybrid between bagging and boosting. There is a loss of information from returning a discrete weak classifier $f_m(x;\theta)$ at each step. To circumvent this problem, Friedman, Hastie, and Tibshirani (2000) proposed a Gentleboost algorithm, which updates the probabilities instead of the classifier. Multiclass problems have also been studied by treating boosting as an algorithm that fits an additive multinomial model.

Associated with each loss function J are model implied probabilities, but additional assumptions are necessary to turn the probabilities into classification. These are determined by the functions $c(\cdot)$ and $C(\cdot)$. Both Logitboost and Adaboost label the weak and strong learners using the sign function; that is, $c(z) = \text{sign}\,(z)$ and $C(z) = \text{sign}\,(z)$. As a consequence, these are Bayes classifiers that applies a threshold of one-half to the posterior probability. It might be desirable to choose a different threshold, especially when the two classes have uneven occurrence in the data. Let τ be the cost of misclassifying Y to 1 and $1-\tau$ be the cost of misclassifying Y to zero. Minimizing the misclassification risk of $(1-P)\tau + P(1-\tau)$ leads to a cost-weighted Bayes rule that labels y to one when $(1-P)\tau < P(1-\tau)$. In this case, step (3) would return

$$C(P(F_M)) = 1 \quad \text{if } P > \tau,$$

which is a form of quantile classification (Mease, Wyner, and Buja 2007).

In summary, boosting can be motivated from the perspective of minimizing an exponential loss function, or fitting an additive logit model using the method of functional gradient descent that implicitly reweights the data at each step. Regularization, subsampling, and cross-validation are incorporated into R packages (such as MBOOST, ADA, and GBM). The analysis to follow uses the beroulli loss function as implemented in the GMB package of Ridgeway (2007). The package returns the class probability instead of classifier. For recession analysis, the probability estimate is interesting in its own right, and the flexibility to choose a threshold other than one-half is convenient.

4. Application to Macroeconomic Data

My analysis uses the same 132 monthly predictors as in Ludvigson and Ng (2011), updated to include observations in 2011, as explained in Jurado, Ludvigson, and Ng (2013). The data cover broad categories: real output and income, employment and hours, real retail, manufacturing and trade sales, consumer spending, housing starts, inventories and inventory sales ratios, orders and unfilled orders, compensation and labour costs, capacity utilization measures, price indexes, bond and stock market indexes, and foreign exchange measures. Denote

the data available for prediction of y_t by a $(t-h) \times 132$ matrix

$$\mathrm{x}_{t-h} = (\mathrm{x}_{1,t-h}, \ldots, \mathrm{x}_{132,t-h})',$$

where each $x_{j,t-h}$ is a $t-h \times 1$ vector.

In the recession studies reviewed in Section 2 dynamic specification of the model is an important part of the analysis. For example, factor models estimate the reference cycle and its dynamics jointly, while a Markov switching model estimates the transition probability between the recession and non-recession states. The standard logit model is designed for independent data and is basically static. Allowing for dynamics complicates the estimation problem even when the number of predictors is small (Kauppi and Saikkonen 2008). To allow for a dynamic relation between y_t and possibly many predictors, I let d lags of every variable be a candidate predictor. The potential predictor set is then a $t-h$ by N data matrix $\mathbb{X}_{t-h}$, where $N = (133 \times d)$:

$$\mathbb{X}_{t-h} = (\mathrm{x}_{t-h-1}, \ldots, \mathrm{x}_{t-h-d}, y_{t-h-1}, \ldots, y_{t-h-d})'.$$

When $d = 4$, a forecast for y_t uses data at lags $t-h-1$, $t-h-2$, $t-h-3$, and $t-h-4$.[4]

So far, the setup applies for h step ahead prediction of any variable y_t. If y_t was a continuous variable such as inflation or output growth, and the identity of the relevant variables in $\mathbb{X}_{t-h}$ were known, a linear regression model would produce what is known as a *direct* h-period forecast. Even if the relevant predictors were not known, and even with $N > T$, it is still possible to use the variable selection feature of boosting as in Bai and Ng (2009a) to identify the useful predictors. The current problem differs in that the dependent variable is a binary indicator. The log-odds model implies that p is non-linear in x. The next issue is then to decide on the specifics of this non-linear function. A common choice of the base learner is a regression tree defined by

$$f_m(x) = \sum_{m=1}^{M_0} \phi(x_m) + \sum_{j,k} \phi_{jk}(x_j, x_k) + \sum_{i,j,k} \phi_{ijk}(x_i, x_j, x_k) + \ldots.$$

Given that the number of potential predictors is large, I do not allow for interaction among variables. As a consequence, each learner takes on only one regressor at a time, leading to what Buhlmann and Yu (2003) referred to as componentwise boosting. I further restrict the regression tree to have twonodes, reducing the tree to a stump. Effectively, variable m at each t belongs to one of two partitions

4 An alternative set of potential predictors defined by averaging each predictor over $t-h-1$ to $t-h-d$ is also considered. The predictor set is constructed as $\bar{\mathbb{X}}_{t-h} = (\bar{\mathrm{x}}_{1,t-h}, \ldots, \bar{\mathrm{x}}_{132,t-h}, \bar{y}_{t-h})$ with $\bar{\mathrm{x}}_{j,t-h} = \frac{1}{d}\sum_{s=1}^{d} x_{j,t-h-s}$, this $t-h \times 133$. The results are qualitatively similar and hence are not reported.

depending on the value of a data dependent threshold τ_m:

$$f_m(x_t^m) = c_m^L 1(x_t^m \leq \tau_m) + c_m^R 1(x_t^m >, \tau_m),$$

where c_m^L and c_m^R are parameters, possibly the mean of observations in the partition. At each stage m, the identity of x^m is determined by considering the N candidate variables one by one, and choosing the one that gives the best fit. It is important to remark that a variable chosen by boosting at stage m can be chosen again at subsequent stages. Because the same variable can be chosen multiple times, the final additive model is spanned by $n \leq M$ variables.

The relative importance of predictor j can be assessed by how it affects variation in $F_M(x)$. Let $\text{id}(x^m)$, be a function that returns the identity of the predictor chosen at stage m. Friedman (2001) suggests considering

$$I_j^2 = \frac{1}{M} \sum_{m=1}^{M} i_m^2 1(\text{id}(x^m) = j). \tag{5}$$

The statistic is based on the number of times a variable is selected over the M steps, weighted by its improvement in squared error as given by i_m^2. The sum of I_j^2 over j is 100. Higher values thus signify that the associated variable is important. Naturally, variables not selected have zero importance.

4.1. Full Sample Estimation

I consider $h = 3,6$, and 12 months ahead forecasts. Considering the $h = 12$ case may seem ambitious given that many studies consider $h = 0$, and economic data appear to have limited predictiability beyond one year (Berge and Jorda 2011). But the present analysis uses more data than previous studies and the boosting methodology is new. It is thus worth pushing the analysis to the limit. After adjusting for lags, the full sample estimation has $T = 610$ observations from 1961:3 to 2011:12. The size of the potential predictor set is varied by increasing d from (3, 3, 4) to (9, 9, 12) for the three forecast horizons considered respectively. The largest predictor set (for the $h = 12$ model) has a total of 1596 predictors: $12 \times 132 = 1584$ variables in $\mathbb{X}_{t-h}$ as well as 12 lags of y_t. Boosting then selects the relevant predictors from this set.

As an exploratory analysis, a model is estimated using the default parameters in the GBM package which sets the learning rate ν to .001, the subsampling rate (BAGFRAC) for stochastic boosting to .5, with TRAIN = .5, so that half the sample serves as training data. The other half is used for cross validation to determining the optimal stopping parameter M. The purpose of stopping the algorithm at step M is to avoid overfitting. While overfitting is generally less of an issue for classification analysis, a saturated model that includes all variables will emerge if M is allowed to tend to infinity. By early termination of the algorithm, $F_M(\cdot)$ is shrunk towards zero. Ridgeway (2007) argues in favour of setting ν

TABLE 1
Top Variables Chosen by Cross-Validation

$(h, h+d)$	Variable	Variable name	I_j^2	k		
(3, 6)	97	6mo-FF spread	25.641	6	5	4
	96	3mo-FF spread	18.830	4	5	6
	133	lagy	18.651	4		
	98	1yr-FF spread	15.024	6		
	101	Aaa-FF spread	8.099	6		
	105	Ex rate: Japan	2.189	5		
(6, 9)	97	6mo-FF spread	24.584	7	8	
	101	Aaa-FF spread	22.306	8	9	7
	98	1yr-FF spread	19.184	7	9	
	96	3mo-FF spread	5.060	7		
	100	10yr-FF spread	2.503	7		
	114	NAPM com price	2.043	8		
(12, 16)	100	10yr-FF spread	14.326	13		
	95	CP-FF spread	12.021	15	13	16
	101	Aaa-FF spread	10.242	13		
	63	NAPM vendor del	10.020	13		
	35	Emp: mining	6.235	16	15	
	99	5yr-FF spread	5.644	13	16	
	64	NAPM Invent	3.494	14		
	114	NAPM com price	2.358	13		

NOTES: Forecasts for period t are based predictors at lag $t-h-1, t-h-2, \ldots, t-h-d$. The column I_j^2 is an indicator of importance as defined in (5). The last column $k \in [h+1, h+d]$ denotes that lag at which the corresponding predictor is chosen.

small and then determining M by cross-validation, or a criterion such as the AIC.

The default values give an M_{opt} that is slightly over 2000. While a large number of variables are being selected, many have low values of I_j^2. I find that by increasing the learning rate ν to 0.01, M_{opt} can be significantly reduced to around 300, and the number of variables being selected (i.e., n) is even smaller. As an example, the $h = 12$ model has $n = 80$ predictors when $d = 4$, much smaller than M_{opt} because, as noted earlier, boosting can select the same variable multiple times. Furthermore, as N increases from 532 to 1596 when d increases from 4 to 12, n only increases from 80 to =107. This suggests that the longer lags add little information. In view of this finding, I focus on results for $d = (3, 3, 4)$.

Listed in Table 1 are the variables chosen and such that Friedman's importance indicator I_j^2 exceeds two. The table says, for example, that the $h = 3$ model selects lags 6, 5, 4 (ordered by the value of I_j^2) of 6mo-FF (the spread between the 6 months Treasury bill and the federal funds rate). Also selected is lag 6 of 1yr-FF spread, which is the difference between the one year Treasury bond and the federal funds rate. There are fewer than 10 variables that pass the threshold of two. Thresholding I_j^2 at a value bigger than zero allows me to focus on the "reasonably" important predictors and ignore the "barely" important ones. It is reassuring that the important variables identified by boosting listed in Table 1

largely coincide with those used in studies of business cycle chronology. As noted earlier, term and credit/default spreads are generally thought to be useful predictors of recessions. However, previous work tends to consider one spread variable at a time. Here, all spreads in the database are potentially relevant and it is up to boosting to pick out the most important ones. The CP-FF spread lagged up to 16 months is important, while the 10yr-FF and Aaa-FF spreads lagged 13 months have predictive power. It can be argued that the CP-FF spread has an information lead advantage over the other spreads. As seen in Table 1, the data in the different spreads are not mutually exclusive. This is true for all three values of h considered. Take the $h = 12$ case as an example. Conditional on the 10yr-FF spread, the spreads CP-FF and Aaa-FF have additional predictive power.

The characteristics of the relevant predictor set are evidently horizon specific. While there are more variables exceeding the I_j^2 threshold of 2 when $h = 12$ than when $h = 3$ and 6, I_j^2 tends to be lower when $h = 12$. As lags of y_t have no predictive power at either $h = 6$ or 12, only the $h = 3$ model has an autoregressive structure. The 6mo-FF and 1yr-FF spreads are important when $h = 3$ and 6, but none of these variables seems to be important when $h = 12$. The NAPM inventories and vendor delivery time are systematically selected when $h = 12$, but none of these variables is selected for $h = 6$ or $h = 3$. Notably, only nominal variables is selected for $h = 3$ and 6 months ahead forecasts. The only variables common to all three forecast horizons is the Aaa-FF spread. Perhaps the surprise variable in Table 1 is employment in mining. Though not frequently used in business cycles analysis, this variable is robustly countercyclical, as will be seen in results to follow.

Figure 1 plots the (in-sample) posterior probability of recessions denoted $\widehat{P}(y_t = 1|\mathbb{X}_{T-h})$ along with the NBER recession dates. The estimated probabilities for the pre-1990 recessions clearly display spikes around the NBER recession dates. However, the fitted probabilities for the post-1990 recessions are poor, especially when $h = 12$. The three recessions since 1983 came and went and the estimated probabilities barely spiked. The fitted probabilities based on the $h = 3$ and 6 models fare better, but the spikes after 1990 are still not as pronounced as expected.

Parameter instability is a generic feature of economic data and can be attributed to changing sources of economic fluctuations over time. The weak learners used in the analysis are stumps, and hence the model for the log-odds ratio is non-parametric. While parameter instability is not a meaningful notion in a non-parametric setup, the possibility of model instability cannot be overlooked. To see if the composition of the predictor set has changed over time. I construct two sets of recession probabilities as follows. The first is based on estimation over the sample $(t_1^1, t_2^1) = (1962{:}3, 1986{:}8)$, and the second is for the sample $(t_1^2, t_2^2) = (1986{:}9, 2011{:}12)$. The in-sample fitted probabilities spliced together from the estimation over the two samples are plotted as the solid blue line in Figure 2. Compared with the full sample estimates in Figure 1, the post-1990

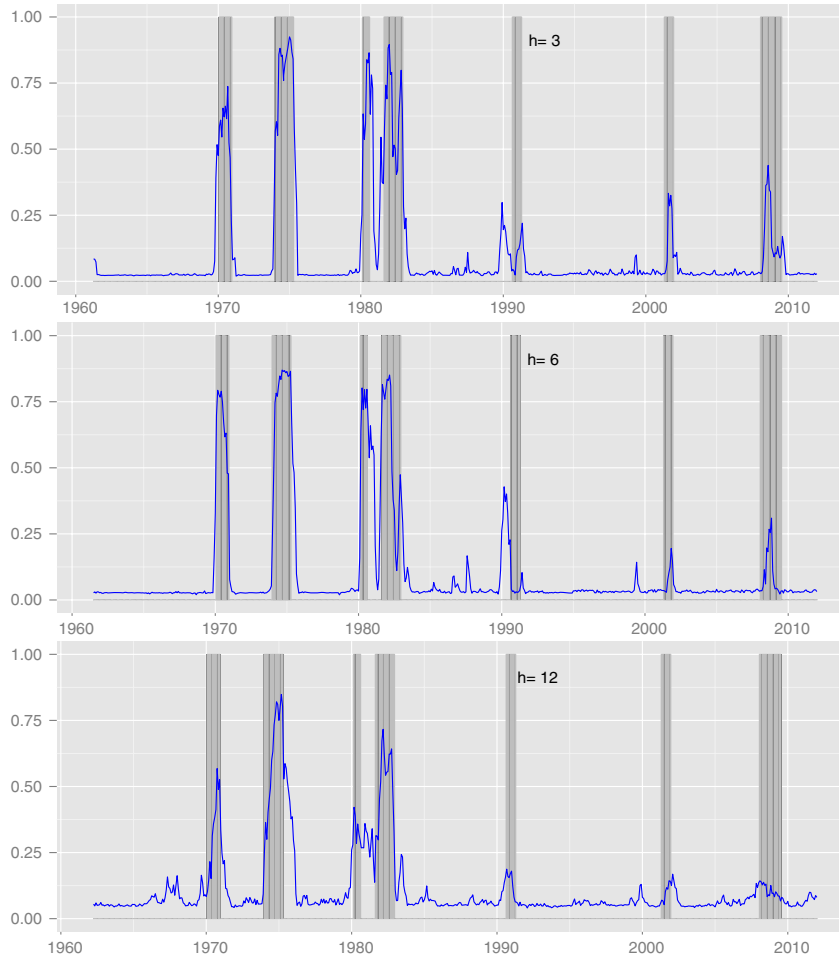

FIGURE 1 Recession Probabilities: In-Sample

recessions are much better aligned with the NBER dates without compromising the fit of the pre-1990 recessions. This suggests that it is possible to fit the data in the two subsamples reasonably well, but different models are needed over the subsamples.

To gain further evidence for model instability, I use the model estimated over the first subsample to compute out-of-sample fitted probabilities for $s \in [t_1^2, t_2^2]$. Similarly, the model estimated for the second subsample is used to compute out-of-sample fitted probabilities for $s \in [t_1^1, t_2^1]$. If the model structure is stable, the out-of-sample fitted probabilities, represented by the dashed lines, should be similar to the in-sample fit, represented by the solid lines. Figure 2 shows that the fitted probabilities based on the model estimated up to 1986:8 do not line

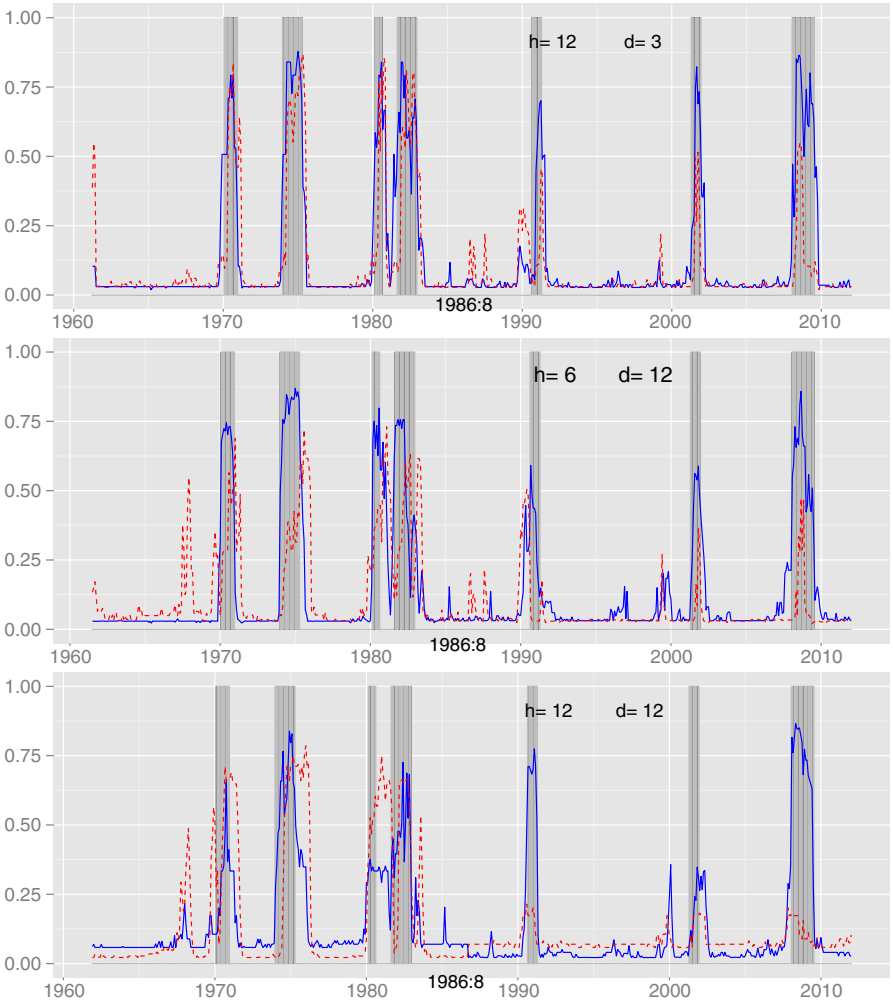

FIGURE 2 Recession Probabilities: Split-Sample, Spliced In-Sample Fit (Solid Line) and Spliced Out-of-Sample Fit (Dashed Line)

up with the recession dates after 1990 recessions. The discrepancy is particularly obvious for $h = 12$. The probabilities based on the model estimated after 1986:8 also do not line up well with the actual recession dates in the first subsample. This confirms that the same model cannot fit the data of both subsamples.

The instability in the predictor set can be summarized by examining which variables are chosen in the two subsamples. This is reported in Table 2. The first impression is that while interest rate spreads are the important predictors in the first subsample for $h = 3$ and 6, many real activity variables become important in the second subsample. For $h = 12$, the real activity variables found to be

TABLE 2
Top Variables Chosen by Cross-Validation: Split Sample Estimation

$(h, h+d)$	Variable	Variable name	I_j^2	k		
	Estimation sample: 1960:3–1986:8					
(3,6)	97	6mo-FF spread	46.162	6	5	
	133	lagy	19.417	4		
	101	Aaa-FF spread	16.997	6		
	98	1yr-FF spread	5.930	6		
	96	3mo-FF spread	5.468	4		
(6, 9)	97	6mo-FF spread	44.165	7	8	
	101	Aaa-FF spread	28.528	8	7	
	98	1yr-FF spread	16.023	7	9	
	114	NAPM com price	3.850	8		
(12, 16)	100	10yr-FF spread	25.842	13		
	101	Aaa-FF spread	23.854	13		
	95	CP-FF spread	18.857	16	15	13
	63	NAPM vendor del	10.229	13		
	61	PMI	9.751	14		
	64	NAPM Invent	3.744	14		
	35	Emp: mining	2.028	16		
	Estimation sample: 1986:9–2011:12					
(3, 6)	133	lagy	64.898	4		
	98	1yr-FF spread	10.252	4	5	
	22	Help wanted/unemp	5.452	5	4	
	15	IP: nondble matls	4.610	5		
	79	DC&I loans	4.267	5		
	62	NAPM new ordrs	3.109	4		
(6, 9)	79	DC&I loans	22.487	7	8	
	99	5yr-FF spread	20.318	9		
	15	IP: nondble matls	16.689	7	9	8
	101	Aaa-FF spread	10.476	9		
	133	lagy	8.865	4		
	62	NAPM new ordrs	5.400	7		
	98	1yr-FF spread	4.870	9		
	36	Emp: const	4.530	7		
	22	Help wanted/unemp	4.456	7		
(12, 16)	99	5yr-FF spread	53.407	15	16	14
	101	Aaa-FF spread	18.519	16		
	79	C&I loans	8.469	14	15	
	102	Baa-FF spread	6.700	16		
	98	1yr-FF spread	2.857	13		
	44	Emp: FIRE	2.774	15		

important in the first sample are no longer important in the second sample. The 5yr-FF spread is important in the second sample but not in the first. Few variables are important in both samples. Even of those that are important, the lags chosen and the degree of importance are different, as seen from the 1yr-FF spread. In general, many of the important predictors identified in the second sample have lower I_j^2.

4.2. Rolling Estimation

The full sample results suggest a change in the dynamic relation between y_t and the predictors, as well as the identity of the predictors themselves. However, the in-sample fit (or lack thereof) does not reflect the out-of-sample predictive ability of the model. Furthermore, the foregoing results are based on the default parameters of the GBM package that randomizes half the sample for stochastic boosting, with M determined by cross-validation as though the data were independent. Arguably, these settings are not appropriate for serially correlated data.

Several changes are made to tailor boosting to out-of-sample predictions that also take into account the time series nature of the data. First, stochastic boosting is disallowed by changing the randomization rate from the default of one-half to zero. Second, rolling regressions are used to generate out-of-sample predictions. These are constructed as follows:

Rolling Forecast. Initialize t_1 and t_2:

1. For $m = 1, \ldots M$, fit $f_m(x^m_{t-h})$ using predictors in $\mathbb{X}_{t-h}$.
2. For $j = 1, \ldots N$, record relative importance $I^m_{t_2,j}$ at each t_2
3. Construct predicted probability $\hat{p}_{t_2} = \widehat{P}(y_{t_2} = 1|\mathbb{X}_{t_2-h})$. Increase t_1 and t_2 by one.

There are 407 of such rolling regressions, each with 180 observations ending in period $t_2 - h$. The first estimation is based on 180 observations from $t_1 - h =$ 1962:3 to $t_2 - h =$ 1977:2. When $h = 12$, the first forecast is made for $t_2 =$ 1978:2. The next forecast is based on estimation over the sample defined by $(t_1 - h, t_2 - h) = (1962{:}4, 1977{:}3)$ and so on.

The final change is that in place of cross-validation, two new indicators are used to determine the set of relevant predictors. The first is the average of relative importance of each variable over the 407 rolling subsamples, constructed for $j = 1, \ldots, N$ as

$$\bar{I}^2_j = \frac{1}{407} \sum_{t_2} I_{j,t_2}.$$

The second indicator is the frequency that variable j is being selected in the rolling estimation:

$$\text{freq}_j = \frac{1}{407} \sum_{t_2} 1(I^2_{j,t_2} > 0).$$

Both statistics are dated according to the period for which the forecast is made: t_2.

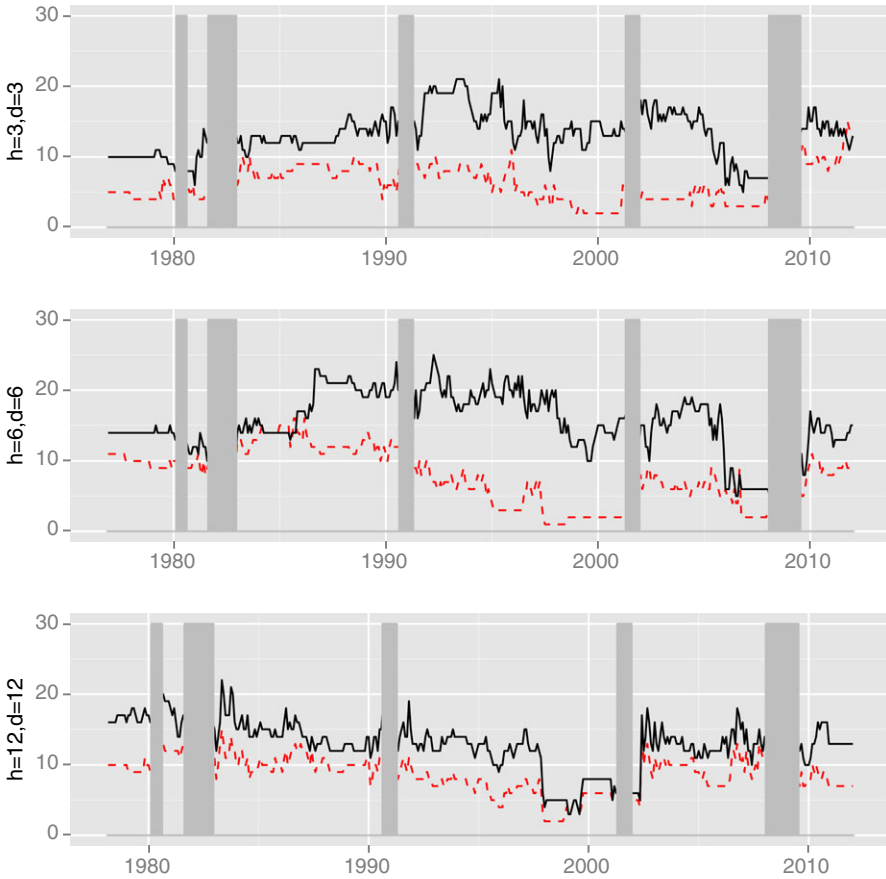

FIGURE 3 Number of Variables Selected in Rolling Estimation, Including Lags (Solid Line) and Unique Variables (Dashed Line)

Figure 3 plots the number of variables with positive importance in forecasting y_{t_2} defined as

$$n_{t_2}(d) = \sum_{j=1}^{N} (I_{j,t_2}^2 > 0).$$

The black solid line indicates the total number of variables selected when lags of the same variable are treated as distinct. The dotted red line indicates the unique number of variables, meaning that variables at different lags are treated as the same variable. On average, the total number of variables selected for $h = 12$ months ahead forecast is between 13 and 16, while the unique number of variables is around 9. These numbers are much smaller than those found in the full sample

TABLE 3
Variables Chosen in Rolling Window Estimation: By Average Importance

$(h, h+d)$	Variable	Variable name	$\bar{I}_j^2$	k		
(3, 6)	133	lagy	18.596	4		
	101	Aaa-FF spread	12.899	6		
	96	3mo-FF spread	11.190	4	6	
	61	PMI	6.449	4		
	97	6mo-FF spread	6.085	6		
	98	1yr-FF spread	5.308	6		
	33	Emp: total	5.087	4		
	95	CP-FF spread	3.566	4		
	102	Baa-FF spread	3.049	6		
(6, 9)	101	Aaa-FF spread	24.155	8	7	9
	98	1yr-FF spread	10.215	7	9	
	102	Baa-FF spread	9.910	7	9	
	99	5yr-FF spread	7.157	9	8	
	100	10 yr-FF spread	6.404	9	8	
	97	6mo-FF spread	5.866	7		
	45	Emp: Govt	4.567	9	7	
	96	3mo-FF spread	4.122	7		
	37	Emp: mfg	2.221	7		
(12, 16)	99	5yr-FF spread	13.993	15	16	
	100	10yr-FF spread	8.958	13		
	102	Baa-FF spread	8.464	13		
	101	Aaa-FF spread	6.977	16	13	
	95	CP-FF spread	6.969	13	16	
	63	NAPM vendor del	6.518	13		
	97	6mo-FF spread	4.600	16		
	64	NAPM Invent	3.106	14		
	61	PMI	2.524	15		
	35	Emp: mining	2.180	15		

analysis. The number of relevant predictors $n_{t2}(d)$ has drifted down since the 1980 recessions and bottomed out in $t_2 = 1999{:}2$, which roughly coincides with the Great Moderation. However, the downward trend is reversed in 2001. The number of relevant predictors for $h = 3$ and 6 generally follow the same pattern as $h = 12$ with the notable difference that since the 2008 recession, the number of important predictors at $h = 3$ has been on an upward trend, when that for $h = 12$ is slightly below the pre-recession level.

Table 3 reports the variables with average importance $\bar{I}_j^2$ exceeding 2. While the term and default spreads are still found to be valuable predictors, there are qualitative differences between the full sample estimates reported in Table 1 and then rolling estimates reported in Table 3. The Aaa spread is the dominant predictor for $h = 3$ and 6 in rolling estimation, but the 6mo-FF spread is better for full sample predictions. For $h = 12$, the 10yr-FF spread is the most important in-sample predictor, but the 5yr-FF spread performs better in rolling estimation. Berge and Jorda (2011) find that the 10yr-FF spread with 18 month

TABLE 4
Variables Chosen in Rolling Window Estimation: By Frequency

$(h, h+d)$	Variable	Variable name	Frequency	k		
(3, 6)	101	Aaa-FF spread	1.330	6	5	4
	97	6mo-FF spread	1.014	6	4	5
	96	3mo-FF spread	0.944	4	6	5
	45	Emp: Govt	0.893	5	4	6
	133	lagy	0.812	4		
	94	Baa bond	0.472	5	4	
	98	1yr-FF spread	0.409	6		
	22	Help wanted/unemp	0.398	4		
	105	Ex rate: Japan	0.386	5		
	33	Emp: total	0.360	4		
	65	Orders: cons gds	0.335	5		
	102	Baa-FF spread	0.333	6		
	93	Aaa bond	0.270	4		
	61	PMI	0.249	4		
(6, 9)	101	Aaa-FF spread	1.761	8	7	9
	45	Emp: Govt	1.248	9	8	7
	98	1yr-FF spread	0.735	7	9	
	102	Baa-FF spread	0.637	7	9	
	96	3mo-FF spread	0.637	7	9	
	95	CP-FF spread	0.607	9	8	
	100	10yr-FF spread	0.585	9	7	
	114	NAPM com price	0.548	8	7	
	99	5yr-FF spread	0.539	9	7	
	97	6mo-FF spread	0.370	7		
	63	NAPM vendor del	0.347	9		
	37	Emp: mfg	0.300	7		
	62	NAPM new ordrs	0.281	7		
	64	NAPM Invent	0.255	7		
	35	Emp: mining	0.246	9		
	66	Orders: dble gds	0.220	7		
	5	Retail sales	0.208	7		
(12,16)	95	CP-FF spread	0.871	16	15	13
	35	Emp: mining	0.859	16	13	15
	81	Inst cred/PI	0.694	16	15	
	99	5yr-FF spread	0.687	15	16	
	100	10yr-FF spread	0.639	13	15	
	101	Aaa-FF spread	0.488	16	13	
	102	Baa-FF spread	0.438	13		
	63	NAPM vendor del	0.371	13		
	61	PMI	0.311	15		
	64	NAPM Invent	0.273	14		
	97	6mo-FF spread	0.270	16		
	117	CPI-U: transp	0.246	13		
	44	Emp: FIRE	0.234	15		
	68	Unf orders: dble	0.218	16		

lag has predictive power. Here, spreads lagged 13–16 months are found to be important.

Table 4 reports the variables whose frequency of being selected (summed over lags) is at least 0.2. The Aaa-FF spread is the most frequently selected among

variables in $\mathbb{X}_{t-h}$ when $h = 3$ and 6, while the CP-spread is most frequently selected when $h = 12$. These variables have high values of freq_j because multiple lags are selected while other variables are selected at only one lag. In this regard, the model chosen by boosting has rather weak dynamics. Tables 3 and 4 together suggest that the Aaa-FF spread is the most robust predictor for $h = 3$ and 6, while the CP and 5yr-FF spreads are most important for $h = 12$. Of the real activity variables, government employment and employment in mining are valuable predictors.

The out-of-sample fitted probabilities are plotted as the solid lines in Figure 4. These probabilities are generally high around recession times, but there are also false positives especially around 1985 during which the key predictor is found to be the Baa-FF spread. Notably, all out-of-sample fitted probabilities are more choppy than the full sample estimates. One explanation is that each rolling window has only $T = 180$ observations, while the full sample estimates use $T = 610$ observations. The more likely reason is that, except for $h = 3$, the other two prediction models have no autoregressive dynamics.

Additional assumptions are necessary to use the estimated probabilities to generate warning signals for recessions. As mentioned earlier, only 15% of the sample are recession months and the Bayes threshold of 0.5 may not be appropriate. Chauvet and Hamilton (2006) suggest declaring a recession when the smoothed probability estimate of the latent state $\hat{P}(s_t|x_t)$ exceeds .65 and the recession ends when the probability falls below 0.35. Chauvet and Piger (2008) require the smoothed probability to move above 0.8 and stay above that level for three consecutive months. Berge and Jorda (2011) searched between 0.3 and 0.6 for the optimal threshold such that the Chauvet and Piger (2008) probabilities would best fit the NBER dates. In contrast to these studies, I analyze $\hat{P}(y_t|\mathbb{X}_{t-h})$, and probabilities estimates are lower the higher is h. Applying the threshold of .65 would lead to the foregone conclusion of no recession. Yet it is clear that the predicted probabilities have distinguishable spikes.

I proceed with the assumption that it is the relative probability over time that matters and use as threshold the mean plus 1.65 times the standard deviation of the fitted probabilities over the sample. This yields thresholds of .435 for $h = 3$, .444 for $h = 6$, and .304 for $h = 12$, respectively. This line is marked in each of the three panels in Figure 4. I then locate those dates t_2 for which the estimated probability exceeds the thresholds to get some idea of whether the threshold-crossing dates are near the NBER recession dates. The most important variable used in the prediction at t_2 is recorded for reference, even though it should be remembered that boosting is an ensemble scheme and prediction is not based on a particular regressor per se.

It is well documented that the risky spreads were successful in predicting recessions prior to 1990 but they completely missed the 1990 recession. The $h = 12$ model produces probability estimates above .2 for three of the five months between 1990:3 and 1990:8, but not on a consecutive basis, as the probabilities are as low as 0.06 in between. The $h = 6$ model also produces heightened probability

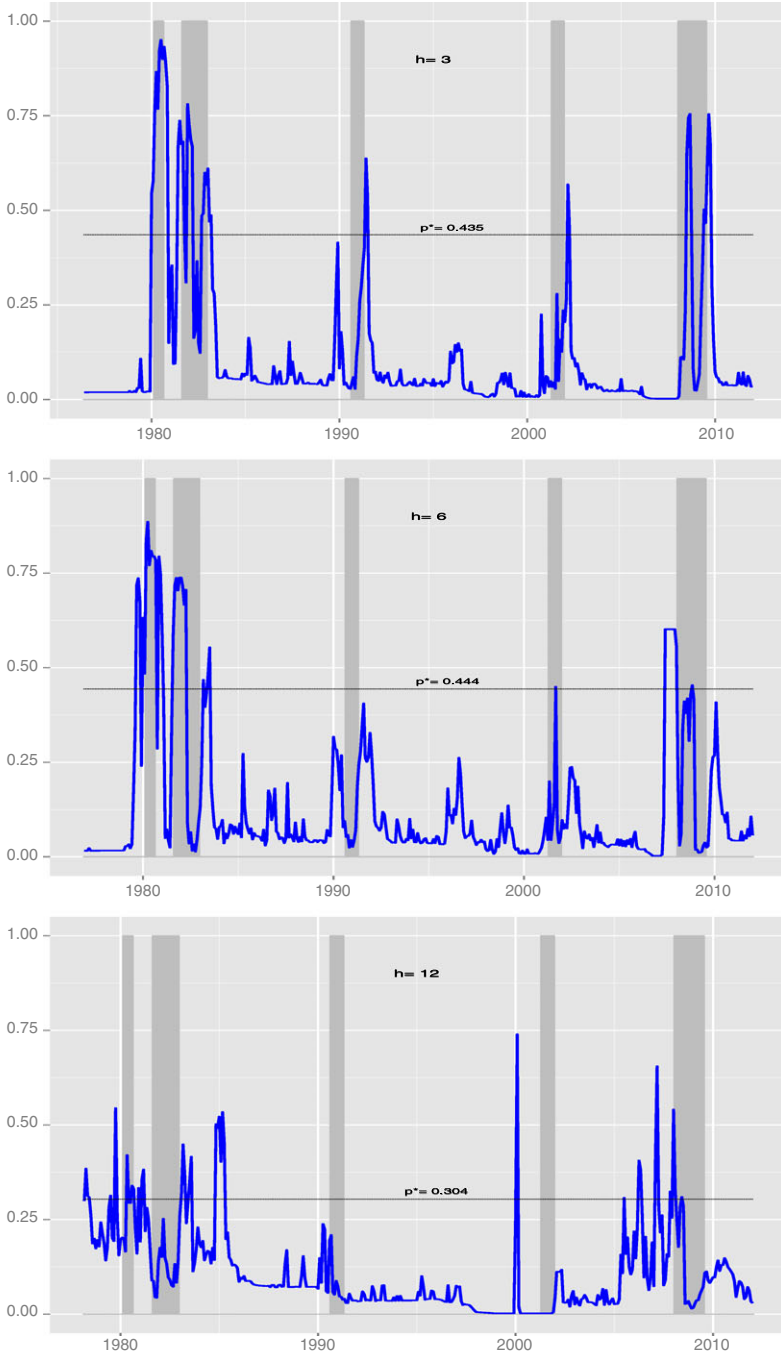

FIGURE 4 Recession Probabilities: Rolling Estimates

estimates, but they fall short of crossing the threshold. That the two models did not strongly signal a recession is perhaps not surprising, as the top predictor is found to be the (Baa-FF) risky spread. The $h = 3$ model predicts a recession probability of .414 for $t_2 = 1989{:}11$, with the 3m-FF spread being the most important predictor. The estimate is still short of the threshold of .436. The probability of recession reaches .534, but only in $t_2 = 1991{:}6$ well after the recession started.

Turning to the 2001 recession, we see in Figure 4 that the probability of recession based on the $h = 3$ model reaches .568 in $t_2 = 2002{:}2$. This is largely due to the lagged recession indicator, since 2001:03 to 2001:10 were identified by the NBER as recession months. The estimated probability based on the $h = 12$ model jumps from .002 in $t_2 = 1999{:}11$ to 0.739 in 2000:1, the 5yr-FF spread identified as being most important. The $h = 6$ model also gives a recession probability of .449 at $t_2 = 2001{:}8$, the most important predictor being the Aaa-FF spread, but both signals of recession are short lived.

There has been much discussion of whether signs of the 2008 recession were missed. The $h = 3$ model gives a probability of .650 in $t_2 = 2008{:}7$. The probability remains high for several months, the most important predictor being the PMI. The probability based on the $h = 6$ model exceeds .6 around $t_2 = 2007{:}5$ for several months, with the most important predictor being the 5yr-FF spread. For the $h = 12$ model, the recession probability is estimated to be .65 for $t_2 = 2007{:}2$ but it returned to lower values before climbing to .54 at $t_2 = 2007{:}12$, the top predictors being the 10yr-FF and the Aaa-FF spreads. Thus, from the data in mid-2006, the models for $h = 6$ and 12 see signs of the 2008 recession a whole year before it occurred, but the signals are sporadic and easy to miss.

Overall, the models seem to give elevated probability estimates around but not always ahead of the actual recessions dates. Instead of eyeballing the probability estimates, I also attempt to automate the start date of the recessions as determined by the model. From the peak of economic activity identified by the NBER, I look within a window of τ-months to see if the predicted probabilities exceed the thresholds and record the variables with the highest $\bar{I}_j$. I set τ to 12 for the recessions before 1990 and to 18 for the ones after to accommodate the observation in hindsight that the signals in the post-1990 recessions appear in the data earlier than the recessions before 1990.[5] These dates are reported in column 3 of Table 5. As a point of reference, Table 5 also lists the recession months identified by the NBER. In parentheses are the dates that the recessions were announced. A noteworthy feature of Table 5 is the changing nature of the predictors in the five recessions. The risky spreads have been important in predicting the two pre-1990 recessions but not the subsequent ones, and employment data have been helpful in predicting the post-1990 recessions but not the pre-1990 ones.

5 If the probability estimates never exceed the threshold, I consider a model with more predictors (larger d) than the base case. This often gives higher probability estimates, but not high enough to cross the thresholds.

TABLE 5
Summary of Model Warnings

Recession	NBER	t_2 (Model)	$\widehat{P}_{t_2}$	Top predictor	Threshold exceed
1: 1980:1–1980:7	1981:7				
$h = 3$		1979:12	0.545	Baa bond	Y
$h = 6$		1979:8	0.718	CP-FF spread	Y
$h = 12$		1979:6	0.313	CP-FF spread	N
2: 1981:7–1981:11	1983:7				
$h = 3$		1980:7	0.899	Baa bond	Y
$h = 6$		1980:7	0.794	CP-FF spread	Y
$h = 12$		1980:7	0.339	CP-FF spread	Y
3: 1990:7–1990:3	1992:12				
$h = 3$		1989:11	.414	3mo-FF spread	N
$h = 6$		1989:12	.317	Govt. Emp	N
$h = 12$		1990:03	.238	Emp. Mining	N
4: 2001:3–2001:11	2003:7				
$h = 3$		2000:09	.224	Govt. Emp	N
$h = 6$		2001:02	.104	Govt. Emp	N
$h = 12$		2000:01	.739	5yr-FF spread	Y
5: 2007:12–2009:6	2010:9				
$h = 3$		2006:06	.003	HWI	N
$h = 6$		2007:05	.606	Govt. Emp	Y
$h = 12$		2007:01	.416	Emp. Mining	Y

NOTES: The business cycle dates are taken from the website http://www.nber.org/cycles/cyclesmain.html. The NBER column denotes the date that the trough was announced by the NBER. The t_2 column is the date within a window since the last peak of economic activity that the probability of recession estimated by the model exceeds the threshold of mean + 1.65 standard deviations. If the thresholds of .435, .444, .304 are not crossed, a model with more lagged predictors is considered. For the first two recessions, the window is 12 months. For the last three recessions, the window is 18 months.

Table 5 shows that all models failed to call the 1990 recession, in agreement with the informal analysis. For the 2008 recession, the $h = 3$ model does not produce any probability estimate in the 18 months prior to 2007:12 that exceeds the threshold. However, the $h = 12$ month model reports a recession probability of .416 in for $t_2 = 2007{:}01$, and the $h = 6$ model reports a recession probability of .606 for $t_2 = 2007{:}05$. That the $h = 12$ model gives earlier warning than the $h = 3$ ones is interesting. The bottom line conclusion is that signals of the 2008 recession were in the data as early as mid-2006, but there is a lack of consensus across models, making it difficult to confidently make a recession call.

5. Conclusion

This analysis sets out to explore what boosting has to say about the predictors of recessions. Boosting is a non-parametric method with little input from economic theory. It has two main features. First, the fit of the model is based on an ensemble scheme. Second, by suitable choice of regularization parameters, it enables

joint variable selection and estimation in a data-rich environment. The empirical analysis finds that even though many predictors are available for analysis, the predictor set with systematic and important predictive power consists of only 10 or so variables. It is reassuring that most variables in the list are already known to be useful, though some less obvious variables are also identified. The main finding is that there is substantial time variation in the size and composition of the relevant predictor set, and even the predictive power of term and risky spreads are recession specific. The full sample estimates and rolling regressions give confidence to the 5yr-FF spread and the Aaa and CP spreads (relative to the Fed funds rate) as the best predictors of recessions. The results echo the analysis of Ng and Wright (2013) that business cycles are not alike. This, in essence, is why predicting recessions is challenging.

Few economic applications have used boosting thus far, probably for the reason that the terminology and presentation are unfamiliar to economists. But binomial boosting is simply an algorithm for constructing stagewise additive logistic models. It can potentially be used to analyze discrete choice problems, such as whether to retire or which brand to use. In conjunction with other loss functions, boosting can also be used as an alternative to information criterion as a variable selection device. Bai and Ng (2009a,b) exploited these properties to choose instruments and predictors in a data rich environment. This paper has used boosting in the less common context of serially correlated data. The method is far from perfect, as there were misses and false positives. A weakness of boosting in recession analysis is that it produces fitted probabilities that are not sufficiently persistent. This is likely a consequence of the fact that the model dynamics are now entirely driven by the predictors. The autoregressive dynamics needed for the estimated probabilities to be slowly evolving are weak or absent altogether. Furthermore, the predictors are often selected at isolated but not consecutive lags. My conjecture is that improving the model dynamics will likely lead to smoother predicted probabilities without changing the key predictors identified in this analysis. How richer dynamics can be incorporated remains very much a topic for future research.

Appendix: A Toy Example

This Appendix provides an example to help in understanding the Adaboost algorithm. Consider classifying whether the 12 months in 2001 using three-month lagged data of the help-wanted index (HWI), new orders (NAPM), and the 10yr-FF spread (SPREAD). The data are listed in columns 2–4 of Table A1. The NBER dates are listed in column 5, where 1 indicates a recession month. I use a stump (two-node decision tree) as the weak learner. A stump uses an optimally chosen threshold to split the data into two partitions. This requires setting up a finite number of grid points for HWI,

TABLE A1
Toy Example

	Data: Lagged 3 Months				$F_1(x)$	$F_2(x)$	$F_3(x)$	$F_4(x)$	$F_5(x)$
Date	HWI −.066	NAPM 48.550	SPREAD 0.244	y	HWI < −.044	NAPM <49.834	HWI < −.100	SPREAD > −.622	NAPM < 47.062
2001. 1	0.014	51.100	−0.770	−1	−1	−1	−1	−1	−1
2001. 2	−0.091	50.300	−0.790	−1	1	−1	−1	−1	−1
2001. 3	0.082	52.800	−1.160	−1	−1	−1	−1	−1	−1
2001. 4	−0.129	49.800	−0.820	1	1	1	1	1	1
2001. 5	−0.131	50.200	−0.390	1	1	−1	1	1	1
2001. 6	−0.111	47.700	−0.420	1	1	1	1	1	1
2001. 7	−0.056	47.200	0.340	1	1	1	1	1	1
2001. 8	−0.103	45.400	1.180	1	1	1	1	1	1
2001. 9	−0.093	47.100	1.310	1	1	1	1	1	1
2001. 10	−0.004	46.800	1.470	1	−1	1	−1	1	1
2001. 11	−0.174	46.700	1.320	1	1	1	1	1	1
2001. 12	−0.007	47.500	1.660	−1	−1	1	−1	1	−1
α					.804	1.098	.710	.783	.575
Error rate					.167	.100	.138	.155	0

NAPM, and SPREAD, respectively, and evaluating the goodness of fit in each partition.

The algorithm begins by assigning an equal weight of $w_t^{(1)} = 1/T$ to each observation. For each of the grid points chosen for HWI, the sample of Y values is partitioned into two parts depending on whether HWI_t exceeds the grid point or not. The grid point that minimizes classification error is found to be −.044. The procedure is repeated with NAPM as a splitting variable, and again for SPREAD. A comparison of the three sets of residuals reveals that splitting on the basis of HWI gives the smallest weighted error. The first weak learner thus labels Y_t to 1 if $\text{HWI}_t < -.044$. The outcome of the decision is given in Column 6. Compared with the NBER dates in column 5, we see that months 2 and 10 are mislabelled, giving a misclassification rate of $2/12 = .167$. This is ϵ_1 of step (2a). Direct calculations give $\alpha_1 = .5\log(\frac{1-\epsilon}{\epsilon})$ of .804. The weights $w_t^{(2)}$ are updated to complete step (2b). Months 2 and 10 now each have a weight of 0.25, while the remaining 10 observations each have a weight of 0.05. Three thresholds are again determined for each of the three variables and the weighted residuals are computed using weights $w^{(2)}$. Of the three, the NAPM split gives the smallest weighted residuals. The weak learner for step 2 is identified. The classification based on the sign of $F_2(x) = .804 \cdot 1(HWI < -0.044) + 1.098 \cdot 1(NAPM < 49.834)$ is given in column 7. Compared with column 5, we see that months 5 and 12 are mislabelled. The weighted misclassification rate is decreased to .100. The new weights $w_t^{(3)}$ are .25 for months 5 and 12, .138 for months 2 and 10, and .027 for the remaining months. Three sets of weighted residuals are again determined using new thresholds. The best predictor is again HWI with a threshold of –.100. Classification based on the sign of $F_3(x)$ is given in column

8, where $F_3(x) = .804 \cdot 1(\text{HWI} < -.044) + 1.098 \cdot 1(\text{NAPM} < 48.834) + .710 \cdot 1(\text{HWI} < -.100)$. The error rate after three steps actually increases to .138. The weak learner in round four is 1(SPREAD$>-.622$). After NAPM is selected for one more round, all recession dates are correctly classified. The strong learner is an ensemble of five weak learners defined by sign($\hat{Y}$), where

$$\begin{aligned}\hat{Y} = {} & .804 \cdot 1(\text{HWI} < -.044) + 1.098 \cdot 1(\text{NAPM} < 49.834) \\ & + .710 \cdot 1(\text{HWI} < -.100) + .783 \cdot 1(\text{SPREAD} > -.622) \\ & + .575 \cdot 1(\text{NAPM} < 47.062).\end{aligned}$$

The key features of Adaboost are (i) the same variable can be chosen more than once, (ii) the weights are adjusted at each step to focus on the misclassified observations, and (iii) a final decision is based on an ensemble of models. No single variable can yield the correct classification, which is the premise of an ensemble decision rule.

References

Aruoba, S.B., F.X. Diebold, and C. Scotti (2009) "Real-Time Measurement of Business Conditions." *Journal of Business and Economic Statistics* 27, 417–47

Bai, J., and S. Ng (2009a) "Boosting Diffusion Indices." *Journal of Applied Econometrics* 24, 607–29

— (2009b) "Selecting Instrumental Variables in a Data Rich Environment." *Journal of Time Series Econometrics* 1(1)

Berge, T., and O. Jorda (2011) "Evaluating the Classification of Economic Activity into Recessions and Expansions." *American Economic Journal: Macroeconomics* 3, 246–77

Breiman, L. (1996) "Bagging Predictors." *Machine Learning* 24, 123–40

— (1998) "Arcing Classifiers." *Annals of Statistics* 26, 801–49

— (1999) "Prediction Games and Arcing Algorithms." *Neural Computation* 11, 1493–517

Bry, G., and C. Boschan (1971) *Cyclical Analysis of Time Series: Procedures and Computer Programs*. New York: NBER

Buhlmann, P., and B. Yu (2003) "Boosting with the L2 Loss: Regression and Classification." *Journal of the American Statistical Association* 98, 324–39

Burns, A., and W. Mitchell (1946) *Measuring Business Cycles*. New York: NBER

Chauvet, M. (1998) "An Econometric Characterization of Business Cycle Dynamics with Factor Structure and Regime Switches." *International Economic Review* 39, 969–96

Chauvet, M., and J. Hamilton (2006) "Dating Business cycle Turning Points." In *Nonlinear Time Series Analysis of Business Cycles*, ed. C. Milas, P. Rothman, and D. van Dijk. Boston: Elsevier

Chauvet, M., and J. Piger (2008) "A Comparison of the Real Time Performance of Business Cycle Dating Methods." *Journal of Business and Economic Statistics* 26, 42–9

Committee, N.B.C.D. (2008) "Determination of the December 2007 Peak in Economic Activity." www.nber.org/cycles/dec2008.html

Cross, P., and P. Bergevin (2012) "Turning Points: Business Cycles in Canada Since 1926." *C.D. Howe Institute Commentary* 366. http://www.cdhowe.org/pdf/Commentary_366.pdf

Dasgupta, S., and K. Lahiri (1993) "On the Use of Dispersion Measures from NAPM Surveys in Business Cycle Forecasting." *Journal of Forecasting* 12, 239–53
Dettling, M. (2004) "Boosting for Tumor Classification with Gene Expression Data." *Bioinformatics* 20, 3583–93
Estrella, A., and F. Mishkin (1998) "Predicting US Recessions: Financial Variables as Leading Indicators." *Review of Economics and Statistics* 80, 45–61
Freund, Y. (1995) "Boosting a Weak Learning Algorithm by Majority." *Information and Computation* 121, 256–85
Freund, Y., and R. Schapire (1996) "Experiments with a New Boosting Algorithm." *Proceedings of ICML* 13, 148–56
Friedman, J. (2001) "Greedy Function Approximation: A Gradient Boosting Machine." *Annals of Statistics* 29, 1189–232
— (2002) "Stochastic Gradient Boosting." *Computational Statistics and Data Analysis* 38, 367–78
Friedman, J., T. Hastie, and R. Tibshirani (2000) "Additive Logistic Regression: A Statistical View of Boosting." *Annals of Statistics* 28, 337–74
Gilchrist, S., and E. Zakrajsek (2012) "Credit Spreads and Business Cycle Fluctuations." *American Economic Review* 102, 1692–720
Hamilton, J. (2011) "Calling Recessions in Real Time." *International Journal of Forecasting* 27, 1006–26
Hamilton, J.D. (1989) "A New Approach to the Economic Analysis of Non-Stationary Time Series and the Business Cycle." *Econometrica* 57, 357–84
Harding, D., and A. Pagan (2006) "Synchronization of Cycles." *Journal of Econometrics* 132, 59–69
Jurado, K., S. Ludvigson, and S. Ng (2013) "Measuring Macroeconomic Uncertainty." Mimeo, Columbia University
Kauppi, H., and P. Saikkonen (2008) "Predicting U.S. Recessions with Dynamic Binary Response Models." *Review of Economics and Statistics* 90, 777–91
Klein, P., and G. Moore (1991) "Purchasing Management Survey Data: Their Value as Leading Indicators." In *Leading Economic Indicators: New Approaches and Forecasting Records*, ed. K. Lahiri and G. Moore. Cambridge: Cambridge University Press
Lahiri, K., and L. Yang (2013) "Forecasting Binary Outcomes." In *Handbook of Forecasting*, ed. G. Elliott and A. Timmermann. Vol. 2. Amsterdam: North-Holland
Littlestone, N., and M. Warmuth (1994) "The Weighted Majority Algorithm." *Information and Computation / Information and Control* 108, 212–61
Ludvigson, S., and S. Ng (2011) "A Factor Analysis of Bond Risk Premia." In *Handbook of Empirical Economics and Finance*, ed. D. Gilles and A. Ullah. Dordrecht: Chapman and Hall
Marcellino, M. (2006) "Leading Indicators." In *Handbook of Forecasting*, ed. G. Elliott, C. Granger, and A. Timmermann. Vol. 1. Amsterdam: Elsevier
Mease, D., A. Wyner, and A. Buja (2007) "Boosted Classification Trees and Class Probability / Quantile Estimation." *Journal of Machine Learning* 8, 409–39
Moench, E., and H. Uhlig (2005) "Towards a Monthly Business Cycle Chronology for the Euro Area." *Journal of Business Cycle Measurement and Analysis* 2, 43–69
Ng, S., and H. Wright (2013) "Facts and Challenges from the Recession for Forecasting and Macroeconomic Modeling." *Journal of Economic Literature*, December
Ridgeway, G. (2007) "Generalized Boosted Models: A Guide to the GBM Package." www.code.google.com/p/gradientboostedmodels
Rudebusch, G., and J. Williams (2009) "Forecasting Recessions: The Puzzle of the Enduring Power of the Yield Curve." *Journal of Business and Economic Statistics* 27, 492–503

Schapire, R.E. (1990) "The Strength of Weak Learnability." *Machine Learning* 5, 197–227
Stock, J., and M. Watson (2010a) "Estimating Turning Points Using Large Data Sets." Mimeo, Princeton University
— (2010b) "Indicators for Dating Business Cycles: Cross-History Selection and Comparison." *American Economic Review: Papers and Proceedings* 100, 16–19
Stock, J.H., and M. Watson (1989) "New Indexes of Coincident and Leading Economic Indications." In *NBER Macroeconomics Annual 1989*, ed. O.J. Blanchard and S. Fischer. Cambridge, MA: MIT Press
Stock, J.H., and M.W. Watson (2001) "Forecasting Output and Inflation: The Role of Asset Prices." *Journal of Economic Literature* 47, 1–48
Wright, J. (2006) "The Yield Curve and Predicting Recessions." Mimeo, Federal Reserve Board
Zarnowitz, B. (1985) "Recent Work on Business Cycles in Historical Perspective." *Journal of Economic Literature* 22, 523–80