ISSN 1936-5330

THE FEDERAL RESERVE BANK of KANSAS CITY
RESEARCH WORKING PAPERS

Predicting Recessions with Leading Indicators: Model Averaging and Selection Over the Business Cycle

Travis Berge

April 2013; Revised January 2014
RWP 13-05

Original: April 2013
Current: January 2014

Predicting recessions with leading indicators: model averaging and selection over the business cycle*

Abstract

Four model selection methods are applied to the problem of predicting business cycle turning points: equally-weighted forecasts, Bayesian model averaged forecasts, and two models produced by the machine learning algorithm boosting. The model selection algorithms condition on different economic indicators at different forecast horizons. Models produced by BMA and boosting outperform equally-weighted forecasts, even out of sample. Nonlinear models also appear to outperform their linear counterparts. Although the forecast ability of the yield curve endures, additional conditioning variables improves forecast ability. The findings highlight several important features of the business cycle.

- JEL: C4, C5, C25, C53, E32
- Keywords: Business cycle turning points; recessions; variable selection; boosting; Bayesian model averaging; probabilistic forecasts.

Travis Berge
Federal Reserve Bank of Kansas City
One Memorial Drive
Kansas City, MO 64198
Email: travis.j.berge@kc.frb.org

*The views herein do not necessarily reflect the views of the Federal Reserve Bank of Kansas City or the Federal Reserve System.

1 Introduction

A common view of the behavior of modern economies is that they oscillate around a trend rate of growth, alternately experiencing phases of expansion and recession. Economic activity grows during expansions, generally increasing standards of living. These periods of expansion are followed by sudden and rapid declines in activity, observed across a large number of sectors in the economy. Most obviously, economic recession suggests a higher probability of unemployment and lower wage growth, but a growing literature documents many other pernicious impacts of recessions, including decreases in lifetime earnings and negative consequences for individuals' health and educational outcomes. For businesses, sluggish growth reduces demand for their wares, decreasing the availability of profitable economic opportunities and inducing the reduction of payrolls.

Given these observations, the enthusiasm with which households, businesses and policymakers attempt to infer the current and future states of the economy comes as no surprise. Classifying economic variables into variables that lead, are coincident to, and lag economic downturns is a long-lived tradition in economic research, going back to at least Burns & Mitchell (1946). The yield curve is probably the most recognized leading indicator of business cycle turning points (see, e.g., Estrella & Mishkin, 1998; Wright, 2006; Kauppi & Saikkonen, 2008; and Rudebusch & Williams, 2009), but the financial press reports on a wide-range of economic indicators. There is also an active academic research agenda focused on producing a statistic that summarizes the state of the aggregate economy, as in Stock & Watson (1999)—implemented as the CFNAI at the Federal Reserve Bank of Chicago—Aruoba, Diebold & Scotti (2009), and others. In any event, practitioners follow a wide range of economic indicators, many of which almost certainly contain additional useful information for the identification of the current or future states of the economy (though many may not).

A primary concern of this paper is the evaluation of the predictive content of a number of commonly followed macroeconomic variables. The analysis then focuses on the problem of combining disparate signals of recession from many commonly-followed economic indicators. A natural method for combining information from a wide range of sources is to use a factor model, as in Stock & Watson (1989, 1999), Chauvet (1998), Chauvet & Piger (2008) or Chauvet & Senyuz (2012). However, forecasting recessions is fundamentally a problem of classification and *a priori* there is no

clear reason to believe that the unobserved factor that best captures the cross-sectional variation in a panel of data (for example) will forecast future states of the economy well. Hamilton (2011) argues that while factor models such as the early example of Stock & Watson (1993) are accurate descriptions of the state of the economy in-sample, their usefulness as forecasting tools may be more limited if they incorporate misleading indicators, or if the behavior of the incorporated indicators has changed.

Although there are many economic indicators that provide useful signals of the present and future states of the economy, the relationships between these indicators and the state of the economy is likely not stable over time (Ng & Wright 2013). For example, the slope of the yield curve has been a useful tool for forecasting turning points in the past, but Chauvet & Potter (2002, 2005) find evidence of breaks in the relationship between the yield curve and economic activity. In addition, the financialization of the U.S. economy and decline of manufacturing likely altered the predictive content of many indicators of the economy. Further complicating matters is the well-documented asymmetry of many economic indicators around business cycle turning points (Hamilton, 2005; Morley & Piger, 2012).

The analysis extends the methodology of forecasting binary time series, focusing on the application of model selection over a set of standard macroeconomic variables. The baseline model relates each possible covariate to the state of the economy up to 24 months ahead. The analysis compares two distinct approaches to performing model selection, using the best-performing univariate model for each forecast horizon as a baseline forecast. The first approach is two applications of forecast combination. First, suite of univariate models are given equal weights to produce a probabilistic forecast. In addition, a Bayesian Model Averaging approach weights each model forecast by its implied posterior probability. Because BMA averages over all possible combinations of covariates, it allows for a richer forecast model that conditions on multiple economic indicators.

The second approach is the application of the boosting algorithm. Boosting originated in the machine learning literature focusing primarily on problems of classification, such as medical diagnostics, spam detection, and handwriting identification. Boosting is increasingly applied to empirical problems in economics.[1] The method can be specified non-parametrically—though the

[1] See, e.g., Bai & Ng (2009), Khandani, Kim & Lo (2010), Berge (2014). Ng (2014) provides a useful summary of the algorithm and also applies the algorithm to the problem of identifying business cycle turning points.

approach taken here is akin to a logistic regression—is highly efficient (and therefore able to handle large quantities of data), yet is resistant to overfitting.

The analysis investigates the in-sample and out-of-sample forecasting performance of the methods. Many of the indicators included in the analysis contain information that can be exploited to make forecasts of future states of the economy. Real economic variables most accurately describe the current state of the economy, especially indicators of the labor market. The results also point to a strong relationship between the bond market and real economic activity, but one that occurs only with a lag. The yield curve is shown yet again to be a robust predictor of future turning points endures, with the caveat that its predictive power is limited to forecasts made for the medium-term. Other interest rate spreads such as those of corporate bonds also contain information useful for forecasting, especially at very long horizons. Of the methods pursued here, BMA and the boosting algorithm produce useful probabilistic forecasts of recession. Each of the three methods produced signals of recession during the most recent 2007-2008 recession at short horizons, although the warning signals produced ahead of time were less dramatic.

The plan for the paper is as follows. The next section describes the methods used for model combination and model selection. Section 3 describes the data and the evaluation of probabilistic forecasts of a discrete outcome. Sections 4 and 5 present the empirical results.

2 Empirical setup

2.1 A baseline model

Recession forecasts are complicated by the fact that the state of the economy is an unobserved variable. To that end, this paper takes the recession dates of the business cycle dating committee of the NBER as a gold-standard chronology of the unobserved state of the economy.[2] Let Y_t denote the state of the business cycle as determined by the NBER, where $Y_t = 1$ denotes that month t is an NBER-defined recession and $Y_t = 0$ indicates an expansion instead. The model assumes that Y_t

[2] See www.nber.org/cycles. Berge & Jordà (2011) evaluate the NBER chronology itself, and conclude that the chronology is a useful classification of economic activity, and that the NBER's classification is superior to other commonly applied rules-of-thumb used to define recessions. Hamilton (2011) and Giusto & Piger (2013) note that NBER decisions regarding turning points can be delayed for quite some time, often up to a year. This is less of an issue here given that the interest is in forecasting future NBER recession dates.

is related to an unobserved variable, y_t,

$$Y_t = \begin{cases} 1 & \text{if } y_t \geq 0 \\ 0 & \text{if } y_t < 0. \end{cases}$$

and that y_t is determined by a vector of observables x

$$y_t = f(x_{t-h-1}) + \varepsilon_t. \tag{1}$$

x is a $(K+1) \times 1$ vector of K observables plus a constant, $f(.)$ is a function that maps $\mathbb{R}^{K+1} \to \mathbb{R}$, the subscript h denotes the forecast horizon, and ε_t is an iid shock with unit variance.[3] A typical probit or logit model would specify $f(x)$ as a linear function, but equation (1) is written more generally to encompass a variety of possible specifications.

The objective in this paper is to forecast the state of the economy conditional on a set of covariates. Let E be the expectations operator, and let $p_{t|t-h-1}$ denote the conditional probability of recession, so that

$$E\left[Y_t = 1 | x_{t-h-1}\right] \equiv p_{t|t-h-1} = \Lambda(y_t), \tag{2}$$

where $\Lambda(.)$ is a twice-differentiable cumulative distribution function.

A large literature has used a model similar to that in (1) and (2) to relate future economic activity to the term structure of interest rates. Estrella & Mishkin (1998), Wright (2006) and Rudebusch & Williams (2009) focus on simple limited dependent models, conditioning on the slope and level of the yield curve. Each finds that simple probit or logit specifications work quite well for predicting future turning points. There are many ways that one could extend the specification described in (1) and (2). Dueker (2005), Chauvet & Potter (2005) and Kauppi & Saikkonen (2008) focus on dynamic specifications of the same basic setup. Chauvet & Senyuz (2012) provide a more modern specification, as they relate the state of the economy to a dynamic factor model of the yield curve. Recently, Giusto & Piger (2013) use tools from machine learning to evaluate how quickly one can call a recession in real-time (that is, they focus on nowcasting instead of forecasting).

[3] One shortcoming of this approach is that it does not attempt to account for the well-documented decline in volatility of macroeconomic aggregates. However, models that focus on predicting a binary indicator of the state of the economy tend to be more robust to changes over time than models that forecast a continuous outcome (Estrella, Rodrigues & Schich 2000). In addition, Chauvet & Potter (2005) estimate a probit model that allows for business cycle specific variance and an autoregressive component. While they find that the inclusion of an AR component and changing volatility improves the in-sample fit of the model, the improvement in forecast ability due to changing volatility is less clear.

While the yield curve on average is a leading indicator of economic downturns, there are many reasons why conditioning on additional economic indicators is likely to improve forecast ability. First, the yield curve may not be a stable predictor of recessions. Time-variation in risk and term premia complicate the relationship between the yield curve and macroeconomic variables. Institutional changes to the market for Treasury debt may also impact the relationship between the yield curve and economic activity. Secondly, since the yield curve is a summary statistic of market expectations for the future path of short-term interest rates (and implicitly the monetary policy reaction function) it likely captures the impact of monetary policy shocks quite well. To the extent that shocks to the real economy also alter the probability of future recession, inclusion of other variables that describe the real economy will improve the forecast ability of a particular model. This complicates the model selection problem substantially. Practitioners follow a wide-range of economic indicators so that the choice of which variables to include in the forecasting model is not obvious. Moreover, the nonlinear behavior of real variables across the business cycle complicates model specification.

The remainder of this section introduces the two methods that are used to combine information from different leading indicators, forecast combination and model selection.

2.2 Averaging model forecasts

Model averaging has long been recognized as a useful method of combining information from a given set of models. Previous applications have shown that model averaging tends to improve forecast accuracy, either because the combination either combines information from partially overlapping information sets (Bates & Granger 1969), or because the combination alleviates possible model misspecification (Hendry & Clements, 2004; Stock & Watson, 2004; Timmermann, 2006). In addition, model weights themselves can be of interest if they are constructed so that they give the posterior probability that a given model produced the observed data. In the current application, these posteriors reveal information regarding the utility of particular indicators at various forecast horizons.

Model averaging has a solid statistical foundation and is straightforward to implement. Each of a set of K covariates is estimated as a univariate model to produce a forecast of some event y_t, resulting in $\{\hat{y}_{1t}, \hat{y}_{2t}, ..., \hat{y}_{Kt}\}$. In the current application, recall that y_t is the latent variable

that relates covariates to the aggregate state of the economy. The combination problem is to find weights w_k for each forecast to combine the individual forecasts into a single forecast $\hat{y}_t^C = C(\hat{y}_{1t}, ..., \hat{y}_{Kt}, w_1, w_2, ..., w_K)$. In principle, because forecasts are useful only to the extent that they impact the actions of policymakers and other economic agents, the weights are the outcome of the minimization of a loss function, which in turn could reflect the underlying utility of decision makers, as in Elliott & Lieli (2009). However, in the current application the tradeoff between true and false positives or true and false negatives for recession forecasts is unclear. Thus, instead of focusing on the loss function, the empirical fit of the models is used to produce weighting schemes for the recession forecasts.

The recession forecast first is the most basic application of model averaging: assume each model is equally useful and give each of the K forecasts the same weight. The equally weighted forecast of the latent is then

$$\hat{y}_t^{EW} = \frac{1}{K} \sum_{i=1}^{K} \hat{y}_{it}. \tag{3}$$

Producing unweighted averages is a simple method to combine information from several different models. However, the application is limited to univariate regressions. In order to more fully explore the possible model-space, a Bayesian Model Averaging (BMA) framework is also used to produce forecasts. In BMA, each model (which may include a set of covariates) receives a weight and the final estimated model is a weighted average of that set of models. Since each model implies a forecast for the future state of the economy, the BMA-implied forecast is a weighted forecast, where each models' weight is determined by its posterior probability.

There are a set of $M = 2^K$ models, where each model M_i is parameterized by Θ_i. Note that in contrast to the unweighted scheme implemented above, each Θ_i may have differing dimensions. Once the model space is constructed, the Bayesian model averaged forecast is the probability-weighted sum of the model-specific forecasts:

$$\hat{y}_t^{BMA} = \sum_{i=1}^{M} \hat{y}_{it} \Pr(M_i | D_{t-h-1}) \tag{4}$$

where $\hat{y}_{it}$ is the forecast from model i and $Pr(M_i|D_{t-h-1})$ denotes posterior probability of model i conditional on the data available at the time the forecast is made.

By Bayes' Law, the posterior probability of model i is proportional to that model's marginal

likelihood multiplied by its prior probability. Let $P(M_i)$ denote the prior belief that model i is true. Given the a set of priors, a researcher observes the data D, then updates beliefs in order to compute the posterior probability of model i:

$$Pr(M_i|D) = \frac{Pr(D|M_i)Pr(M_i)}{\sum_{j=1}^{M} P(D|M_j)P(M_j)} \tag{5}$$

where $Pr(D|M_i) = \int Pr(D|\theta_i, M_i)Pr(\theta_i|M_i)\partial\theta_i$. Directly implementing (5) requires the calculation of a marginal likelihood. This is conceptually and computationally demanding since the average is taken over the prior distribution for all model's parameters.

However, the Bayesian information criterion is a consistent estimate of the marginal likelihood of a model (Raftery 1995). This estimate of the marginal likelihood is commonly used in applied work, and is advantageous since its it requires only a maximum likelihood estimate and allows the researcher to set aside the production of priors for each model's parameters (see, e.g., Sala-I-Martin, Doppelhofer & Miller (2004), Brock, Durlauf & West (2007) and Morley & Piger (2012)). Each model is assumed equally likely *a priori*. Given these assumptions, model posterior probabilities are calculated as model fit relative to the fit of all models, or

$$Pr(M_i|D) = \frac{\exp(\widehat{BIC_i})}{\sum_{i=1}^{M} \exp(\widehat{BIC_i})}.$$

2.3 Model selection via the boosting algorithm

Model averaging is one solution to the problem of model specification. An alternative solution to the problem is to perform model selection. This section introduces a model selection algorithm as a methodology that can produce empirically-driven forecasting models of the business cycle. We wish to model the relationship between the observed discrete variable Y_t and a vector of potential covariates, $x_t = (x_{1t}, x_{2t}, ..., x_{Kt})$. In the section above, this was achieved by weighting different models, either equally or based on a measure of in-sample fit. The approach here is more general, in the sense that we wish to endogenously model the choice of covariates to be included in the model (that is, forecasts may be made using only a subset of $x_1, x_2, ..., x_K$). The method allows for a potentially non-linear relationship between the latent and the covariates.

These goals are accomplished by estimating a function $F : \mathbb{R}^K \to \mathbb{R}$ that minimizes the expected

loss $\mathcal{L}(Y, F)$, i.e.,

$$\hat{F}(x) \equiv \arg\min_{F(x)} E\left[\mathcal{L}(Y, F(x))\right]. \tag{6}$$

We require only that the loss function is differentiable with respect to the function F. The setup encompasses many different types of problems. For example, if Y were a continuous outcome, then specifying $\mathcal{L}(Y, F(x))$ as squared-error loss and $F(x)$ as a linear function results in a problem analogous to a standard OLS regression. For models of binomial variables, the loss function is typically specified as the negative of the log-likelihood of the error's distribution. F can be specified very generally—in the machine-learning literature it is common to specify $F(x)$ as decision trees, a non-parametric method. Smoothing splines are another common choice.

The following algorithm minimizes the empirical counterpart to (6) by specifying that $F(x)$ is an affine combination of so-called 'weak learners,' each of which are specified separately. The algorithm is due to Friedman (2001) and can be summarized as follows. First, initialize the learner in order to compute an approximate gradient of the loss function. Step 3 fits each weak learner to the current estimate of the negative gradient of the loss function. Step 4 searches across each weak learner and chooses the one that most quickly descends the function space and then chooses the step size. In step 5 we iterate on 2-4 until iteration M, which will be endogenously determined.

Functional Gradient Descent.

1. *Initialize the model.* Choose a functional form for each weak learner, $f^{(k)}, k = 1, ...K$. Each weak learner is a regression estimator with a fixed set of inputs. Most commonly each covariate receives its own functional form, which need not be identical across each variable.

   Let m denote iterations of the algorithm, and set $m = 0$. Initialize the strong learner F_0. It is common to set F_0 equal to the constant c that minimizes the empirical loss.

2. *Increase m by 1.*

3. *Projection.* Compute the negative gradient of the loss function evaluated at the current estimate of F, $\hat{F}_{m-1}$. This produces:

$$\mathbf{u_m} \equiv \{u_{m,t}\}_{t=1,...,T} = -\frac{\partial \mathcal{L}(Y_t, F)}{\partial F}|_{F=\hat{F}_{m-1}(x_t)}, t = 1, ..., T$$

   Fit each of the K weak learners separately to the current negative gradient vector $\mathbf{u_m}$, and produce predicted values from each weak learner.

4. *Update F_m.* Let $\hat{f}_m^{(\kappa)}$ denote the weak learner with the smallest residual sum of squares among the K weak learners. Update the estimate of F by adding the weak learner κ to the estimate of F:

$$\hat{F}_m = \hat{F}_{m-1} + \rho \hat{f}_m^{\kappa}$$

Most algorithms simply use a constant but sufficiently small shrinkage factor, ρ. Alternatively, one can solve an additional minimization problem for the best step-size.

5. *Iterate.* Iterate on Steps 2 through 4 until $m = M$.

The two parameters ρ and M jointly determine the number of iterations required by the algorithm in order to converge. Small values of ρ are desirable to avoid overfitting, since the computational cost of additional iterations is low. In the applications below, M is chosen to minimize the Schwarz information criterion; i.e. $M \equiv \arg\min_m BIC(m)$. A weak learner can be selected many times throughout the course of the boosting algorithm, or not at all. This data-driven approach of model selection is very flexible; different specifications of the loss function $\mathcal{L}$ and the weak learners lead to approximations of different models. When used in a forecasting exercise, re-estimating the model at each point in time also allows the relationship between covariates and the dependent variable to change.

3 Data and evaluation

3.1 Data

The implementation of either the forecasting schemes described in the previous section requires that the model-space to be defined. In the interest of parsimony, and following the seminal work of Estrella & Mishkin (1998), the analysis is limited to the commonly followed financial and macroeconomic indicators listed in table 1. While the majority of the literature focuses on the slope of the yield curve as a predictor of future economic activity, in principle other features of the curve, such as its level and curvature, may also be important predictors. The level, slope and curvature of the yield curve are constructed using monthly averages of the daily yields of zero-coupon 3-month, 2-year and 10-year yields compiled by Gurkaynak, Sack & Wright (2007). Specifically, the level of the curve is calculated by taking the mean of the 3-month, 2-year and 10-year yields; the slope of the curve is constructed as the difference between the 10-year and the 3-month yields; and curvature is measured by taking the difference between two times the 2-year yield and the sum of the 3-month and 10-year yields. In addition, the TED spread and two corporate bond spreads measure

the degree of credit risk in the economy. Other financial indicators in the empirical model include money growth rates (both nominal and real), and, because they are forward looking, changes in a stock price index and the value of the U.S. dollar. The VIX is included in the model search in recognition that financial volatility may presage a decline in real economic activity.[4]

Several variables that describe the real economy are also included as predictors in the model. Industrial production serves as a proxy for output. Housing permits proxy for the housing market. In order to gauge the health of the labor market, the four-week moving average of initial claims for unemployment insurance and a measure of hours worked are included in the model. The purchasing managers index, a commonly followed leading indicator, is also included in the model. The data span the period January 1973–June 2013 at a monthly frequency. Table 1 lists the indicators in detail.

[Table 1 about here.]

3.2 Evaluating forecasts of discrete outcomes

Three metrics are used to perform model evaluation. The first measures in-sample fit and is analogous to the R^2 statistic of a standard linear regression. The other two measures focus on the predictive ability of each model, with one measure focusing on each model's ability to classify future dates into recessions and expansions.

The pseudo-R^2 developed by Estrella (1998) measures the goodness-of-fit of a model fit to discrete outcomes. The pseudo-R^2 can be written

$$\text{pseudo}R^2 = 1 - \left(\frac{\log L_u}{\log L_c}\right)^{-(2/n)\log L_c}, \tag{7}$$

where L_u is the value of the likelihood function and L_c is the value of the likelihood function under the constraint that all coefficients are zero except for the constant. As with it's standard OLS counterpart, a value of zero indicates that the model does not fit the data whereas a value of one indicates perfect fit.

The test statistic of Giacomini & White (2006) is used to evaluate predictive ability. The GW test assesses the difference between the loss associated with the predictions of a prospective model

[4] VIX is taken from CBOE and is available from 1990 to present. Prior to1990, VIX is proxied by the within-month standard deviation of the daily return to the S&P 500 index, normalized to the same mean and variance as the VIX for the period when they overlap (1990-2012). These two series are highly correlated: $\rho = 0.869$.

to those of a null model. Importantly, given the myriad of methods used to make forecasts here, the statistic compares *conditional* forecast ability, and does not depend on the limiting distribution of model parameters. The method also properly evaluates nested models. Let $L(\hat{\varepsilon}_t^i)$ be a loss function, where $i \in \{0, 1\}$ denotes the model used to produce the forecast, and $\hat{\varepsilon}_t^i \equiv \hat{y}_t^i - y_t$ is the forecast error. The test statistic that evaluates P predictions can be written as:

$$GW^{1,0} = \frac{\Delta\bar{L}}{\hat{\sigma}_L/\sqrt{P}} \rightarrow N(0,1), \text{where}$$
$$\Delta\bar{L} = \frac{1}{P}\sum_t \left(L(\hat{\varepsilon}_t^1) - L(\hat{\varepsilon}_t^0)\right); \quad \hat{\sigma}^2 = \frac{1}{P}\sum_t \left(L(\hat{\varepsilon}_t^1) - L(\hat{\varepsilon}_t^0)\right)^2$$

The GW test is evaluated using two loss functions: absolute error and squared error. These measures measure the distance between the probabilistic forecast and the 0-1 outcome. Squared-error loss is closely related to the commonly-used Quadratic Probability Score (Brier & Allen 1951), another popular evaluation test for probabilistic forecasts. Tables display p-values from GW tests that are robust to heteroskedasticity and autocorrelation.

One well-known drawback of the QPS is that it does not measure resolution or discrimination. For this reason, the final tool used to evaluate the forecasts explicitly recognizes that the problem is one of classification. Classification is a distinct measure of model performance since two models could have different model fit but still classify the discrete outcome in the same way (Hand & Vinciotti 2003). Specifically, the area under the Receiver Operating Characteristics (ROC) curve is used to evaluate each model's ability to distinguish between recessions and expansions. The ROC curve describes all possible combinations of true positive and false positive rates that arise as one varies the threshold used to make binomial forecasts from an real-valued classifier. As a threshold c is varied from 0 to 1, a curve is traced out in $\{TP(c), FP(c)\}$ space that describes the classification ability of the model. The area underneath this curve (AUC) is a well-known summary statistic that describes the classification ability of a given model.[5] The statistic has a lower bound of 0.5 and an upper bound of 1 where a higher value indicates superior classification ability. The statistic has standard asymptotic properties, although for inferential purposes standard errors are found with the bootstrap.

An important advantage of ROC curves relative to the Giacomini-White test described above is

[5] For a complete introduction to ROC curves, see Pepe (2003).

that it does not require the specification of a loss function. Instead, the ROC curve is a description of the tradeoffs between true positives and false negatives produced by a forecasting model. In the current application, this is advantageous since it is hard to know how to weigh true and false positives against true and false negatives when forecasting states of the business cycle.

4 In-sample results

Although the primary interest is in the out-of-sample performance of the two models, this section first evaluates their in-sample performance. As an initial investigation into the statistical and predictive power of commonly acknowledged leading indicators, table 2 presents estimates of univariate logit models at each horizon. For each variable, the full sample of available data is used to estimate model parameters, which are then used to produce predicted recession probabilities. The predicted probabilities are then compared to NBER-defined recessions for evaluation. Table 2 displays three summary statistics that measure model fit for each variable and at each forecast horizon. The first statistic contains the pseudo-R^2 that provides a measure of model fit. The second row is the t-statistic that tests the null hypothesis that the model coefficient from the logit regression is zero. The final row is the AUC statistic. In the interest of concision, tests of forecast accuracy are not presented for in-sample results.

[Table 2 about here.]

Variables that describe real economic activity perform best in the very short-term. For variables that describe economic activity, both model fit and classification ability decline as the forecast horizon grows, an observation that is not unsurprising since the NBER recession dates themselves are based on the contemporaneous behavior of highly correlated variables. Industrial production, monthly employment gains, and initial claims are useful indicators of the state of the economy at short horizons, each with high pseudo-R^2 and AUC statistics approaching 0.90. In both instances, however, the predictive power of the indicator is limited at horizons longer than 12 months.

In contrast, many of the financial indicators exhibit forward-looking behavior. The slope of the yield curve performs best at the horizon of 12 months, and the other yield spreads also appear to be forward-looking. Unconditionally, the level and curvature of the yield curve appear to contain only modest information about the state of the economy. Variables that reflect turmoil in financial

markets give information that reflects the probability of a recession, particularly in the near-term. Changes in stock prices also appear to contain modest explanatory power, although only at short forecast horizons. Neither nominal nor real money growth appear to consistently contain useful information across forecast horizons, with very low pseudo-R^2 values and modest forecast ability. Finally, although model fit and classification ability are distinct features of a model, in the application here the two align very closely. Models that have a better fit also display superior classification ability.

4.1 Model averaging

Table 3 evaluates the forecasts produced by the two weighting schemes described in section 2.2. The top panel presents results from the equally-weighted forecasts of each univariate model using covariate K; the bottom panel weights each forecast model according to its posterior probability as in equation (4). For each weighting scheme, three statistics are presented: the AUC statistic and the Giacomini-White statistic. The null model for each forecast horizon is the best performing univariate model as measured by the AUC in Table 2.

Each forecasting method produces accurate classification of NBER recession dates. The AUC for each method is close to unity when producing nowcasts, and tends to deteriorate as the forecast horizon increases. Combining forecasts with equal weights produces forecasts that do not perform much worse than their univariate counterparts. Combining forecasts often improves forecast ability, especially when the forecasts are weighted by their in-sample posterior probability. However, the Giacomini-White test statistics show that model combination does not necessarily improve forecast ability beyond the best-performing univariate logit model. Equally weighted forecasts on average perform worse than the best-performing univariate model (although they sidestep the need to elicit the best model). The BMA-weighted forecasts do tend to perform better than the best-performing univariate model. The exception is the 12-month ahead forecast—it is difficult to improve on the forecast produce by a univariate model of the slope of the yield curve at this horizon.

[Table 3 about here.]

Figure 1 gives further insight into the performance of the Bayesian Model Average models. The figure displays the posterior inclusion probabilities (PIP) for each covariates and each forecast model. The PIP for is the probability that a particular covariate is included in a model. It can be

thought of as a weighted average of the poster probabilities for each of the 2^K models that includes covariate j.[6] The posterior probabilities of the models strongly favor only a handful of variables to produce the model forecasts. Contemporaneously, both employment data, the Ted spread and housing data are included in the forecast model. The slope of the yield curve is strongly preferred at forecast horizons of six and twelve months; at a forecast of 12 months, no other covariate has a PIP greater than 10 percent. The models at longer horizons–18 and 24 months–are less intuitive. Table 2 revealed that corporate yield spreads dominate in terms of model fit, but the PIPs for the trade-weighted dollar and initial claims of unemployment receive large weight in these models as well.

[Figure 1 about here.]

These results highlight the stark difference between equal weights and weighting according to model fit. On the one hand, equal weights allows for a broad range of information to enter into the forecast model but does not differentiate between information useful for the forecasting problem. In contrast, the BMA methodology gives a higher weight to models that have a better in-sample fit. The BMA methodology highlights that different economic indicators carry very different information at various forecast horizons.

4.2 Forecasts from the boosted models

4.2.1 A linear model

This section evaluates the empirical performance of the boosting procedure described in section 2.3. Equation (6) is estimated using a loss function of negative one-half times the Bernoulli log likelihood function. Each indicator listed in Table 1 is included in the model search. The initial weak learner specified as a univariate linear function; this is equivalent to the logit models used in the forecast averaging exercise. At each iteration m, the covariate that minimizes the empirical

[6] That is, $PIP(\beta_j) = Pr(\beta_j \neq 0) = \sum_{M_i:\beta_j \in M_i} p(M_i|D)$.

loss at that iteration is included in the forecast model. After the M iterations, the final model is:

$$\begin{aligned} \hat{F}_M(x) &= \sum_{m=1}^{M} \rho_m \hat{f}_m(x) \\ \hat{f}_m(x) &= f^{\kappa}(x) \\ \kappa &= \arg\min_k \sum_{t=1}^{T} (u_t - \hat{f}^k(x_t))^2 \end{aligned} \tag{8}$$

I set $\rho_m = \rho = 0.1$, as is common in the boosting literature. The number of boosting iterations M is chosen so that the final boosted model has minimum BIC[7] .

Table 4 presents the in-sample estimates of equations (6) and (8). As the number of iterations grows large, the boosted model is equivalent to a 'kitchen-sink' logit model. Thus as a method of comparison, the table presents the ratio of the coefficient from the boosted model to its unrestricted 'kitchen-sink' logit counterpart for each variable included in the model search (i.e., $\beta_{boost}/\beta_{kitchensink}$).[8] The forecasting models produced by the boosting model differ from their BMA counterpart in that there are many more indicators included in the forecast model at each horizon on average. However, the coefficients for the variables included in the forecasting model are shrunk significantly towards zero. The pattern revealed by the in-sample BMA analysis is seen again here: at short horizons the method relies largely on indicators of real economic activity. The slope of the yield curve dominates the models used to produce forecasts into the medium term. The model forecasting at longer horizons again include many indicators not included in the models that forecast at shorter horizons: the level of the yield curve, dollar depreciation and corporate yield spreads.

[Table 4 about here.]

The left panel of figure 2 below presents the model selected for each forecast horizon in a slightly different way. The figure shows the fraction of iterations that the boosting algorithm selected a particular covariate for both the linear and non-linear variants of the model.[9] Although there are

[7] Throughout, BIC is defined as $-2\times\ln(L) + 2\times\log(N)$. Thus, smaller values of the BIC indicates better (penalized) fit.

[8] Since the boosted model minimizes one-half the log of the odds-ratio, coefficients from the boosted model are doubled to facilitate comparison.

[9] The plot is of ψ_k^h for each covariate k. h denotes forecast horizon and κ denotes the covariate with minimum mean squared error at iteration m.

$$\psi_k^h = \frac{1}{M} \sum_{m=1}^{M} I(k = \kappa); \tag{9}$$

many more variables included in the boosted model, they are selected are qualitatively similar to the forecasting models produced by Bayesian Model Averaging. At short forecast horizons, the forecast model primarily relies on measures of real activity. When forecasting a six and twelve months ahead, the model relies on the slope of the yield curve to forecast business cycle turning points. Only at some horizons do other elements of the yield curve enter the model—curvature enters into models forecasting into the medium-term—and the level of the curve is informative for forecasts into the very distant future. Corporate yield spreads are included in the models that forecast into the distant future. Finally, it is worth noting that several variables are never or only marginally included in the final forecasting models. The measures of money growth, the curvature of the yield curve and VIX rarely enter the model.

The value of the maximized BIC is shown in the table. In the interest of brevity I do not report the BIC models from the univariate models or for the kitchen-sink logit models, but the BIC is as expected. The sparse boosted models are strongly preferred to the kitchen-sink logit models; for example, at $h = 12$, the BIC of a kitchen-sink logit model is 369.5, which is clearly dominated by the boosted model. The BICs of the boosted models also tend to dominate the best-fitting univariate models of table 2. The exception is at the 12-month horizon. The BIC of the univariate model using the slope of the yield curve is 289.4, similar to the boosted model's BIC of 291.3.

Finally, the bottom half of table 4 describes the in-sample forecast ability of each model. In terms of classification, the models produce predicted probabilities that track NBER recession dates very well, particularly at short horizons. When nowcasting, the linear model achieves an AUC statistic of 0.97, indicating that model achieves near-perfect classification of economic activity contemporaneously. At longer horizons the AUC decreases—at a forecast horizon of two years the AUC falls to 0.84.

4.2.2 A nonlinear model

In a standard logistic model, the unobserved latent variable depends on the covariates in a linear fashion, as in equation (1). However, this functional form is used only because it is convenient, and there are good reasons to believe that a non-linear specification is appropriate. Financial market indicators are often erratic and behave in a non-linear fashion. Real economic variables also move non-linearly across the business cycle, as documented by Hamilton (2005) and Morley & Piger

(2012).

For these reasons, non-linearity is introduced into the forecast model using the smoothing splines of Eilers & Marx (1996). Incorporating smoothing splines into the boosting algorithm is straightforward. The weak learner for each covariate k within the boosting algorithm is specified to minimize the penalized sum of squared error:

$$PSSE(f^k, \lambda) = \sum_{t=1}^{T} [y_t - f^k(x_t)]^2 + \lambda \int [f^{k''}(z)]^2 dz \tag{10}$$

where the smoothing parameter λ determines the magnitude of the penalty for functions with a large second derivative. Splines have several attractive features: they are estimated globally, conserve moments of the data and are computationally efficient as they are extensions of generalized linear models.[10] As with the linear case, at each iteration m the covariate that best minimizes the empirical loss at that iteration is included in the forecast model. Let f_m^k be the smoothing spline fit to indicator k at iteration m, then at each iteration the weak learner can be expressed:

$$\begin{aligned} \hat{f}_m(x) &= f^{\kappa}(x) \\ \hat{f}^k(x) &= \arg\min_{f(x)} \Big[PSSE(f, \lambda, x^k) \Big] \\ \kappa &= \arg\min_{k} \sum_{t=1}^{T} (u_t - \hat{f}^k(x_t))^2 \end{aligned} \tag{11}$$

The number of boosting iterations M is chosen to be the one that minimizes the BIC of the final boosted model.

The performance of the non-linear models in-sample is impressive. Table 5 displays the results of the non-linear model fit to the full sample. Interestingly, comparing information criteria to that from the linear model presented in table 4, it is not clear whether the data prefer a non-linear version of the forecasting model. The BIC appears to prefer a nonlinear at medium horizons, but the linear model appears to be a suitable specification at very short and very long forecast horizons. The non-linear models classify the data slightly more accurately than their linear counterparts. For each forecast horizon, the AUC indicates that the non-linear specification improves the classification ability of the indicators. The difference between forecasting models is most notable at horizons of

[10] Eilers & Marx approximate the penalty term in (10) by constraining the difference in parameter values of the spline in neighboring regions of the data, transforming the problem into a modified regression equation that is computationally very efficient. Buhlmann & Yu (2003) recommend setting $\lambda = 4$, which is the value used here.

12 months and higher.

[Table 5 about here.]

The right panel of figure 2 presents the fraction of time the algorithm included a particular covariate at each iteration for each non-linear forecasting model. It is clear that no covariate dominates a forecasting model as was the case when performing model averaging. Nonlinearity allows some covariates to enter into the forecasting model when they were excluded from the linear case. For example, VIX, which was rarely included in the linear model, enters into the nonlinear model at all forecast horizons, and gets a relatively heavy weight contemporaneously.

[Figure 2 about here.]

4.3 Comparison to NBER recession dates

A key lesson from the analysis above is that the information carried by economic indicators varies widely by forecast horizon. When forecasting the contemporaneous probability of recession, many of the indicators were found to carry little or no information. While model averaging can alleviate concerns of model misspecification, the average still depends critically on the underlying indicators. This can be seen in figure 3, which displays the recession probabilities from each of the methods estimated at a forecast horizon of zero months. Averaging many indicators—some of which contain only noise regarding the current state of the economy—results in a forecast that hovers around the unconditional probability of recession (upper-left figure).

The BMA forecast (panel b) highlights the opposite extreme, in the sense that the averaged model relies on only a handful of indicators. The forecasts are clearly an improvement over the unweighted average, probability of recession clearly crosses the relevant threshold during recession. The bottom half of the figure shows the in-sample forecasts from the boosted models. The BMA and boosted models produce in-sample forecasts that are quite similar. The probabilities from each cross the threshold during NBER recessions, and conditional on the use of the optimal threshold, there are no false positives in the sample. The 2001 recession was quite difficult to recognize in real-time. Although each model produces a predicted probability that crosses that model's relevant threshold, only the non-linear boosted model produces predicted probability that stays above the threshold for more than three consecutive months.

[Figure 3 about here.]

The in-sample probabilities of recession can be combined with the threshold value to produce a chronology of business cycle turning points for the U.S. economy. The threshold displayed in figure 3 is produced under the assumption of symmetric utility/disutility from true/false positives. Under this assumption, the utility of a classifier is the difference between the true and false positive rates (Berge & Jordà 2011). The threshold shown for each forecast maximizes this difference. Table 6 uses this threshold to produce peaks and troughs from each of the four models, and displays those dates relative to NBER recession dates. For each model, peaks and troughs are the first and final month for which the recession probability is equal to or greater than the threshold, with the additional natural restriction that each phase of the business cycle lasts more than three months. The table displays those months relative to NBER recession peaks/troughs.

Generally, the recession dates align with the NBER dates quite closely. Chronologies from the unweighted and BMA-implied models alight somewhat less closely than dates from the boosted models. Each has a relatively large miss: the unweighted average misses the beginning of the 2007 recession by 8 months and the BMA model has a difficult time clearly identifying the start of the 1973 recession. The model averaging schemes also produce a false negative event as they do not identify the 2001 recession. The linear boosted event also produces a false negative event for the 2001 recession. In the case of the BMA forecast and the linear boosted model forecast, the probability of recession does cross the threshold, but does not stay above the threshold for 3 consecutive months so that the additional rule that a recession last more than three months filters these signals out. Finally, each of the models appear to have an easier time identifying trough dates rather than peak; for example, the trough dates from the non-linear boosted model are never more than one month different than the NBER-defined dates.

[Table 6 about here.]

5 Out-of-sample performance

Both BMA and the boosting model selection algorithm produce highly accurate probabilistic forecasts of recession in-sample. However, aggressive model search may produce models that overfit the data, which would reduce the out-of-sample forecast ability of the method. Weighting forecasts

with a statistical measure of fit carries the same risk. This section considers the out-of-sample performance of the forecasts produced by the four methods presented above.

Forecasts are produced using an expanding window, and the initial out-of-sample forecast is made for May 1985. From this point forward, at each point in time, a total of 20 forecast models are estimated: each of the four different model produces forecasts of recession at five horizons, zero, six, 12, 18 and 24 months. After the forecasts have been produced, an additional data point is added to the model and the process is repeated.

Table 7 displays the results of the out-of-sample exercise. Although the forecast performance is diminished relative to the in-sample forecasts, each method appears to produce valuable recession forecasts. With the exception of the equally weighted forecasts, the AUC tends to exceed 0.85. The nonlinear boosted model performs very well out-of-sample, maintaining AUCs exceeding 0.90 for each but the longest forecast horizon.

As before, the null forecast used for comparison in the Giacomini-White test is the best-performing univariate forecast at that point in time, as measured by the in-sample AUC statistic. Using the best-performing univariate model is a high-hurdle since it itself involves some degree of model search, but it is intended to measure the ability of the various model combination schemes to combine disparate information. The equally-weighted model averaging scheme performs about as well as the best-performing univariate model. The more sophisticated model averaging and model selection schemes are more successful than the best-performing univariate model, especially at short horizons. The nonlinear boosting algorithm outperforms each of the other two the model combination schemes, providing evidence that the nonlinear specification is helpful when producing forecasts of the state of the economy.

[Table 7 about here.]

The use of model selection schemes out-of-sample may produce unstable models. In particular, Estrella et al. (2000) and Chauvet & Potter (2002) have argued that the relationship between the macroeconomy and the slope of the yield curve may not be stable. Similarly, Chauvet & Potter (2005) argue that a predictive model that allows for structural breaks improves the forecast ability of the yield curve. As a check on the stability of the forecasting models selected, figures 4 and 5 display the PIPs and selection frequency of models used to forecast out of sample. Figure 4 gives

the results of the BMA method when used out-of-sample for the forecasting horizons of zero (left) and 12 (right) months.Figure 4 shows the PIPs for the five variables found to have the largest PIP, on average, for the out-of-sample exercise.

The left panel of figure 4 is suggestive of regime changes in the nowcasting model, especially following the 1991 and 2001 recessions. Interestingly, the behavior does not appear to have changed dramatically following the most recent recession, although there does appear to be evidence of a regime shift occurring following the 2001 recession. In contrast, the posterior probabilities of the model forecasting 12 months ahead do appear to have changed following the most recent recession but were more stable surrounding prior recessions. Prior to 2007, the slope of the yield curve and housing permits both had PIPs greater than 70 percent. Following the 2007 recession, however, the PIPs for housing permits, the curvature of the yield curve, and the AAA-10Y Treasury spread all fell to zero and only the slope of the yield curve received a high weight. The PIPs for variables not shown do not show a significant change (and are generally well below 10 percent).

The boosted models display a much higher degree of flexibility. Figure 5 is the out-of-sample analogue of figure 2; it displays the frequency of times for which a given covariate is selected by the boosting algorithm for the linear model that forecasts the twelve month ahead probability of recession. As in figure 4, the figure shows the variables with the highest inclusion probability, on average.

The figure shows that the method includes many more indicators on average than the BMA-produced models. Since there are 18 possible covariates, if each covariate were purely noise, one would expect each covariate to be chosen approximately 5 percent of the time. That payroll employment is included in the model much more frequently than 5 percent confirms that the boosted model relies on that variable contemporaneously. Similarly, the slope of the yield curve is included in the forecast model frequently for the model forecasting a year ahead. In addition, the boosted models appears quite stable throughout sample period, relying primarily on employment indicators and the slope of the yield curve, respectively.

[Figure 5 about here.]

Figures 6 and 7 display the out-of-sample forecasts surrounding the 2001 and 2007 recession events. Specifically, each figure displays the nowcast (left panel) and 12-month ahead forecast (right

panel) out-of-sample probabilistic forecast of recession (i.e, $P(NBER_t|D_{t-h-1})$ in three month increments, beginning one year prior to the NBER-defined recession. Of course, probabilistic forecasts need to be considered within the context of a threshold with which to make binomial classifications. Each model's in-sample optimal threshold are shown in figure 3 for the full-sample. The thresholds change little when estimated in a true, out-of-sample way and so are suppressed in the interest of simplicity.

Focusing first on the 2001 recession, table 6 already revealed that the models were hard-pressed to identify this recession, even ex-post. Out-of-sample, this continues to be true, although three of the models—BMA and the two boosted models—do provide strong signals of recession in fall 2001, with estimated probabilities of recession exceeding 0.85. The signals, however, are short-lived: only the probabilities from the non-linear boosted model exceed their threshold for more than three consecutive months. When forecasting 12 months ahead, only the nonlinear boosted model provided a signal of upcoming recession. The recession probability estimates for September 2001–January 2002 all exceed 65 percent, indicating that the model was signaling recession for a brief period of time during the fall of 2000. The only other model to produce a recession signal of greater than 50 percent is the BMA model, which forecasted a forecast exceeding 0.5 for one month (December 2001).

[Figure 6 about here.]

The forecast models were somewhat more successful identifying the 2007 recession event. The recession probabilities are shown in figure 7. Each of the forecast models signals a substantial recession risk contemporaneously. The equally weighted forecast exceeded its threshold value of 0.21 in September of 2008, and achieves a maximum value of 0.48 in October 2008. The BMA and boosted models sent a more clear signal, as the BMA forecasts and the forecasts produced by the boosted models signal a high risk of recession beginning in September of 2007. Each signaled a very high risk of recession, with probabilistic forecasts approaching 1, beginning in September of 2008. The models signaled a high risk of recession until the middle of 2000: the nonlinear boosted model fell below its threshold value beginning in June 2009, the BMA and linear boosted model followed suit in July. This aligns with the behavior of many of the real economic variables: initial claims for unemployment insurance peaked in March 2009, the rate of job losses in the economy bottomed in

March 2009 (though employment continued to shrink throughout 2009; employment gains turned positive consistently in March 2010), and industrial production bottomed in June 2009. Quarterly variables not used in the forecasting model such as GDP and GDI also troughed in mid-2009.

On the other hand, the right-side panel of figure 7 shows that the forecasting models did not produce strong signals of recession. The strongest signal sent by a model forecasting 12 months ahead during this period was produced by the BMA forecasts, which signaled a 45 percent probability of recession 12-months ahead in February of 2008. The signal was relatively short-lived, however, as the forecast probability exceeded 25 percent only between November 2011 and June 2008. That the signal was not stronger is perhaps somewhat surprising: the yield curve did invert for a brief period in late 2006 and early 2007 and it was clear that the housing market was in trouble at that point: housing permits, the other signal used by the BMA model fell dramatically in late 2006. However, neither signal was very strong. The yield curve inverted, but at a minimum of -40 bps, which, relative to prior recessions is not a strong signal. Similarly, while housing permits were negative throughout the fall of 2006, they turned briefly positive in early 2007 before the bottom fell out in late 2007 and 2008. These observations highlight the difficulty of recognizing recessions ahead of time: recession signals can come from disparate sources and important economic relationships can change over time.

[Figure 7 about here.]

6 Conclusion

There is an intense interest in establishing the current and future states of the business cycle, yet even a casual consumer of macroeconomic news would observe the spotty record economists have when identifying economic downturns. This paper evaluates the information content of many commonly cited economic indicators. The methods provide evidence that many economic indicators contain information that can be exploited to identify and forecast business cycle turning points, but that different indicators provide valuable information about different forecast horizons. This observation complicates the modeling decision faced by forecasters, and can help to explain the difficulty of forecasting business cycle turning points. Every time one economic indicator signals "recession," there are likely many more signaling "no recession." The obvious difficulty, then, is sorting through indicators that have predictive power and those that do not.

The compared two distinct methods methods—model averaging and model selection—to highlight these difficulties. Since many commonly followed indicators are valuable only at particular forecast horizons, or have only modest predictive value at any horizon, a simple model average dilutes useful information. Empirically-driven model selection algorithms—Bayesian Model Averaging and boosting—are more successful. Forecasts of the current state of the economy rely on measures of real economic activity: industrial production and initial claims. In contrast, forecasts into the medium-term (six or 12 months) rely on signals originating from the bond market. The results also indicate that a model incorporating the well-known nonlinear behavior of real economic variables around business cycle turning points more accurately identifies turning points.

Overall, the results indicate that there is no sufficient summary statistic for identifying or forecasting business cycle turning points. Indeed, this is the approach taken by the Business Cycle Dating Committee of the NBER, who consider a wide-variety of economic indicators when making pronouncements regarding the state of the economy. At the same time, the power of the yield curve as a predictor of future economic activity endures. Models selected to forecast recessions one-year ahead rely heavily on this indicator, althoughs the best-performing models combine the slope of the yield curve with other macroeconomic information.

References

Aruoba, S. B., Diebold, F. X. & Scotti, C. (2009), 'Real-time measurement of business conditions', *Journal of Business & Economic Statistics* **27**(4), 417–427.

Bai, J. & Ng, S. (2009), 'Boosting diffusion indices', *Journal of Applied Econometrics* **24**(4), 607–629.

Bates, J. & Granger, C. (1969), 'The combination of forecasts', *Operations Research Quarterly* **20**, 451–468.

Berge, T. J. (2014), 'Forecasting disconnected exchange rates', *Journal of Applied Econometrics* .

Berge, T. J. & Jordà, O. (2011), 'Evaluating the classification of economic activity', *American Economic Journal: Macroeconomics* **3**.

Brier, G. & Allen, R. (1951), Verification of weather forecasts, *in* 'Compendium of Meteorology', American Meteorology Society, pp. 841–848.

Brock, W. A., Durlauf, S. N. & West, K. D. (2007), 'Model uncertainty and policy evaluation: Some theory and empirics', *Journal of Econometrics* **136**(2), 629–664.

Buhlmann, P. & Yu, B. (2003), 'Boosting with the l2 loss: Regression and classification', *Journal of the American Statistical Association* **98**, 324–340.

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

Chauvet, M. (1998), 'An econometric characterization of business cycle dynamics with factor structure and regime switching', *International Economic Review* **39**(4), 969–96.

Chauvet, M. & Piger, J. (2008), 'A comparison of the real-time performance of business cycle dating methods', *Journal of Business & Economic Statistics* **26**, 42–49.

Chauvet, M. & Potter, S. (2002), 'Predicting a recession: evidence from the yield curve in the presence of structural breaks', *Economics Letters* **77**(2), 245–253.

Chauvet, M. & Potter, S. (2005), 'Forecasting recessions using the yield curve', *Journal of Forecasting* **24**(2), 77–103.

Chauvet, M. & Senyuz, Z. (2012), A dynamic factor model of the yield curve as a predictor of the economy, Technical Report 2012-32, Federal Reserve Board.

Dueker, M. (2005), 'Dynamic forecasts of qualitative variables: A qual var model of u.s. recessions', *Journal of Business & Economic Statistics* **23**, 96–104.

Eilers, P. H. & Marx, B. D. (1996), 'Flexible smoothing with b-splines and penalties', *Statistical Science* **11**(2), 89–121.

Elliott, G. & Lieli, R. P. (2009), Predicting binary outcomes. U.C.S.D. mimeograph.

Estrella, A. (1998), 'A new measure of fit for equations with dichotomous dependent variables', *Journal of Business & Economic Statistics* **16**(2), 198–205.

Estrella, A. & Mishkin, F. S. (1998), 'Predicting u.s. recessions: Financial variables as leading indicators', *The Review of Economics and Statistics* **80**(1), 45–61.

Estrella, A., Rodrigues, A. P. & Schich, S. (2000), How stable is the predictive power of the yield curve? evidence from germany and the united states, Staff reports 113, Federal Reserve Bank of New York.

Friedman, J. (2001), 'Greedy function approximation: A gradient boosting machine', *The Annals of Statistics* **29**(5).

Giacomini, R. & White, H. (2006), 'Tests of conditional predictive ability', *Econometrica* **74**(6), 1545–1578.

Giusto, A. & Piger, J. (2013), Nowcasting u.s. business cycle turning points with vector quantization. Working paper, University of Oregon.

Gurkaynak, R. S., Sack, B. & Wright, J. H. (2007), 'The u.s. treasury yield curve: 1961 to the present', *Journal of Monetary Economics* **54**(8), 2291–2304.

Hamilton, J. D. (2005), 'What's real about the business cycle?', *Review* pp. 435–452.

Hamilton, J. D. (2011), 'Calling recessions in real time', *International Journal of Forecasting* **27**(4), 1006–1026.

Hand, D. J. & Vinciotti, V. (2003), 'Local versus global models for classification problems: Fitting models where it matters', *The American Statitician* **57**, 124–131.

Hendry, D. F. & Clements, M. P. (2004), 'Pooling of forecasts', *The Econometrics Journal* **7**.

Kauppi, H. & Saikkonen, P. (2008), 'Predicting u.s. recessions with dynamic binary response models', *The Review of Economics and Statistics* **90**(4), 777–791.

Khandani, A., Kim, A. J. & Lo, A. W. (2010), Consumer credit risk models via machine-learning algorithms, Technical report, MIT Working Paper Series.

Levanon, G., Manini, J.-C., Ozyildirim, A., Schaitkin, B. & Tanchua, J. (2011), Using a leading credit index to predict turning points in the u.s. business cycle, Economics Program Working Papers 11-05, The Conference Board, Economics Program.
URL: *http://ideas.repec.org/p/cnf/wpaper/1105.html*

Morley, J. & Piger, J. (2012), 'The asymmetric business cycle', *The Review of Economics and Statistics* **94**(1), 208–221.

Ng, S. (2014), 'Boosting recessions', *Canadian Journal of Economics* .

Ng, S. & Wright, J. H. (2013), Facts and challenges from the great recession for forecasting and macroeconomic modeling, NBER Working Paper 19469, National Bureau of Economic Research.

Pepe, M. S. (2003), *The Statistical Evaluation of Medical Test for Classification and Prediction*, Oxford: Oxford University Press.

Raftery, A. E. (1995), 'Bayesian model selection in social research', *Sociological Methodology* **25**, 111–163.

Rudebusch, G. D. & Williams, J. C. (2009), 'Forecasting recessions: the puzzle of the enduring power of the yield curve', *Journal of Business & Economic Statistics* **27**(4).

Sala-I-Martin, X., Doppelhofer, G. & Miller, R. I. (2004), 'Determinants of long-term growth: a bayesian averaging of classical estimates (bace) approach', *The American Economic Review* **94**(4), 813–835.

Stock, J. H. & Watson, M. (2004), 'Combination forecasts of output growth in a seven-country data set', *Journal of Forecasting* **23**(6), 405–430.

Stock, J. H. & Watson, M. W. (1993), A procedure for predicting recessions with leading indicators: Econometric issues and recent experience, *in* 'Business Cycles, Indicators and Forecasting', NBER Chapters, National Bureau of Economic Research, Inc, pp. 95–156.

Stock, J. H. & Watson, M. W. (1999), 'Forecasting inflation', *Journal of Monetary Economics* **44**(2), 293–335.

Timmermann, A. (2006), Forecast combinations, *in* G. Elliott, C. Granger & A. Timmermann, eds, 'Handbook of Economic Forecasting'.

Wright, J. H. (2006), The yield curve and predicting recessions, Technical report, Federal Reserve Board.

Figures and tables

Variable	Definition	Transformation
	Interest rates and interest rate spreads	
Level of yield curve	Average of 3-mo, 2- and 10-year yields	–
Slope of yield curve	10-yr less 3-mo yield	–
Curvature of yield curve	2x2-yr minus sum of 3-mo and 10-yr yield	–
TED spread	3-mo. ED less 3-mo. treasury yield	–
BAA corporate spread	BAA less 10-yr. treasury yield	–
AAA corporate spread	AAA less 10-yr. treasury yield	–
	Other financial variables	
Change in stock index	Dow Jones Industrial Average	3-month log difference
Money growth	M2	3-month log difference
Real money growth	M2 deflated by CPI	3-month log difference
U.S. dollar	Trade-weighted dollar	3-month log difference
VIX	VIX from CBOE and extended following Bloom	–
	Macroeconomic indicators	
Output	Industrial production (s.a.)	3-month log difference
Income	Real personal income (s.a.)	3-month log difference
Housing permits	–	3-month log difference
Total employment	Payroll employment	3-month log difference
Initial claims	4-week moving average (s.a.)	3-month log difference
Weekly hours, manufacturing	–	3-month log difference
Purchasing managers index	–	3-month log difference

Table 1: Variables included in forecasting models.

		Nobs	h=0	h=6	h=12	h=18	h=24
Level	Pseudo-R^2	483	0.04	0.06	0.05	0.04	0.01
	t-statistic		4.09	5.06	4.62	4.16	2.41
	AUC		0.61	0.65	0.65	0.65	0.59
Slope	Pseudo-R^2	483	0.03	0.23	0.27	0.20	0.10
	t-statistic		-3.83	-8.48	-8.63	-7.98	-6.40
	AUC		0.62	0.83	0.89	0.86	0.78
Curve	Pseudo-R^2	483	0.01	0.00	0.00	0.02	0.02
	t-statistic		2.21	0.78	1.50	2.84	2.75
	AUC		0.52	0.50	0.54	0.61	0.60
Yield spread (AAA-10Y)	Pseudo-R^2	483	0.02	0.14	0.21	0.23	0.13
	t-statistic		-2.94	-7.02	-7.85	-7.92	-6.67
	AUC		0.60	0.78	0.85	0.87	0.80
Yield spread (BAA-10Y)	Pseudo-R^2	483	0.00	0.10	0.19	0.22	0.13
	t-statistic		-0.11	-6.21	-7.54	-7.65	-6.53
	AUC		0.52	0.75	0.83	0.86	0.79
Ted spread	Pseudo-R^2	483	0.27	0.19	0.07	0.04	0.01
	t-statistic		9.21	8.30	5.50	4.62	2.10
	AUC		0.83	0.82	0.76	0.72	0.62
VIX	Pseudo-R^2	483	0.09	0.01	0.00	0.01	0.04
	t-statistic		5.83	2.68	-0.29	-1.46	-3.43
	AUC		0.77	0.64	0.46	0.53	0.64
S&P 500	Pseudo-R^2	483	0.11	0.05	0.00	0.00	0.00
	t-statistic		-6.62	-4.91	-1.49	0.60	1.17
	AUC		0.74	0.71	0.58	0.52	0.54
M2	Pseudo-R^2	483	0.01	0.00	0.01	0.00	0.00
	t-statistic		1.87	1.30	1.73	0.07	0.04
	AUC		0.57	0.57	0.60	0.52	0.51
Real M2	Pseudo-R^2	483	0.02	0.04	0.02	0.04	0.02
	t-statistic		-2.71	-4.01	-3.16	-3.90	-3.05
	AUC		0.62	0.65	0.61	0.64	0.61
Trade-weighted dollar	Pseudo-R^2	483	0.01	0.01	0.00	0.00	0.00
	t-statistic		2.43	2.06	1.19	-1.34	-0.63
	AUC		0.57	0.55	0.54	0.55	0.53
Industrial production	Pseudo-R^2	483	0.32	0.04	0.00	0.00	0.00
	t-statistic		-8.60	-4.13	-0.93	0.05	-0.18
	AUC		0.88	0.71	0.59	0.48	0.53
Real personal income	Pseudo-R^2	483	0.14	0.03	0.00	0.00	0.00
	t-statistic		-7.04	-3.76	0.15	0.54	1.01
	AUC		0.78	0.65	0.50	0.52	0.55
New private housing permits	Pseudo-R^2	483	0.15	0.13	0.02	0.02	0.00
	t-statistic		-7.49	-6.95	-3.34	-3.13	-1.32
	AUC		0.76	0.76	0.68	0.67	0.59
ISM Purchasing manager index	Pseudo-R^2	483	0.10	0.04	0.00	0.00	0.00
	t-statistic		-6.03	-4.32	-1.21	-1.08	0.29
	AUC		0.69	0.68	0.62	0.58	0.50
Average weekly hours	Pseudo-R^2	483	0.11	0.02	0.01	0.00	0.00
	t-statistic		-6.17	-3.07	-2.01	-0.16	0.15
	AUC		0.79	0.65	0.61	0.50	0.52
Monthly employment gain	Pseudo-R^2	483	0.26	0.02	0.00	0.00	0.01
	t-statistic		-8.86	-2.93	0.20	1.38	2.25
	AUC		0.86	0.65	0.48	0.53	0.57
Initial claims (4 wk ma)	Pseudo-R^2	483	0.33	0.06	0.01	0.00	0.00
	t-statistic		8.78	5.19	2.59	0.76	-0.69
	AUC		0.88	0.74	0.65	0.55	0.54

Table 2: In-sample statistics for univariate forecasts for each indicator. The first row presents the pseudo-R^2, the second row is the t-statistic for the test that slope coefficient is different from zero, and the final row gives the classification ability of the logit model as measured by the AUC.

		Forecast horizon				
		h=0	h=6	h=12	h=18	h=24
Equally weighted	N	482	476	470	464	458
	AUC	0.950	0.903	0.856	0.861	0.811
		(0.01)	(0.01)	(0.02)	(0.02)	(0.02)
	GW (abs. error)	+0.00	+0.00	+0.00	+0.00	+0.03
	GW (sq. error)	+0.02	+0.06	+0.02	+0.05	+0.14
BMA weights	N	482	476	470	464	458
	AUC	0.983	0.925	0.890	0.898	0.844
		(0.00)	(0.01)	(0.02)	(0.02)	(0.03)
	GW (abs. error)	-0.00	-0.00	0.48	-0.03	-0.00
	GW (sq. error)	-0.00	-0.02	0.44	-0.08	-0.00

Table 3: In-sample forecast accuracy of weighted recession forecasts. The top panel gives summary statistics for equally weighted forecasts, while the bottom panel gives summary statistics from forecasts weighted by each model's relative fit. Standard errors of AUC statistic in parentheses. GW test statistics are p-values from a two-sided test that the model forecast outperforms the null. The sign indicates the direction of relative loss: a negative sign indicates that the model averaged forecast outperforms the best performing univariate model. P-values are robust to heteroskedasticity and autocorrelation.

	Forecast horizon				
	h=0	h=6	h=12	h=18	h=24
Level	–	–	–	8.0	23.5
Slope	–	65.5	180.3	118.0	4.6
Curve	–	–	43.1	–	–
Ted spread	56.4	54.0	–	–	19.2
Yield spread (BAA-10Y)	–	–	–	–	24.4
Yield spread (AAA-10Y)	–	–	–	25.2	1.0
S&P 500	60.0	61.3	–	61.2	27.8
M2	–	–	22.6	–	–
Real M2	5.4	–		–	–
Trade weighted dollar	–	–	–	66.3	66.5
VIX	29.3	–	–	–	–
Industrial production	40.1	–	–	–	19.3
Real personal income	55.4	39.1	–	–	–
New private housing permits	50.1	58.8	–	–	2.4
ISM purchasing managers index	–	–	–	18.3	53.0
Average weekly hours	–	–	4.0	21.6	–
Monthly employment gain	66.3	56.1	3.2	–	–
Initial claims (4 wk ma)	49.5	30.5	–	–	38.4
BIC	164.1	288.8	291.3	278.4	303.9
N	482	476	470	464	458
AUC	0.982	0.931	0.900	0.897	0.845
	(0.01)	(0.01)	(0.02)	(0.02)	(0.03)
GW (abs. error)	-0.00	-0.00	+0.04	+0.49	-0.29
GW (sq. error)	-0.00	-0.03	-0.37	-0.08	-0.02

Table 4: Summary of in-sample linear boosted model. For each regressor, the table shows the ratio of the coefficient of the boosted model to the coefficient from an unrestricted kitchen-sink logit regression, expressed as a percentage. "–" indicates that the coefficient was not selected by the boosted model. Standard errors of AUC statistic in parentheses. GW test statistics are p-values from a two-sided test that the model forecast outperforms the null. The sign indicates the direction of relative loss: a negative sign indicates that the boosted model outperforms the best performing univariate model. P-values are robust to heteroskedasticity and autocorrelation.

	Forecast horizon				
	h=0	h=6	h=12	h=18	h=24
N	482	476	470	464	458
BIC	185.6	284.2	275.9	286.8	309.8
AUC	0.990	0.970	0.953	0.941	0.910
	(0.00)	(0.01)	(0.01)	(0.01)	(0.02)
GW (abs. error)	-0.00	-0.00	-0.00	-0.00	-0.00
GW (sq. error)	-0.00	-0.00	-0.00	-0.00	-0.00

Table 5: In-sample performance of boosted model fit with smoothing splines as weak learners. Standard errors of AUC statistic in parentheses. GW test statistics are p-values from a two-sided test that the model forecast outperforms the null. The sign indicates the direction of relative loss: a negative sign indicates that the boosted model outperforms the best performing univariate model. P-values are robust to heteroskedasticity and autocorrelation.

Peak dates				
NBER	Unweighted average	BMA	Linear boost	Non-linear boost
1973:11	2	8	2	-1
1980:1	-2	-2	-3	-3
1981:7	0	0	0	0
1990:7	3	3	3	2
2001:3	–	–	–	0
2007:12	8	1	0	1
Trough dates				
NBER	Unweighted average	BMA	Linear boost	Non-linear boost
1975:3	1	1	2	1
1980:7	1	0	1	0
1982:11	0	0	0	0
1991:3	-1	0	0	1
2001:11	–	–	–	1
2009:6	-1	0	0	0

Table 6: Business cycle turning points, 1973-2012. Value shown is the model-implied peak/trough calculated using the optimal threshold and a rule that each phase of the business cycle last more than three months. The value listed is the model-implied peak/trough date relative to the NBER-defined peak/trough date. For example, a positive value of 3 indicates that the model dated the peak or trough 3 months later than the NBER (e.g., May instead of February). '–' indicates that the model did not generate a business cycle phase lasting 3 months that corresponds to the NBER recession.

		Forecast horizon				
		h=0	h=6	h=12	h=18	h=24
	N	338	338	338	338	338
Equally weighted	AUC	0.819	0.739	0.746	0.808	0.722
		(0.04)	(0.04)	(0.03)	(0.03)	(0.04)
	GW (abs. error)	+0.00	+0.00	+0.00	+0.00	+0.00
	GW (sq. error)	+0.03	+0.04	+0.00	+0.00	+0.01
BMA	AUC	0.928	0.871	0.879	0.877	0.812
		(0.03)	(0.02)	(0.03)	(0.03)	(0.04)
	GW (abs. error)	-0.00	0.21	0.08	0.42	-0.01
	GW (sq. error)	-0.04	-0.24	0.44	0.40	0.47
Linear boost	AUC	0.965	0.908	0.868	0.883	0.808
		(0.01)	(0.02)	(0.03)	(0.03)	(0.04)
	GW (abs. error)	-0.00	0.35	0.27	0.06	0.41
	GW (sq. error)	-0.00	-0.17	0.12	0.32	0.-33
Nonlinear boost	AUC	0.975	0.923	0.943	0.926	0.863
		(0.01)	(0.02)	(0.01)	(0.02)	(0.03)
	GW (abs. error)	-0.00	-0.00	-0.00	-0.00	-0.00
	GW (sq. error)	-0.00	-0.06	-0.10	-0.19	-0.15

Table 7: Out-of-sample forecast performance. Standard errors of AUC statistic in parentheses. GW test statistics are p-values from a two-sided test that the model forecast outperforms the null. The sign indicates the direction of relative loss: a negative sign indicates that the model outperforms the null of best performing univariate model. P-values are robust to heteroskedasticity and autocorrelation.

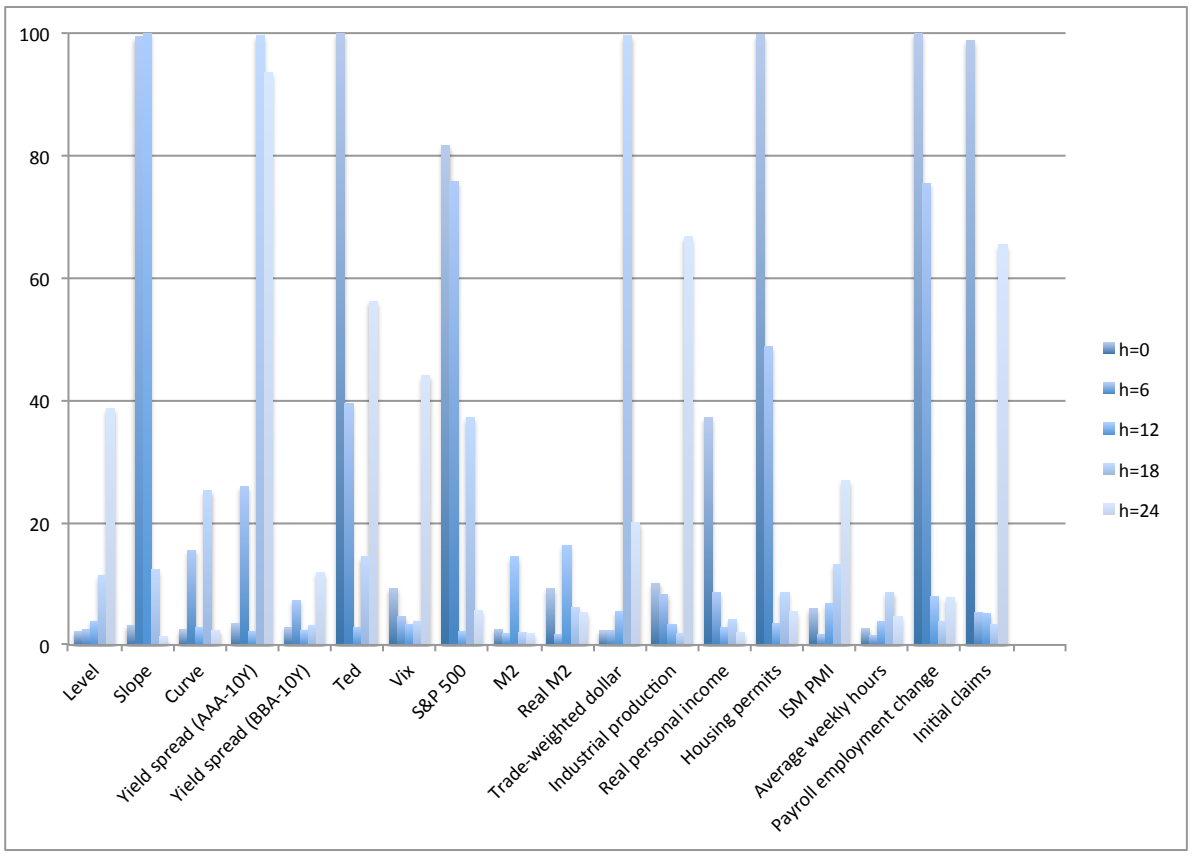

Figure 1: Posterior inclusion probabilities of each forecast model, in-sample BMA exercise. All models contain an intercept. See text for details.

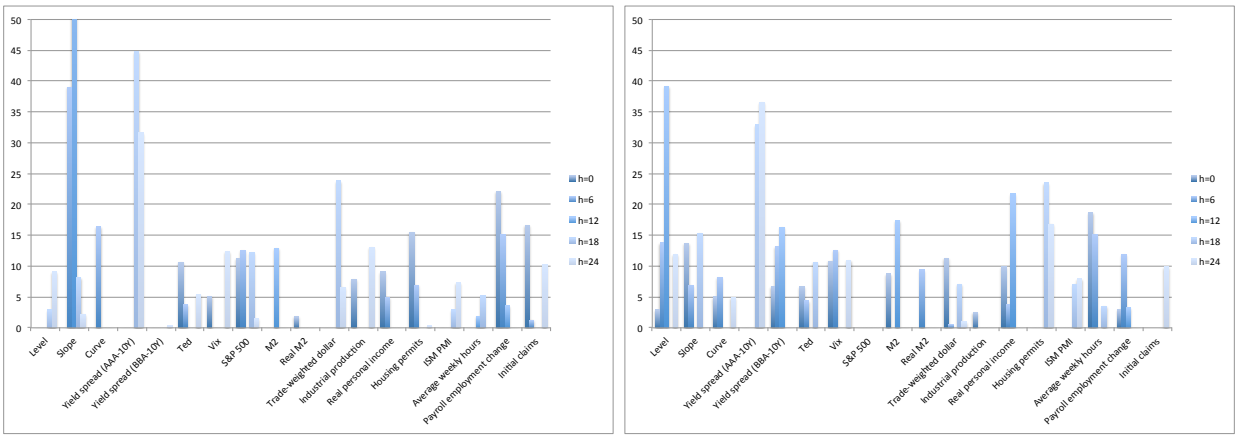

Figure 2: In-sample model selection frequency, linear and non-linear models. The bar graph presents the fraction of iterations for which a particular covariate was selected by the model selection procedure for each model, linear (left) and non-linear (right), and for each forecast horizon.

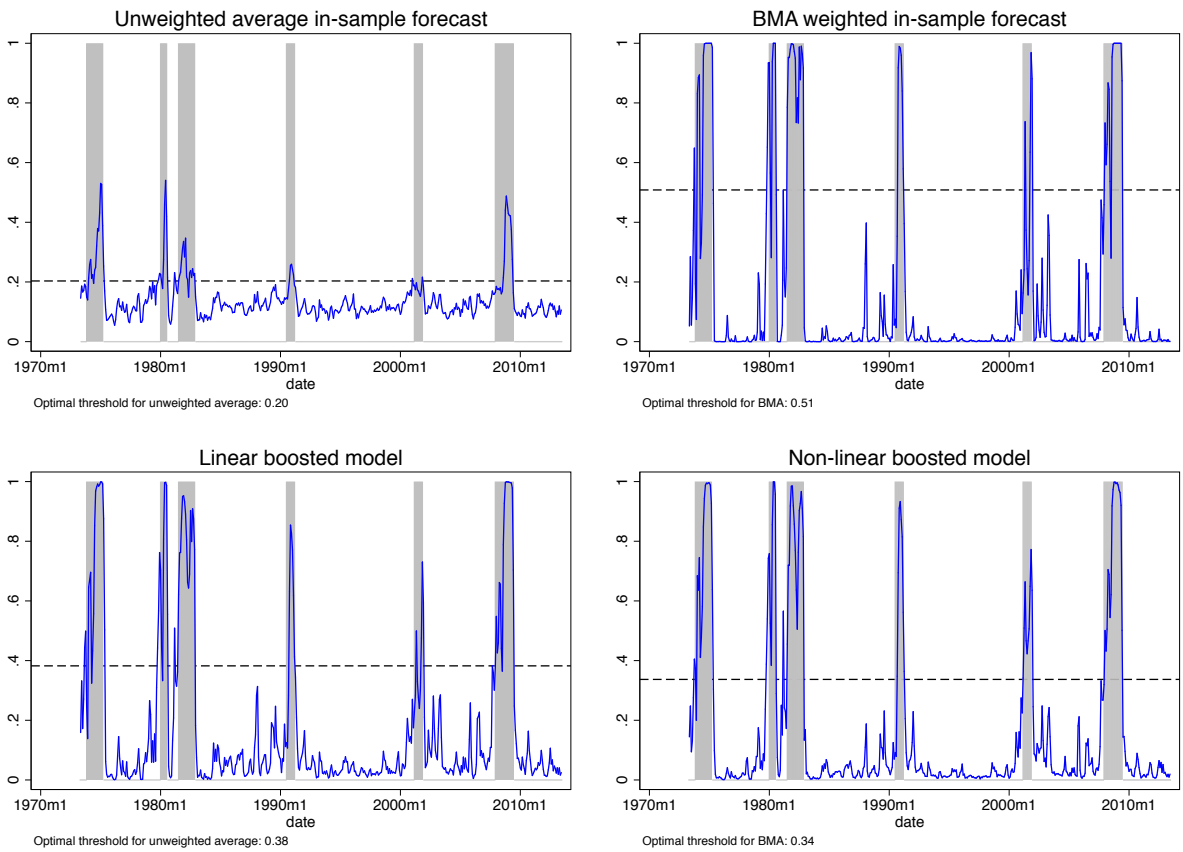

Figure 3: In-sample recession probabilities forecast at horizon of zero months. The top panels contain forecasts from model combination schemes, both unweighted (left) and Bayesian model averaged (right). The two bottom panels contain recession probabilities from linear (left) and non-linear (right) boosted forecast models. Grey shading indicates NBER-defined recession dates and dashed line denotes optimal threshold.

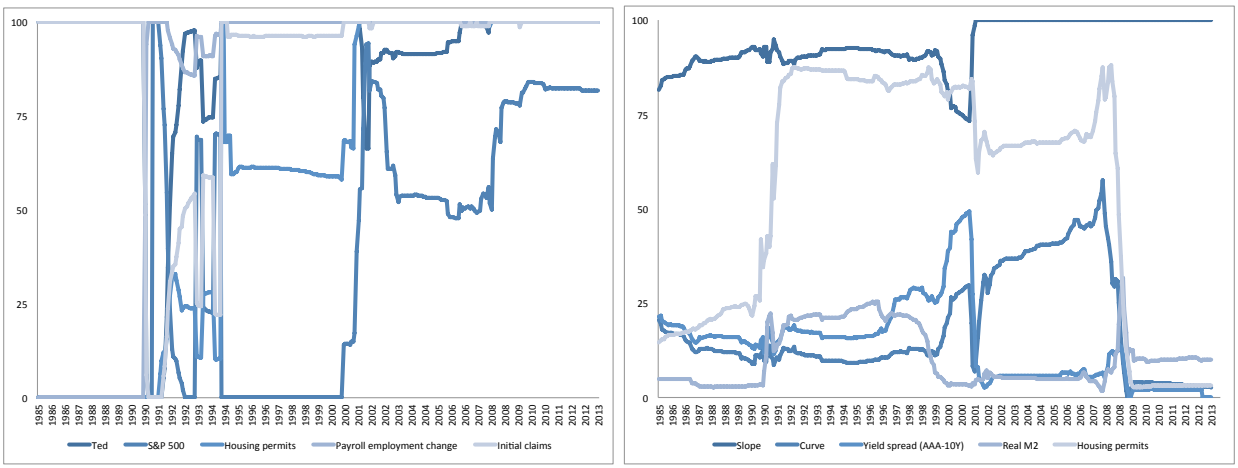

Figure 4: Posterior inclusion probabilities for BMA models used to produce out-of-sample forecasts. The figure shows the PIP for the five indicators with the highest average PIP as they evolved through the expanding window out-of-sample forecasting exercise. The figure on the left is for the model producing nowcasts while the figure on the right shows the model forecasting 12 months ahead. See text for details.

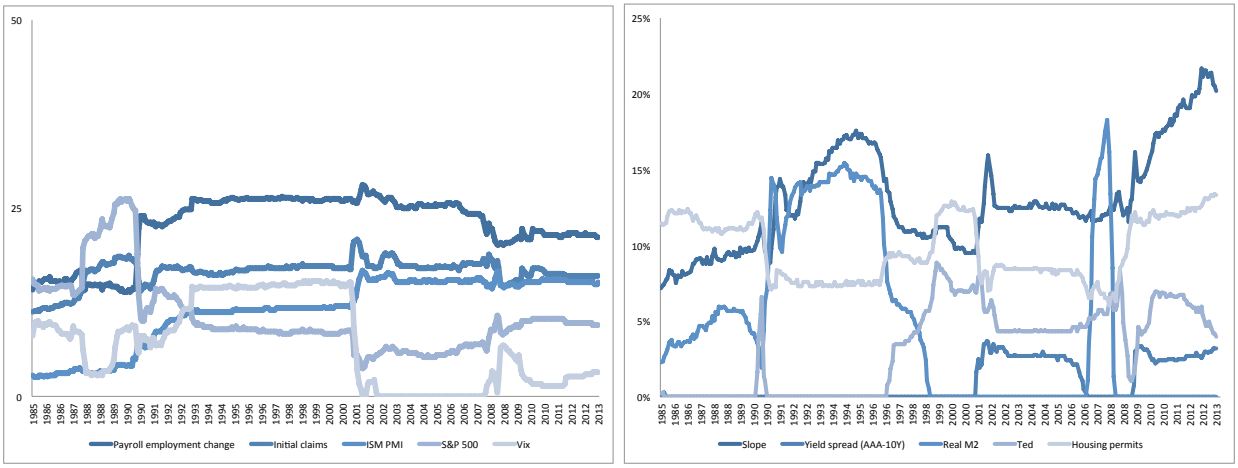

Figure 5: Out of sample selection frequency of covariates for linear boosted model. The figure shows the fraction of time a covariate was selected for the model forecasting 0- (left) and 12-months ahead (right) at time t, as forecast model is run through the sample.

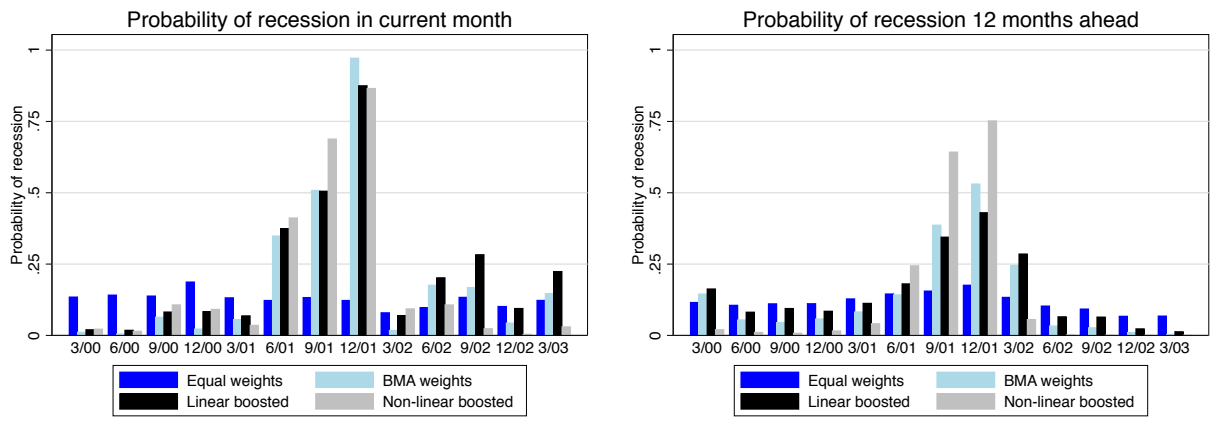

Figure 6: Out-of-sample recession probabilities surrounding 2001 recession, $P(NBER_t|D_{t-h-1})$. Nowcasts (h=0) of the probability of recession are shown in the figure on the left. The figure on the right is the one-year ahead forecast probability (h=12). Official NBER recession dates March 2001–November 2001.

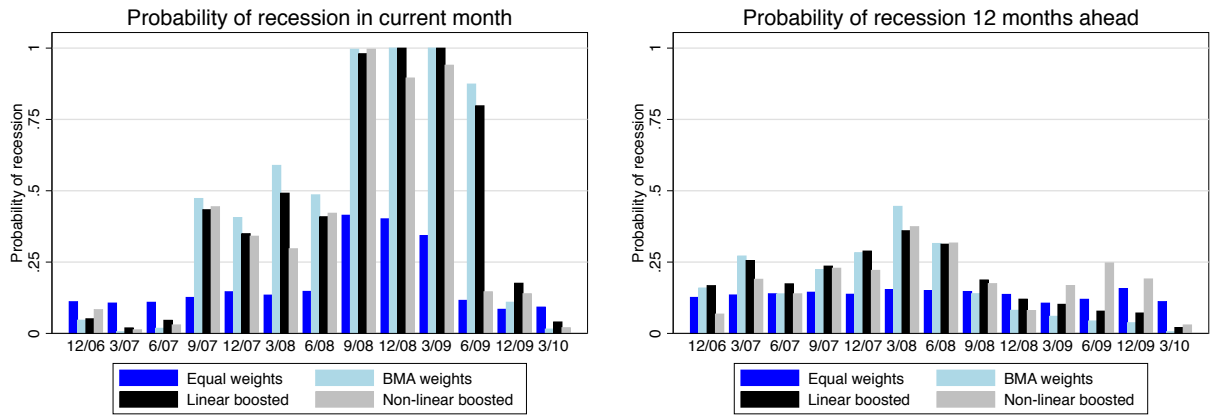

Figure 7: Out-of-sample recession probabilities surrounding 2007 recession, $P(NBER_t|D_{t-h-1})$. Nowcasts (h=0) of the probability of recession are shown in the figure on the left. The figure on the right is the one-year ahead forecast probability (h=12). Official NBER recession dates December 2007–June 2009.

Appendix

Data sources

Table 8: Data sources

Variable	Available dates	Source
	Interest rates	
Effective Fed Funds Rate	1954m7-2013m6	FRB H.15 release
3-month Eurodollar deposit rate	1971m1-2013m6	FRB H.15 release
3-month Treasury yield (secondary market)	1934m1-2013m6	Gurkaynak, Sack and Wright (2006)
1-year Treasury yield (constant maturity)	1953m4-2013m6	Gurkaynak, Sack and Wright (2006)
10-year Treasury yield (constant maturity)	1953m4-2013m6	Gurkaynak, Sack and Wright (2006)
Moody's BAA corporate bond yield	1947m1-2013m6	FRB H.15 release
	Other financial variables	
Dow Jones Industrial Average	1947m1-2013m6	Dow Jones & Company
M2 (seasonally adj.)	1959m1-2013m6	FRB H.6 release
Trade-weighted US Dollar: major currencies	1973m1-2013m6	FRB H.10 release
	Macroeconomic indicators	
CPI index (seasonally adj.)	1947m1-2013m6	U.S. Dept of Labor: BLS
Industrial production (seasonally adj.)	1947m1-2013m6	FRB G.17 release
Real personal income (seasonally adj.)	1959m1-2013m6	U.S. Dept of Commerce
ISM manufacturing PMI index	1948m1-2013m6	Institute for Supply Management
Housing permits	1960m1-2013m6	U.S. Census Bureau
Average weekly hours: manufacturing	1947m1-2013m6	U.S. Dept of Labor: BLS
All employees (total nonfarm)	1939m1-2013m6	U.S. Dept of Labor: BLS
Initial claims for unemployment insurance	1967m1-2013m6	U.S. Dept of Labor: BLS