Heikki Kauppi

Recession Prediction with Optimal Use of Leading Indicators

Aboa Centre for Economics
Discussion paper No. 125
Turku 2019

The Aboa Centre for Economics is a joint initiative of the economics departments of the University of Turku and Åbo Akademi University.

ISSN 1796-3133

Printed in Uniprint
Turku 2019

Heikki Kauppi

Recession Prediction with Optimal Use of Leading Indicators

Aboa Centre for Economics
Discussion paper No. 125
April 2019

ABSTRACT

We use the gradient boosting estimation technique and the ROC curve to non-parametrically measure and exploit the maximal predictive power of leading indicators for the future state of the business cycle. We develop novel procedures for finding the best performing transformations of individual indicators, for combining them to form an optimal recession prediction model and for assessing which predictors are contributing in the model. Among our empirical findings with US data are that the predictive impact of various indicators is non-monotone and that recession predictions based on our nonparametric procedures clearly outperform the ones based on a conventional probit model.

JEL Classification: C22, C25, C53, E37

Keywords: gradient boosting, leading indicators, non-parametric estimation, optimal binary prediction, recession prediction

Contact information

Heikki Kauppi
Department of Economics
University of Turku
FI-20014, Finland
Email: heikki.kauppi (at) utu.fi

1 Introduction

We address the problem how leading economic indicators are to be combined in order to exploit their joint predictive power for business cycle turning points maximally. In a related paper, Berge and Jordà (2011) demonstrate how the so called receiver operating characteristics (ROC) curve can be used to assess which individual indicators are best at predicting future turning points, and at what horizons. We extend that analysis in at least two ways. First, we provide a framework and methods for assessing the optimal predictive performance of an individual indicator under more general conditions than is implicit in the conventional ROC curve analysis. Second, we review the general principle by which several indicators can be combined to predict future turning points optimally and we apply a nonparametric procedure to handle the underlying problem in practise. We demonstrate the performance of the proposed procedures with US data.

Basically, our goal is to use a vector of variables Z_t to predict, for a given horizon τ, the value of $R_{t+\tau}$, where $R_t = 1$, if the economy is in recession at month t, and $R_t = 0$, otherwise. It is straightforward to show that an optimal prediction rule (minimizing anybody's expected loss) is generally of the form $D_{t+\tau} = I(p^*(Z_t) > c)$, where $I(\cdot)$ is the indicator function, c is a constant and $p^*(z)$ is the conditional probability function $\Pr(R_{t+\tau} = 1|Z_t = z)$. This provides a clear guideline how to search for an optimal prediction of future turning points given the indicators. We just need to identify and estimate $\Pr(R_{t+\tau} = 1|Z_t)$.

One standard approach is to assume a parametric model for $\Pr(R_{t+\tau} = 1|Z_t)$ and estimate it by maximum likelihood. In fact, a number of papers on recession prediction have done this even if these papers do not refer to the notion of optimal prediction (e.g., Estrella and Mishkin (1998) and Kauppi and Saikkonen (2008)). Nevertheless, an issue with such an approach is that the assumed parametric model, which is typically a probit model with a linear index function, need not be the same as the true conditional probability. To overcome this problem we estimate $\Pr(R_{t+\tau} = 1|Z_t)$ nonparametrically by the so called boosting estimator developed in the machine learning literature.

The boosting estimator for $\Pr(R_{t+\tau} = 1|Z_t)$ has many desirable properties and we will

review these in the paper. However, the potential drawback of the boosting estimator, like any nonparametric estimation technique, is that the estimated model may seem like a "black box." In particular, it tends to be cumbersome to judge what variables in the model are really important or "significant." To overcome this problem we propose a model specification procedure, where the ROC curve is used as the main device for assessing the importance of different variables in the estimated model.

In the first step of our proposed approach, we assess the predictive power of individual indicators at different horizons. This is as in Berge and Jordà (2011) except that, in addition to the conventional ROC curve for a given variable Z_{it} (a component in Z_t), we consider the one for the transformed variable $p(Z_{it})$, where $p(z)$ is the boosting estimate of $\Pr(R_{t+\tau} = 1|Z_{it} = z)$. The latter ROC curve is regarded as an estimate for the true maximal predictive power of Z_{it} and its reliability rests on the robustness of the boosting estimator.

The examination of the predictive performance of individual variables help us to assess whether a particular original variable should be used as such, as month-to-month or year-to-year difference. Using the results of this initial analysis, we select candidate variables for a prediction model that exploits them jointly. As noted above, we can use the boosting method to non-parametrically estimate the optimal prediction model for any given set of predictors.

To find which predictors are really needed in the prediction model, we undertake a general-to-specific modeling strategy, where we drop a variable at the time based on a particular variable importance estimation procedure developed in the boosting literature. By using the ROC curve of the estimated model fits for different indicator sets we can assess what indicators are really carrying marginal predictive power and which have little or nothing to add to the accuracy of the prediction model. Altogether, the stepwise procedure yields a transparent picture of what indicators have predictive content for the future state of the business cycle at different horizons.

Our empirical analysis yields the following main findings. For most leading indicators, a year-to-year difference contains more predictive power than the commonly applied month-to-month difference. We also find that for more than half of the indicators the

predictive relationship is not monotone. In most of such cases, the conditional risk of a future recession is first increasing and then decreasing, or vice versa, as a function of the indicator. These findings suggest that we need a nonparametric modeling procedure to extract maximal predictive power from the leading indicators. In line with this, we find that our proposed nonparametric estimation approach yields clearly more accurate recession predictions than what can be achieved by applying a conventional recession prediction procedure based on a probit model.

Prior to this paper at least Ng (2014) and Berge (2015) have applied the boosting estimator to predict recessions. The present paper contributes to these studies by putting the method into a general optimality framework and by demonstrating how we can apply the ROC curve in the modeling process so as to clarify which variables are really needed in the optimal prediction model. Our analysis is also relevant to a number of earlier papers that predict recessions merely by using the probit model (e.g., Estrella and Hardouvelis (1991), Stock and Watson (1993), Dueker (1997), Bernard and Gerlach (1998), Estrella and Mishkin (1998), Filardo (1999), Estrella, Rodrigues, and Schich (2003), Wright (2006), Kauppi and Saikkonen (2008), Rudebusch and Williams (2009), Chen, Iqbal and Lai (2011)). The majority of these papers conclude that once the yield curve is used as a predictor, there is hardly any other predictor that can improve the prediction of future recessions. In contrast, we find that several predictors have clear marginal predictive power for recessions provided we apply a flexible enough model.

The plan of the rest of the paper is as follows. In Section 2, we conduct a comprehensive population level analysis that explains the notion of optimal binary prediction and how the ROC curve can be related to it. The section also reviews alternative modeling options and justifies the choice made in this paper. Section 3 presents the estimation method, while Section 4 conducts the empirical analysis. Section 5 concludes.

2 Population Level Analysis

2.1 The Starting Point

Let R_t indicate the phase of the business cycle at time t. In our empirical application below, R_t is a dummy variable that equals 1 (0), if the US economy is in recession (in expansion) at month t, as defined by the NBER business cycle chronology.

The goal is to use an n-vector of variables Z_t to predict the value of $R_{t+\tau}$, where τ is the selected forecast horizon. Any prediction rule for the value of $R_{t+\tau}$ given Z_t can be expressed as

$$D_{t+\tau}(Z_t) = I(h(Z_t) > c) \tag{1}$$

where h is a function from $\mathbb{R}^n$ to $\mathbb{R}$ and c is a constant.

The so called leading economic index (LEI) of the Conference Board can be seen as an approach to specify $h(Z_t)$ even if LEI has somewhat broader scope than making a "point forecast" for $R_{t+\tau}$ at particular τ. The methodology behind LEI uses a rich array of judgement and informal evaluations to find the components in Z_t.[1] In our empirical analysis, we largely use the same candidate series as in LEI, but we reassess the transformations by which the actual component series, called leading indicators, are obtained from the original "raw" data.

The applied function h in LEI is basically a weighted linear combination of the selected components, where the weight of a component is the inverse of its standard deviation. This "symmetric" weighting scheme is designed to give each component a similar opportunity to contribute to the index. The actual LEI is a cumulative sum of $h(Z_t)$ over t and evolves like a stochastically trending (integrated) time series. An objective in the methodology is that the turning points of the index occur in advance to the business cycle turning points.

The purpose of the present paper is not to give a better version of LEI, but rather we attempt to develop complementary procedures that (i) measure and (ii) utilize the maximal predictive content of the indicators when the goal is to predict the state of the business cycle at a desired horizon. When we have an idea about (i) and (ii) for several

[1] For a detailed description of the methodology see the Conference Board (2001).

horizons this is certainly useful for business cycle analysis. The next section explains the general principles by which we approach the two objectives (i) and (ii).

2.2 The Optimal Prediction

When we think about the maximal predictive content of Z_t for $R_{t+\tau}$ we have in mind a situation where the applied prediction rule in (1) is the best possible. What is best depends on one's preferences. We benefit from correct predictions $D_{t+\tau} = R_{t+\tau}$, while wrong predictions $D_{t+\tau} \neq R_{t+\tau}$ incur losses. Let $L(d) > 0$, $d \in \{0, 1\}$, be the loss that results in when we predict $D_{t+\tau} = d$ and $R_{t+\tau} = 1 - d$ realizes.[2] The expected loss from applying the prediction rule $D_{t+\tau}$ given Z_t is

$$\begin{aligned} E(L(D_{t+\tau}(Z_t))|Z_t) &= L(0)\Pr(R_{t+\tau} = 1|Z_t) \\ &\quad + D_{t+\tau}(Z_t)\left\{L(1) - \Pr(R_{t+\tau} = 1|Z_t)\left[L(0) + L(1)\right]\right\} \end{aligned}$$

To minimize this, we set $D_{t+\tau}(Z_t) = 1$, if (and only if) the term in the curly brackets is negative. This yields the optimal prediction

$$D^*_{t+\tau}(Z_t) = I\left(\Pr(R_{t+\tau} = 1|Z_t) > \frac{L(0)}{L(0) + L(1)}\right) \tag{2}$$

Sometimes the goal is just to choose a prediction rule that minimizes the probability of $D_{t+\tau} \neq R_{t+\tau}$ (the classification error) given Z_t

$$\Pr(D_{t+\tau} \neq R_{t+\tau}|Z_t) = D_{t+\tau}(Z_t)(1 - 2\Pr(R_{t+\tau} = 1|Z_t)) + \Pr(R_{t+\tau} = 1|Z_t)$$

The optimal solution to this problem is

$$D^*_{t+\tau}(Z_t) = I\left(\Pr(R_{t+\tau} = 1|Z_t) > \frac{1}{2}\right)$$

This rule is often called the Bayes rule and is obtained from (2) under the "symmetric loss" $L(0) = L(1)$. Clearly, the above rules are also optimal under an (unconditional) expectation with respect to the distribution of Z_t.

[2] You may think $L(d)$ as the difference $u(d, d) - u(d, 1 - d)$, where $u(d, r)$ is the utility or benefit that one obtains when $D_{t+\tau} = d$ and $R_{t+\tau} = r$ occur. We assume $u(d, r)$ are finite for $d, r \in \{0, 1\}$ and $u(d, d) > u(d, 1 - d)$ for $d \in \{0, 1\}$.

The preceding considerations show that we can minimize everybody's expected loss, if we know the conditional probability $\Pr(R_{t+\tau} = 1|Z_t)$. Note also that the optimal rule in (2) amounts to the rule (1), where $h(Z_t) = \eta(\Pr(R_{t+\tau} = 1|Z_t))$, $c = \eta(L(0)/(L(0)+L(1)))$ and η is any strictly monotone increasing function from $\mathbb{R}$ to $\mathbb{R}$.

2.3 Assessing and Comparing Prediction Rules

In practise, the applied prediction rule may be based on a model that is misspecified or only an approximation to the true conditional probability. This means that the applied function h in (1) cannot be expressed as a strictly monotone function of $\Pr(R_{t+\tau} = 1|Z_t)$. How do we assess and compare the performance of rules, whether they are optimal or not? The ROC curve is useful for this in the sense that it is not directly tied to a particular loss function. In other words, a user (with any loss function) can apply the ROC curve to judge which model out of alternatives is the best one for her.

Henceforth, we call the function h in (1) as the (prediction) model. In what follows, we understand that h is unique only up to a strictly monotone increasing transformation.

Consider the prediction rule in (1) with the model h and the threshold c being fixed. The ROC curve derives from the following conditional probabilities

$$\begin{aligned} \alpha &= \Pr(D_{t+\tau}(Z_t) = 1|R_{t+\tau} = 0) \\ \beta &= \Pr(D_{t+\tau}(Z_t) = 1|R_{t+\tau} = 1) \end{aligned}$$

In biometrics, α is called *false positive fraction* and β *true positive fraction* (Pepe (2003)). Importantly, we can interpret the rule $D_{t+\tau}$ as a simple hypothesis test with $D_{t+\tau} = 0$ indicating that the null hypothesis $R_{t+\tau} = 0$ is accepted and $D_{t+\tau} = 1$ that it is rejected in favor of the alternative $R_{t+\tau} = 1$. Hence, α can be regarded as the "size" and β as the "power" of the prediction rule.

The parameters α and β depend monotonously on the threshold c. Let F_h (G_h) denote the cumulative distribution function of $h(Z_t)$ conditional on $R_{t+\tau} = 0$ ($R_{t+\tau} = 1$). Then

$$\begin{aligned} \alpha(c) &= 1 - F_h(c) \\ \beta(c) &= 1 - G_h(c) \end{aligned}$$

We have $\alpha(c), \beta(c) \in [0,1]$ for all $c \in \mathbb{R}$, $\alpha(c), \beta(c) \to 0$, as $c \to \infty$, and $\alpha(c), \beta(c) \to 1$, as $c \to -\infty$.

The ROC curve for the model h is

$$ROC_h(\alpha) = 1 - G_h(F_h^{-1}(1-\alpha)) \tag{3}$$

Note that ROC_h is a monotone increasing function from $[0,1]$ to $[0,1]$. It tells how $\alpha(c)$ and $\beta(c)$ are linked when we let c go from $-\infty$ to ∞. Hence, a rule based on the model h in (1) attains the set $\{(\alpha, \beta),\ \beta = ROC_h(\alpha),\ \alpha \in [0,1]\}$. For clarity, we may write $ROC_{h(Z)}$ to indicate that the underlying model h uses predictors Z_t.

Using the ROC curve we can write the expected (unconditional) loss as

$$E(L(D_{t+\tau}(Z_t))) = (1-\delta)L(1)\alpha - \delta L(0)ROC_h(\alpha) + \delta L(0) \tag{4}$$

where $\delta = \Pr(R_t = 1)$. For a given model h, the expected loss depends on α (or equivalently on the threshold c in (1)). When $ROC_h(\alpha)$ is concave and differentiable, the optimal choice of α follows from the first order condition

$$ROC_h'(\alpha) = \frac{1-\delta}{\delta}\frac{L(1)}{L(0)} \tag{5}$$

and the corresponding optimal threshold c in (1) is $F_h^{-1}(1-\alpha)$.

From equation (4) we can conclude that with predictors Z_t we can please anybody (with any loss function) maximally by a model that yields the largest $ROC_h(\alpha)$ for any value of α. But, we already know from above that the conditional probability function given Z_t yields the best rule for any loss function. Hence, if the model h^* can be written as a strictly monotone increasing function of $\Pr(R_{t+\tau} = 1|Z_t)$, then for any other model h (that is based on the same predictors Z_t) we have

$$ROC_{h(Z)}(\alpha) \leq ROC_{h^*(Z)}(\alpha) \text{ for all } \alpha \in [0,1] \tag{6}$$

It can be shown that $ROC_{h^*(Z)}$ is a strictly concave function from $[0,1]$ to $[0,1]$.[3] In what follows, we call $ROC_{h^*(Z)}$ the optimal ROC curve based on the predictors Z_t and h^* the optimal model.

[3] Above, we noted that the underlying prediction problem is equivalent to the one of a simple hypothesis testing. It follows that we can derive (6) also by appealing to the Neyman Pearson (NP) lemma (Neyman and Pearson (1933)). By the NP lemma, an optimal test that yields the largest power (β) for any given

Figure 1 depicts ROC curves for three models that use the same predictors. The black curve is the optimal ROC curve, while the blue and red curves are ROC curves for models of which neither can be expressed as a monotone increasing function of the conditional probability. In line with (6), the blue and red curves are uniformly below the black curve.

Let λ denote the right hand side of (5), that is, $\lambda = (1-\delta)L(1)/\delta L(0)$. For fixed δ (the frequency of the event $R_t = 1$), λ can be regarded as an index for alternative preferences.

The crossed circles in Figure 1 denote the "optimal operating points" determined by the condition (5) for each model when $\lambda = 1.5$. The dashed lines indicate the tangents to the ROC curves at the operating points. For a given value of λ, the operating point whose tangent line is above the other tangent lines yields the smallest expected loss.[4] Hence, one prefers the red operating point (model B) to the blue one (model A) when $\lambda = 1.5$. It should be clear from the figure that that the ranking of the models A and B is eventually reversed when λ becomes smaller.

In general, if the ROC curves of two models cross (as the blue and the red ones in Figure 1), then the ranking of the models depends on one's loss function. This fact means that there is no universal ranking between models whether they use the same or different predictors. There is one exception. Take an optimal model based on two predictors, $h^*(Z_{1t}, Z_{2t})$, and another optimal model based on Z_{1t} alone, $h^*(Z_{1t})$ (Z_{1t}, Z_{2t} may be vector-valued). Then[5]

$$ROC_{h^*(Z_1)}(s) \leq ROC_{h^*(Z_1,Z_2)}(s) \text{ for all } s \in [0,1] \tag{7}$$

This condition says that more predictors is better provided one knows the corresponding

size (α) is generally as in (1), with h being of the "likelihood ratio" form. In Section 2.5, we show (see (8)) that $\Pr(R_{t+\tau} = 1|Z_t)$ is linked to the likelihood ratio through a strictly monotone increasing transformation. Hence, what we have shown here agrees with the NP lemma. By analyzing the NP lemma one can see that $ROC_{h^*(Z)}(s)$ has to be concave (see Lehmann and Romano (2005)).

[4]To see this note that one is indifferent between any points on a tangent line. Draw a vertical line crossing the operating point of the upper most tangent line. This way you see that for the α of this operating point, the other operating points correspond to a smaller $ROC(\alpha)$ and hence return a higher expected loss (see the formula (4)).

[5]Note that by the definition of $h^*(Z_{1t}, Z_{2t})$ we have $ROC_{h(Z_1,Z_2)}(s) \leq ROC_{h^*(Z_1,Z_2)}(s)$, where $h(Z_{1t}, Z_{2t})$ is any function of Z_{1t} and Z_{2t}, including $h(Z_{1t}, Z_{2t}) = h^*(Z_{1t})$.

optimal model. To look at it from another angle, more predictors yield more predictive power at the population level. Finally, note that if $R_{t+\tau}$ is conditionally independent of Z_{2t} given Z_{1t}, then the weak inequality in (7) is replaced by equality.

2.4 Points on AUC

It is quite common, especially in biometrics, to apply the area under the ROC curve (AUC) to assess and compare the performance of alternative predictors or models. For a given model h, this measure is defined as

$$AUC_h = \int_0^1 ROC_h(s)ds$$

By conditions (6) and (7), it is clear that

$$AUC_{h^*(Z)} \geq AUC_{h(Z)} \text{ for all } h$$

$$AUC_{h^*(Z_1,Z_2)} \geq AUC_{h^*(Z_1)}$$

Hence, the optimal model has larger AUC than any other model based on the same predictors and in terms of the optimal model more predictors yields a larger AUC.

We emphasize, however, that AUC does not in general indicate which model is the best one, because models may have crossing ROC curves (as in Figure 1). In particular, when we compare two models neither of which is optimal, then the one with a larger AUC does not need to be better for everybody. Hence, it is better to use the ROC curve than the AUC to compare the performance of alternative models.[6] Unlike with AUC, by the ROC curve we can always see which model is better for any loss function.

2.5 Modeling Options

The above analysis makes it clear that we can attain the largest predictive power from Z_t by finding the optimal model h^*, any monotone increasing function of $\Pr(R_{t+\tau} = 1|Z_t)$. For us to have the capacity to estimate h^* from the data, our estimation method should cover a broad enough class of models such that h^* belongs to it. This is often easier to

[6]For a related point against the use of AUC in biometrics see Pepe (2003, p. 228).

do in terms of the corresponding class of models for $\Pr(R_{t+\tau} = 1|Z_t)$. Here we present a general model framework that allows us to consider alternative modeling approaches.

Let F (G) denote the cumulative distribution function of Z_t conditional on $R_{t+\tau} = 0$ ($R_{t+\tau} = 1$). Assume that F and G are absolutely continuous and let f and g denote the corresponding densities. Then, we have

$$\Pr(R_{t+\tau} = 1|Z_t = z) = \frac{\delta g(z)}{\delta g(z) + (1-\delta)f(z)} = \frac{\delta LR(z)}{\delta LR(z) + 1 - \delta} \tag{8}$$

where LR is the likelihood ratio

$$LR(z) = \frac{g(z)}{f(z)}$$

The modeling of h^* is essentially the same as the modeling of the conditional probability or the likelihood ratio.

Given the presentation in (8), we can always write

$$\Pr(R_{t+\tau} = 1|Z_t) = \Lambda(h^*(Z_t)) \tag{9}$$

where Λ is a strictly monotone increasing function from $\mathbb{R}$ to $[0, 1]$ and h^* is a monotone increasing function of LR.

We specify a model class by making assumptions on the "link function" Λ and the "index function" h^* in (9). In the recession prediction literature, it has been common to apply the probit (the logit) model, where Λ is the standard normal (the logistic) distribution function and h^* is assumed to belong to a simple parametric (typically linear) class of functions. This approach leads to maximum likelihood estimation and yields a consistent estimate for $\Pr(R_{t+\tau} = 1|Z_t = z)$ provided it belongs to the assumed parametric model class.

A parametric model tends to be rather restrictive and may not capture the true conditional probability. For example, for a given parametric index function, the assumed link function Λ (probit or logit) may not be of the right shape. To obtain a more general model than probit and logit we may treat Λ nonparametrically, the only restriction being that it belongs to the set of strictly monotone increasing functions. If h^* is assumed to be a linear function, then such a semiparametric model can be estimated consistently, e.g., by the maximum rank correlation (MRC) estimator (see Han (1987) and Sherman

(1993)). Again, even if there are no restrictions on Λ (except that it must be monotone increasing), the assumed parametric model for h^* may be too special for capturing the true conditional probability in (9).

The third possibility for modeling (9) is to let Λ be a known monotone increasing function and assume that h^* belongs to a nonparametric class of functions. Even if Λ is fixed, this approach is essentially nonparametric and thereby allows us to estimate arbitrarily complex conditional probability functions. To concretize, suppose the true conditional probability can be expressed as the probit model

$$\Pr(R_{t+\tau} = 1|Z_t) = \Phi(\gamma' Z_t) \tag{10}$$

where Φ is the standard normal distribution and γ is a $n \times 1$ parameter vector.[7] We can express this model with any (monotone) link function in (9) by choosing $h^*(Z_t) = \Lambda^{-1}(\Phi(\gamma' Z_t))$, where Λ^{-1} is the inverse of Λ. For example, if Λ in (9) is the logistic distribution function (that is, $\Lambda(h) = 1/(1 + \exp(-h))$), then

$$h^*(Z_t) = \log\left(\frac{\Phi(\gamma' Z_t)}{\Phi(-\gamma' Z_t)}\right) \tag{11}$$

Even if this is a highly nonlinear function, we have no problem, if the assumed nonparametric class for h^* is general enough.

As a fourth modeling option we could treat both Λ and h^* nonparametrically. Along this direction, Matzkin (1992) proposes an estimation method, where Λ is treated nonparametrically as in the second approach above, while h^* is assumed to belong to the set of monotone increasing, concave, and homogenous of degree one functions. The model of Matzkin (1992) is more general than typical conditional probability models in microeconometric literature. However, by assuming a fixed Λ and by letting h^* be flexible enough we have the capacity to cover richer classes of models than the one implicit in the Matzkin's (1992) framework. Hence, the third modeling option above should be sufficiently general for our purposes. The subsequent section shows how we handle such a model empirically in practice.

[7] Here it is natural to assume that Z_t includes a constant.

3 Estimation

In the previous section, we argued that we can capture arbitrarily complex forms of $\Pr(R_{t+\tau} = 1|Z_t)$ by a model, where the link function Λ in (9) is fixed and the index function h^* is treated nonparametrically. To estimate such a model from data we use the gradient boosting estimator from the machine learning literature (see Friedman (2001)). Here we present the estimation method and its special features.

At the population level, the goal of the boosting estimator is to use observations $(R_{t+\tau}, Z_t)$, $t = 1, ..., T$, to find a function h from $\mathbb{R}^n$ to $\mathbb{R}$ such that it minimizes

$$E(C(R_{t+\tau}, h(Z_t))) \tag{12}$$

where $C(r, h)$ is a criterion function from $\mathbb{R} \times \mathbb{R}$ to $\mathbb{R}^+$.

For our goal, the criterion $C(r, h)$ must be such that h^* is the minimizer of (12). For a given Λ in the model (9), a suitable choice is the "negative binomial likelihood"

$$C(r, h) = -\left(r \log(\Lambda(h)) + (1 - r) \log(1 - \Lambda(h))\right) \tag{13}$$

From the maximum likelihood theory we know that under this criterion (12) is minimized by h^*.

In our modeling strategy h^* is what "adjusts," and we can basically fix Λ arbitrarily. In the boosting technique (see Friedman, Hastie and Tibshirani (2000)), it is common to use the logistic distribution function in the form

$$\Lambda(h) = \frac{1}{1 + e^{-2h}}$$

Under this model, the criterion in (13) can be written as

$$C(r, h) = \log(1 + \exp(-2(2r - 1)h)) \tag{14}$$

and the corresponding population minimizer of (12) is the half of the log-odds ratio

$$h^*(z) = \frac{1}{2} \log\left(\frac{\Pr(R_{t+\tau} = 1|Z_t = z)}{1 - \Pr(R_{t+\tau} = 1|Z_t = z)}\right) \tag{15}$$

As is shown in Friedman et al. (2000), the same population minimizer of (12) is attained by the "exponential loss"

$$C(r, h) = \exp(-(2r - 1)h)) \tag{16}$$

When the boosting estimator uses (14), it is often called the LogitBoost or the Binomial-Boosting, while when (16) is used, it is common to refer to the Adaboost estimator (see Friedman et al. (2000)).

For a given criterion function, the actual boosting estimator is concerned with minimization of the empirical version of (12)

$$\frac{1}{T}\sum_{t=1}^{T} C(R_{t+\tau}, h(Z_t)) \tag{17}$$

Bühlmann and Hothorn (2007) show that the bulk of the boosting estimators can be represented by the following generic algorithm.

1. Choose an initial estimate $\widehat{h}_0(\cdot)$. Common choices include $\widehat{h}_0(\cdot) = 0$ and $\widehat{h}_0(\cdot) = \overline{h}$, where $\overline{h}$ is a constant that minimizes (17). Set $m = 0$.

2. Increase m by 1. Compute the negative gradient $-\partial C(r,h)/\partial h$ and evaluate it at $\widehat{h}_{m-1}(Z_t)$:
$$U_t = -\frac{\partial}{\partial h} C(R_{t+\tau}, h)\Big|_{h=\widehat{h}_{m-1}(Z_t)}, \ t = 1, ..., T$$

3. Fit $(U_1, ..., U_T)$ to $(Z_1, ..., Z_T)$ by a real-valued "base procedure" so as to obtain the estimate $\widehat{b}_m(\cdot)$. Here, $\widehat{b}_m(\cdot)$ can be regarded as an approximation of the negative gradient vector.

4. Update $\widehat{h}_m(\cdot) = \widehat{h}_{m-1}(\cdot) + \nu \times \widehat{b}_m(\cdot)$, where $\nu \in (0, 1]$ is a "shrinkage" parameter.

5. Iterate steps 2 to 4 until $m = m_{stop}$, where m_{stop} meets a given stopping rule.

As a whole, the above algorithm can be understood to minimize (17) by iterative steepest descent in function space. Bühlman and Hothorn (2007) call it the functional gradient descent algorithm. For the performance of the boosting estimator, the base procedure in step 3 and the stopping rule in step 5 are most important. These and tuning methods of the estimator are discussed in the following two subsections.

3.1 Base Procedure

Realize from step 4 of the algorithm that the boosting estimate after m iterations is

$$\widehat{h}_m(\cdot) = \widehat{h}_0(\cdot) + \nu \sum_{j=1}^{m} \widehat{b}_j(\cdot)$$

We see that the structure of the boosting estimate for h^* is given by a linear combination of the structural features of the base procedure estimates $\widehat{b}_j(\cdot)$.

The base procedure consists of a base model and a fitting method by which it is fitted to predict $(U_1, ..., U_T)$ using $(Z_1, ..., Z_T)$ (step 3). The base model is commonly a parametric function $b(\cdot; \theta)$, where the parameter θ is finite dimensional. If θ estimated by least squares (LS), as is common, then at each iteration m one sets $\widehat{b}_m(\cdot) = b(\cdot; \widehat{\theta})$, where $\widehat{\theta}$ is a solution to the problem

$$\min_{\theta} \sum_{t=1}^{T} (U_t - b(Z_t; \theta))^2$$

In a "componentwise" base procedure, the estimate $\widehat{b}_m$ uses just a single predictor. For example, one may fit (by LS) a single variable model $q(\cdot; \theta)$ for each predictor so as to obtain the estimates $q(\cdot; \widehat{\theta}_i),\ i = 1, ..., n$, and then set

$$\widehat{b}_m(z) = \widehat{b}_m(z_1, ..., z_n) = q(z_\upsilon; \widehat{\theta}_\upsilon)$$

where

$$\upsilon = \arg \min_{1 \le i \le n} \sum_{t=1}^{T} (U_t - q(Z_{it}; \widehat{\theta}_i))^2$$

In a componentwise base procedure, the final boosting estimate can always be expressed in the form

$$\widehat{h}_m(z) = \widehat{h}_m(z_1, ..., z_n) = \sum_{i=1}^{n} \widehat{\psi}_{i,m}(z_i) \tag{18}$$

If either (14) or (16) is used as the criterion, the corresponding estimate for the conditional probability $\Pr(R_{t+\tau} = 1 | Z_t = z)$ is obtained via (15). Hence, the estimator is implicitly estimating a model of the form

$$\Pr(R_{t+\tau} = 1 | Z_t = z) = \frac{1}{1 + \exp\left(-2 \sum_{i=1}^{n} \psi_i(z_i)\right)} \tag{19}$$

where $\psi_i,\ i = 1, ..., n$, are functions from $\mathbb{R}$ to $\mathbb{R}$. The flexibility of the model in (19) depends on how general functions ψ_i can be approximated by the applied base model.

A popular choice for a base model is a "decision tree." In the componentwise base procedure, such a model reduces to a single-step function and is often called a "stump tree." Despite its simplicity, the stump tree has, with a large enough number of iterations m, the capacity to approximate ψ_i that is arbitrarily complicated. The componentwise boosting estimator with a stump tree base model has turned out to perform well in numerous simulation studies and empirical applications (see Hastie, Tibshirani, and Friedman (2009)). We can understand this as follows. When a conventional parametric (like logit) model is applied, one must often figure out whether the predictors be included in the model as such or whether one should instead use their nonlinear transformations as predictors. The advantage of the boosting estimator is that it can estimate the right transformations, the functions ψ_i, automatically. Moreover, the method has been found to be very successful in selecting the best performing predictors, even if a large number of variables is given to the estimator (see Bühlmann (2006)). Finally, even if the assumed link function (typically the logit function) may not be the right one, this is not necessarily so critical, because it may be largely compensated by the adjustment of the transformations of the predictors.

While the additive separable structure implicit in the componentwise base procedure may yield accurate approximations for the true h^* under the assumed link function, sometimes achieving appropriate accuracy requires a more general base procedure.

A generalization to the componentwise base procedure is obtained by specifying the base model $b(\cdot;\theta)$ so that it uses $k \leq n$ predictors at an iteration m. Denote the different subsets of k members of the integers $\{1, ..., n\}$ by $K_j,\ j = 1, ..., J\ (J = n!/k!(n-k)!)$. Then, the boosting estimate after m iterations can be expressed as

$$\widehat{h}_m(z_1, ..., z_n) = \sum_{j=1}^{J} \widehat{\psi}_{j,m}(z_{j_1}, z_{j_2}, ..., z_{j_k}), \tag{20}$$

where $j_1, ..., j_k \in K_j$. Realize that the actual estimate $\widehat{\psi}_{j,m}(z_{j_1}, z_{j_2}, ..., z_{j_k})$ does not need to depend on each k arguments and hence may be a function of fewer than k predictors. In practice, it is likely (especially if n is large) that $\widehat{\psi}_{j,m}(z_{j_1}, z_{j_2}, ..., z_{j_k}) = 0$ for a majority of the indices j.

How does one choose k, "the interaction depth"? The estimate in (20) is implicitly estimating a model of the form

$$p(z_1, ...z_n) = \frac{1}{1 + \exp\left(-2\sum_{j=1}^{J} \psi_j(z_{j_1}, z_{j_2}, ..., z_{j_k})\right)}$$

It is clear that with a larger k this model can cover a larger class of conditional probability functions. We can think of situations where the perfect population level fit to the true model entails using the maximum interaction depth $k = n$. For example, if the true conditional probability can be expressed with a probit link and a linear index function, then we know that under the logit link the right index function is given by (11), where the effective number of interactions may be n. However, several reasons speak against assuming the maximum interaction depth. First, the computational burden increases quickly as we increase the interaction depth and often it is simply infeasible to apply the maximum interaction depth. Secondly, already a small increase in the interaction depth may yield sufficient improvement in accuracy for practical purposes. Most importantly, assuming a relatively small interaction depth is motivated, because the boosting estimator tends to perform best when the base procedure is kept simple. We turn to this in more detail in the subsequent section.

3.2 Preventing Overfitting

Even if the base model may be very simple (but still nonlinear), the boosting estimate can become exceedingly complicated as the number of iterations m grows large. As a result, when the boosting estimate is plugged into the empirical criterion in (17), we get in return smaller (or never larger) number the larger m is. Hence, the boosting estimator can overfit and does so eventually as m gets larger. The success of the boosting estimator rests largely on techniques that prevent overfitting from happening.

The goal is choose the final iteration number m_{stop} (in step 5 of the algorithm) such that the resulting model estimate minimizes the population level criterion in (12). In other words, we want that the estimated prediction rule is the best possible when it is applied to new observations, which are not in our estimation sample.

Most commonly, the stopping rule (in step 5 of the algorithm) is based on a cross-validation (CV) procedure, where the value of (12) is estimated for a sequence of values of m. In a K-fold CV, the original data are randomly divided into K portions of equal size. One of the folds is put aside at the time and the remaining $K-1$ folds are used to obtain the boosting estimate for a given m. For each of the K such boosting estimates, an out-of-sample estimate for (12) is computed by using data on the single fold that was put aside. The resulting K estimates are averaged so as to obtain the final out-of-sample estimate for (12) based on m iterations. This is run for a range of values of m and the one yielding the best out-of-sample accuracy will serve as the estimate m_{stop} for the optimal number of iterations. In the application, we will use ten-fold CV.

The performance of the CV procedure for estimating the optimal m is better, if the true out-of-sample accuracy of the boosting estimator changes only slowly as a function of m. Bühlmann and Hothorn (2007) call this "slow overfitting behavior."

In general, slow overfitting is more likely when the base procedure is "simple." Sufficient simplicity is attained by specifying the base model so that it belongs to a narrow class of functions and that it applies a small number of predictors at an iteration. The componentwise base procedure (discussed earlier) takes parsimonity into an extreme and has turned out to be very competitive approach in various studies. When one has to make the base model more flexible by increasing its interaction depth, it is good to do this cautiously.

As to a simple functional form of the base model, the decision tree (discussed in the previous section) is a good option as it amounts to a piecewise constant function with a chosen maximum number of discrete jumps. A linear base model is also simple, but has the disadvantage that the final boosting estimate is restricted to be linear. Among nonlinear parametric functions, smoothing splines have turned out to be particularly well performing (see Bühlmann and Yu (2003)). In our application, we use the smoothing splines as the main method, but also consider estimates based on the decision tree.

The shrinkage parameter, ν, in step 4 is an additional device for preventing the boosting algorithm from fitting too quickly. When ν is small, the contribution of an iteration to the fit is small so that it takes more iterations to improve the fit. In this paper, we

apply the value $\nu = 0.01$, which agrees with common recommendations in the literature (e.g., Friedman (2001) and Bühlmann and Yu (2003)).

We end this section by noting that with suitable choices for the base procedure, the stopping rule and tuning parameters, the boosting estimator is consistent for the optimal h^* (see Bühlmann and Hothorn (2007) for a detailed analysis). The particular advantages of the boosting estimator include the following features. It can be configured to be as general as alternative nonparametric estimation techniques and hence can estimate optimal models that do not belong to conventional parametric and semiparametric model classes. Moreover, it has been shown that the boosting estimator can attain the optimal minimax rate of convergence of nonparametric estimators under more general conditions than conventional nonparametric estimators (see Bühlmann and Yu (2003)).

4 Empirical Analysis

4.1 Variables

As was described in Section 2, in our empirical illustration R_t is a dummy variable indicating whether the US economy is in recession at month t (according to the NBER). Our candidate predictors, the components of Z_t, derive from variables given in Table 1. See the Conference Board (2001, pp. 50-51) for detailed arguments for why these variables are assumed to have predictive content for future business cycle turning points.

For all variables in Table 1, we examine the performance of alternative transformations as indicated by crosses in the table. If an original variable is regard as stationary, it is applied as such ("L"), as month-to-month percentage change ("M-M") and as year-to-year percentage change ("Y-Y"). If there is evidence from unit root tests that the original series (or its logarithmic form) is integrated of order one so that it is stationary only after differencing, then we consider only its M-M and Y-Y transformations as candidate predictors.

All data on the original series are downloaded from Macrobond database and constitute observations on the period 1959M1-2017M12. For a given predictor, we use the

observation that is available at the end of the month. When we estimate a 12 month ahead prediction model conditional on a predictor in Y-Y form, the available sample starts from 1961M1. We use the same starting month of the sample also with other (shorter) prediction horizons and with other transformations of the predictors.

4.2 Assessing Individual Predictors

In this section, we analyze the predictive performance of the candidate predictors in Table 1. We take a variable at the time and assess the predictive content of its alternative transformations for prediction horizons $\tau = 3, 6, 9, 12$. The idea is to examine what transformation is best for a given variable and at what horizon the variable has the strongest predictive power.

For a given predictor Z_t and horizon τ, let $\widehat{p}_\tau(z)$ be the boosting estimate for $\Pr(R_{t+\tau} = 1|Z_t = z)$. We assess the predictive power of Z_t for $R_{t+\tau}$ by the standard nonparametric ROC curve estimate for $\widehat{p}_\tau(Z_t)$ (where we use the sample observations on $R_{t+\tau}$ and $\widehat{p}_\tau(Z_t)$). The underlying population level target is the optimal ROC curve based on Z_t, that is, the ROC curve for $h^*(Z_t)$, where h^* is $\Pr(R_{t+\tau} = 1|Z_t = z)$ or its monotone increasing transformation.

When one knows that the true conditional probability is monotone increasing (decreasing), the optimal ROC curve is obtained by letting h in (3) be $h(z) = z$ ($h(z) = -z$). In this situation, we can estimate the optimal ROC curve directly by the empirical ROC curve for Z_t (or $-Z_t$), and there is no need to estimate $\Pr(R_{t+\tau} = 1|Z_t)$.

For most predictors prior reasonings suggest that the underlying predictive relationship is monotone (see the Conference Board (2001)). However, we cannot be certain that such monotonicity holds and for some variables we might even be uncertain whether there is a positive or a negative predictive relationship depending on the prediction horizon (see Berge and Jordà (2011)). We evaluate these questions by using the boosting estimate $\widehat{p}_\tau$. In particular, we question the initial hypothesis of monotonicity, if we find that the graph of $\widehat{p}_\tau(z)$ is not monotone and if the empirical ROC curve for $\widehat{p}_\tau(Z_t)$ is uniformly above the empirical ROC curves for Z_t and $-Z_t$. Obviously, there is evidence against monotonicity

of $\Pr(R_{t+\tau} = 1|Z_t = z)$ also, if the empirical ROC curve for Z_t has convex and concave segments.

Table 2 summarizes our findings on the examined variables and their alternative transformations. The table gives answers to four main questions: (1) Does the predictor have predictive power for future recessions and how the predictive power changes by horizon? (2) How alternative transformations of the variable perform? (3) Is the predictive relationship monotone and how stable it is? (4) What is the overall performance of the predictor and how it compares with other predictors?

As an illustration, Figure 2 presents results for the yield spread, the difference between the ten year Treasury bond rate and the Federal funds rate.[8] The four panels in the figure depict the aforementioned ROC curve estimates for horizons $\tau = 3, 6, 9, 12$. The ROC curves indicate that the spread has clear predictive power for recessions and that this power is strongest at the horizon of twelve months. Across the examined horizons, the ROC curves for the boosting estimates do not differ (when a smoothing spline is applied as the base model) or differ only slight (when a stump tree is applied) from the ROC curve estimated for the negative value of the series. This observation suggests that the predictive relationship is monotone decreasing, a drop in the yield curve increases the likelihood of an upcoming recession period. The result is consistent with earlier evidence and reasonings on the underlying relationship (e.g., Estrella and Mishkin (1998)).

As is stated in Table 2, the spread is found to be the most powerful single predictor when the horizon is twelve months. The closest competitor at this horizon is the building permits (in Y-Y form). The ROC curves for this predictor are given in Figure 3. The blue dashed curve is for the negative value of the original variable and hence indicates that a drop in the year-to-year change of building permits increases the risk of a future recession. However, the ROC curves for the boosting estimates for the conditional probabilities at horizons 9 and 12 are somewhat above the corresponding ROC curves for the variable as such. This suggests that the underlying predictive relationships are not monotone.

The aforementioned non-monotonicity gets support from the estimated conditional

[8] We obtain almost identical results when the Fed funds rate is replaced by the three month Treasury bill rate.

probability functions. Indeed, as can be seen from Figure 4, the risk of recession (either 9 or 12 month ahead) is first increasing and then decreasing as a function of the annual change of building permits. This pattern is also consistent with the graph of the annual change of building permits shown in Figure 5. The annual change of building permits declines gradually in advance to recession periods (marked by the shaded area) explaining the decreasing parts of the estimated conditional probability functions. On the other hand, we also observe that the largest annual decreases of building permits occur during recessions and sometimes just towards their ends. Hence, the deepest annual drops of building permits tend to anticipate a turn from a recession to an expansion phase, which is in line with the increasing parts of the conditional probability function estimates in Figure 4. In view of these observations, it is plausible that the underlying predictive relationships are unimodal, not monotone. Hence, we gain more predictive power from building permits by using the boosting estimate of the conditional probability than by using the series (its negative value) as such.

While the year ahead prediction problem seems to be the game of a few predictors, there are many powerful variables for short horizon prediction. The strongest quarter ahead predictors include weekly initial claims (in Y-Y form), consumer expectations index (level), and ISM new orders (level). Figure 6 shows ROC curves for the first of these variables. At horizon $\tau = 3$, the red and green ROC curves for initial claims are almost uniformly better than the ROC curve for the spread (see the case $\tau = 3$ in Figure 2).

The ROC curves in Figure 6 indicate strongly that the conditional risk of a future recession is not a monotone function of initial claims. The boosting estimates in Figure 7 indeed show that the underlying relationship is unimodal at all horizons. Whenever the annual change of initial claims is negative, there is a very small risk that a recession is coming ahead. Otherwise, when claims are going up, the risk of recession is first increasing and then decreasing as a function of the annual change of claims. This unimodal shape is due to the fact that initial claims tend to rise fastest at the end of recession periods and therefore yield signals for an expansion rather than for a continuation of recession. This observation is similar to what we obtained for building permits above.

We complete this section by two summary notes. First, for most variables the annual

difference (Y-Y form) yields the largest predictive power at all four horizons. This is an important finding, because so far most recession prediction models have transformed I(1) predictors into month-to-month rather than year-to-year differences. We interpret that the cyclical patterns of year-to-year differences carry more predictive power for business cycle fluctuations than the often quite erratic month-to-month differences. Second, for at least half of the predictors the estimates $\widehat{p}_\tau$ and their ROC curves indicate that the underlying predictive relationship is not monotone. This finding supports the idea that the predictive power of leading indicators cannot be captured optimally by a model that assumes monotone marginal effects of predictors.

4.3 Combining Predictors

In this section, we use the boosting method to estimate recession prediction models based on the transformations that we found to perform best in the previous section. That is, our predictors in Z_t constitutes the ten variables specified in column 2 of Table 2.

We apply the boosting estimator with a componentwise base procedure to estimate models of the form (19) for horizons $\tau = 3,\ 6,\ 9,\ 12$. We use the smoothing spline as the base model. Initially, the estimated models use all 10 predictors.

To measure the importance of the predictors in a model, we compute how much each predictor contributes to the minimization of the estimation criterion function. Let $\overline{C}_m$ denote the value of (17) after iteration m is completed. For each predictor Z_{it}, $i = 1, ..., n$, compute $V_i = \sum_{m=1}^{m_{stop}} P_{im}(\overline{C}_m - \overline{C}_{m-1})$, where $P_{im} = 1$, if Z_{it} is applied in step m, and $P_{im} = 0$ otherwise. Then, the relative contribution, or "importance," of Z_{it} is $c_i = V_i/V$, where $V = \sum_{i=1}^{n} V_i = \overline{C}_{m_{stop}} - \overline{C}_0$. This is essentially the variable importance procedure available in the "mboost" package for R program (see Hofner, Hothorn, Robinzonov and Scmid (2014)).[9]

Table 3 presents the relative contributions of the predictors for the four horizon specific models using all ten predictors. According to the results, the spread has the largest contribution share in the 12, 9 and 6 month ahead models. Its contribution share is larger

[9] The most recent version of the mboost package is available at https://github.com/boost-R/mboost.

than 50% at horizons 12 and 9, and 40% at horizon 6. At horizon 3, the contribution share of the spread is about 20%, while building permits and initial claims both have 25% shares. Building permits and initial claims have relatively large shares, more than 10%, at horizon 6 as well. M2 has the second largest share, more than 10%, at horizons 12 and 9, while stock shows up with 10% share at horizon 3. According to the estimated shares, initial claims is the weakest indicator at horizon 12, stock at horizon 9, manufacturers' new orders for capital goods at horizon 6, and hours at horizon 3.

In general, the estimated relative contributions as such do not tell which variables are really needed in the model. To assess this question, we apply a two step procedure. In the first step, we estimate a series of submodels obtained from the full model by dropping the predictors one-by-one, as devised by their estimated importance. In the second step, we compare the ROC curves for the estimated sub-models so as to assess which predictors are really carrying marginal predictive power among the ten predictors. To this end, we describe the dropping procedure in more detail and what results it yields for different horizons, while the second step analysis will be reported in Sections 3.2.1 and 3.2.2 below.

Initially, we drop the indicator that has the weakest estimated contribution in the model using all ten indicators and re-estimate the model using the remaining nine indicators. Based on the estimated relative contributions of the indicators in the latter model, we again drop the weakest predictor and re-estimate the model using the remaining eight predictors. We keep on running this loop until there is no predictor left. The resulting dropping orders of indicators are shown for the four horizons in Table 4.

The dropping order of the indicators changes by the horizon. For example, initial claims is the first dropout indicator at horizon 12, but is among the last ones to be dropped when the horizon is either 6 or 3. Building permits is dropped at the fourth place at horizon 12, while at horizons 9, 6, and 3 it is clearly among the top indictors. Hours is clearly the worst performer in that it is dropped among the very first ones at all horizons. As is expected, the spread is the most precious indicator by being the last indicator to be dropped at horizons 12, 9 and 6, and is among the top three at horizon 3. M2 is also among the strongest performers, but its position profile in the dropping order suggests that it is slightly more valuable at longer horizons.

4.3.1 ROC Curve Analysis of Twelve Month Ahead Models

Figure 8 plots ROC curves for ten 12 month ahead models. The red ROC curve corresponds to the full model using all ten predictors. The green curve is for the model without initial claims, which was estimated to be the weakest predictor in the full model. The blue ROC curve is for the model that we estimated after we dropped hours, the weakest predictor in the previous model, and so on.

First of all, Figure 8 suggests that, as a suitable combination, many indicators give more predictive power 12 months ahead than any single indicator alone. Do we need to employ all ten indicators to obtain the most accurate 12 month ahead predictions? From Figure 8 we see that not necessarily, because the red ROC curve is not uniformly above the other ROC curves. In fact, for values of α larger than 0.6, the ROC curve based on the spread alone is not inferior to any other ROC curve in the figure. That is, if one's preferences are such that the right hand side of equation (5) is very small, then according to Figure 8, one might well be fine by using the spread as the sole predictor. However, in practise, it is unlikely that anyone is willing to use a recession prediction model, if it means that more than half of the expansion months a year ahead are incorrectly predicted to be recession months.

Consider what a user with a symmetric loss ($L(0) = L(1)$) might choose to do. Given the observed frequency of recessions ($\widehat{\delta} = 0.12$), the right hand side of equation (5) is now about 7.1. Hence, for any fixed number $b \in [0, 1]$, the user is indifferent between any pairs (α, β) such that $\alpha \in [0, (1 - b)/7.1]$ and $\beta = b + 7.1\alpha$. For each of the ROC curves in Figure 8, the user finds the maximum b under the condition that $b + 7.1\alpha = ROC(\alpha)$. The model whose corresponding ROC curve yields the largest b results in the smallest expected loss. It turns out that the largest b (around 0.61) is attained by the blue ROC curve in Figure 8. The corresponding optimal operating point (indifference line) is marked by a crossed circle (thin black line).

In view of the above considerations, a user with symmetric loss would prefer the model that is estimated without hours and initial claims. Looking at Figure 8, we can also conclude that for the user's expected loss it makes no essential difference, whether

he applies any of the models yielding the four solid ROC curves. In fact, we may draw the same conclusion quite much independent of the slope of the indifference line of the user. Hence, there are at least four models between which most users are indifferent. Among such models it reasonable to choose the most parsimonious one, as parsimonity is likely to result in better estimation accuracy in finite samples. This means that from the full model one could well drop hours, initial claims and manufacturers' new orders for consumer goods.

For the sake of full robustness one might want to compare ROC curves for models estimated with all combinations of the ten predictors. This would be difficult to handle in practice, because there are altogether 1024 different subsets of ten predictors. In fact, the dropping procedure above is a short-cut by which we can avoid the comparison of a vast number of alternative specifications. Based on extensive experimentation with the dropping procedure we are confident that it provides us with a reliable enough device for specifying the alternative models through which we can find what predictors are really carrying marginal predictive content. It is also to be noted that the dropping procedure starts from the most general model including all predictors so that each predictor is given a chance to contribute the prediction.

Nevertheless, if one is uncertain whether the dropping procedure might miss an important subset of predictors, it may be feasible to conduct a targeted robustness check. For example, above we concluded that initial claims, hours and manufacturers' new orders for consumer goods could be dropped from the model. This conclusion rests on the observation that there is no essential difference in accuracy between the full model and the three sub-models that we obtain by dropping the variables in the order they become estimated as the least contributing predictor in a model. If we account for other possible dropping orders of the three variables, we come up with four additional subsets of the three variables. We can check if any of the corresponding alternative models show up more predictive power in some of the three predictors. In the present application, we find that the additional four subsets of the three variables make no difference to the accuracy of the model as measured by the ROC curve. This result makes us more confident that the three variables are not essential for the accuracy of the estimated model.

4.3.2 ROC Curve Analysis of Shorter Horizon Models

We briefly discuss our results on shorter horizon prediction models. We take the dropping orders from Table 4 and consider the ROC curves for the associated conditional probability fits. To enhance the visual comparison of the ROC curves we adjust the figure in two ways. First, for each horizon, we exclude from the figure any ROC curve that is dominated by another ROC curve in the set in that the former curve is uniformly below the latter one. Second, we plot the ROC curves only for $\alpha \in [0, 0.4]$, as for each of the considered ROC curves $ROC(\alpha) = 1$ for $\alpha \in [\overline{\alpha}, 1]$, where $\overline{\alpha} \leq 0.4$.

Figure 9 plots the ROC curves, subject to the adjustments just described, separately for 9, 6 and 3 month ahead models. For the horizon 9 (the top panel in Figure 9), two of the original ten ROC curves are dominated and excluded from the figure. These are the ROC curve for spread alone and the one for the model using spread, M2, manufacturers' new orders for capital goods and consumer expectations. Hence, we may conclude that the aforementioned four predictors cannot capture the full predictive power of the ten predictors at horizon 9. Overall, the ROC curves support the view that all ten predictors, perhaps with the exception of stock, are needed in the 9 month ahead prediction model. This conclusion is clearly different from the 12 month ahead model where it seemed that we could drop three predictors.

In the case of horizon 6 (the middle panel in Figure 9), we find that six ROC curves are dominated by one of the four ROC curves. In view of the figure, it seems fairly clear that all ten predictors play a role in the 6 month ahead prediction model. In the case of 3 month ahead horizon (the bottom panel in Figure 9), we can exclude four of the ten ROC curves. Among the curves in the figure, the red and the green ones are essentially equal, which suggests that hours could be dropped from the 3 month ahead model.

4.4 Increasing Interaction Depth

The results of the previous section are based on the boosting estimates of the model (19), where the index function is restricted to be separable additive. As is explained in Section 3.1, a more general class of models can be accommodated by using a base model

that is a function of several variables. In this section, we explore whether increasing the interaction depth of the base model results in improvements to the prediction accuracy. To save space, we limit the analysis on the 12 month ahead prediction model.

As a first cut, we apply a bivariate extension to the smoothing spline base model. That is, we effectively increase the interaction depth of the smoothing spline base model (of the previous section) from one to two. We include in the estimation the same ten predictors as in the previous section.

Effectively, now the estimation procedure using all ten predictors corresponds to the componentwise procedure where the "candidate models" for $\widehat{b}_m$ include the ten univariate smoothing spline models (as before) and the 45 bivariate smoothing spline models that can be obtained from all possible pairs of the ten variables. Hence, at each iteration m, the base procedure (step 3 of the boosting algorithm) fits 55 different smoothing spline functions, one at the time, and selects for $\widehat{b}_m$ the one that provides the best fit.

As in the previous section, we wish to assess the importance of the predictors. For this task, we now compute by how much each of the 55 predictor combinations (10 individual predictors and 45 pairs of predictors) contributes to the minimization of the estimation criterion (17). We then rank the 55 predictor combinations (smoothing spline models) by their estimated relative contribution. We drop the predictor whose best position in this ranking is the weakest one and re-estimate the model using the remaining nine predictors (with the candidate models for $\widehat{b}_m$ constituting nine univariate and 36 bivariate smoothing spline models). We again rank the underling 45 predictor combinations by their relative contribution and use the ranking to identify the weakest predictor to be dropped. We keep on running this procedure until there is just one predictor left.

The dropping order that results in from the above described procedure is very similar to the one that we obtained based on the univariate base model (see the first column in Table 4). The only clear difference between the two dropping orders concerns the two versions of manufacturers' new orders. When we allow for the bivariate smoothing spline functions in the base model, manufactures' new orders for capital goods is dropped as the third and manufactures' new orders for consumer goods as the ninth indicator. In Table 4, the corresponding dropping positions are essentially the opposite, the former being

dropped as the eight, the latter as the third indicator. Otherwise, the dropping order of predictors is very similar whether we apply the univariate or the bivariate base model.

Again, we use the ROC curve to assess which predictors are really needed in the model, when it is estimated by using the bivariate extension to the smoothing spline base model. We consider ROC curves for the ten boosting fits that we obtained by running the dropping procedure described above. Among the ten ROC curves we consider only ones that are not dominated by another ROC curve in the set. This leaves us with four ROC curves, which are shown as red, blue, cyan and magenta curves in Figure 10. The black solid curve in Figure 10 is the ROC curve for the fit of the univariate componentwise model using all ten predictors (the red line in Figure 8), while the black dashed curve is the ROC curve for the spread variable alone. Finally, the orange line in Figure 10 is the ROC curve for a probit model fit of the 12 month ahead recession dummy on the ten predictors. Various interesting observations arise from Figure 10.

First, the added flexibility of the base model improves the accuracy of the prediction. The ROC curves coming from the bivariate base model dominate rather clearly the black ROC curve, where the underlying model assumes the univariate base model with all ten predictors.

Second, when we allow the bivariate base model, it seems quite clear that we get the same prediction accuracy whether we use all ten predictors or drop initial claims and hours from the predictor set. Moreover, if we drop any other variable (besides claims and hours) then we lose accuracy rather clearly. Overall, Figure 10 should convince us that there are at least seven predictors that carry out significant marginal predictive content beyond the spread.

Third, observe from Figure 10 that the ROC curve for the probit fit using all ten predictors is not uniformly above the ROC curve based on the spread alone. That is, the two ROC curves cross at several places. According to the estimation results on the probit model, the spread variable is clearly significant with a very large t-value. On the part of the other predictors, the estimated probit coefficients and their standard errors suggest that at least stock, manufacturers' new orders for capital goods and building permits are statistically significant in the model and hence these variables should provide

additional predictive power beyond the spread. But, in view of the ROC curves in Figure 10, the probit model cannot combine the predictive power of the predictors in an optimal manner. Instead, the ROC curves for the various boosting fits show that to optimally exploit the joint predictive power of the predictors we need to apply a more flexible modeling approach.

The fact that we get more accurate predictions by assuming the bivariate rather than the univariate base model is quite expected. First, even if the true optimal prediction rule could be expressed as a separable additive model, we have no reason to believe that this holds under the logistic link function (as is assumed in the applied boosting estimation procedure). If the true model is obtained using the probit link function and a separable additive index function, then (as was seen in (11)) the corresponding optimal h^* assumes a large interaction depth. Of course, we have no better reason to believe that the separable additive index function holds under the probit link either. Independent of what link function we choose, by a flexible enough base model, the boosting estimator has the capacity to estimate the true conditional probability model with sufficient approximation. It is then natural to ask whether we should assume a larger interaction depth than two.

There is currently no procedure for running the boosting estimator with smoothing spline functions that are more general than bivariate. As an alternative to a multivariate spline function, one could increase the interaction depth of the base model by allowing for cross-products of one- and bivariate spline functions. However, in the present application such extensions turned out computationally very demanding, even infeasible, to handle. Hence, we instead applied the decision tree as the base model with various higher order interaction terms between the predictors. We found that such a model could not beat the accuracy of the fit that we obtained by using the bivariate smoothing spline model. That is, even if we allowed the interaction depth of the decision tree to be larger than two, the resulting fit did not produce ROC curves that are uniformly above the best curves in Figure 10. The results on these analyses are available upon request.

Finally, we note that the prediction accuracy that we obtain by applying the bivariate smoothing spline base model is already quite impressive. From Figure 10, we observe that the best ROC curve hits 1 at $\alpha = 0.053$. This suggests that we could predict correctly all

recession months a year ahead, if we accept that out of the expansion months occurring a year ahead some 5 percent are incorrectly predicted to be recession months. If there is a prediction model based on the applied indicators that can perform even better, then such a model is not far from making no mistakes at all. We think it is unrealistic to assume that the applied indicators could be transformed to such a perfect prediction model. We are willing to conclude that the boosting based prediction model considered above may well be able to extract the optimal predictive power of the applied indicators for the state of the business cycle a year ahead.

5 Conclusion

This paper reviewed the problem of finding the best way to predict recession periods using one or several leading indicators. To minimize anybody's expected loss at a given forecast horizon, it suffices to find the underlying true conditional probability given the predictors. We reviewed alternative approaches for modeling the conditional probability and demonstrated that a good choice, in terms of generality and practice, is to assume a known parametric link function and a nonparametric index function. For estimation of such a model, we applied the gradient boosting technique from the machine learning literature, as this method allows the underlying index function to belong to a wide nonparametric class and has generally proved to have many advantages over alternative nonparametric estimation techniques.

We showed how the ROC curve for the estimated conditional probability fit allows one to assess the predictive power of a given indicator when the underlying true predictive relationship is not monotone. We applied the boosting method to investigate the predictive performance of alternative time series transformations of individual indicators and at what horizon an indicator carries the most of its predictive power. Based on this analysis we chose the best transformation for each original indicator and then applied the boosting method to estimate prediction models that exploit them jointly. We used a variable importance estimation procedure to drop the weakest indicator from a given prediction model and evaluated which predictors can be dropped without losing predictive accuracy

as measured by the ROC curve of the estimated conditional probability fit. Even if this procedure is not a formal statistical test, it allowed us to obtain a convincing picture of what indicators really carry marginal predictive power in the model.

The main findings from our empirical analysis were as follows. The original indicators should mostly be used as annual (year-to-year) rather than monthly (month-to-month) differences. Secondly, for most of the indicators the underlying predictive relationship is not monotone, and often the non-monotonicity is of the form first-increasing-then-decreasing, or vice versa. Third, our results on models that exploit many predictors indicate that more or less all of the examined ten leading indicators contribute the prediction accuracy at 3, 6 and 9 month horizons, while at the 12 month horizon at least two indicators can be dropped. Finally, our results indicate that our proposed procedures can exploit the predictive power of the applied indicators much more efficiently than a conventional approach based on a parametric model.

There are interesting questions for future research. Some studies (e.g., Ng (2014)) use boosting and other machine learning methods to predict recessions using a very large number of predictors. The procedures applied in this paper provide an approach for investigating which predictors in such settings really contribute the estimated prediction model. The fact that there may be tens or even hundreds of variables is likely to bring in computational and other challenges that call for new automated techniques to rank and drop predictors. Moreover, in general, it would be desirable to develop formal test procedures for assessing whether the ROC curve for a given model is uniformly above the curves based on competing models.

6 References

Berge, T. J. (2015) "Predicting recessions with leading indicators: Model averaging and selection over the business cycle," *Journal of Forecasting*, 34, 455–471.

Berge, T. J., and O. Jordà (2011) "Evaluating the classification of economic activity into recessions and expansions," *American Economic Journal: Macroeconomics* 3, 246–277.

Bernard, H., and S. Gerlach (1998) "Does the term structure predict recessions? The international evidence," *International Journal of Finance and Economics*, 3, 195–215.

Bühlmann, P. (2006) "Boosting for high-dimensional linear models," *Annals of Statistics*, 34, 559–583.

Bühlmann, P., and T. Hothorn (2007) "Boosting algorithms: Regularization, prediction and model fitting," *Statistical Science*, 22, 477–505.

Bühlmann, P., and B. Yu (2003) "Boosting with the L2 loss: Regression and classification," *Journal of American Statistical Association*, 98, 324–339.

Chen, Z, Iqbal, A, and H., Lai (2011) "Forecasting the probability of US recessions: A probit and dynamic factor modelling approach," *Canadian Journal of Economics*, 44, 651–672.

The Conference Board (2001). Business Cycle Indicators Handbook, New York.

Dueker, M. (1997) "Strengthening the case for the yield curve as a predictor of U.S. recessions," Federal Reserve Bank of St. Louis Economic Review 79 (March/April), 41–51.

Estrella, A., and F. S. Mishkin(1998) "Predicting U.S. recedssions: Financial variables ads leading indicators," *Review of Economics and Statistics*, 80, 45–61.

Estrella, A., and G.A., Hardouvelis (1991) "The term structure as a predictor of real economic activity," *Journal of Finance*, 46, 555–576.

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

Estrella, A., Rodrigues, A. P., and S., Schich (2003) "How stable is the predictive power of the yield curve? Evidence from Germany and the United States," *Review of Economics and Statistics*, 85, 629–644.

Filardo, A. J., (1999) "How reliable are recession prediction models?" Federal Reserve Bank of Kansas Economic Review 84 (2nd quarter), 35–55.

Friedman, J. H. (2001) "A greedy function approximation: A gradient boosting machine," *Annals of Statistics*, 29, 1189–1232.

Friedman, J. H., T. Hastie, and R. Tibshirani (2000) "Additive logistic regression: A statistical view of boosting," *Annals of Statistics*, 28, 337–407.

Hastie, T., R. Tibshirani, and J. Friedman (2009) "The elements of statistical learning: Data mining, inference, and prediction (2nd ed.)," Springer.

Hofner, B., T. Hothorn, N. Robinzonov, and M. Scmid (2014) "Model-based boosting in R: A hands-on tutorial using the R package mboost," *Computational Statistics*, 29,3–35.

Han, A. K. (1987) "Non-parametric analysis of a generalized regression model," *Journal of Econometrics*, 35, 303–316.

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

Lehmann, E. L., and J. P. Romano (2005) "Testing statistical hypotheses (3rd ed.)," Springer, New York, USA.

Matzkin, R. L. (1992) "Nonparametric and distribution-free estimation of the binary threshold crossing and the binary choice models," *Econometrica*, 60, 239–270.

Neyman, J., and E. S. Pearson (1933) "On the problem of the most efficient tests of statistical hypothesis," *Philosophical Transactions of the Royal Society of London, Series A*, 231, 289–337.

Ng, S. (2014) "Boosting recessions," *Canadian Journal of Economics*, 47, 1–34.

Pepe, J. L. (2003) "The statistical evaluation of medical tests for classification and prediction," Oxford University Press, Oxford, UK.

Rudebusch, G. D. and J. C., Williams (2009) "Forecasting recessions: The puzzle of the enduring power of the yield curve," *Journal of Business & Economic Statistics*, 27, 492–503.

Sherman, V. K. (1993) "The limiting distribution of the maximum rank correlation estimator," *Econometrica*, 61, 123–137.

Stock J. and M. Watson (1993) "A procedure for predicting recessions with leading indicators: Econometric issues and recent performance," in J. Stock and M. Watson (eds.) *Business cycles, indicators and forecasting*, University of Chicago Press, Chicago, USA.

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

Table 1. Predictor Candidates

Variable	Short name	Transformations		
		L	M-M	Y-Y
Term spread (Treasury Bond 10Y - Fed Funds rate)	spread	x	x	x
Average weekly hours, manufacturing	hours	x	x	x
Average weekly initial claims for unemployment insurance	claims		x	x
Building permits, new private housing units	permits		x	x
Manufacturers' new orders, consumer goods and materials	mnocons		x	x
Manufacturers' new orders, nondefense capital goods	mnocap		x	x
Stock prices, 500 common stocks	stock		x	x
Money supply, M2 (real)	m2		x	x
Index of consumer expectations	consexp	x	x	x
ISM new orders	ism	x	x	x

Notes: The series listed in the table are the same as those applied in the leading economic index (LEI) of Conference Board (see https://www.conference-board.org/data/bcicountry.cfm?cid=1) with the exception of M2, in place of which LEI uses the so called Leading Credit Index (LCI). The LCI is constructed, using principal component analysis, from six series related to financial and credit markets. As the formula of LCI is not available to us and as our purpose is not to try to mimic LEI as such, we prefer to use M2. We also note that LCI equals M2 until 1990 so the actual difference between the two series concerns the period from 1990 on.

Table 2. Performance Summary of Individual Predictors

Variable	Best transf.	Best horizon	Relationship shape	Overall potential
spread	L	all, long	monotone decreasing	best overall performance
hours	Y-Y	medium, long	unimodal	not among best
claims	Y-Y	short, medium	unimodal	among best
permits	Y-Y	all	unimodal	good for all horizons
mnocon	Y-Y	short, medium	unimodal	not among best
mnocap	Y-Y	medium, long	up, down, up	not among best
stock	Y-Y	short	mixed	good for short horizon
m2	Y-Y	medium, long	mixed	moderate performance
conexp	L	short, medium	mixed	moderate performance
ism	L	all	mixed	moderate performance

Notes: For a variable, we take one of its allowable transformations (as indicated by crosses in Table 1) at the time and estimate $\Pr(R_{t+\tau} = 1|Z_t)$ by the boosting method for $\tau = 3, 6, 9, 12$. For example, for the *spread*, Z_t is either the original variable as such (L), its month-to-month difference (M-M), or its year-to-year difference (Y-Y). The chracterizations in the table derive from qualitative analysis of the graphs of the estimated conditional probability functions and of their ROC curves. See text for more details.

Table 3. Predictor Contribution Shares

Indicator	$\tau = 12$	$\tau = 9$	$\tau = 6$	$\tau = 3$
spread, L	.558	.527	.406	.221
m2, Y-Y	.122	.146	.097	.036
mnocap, Y-Y	.080	.061	.046	.009
consexp, L	.068	.063	.087	.063
ism, L	.058	.044	.038	.013
stock, Y-Y	.039	.007	.053	.134
permits, Y-Y	.030	.055	.144	.255
mnocons, Y-Y	.023	.027	.009	.013
hours, Y-Y	.017	.029	.010	.003
claims, Y-Y	.006	.041	.112	.254

Notes: The numbers indicate the relative share of how much each indicator contributes to the minimization of the estimation criterion function of the boosting estimator using the univariate smoothing spline base model.

Table 4. Dropping Order of Predictors

Indicator	$\tau = 12$	$\tau = 9$	$\tau = 6$	$\tau = 3$
claims, Y-Y	1	5	8	9
hours, Y-Y	2	3	2	1
mnocons, Y-Y	3	2	1	4
permits, Y-Y	4	8	9	10
stock, Y-Y	5	1	5	7
consexp, L	6	6	6	6
ism, L	7	4	3	3
mnocap, Y-Y	8	7	4	2
m2, Y-Y	9	9	7	5
spread, L	10	10	10	8

Notes: The numbers indicate the dropping order of indicators that resuls in when the dropping procedure described in the text is applied to prediction models estimated using the boosting estimator based on the univariate smoothing spline base model.

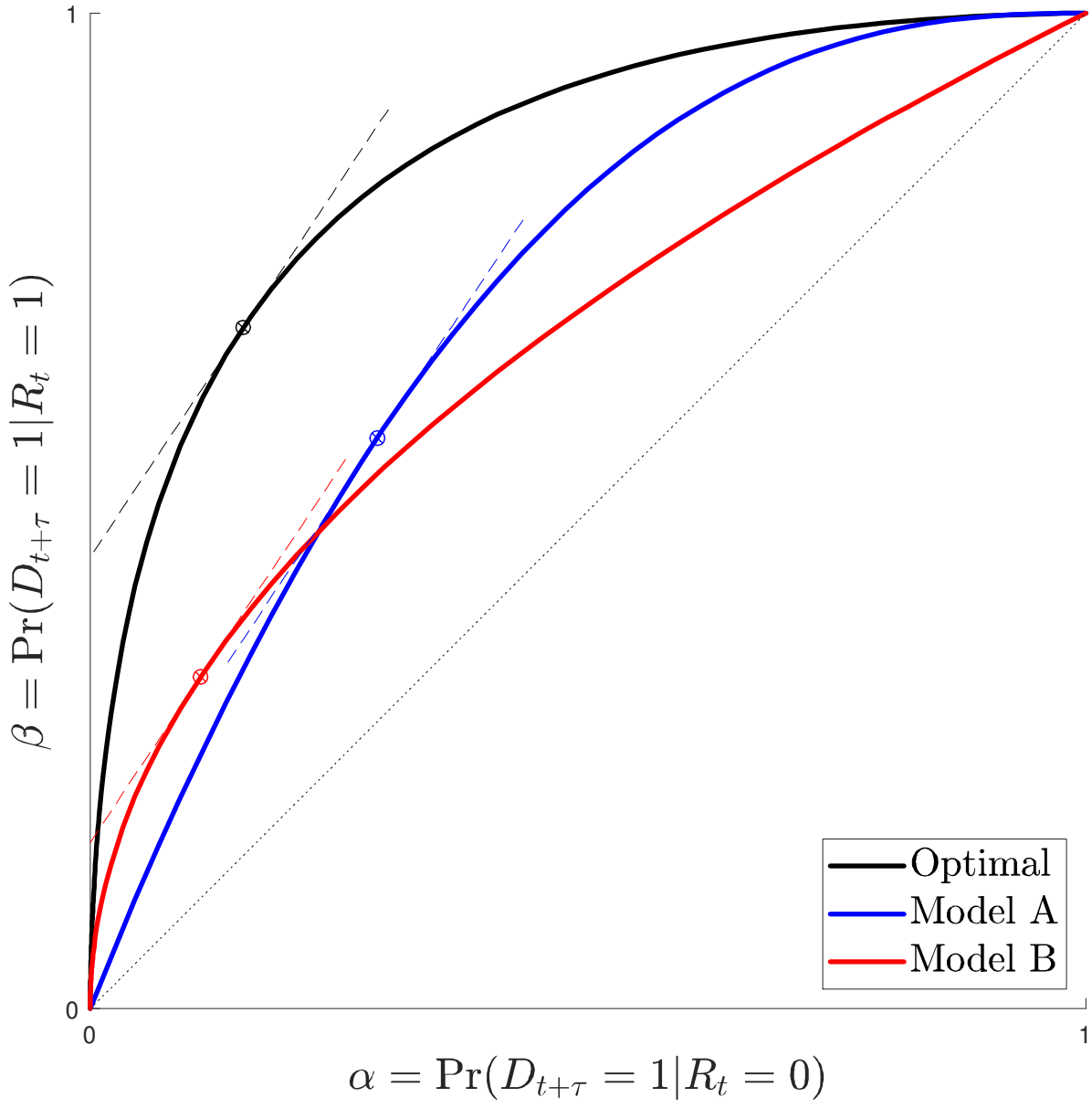


Figure 1: ROC Curves for Three Models

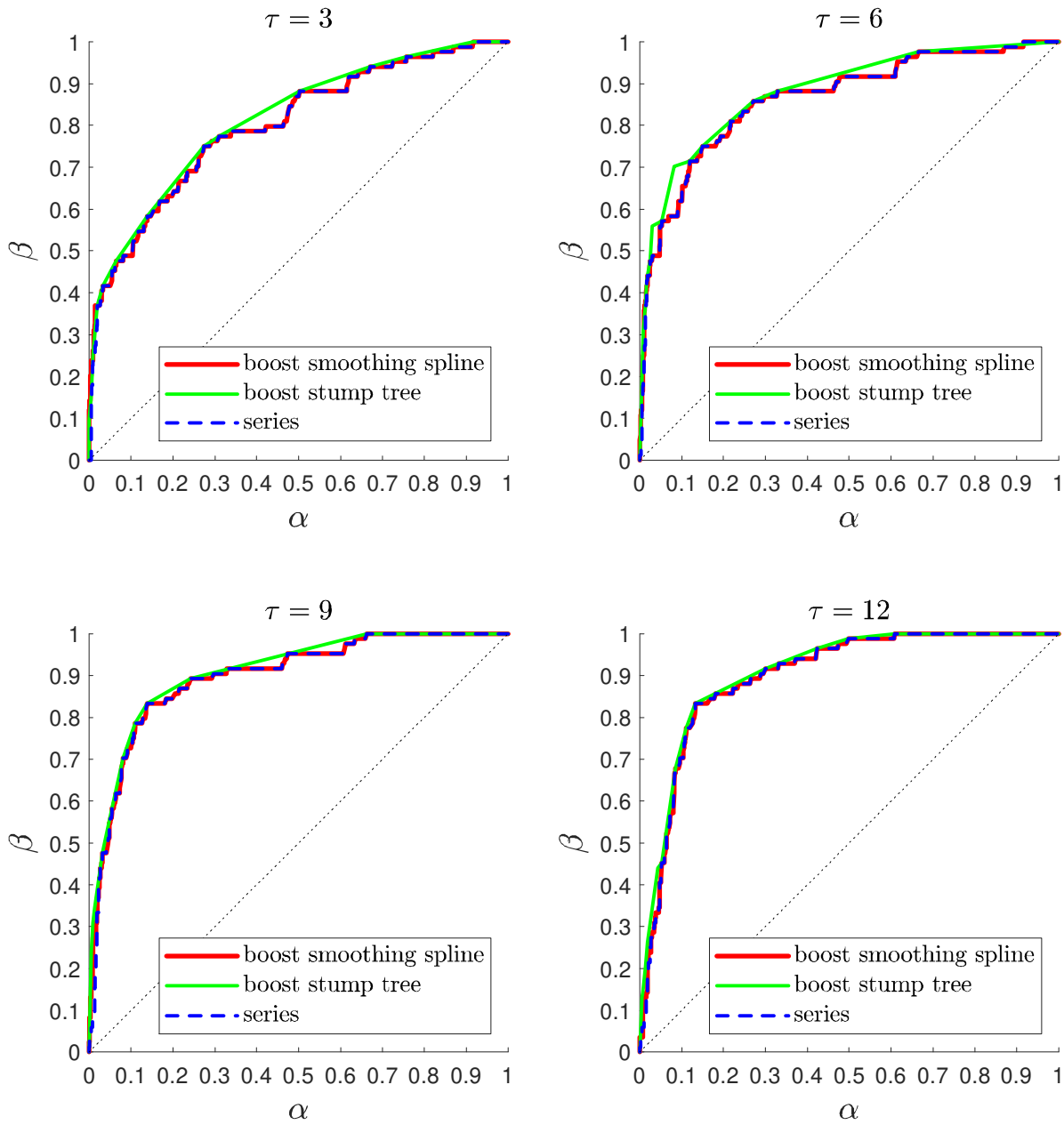


Figure 2: ROC Curves for Spread

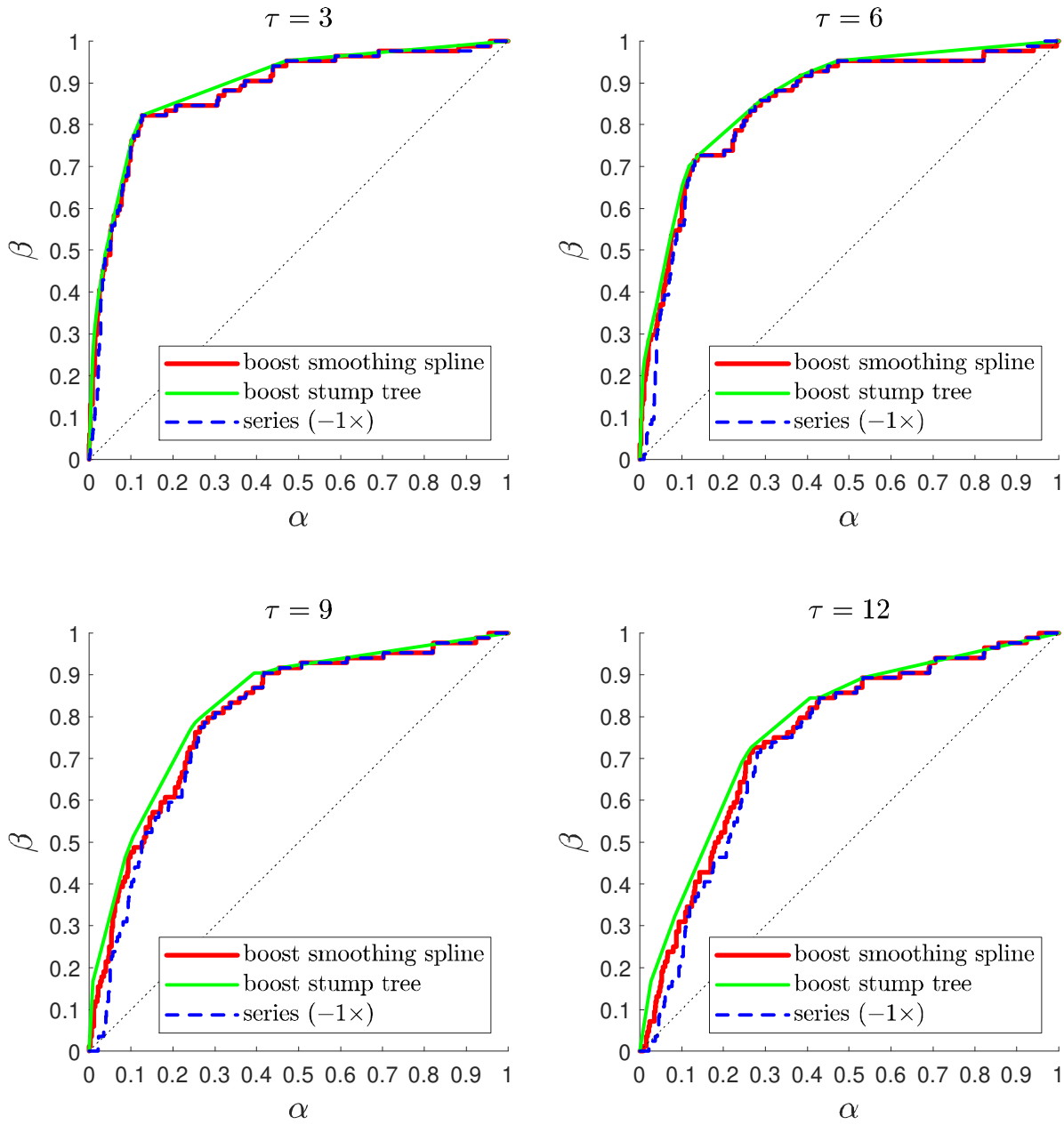


Figure 3: ROC Curves for Building Permits (Annual Percentage Change)

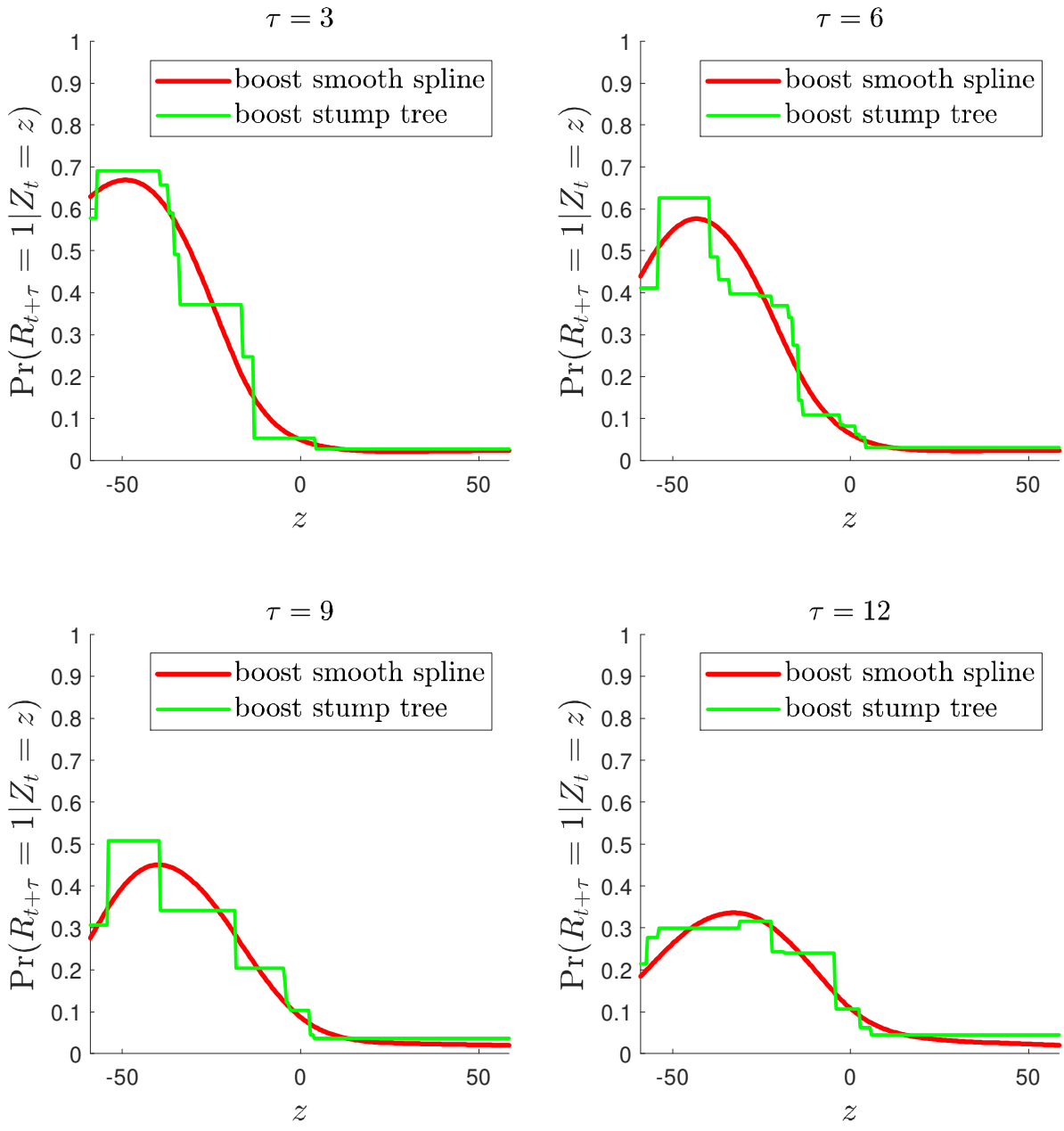


Figure 4: Probability of Recession Conditional on Building Permits (Annual Percentage Change)

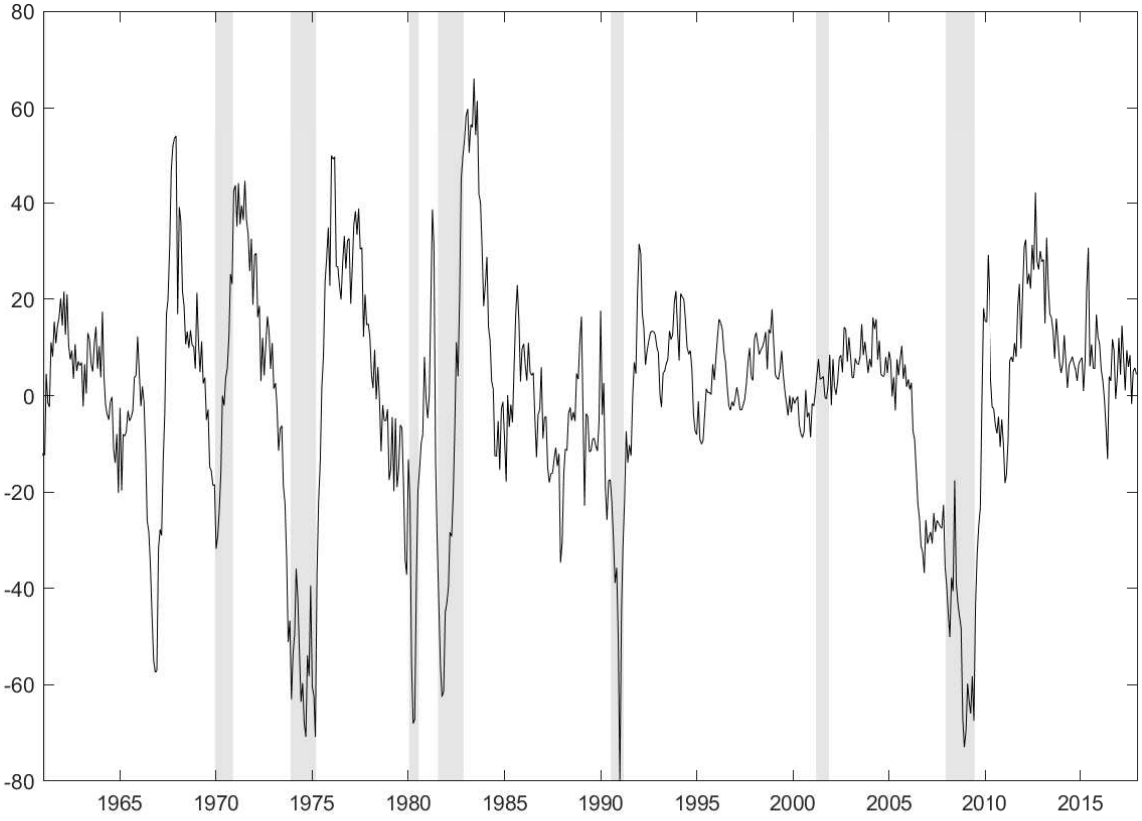


Figure 5: Building Permits (Annual Percentage Change), NBER recession months indiced by shaded area.

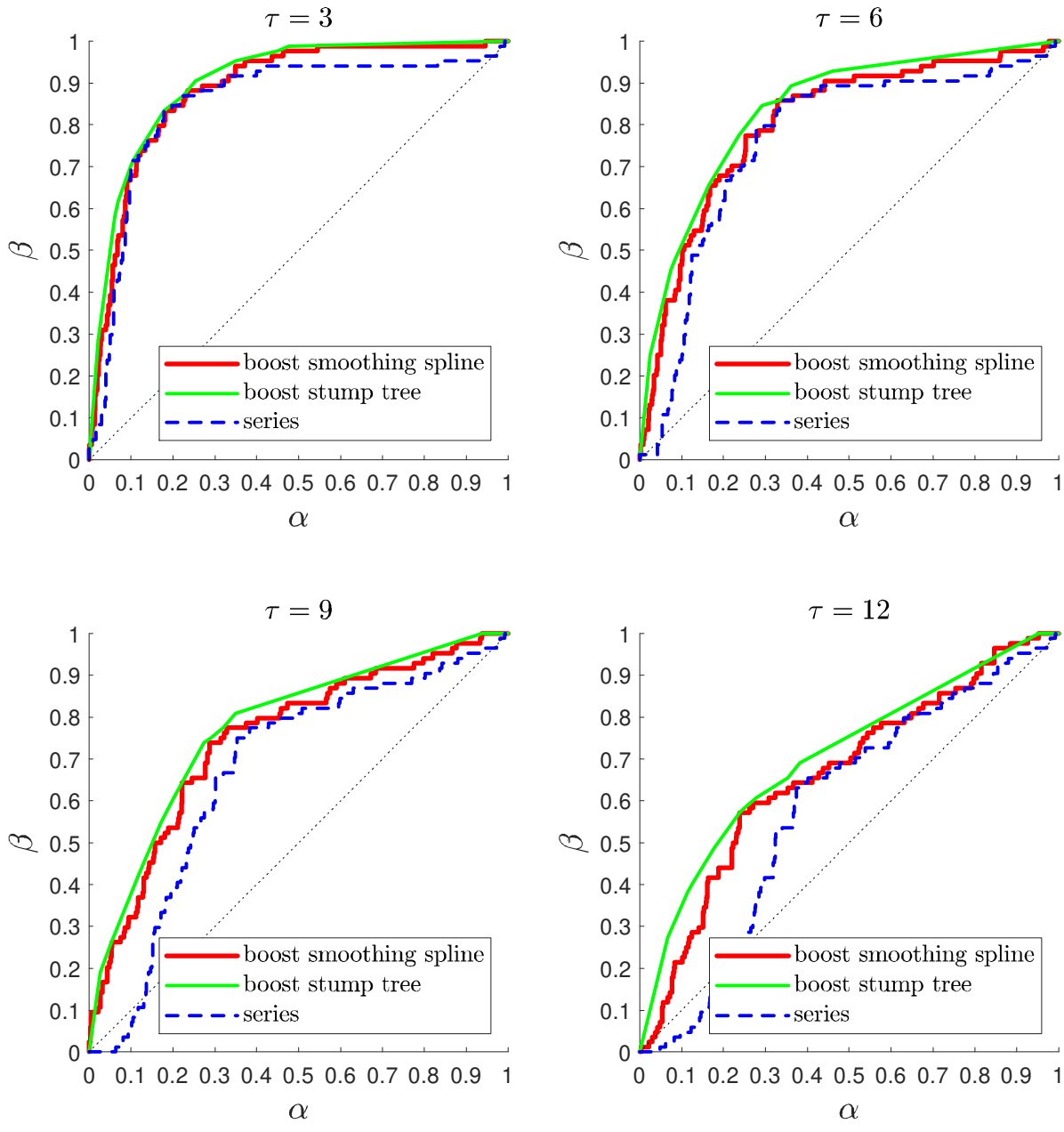


Figure 6: ROC Curves for Weekly Initial Claims (Annual Percentage Change)

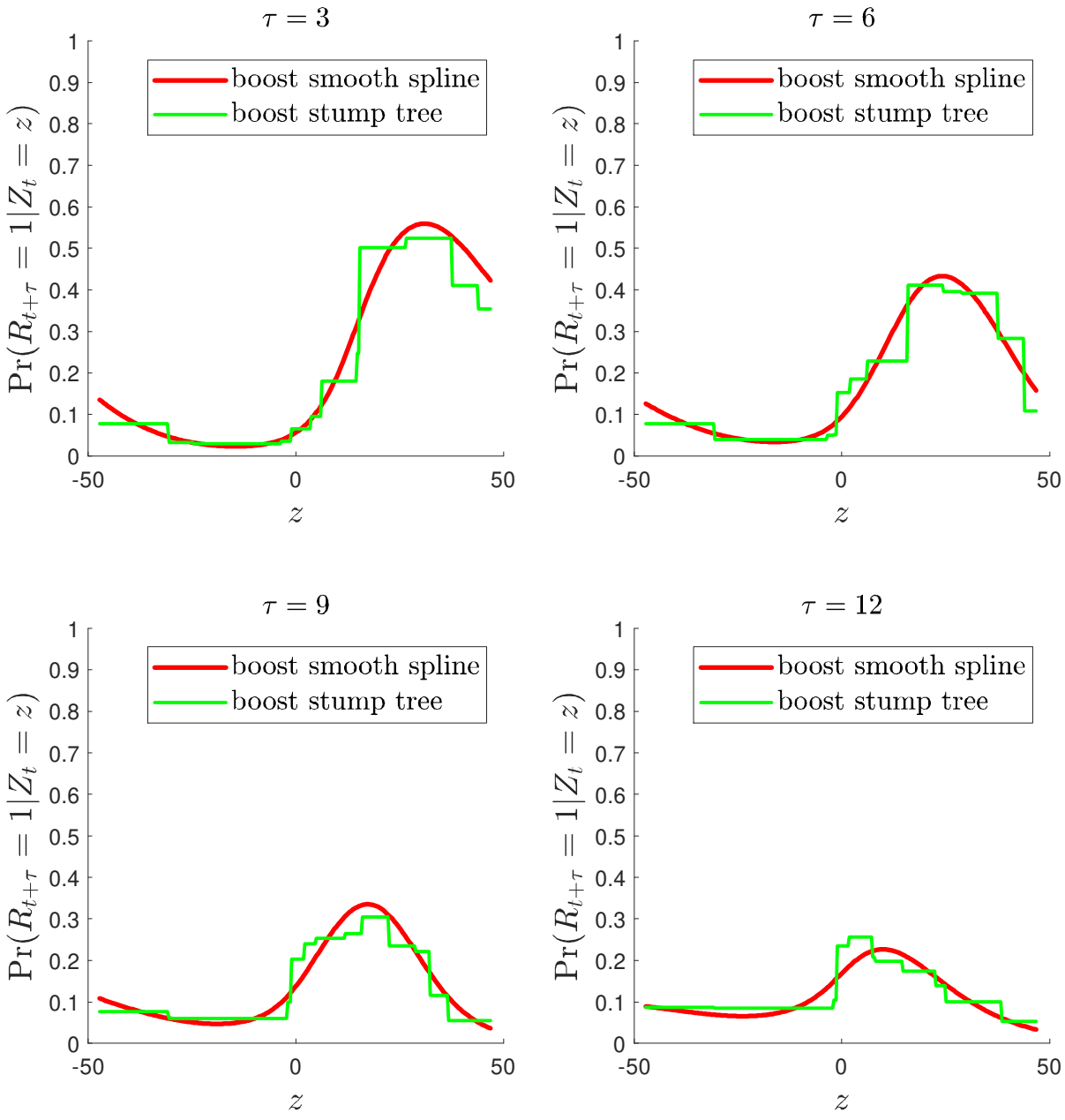


Figure 7: Probability of Recession Conditional on Weekly Initial Claims (Annual Percentage Change)

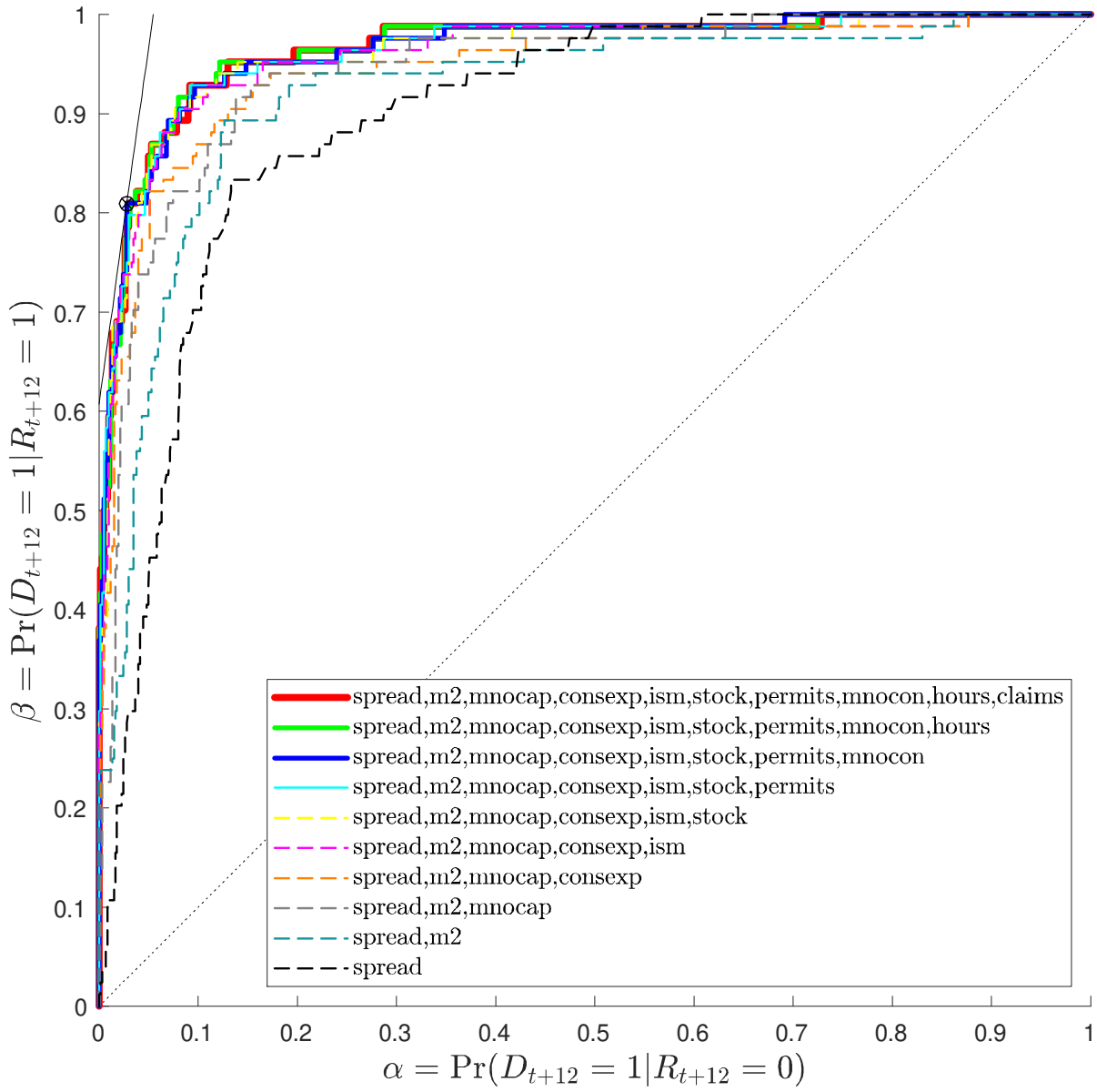
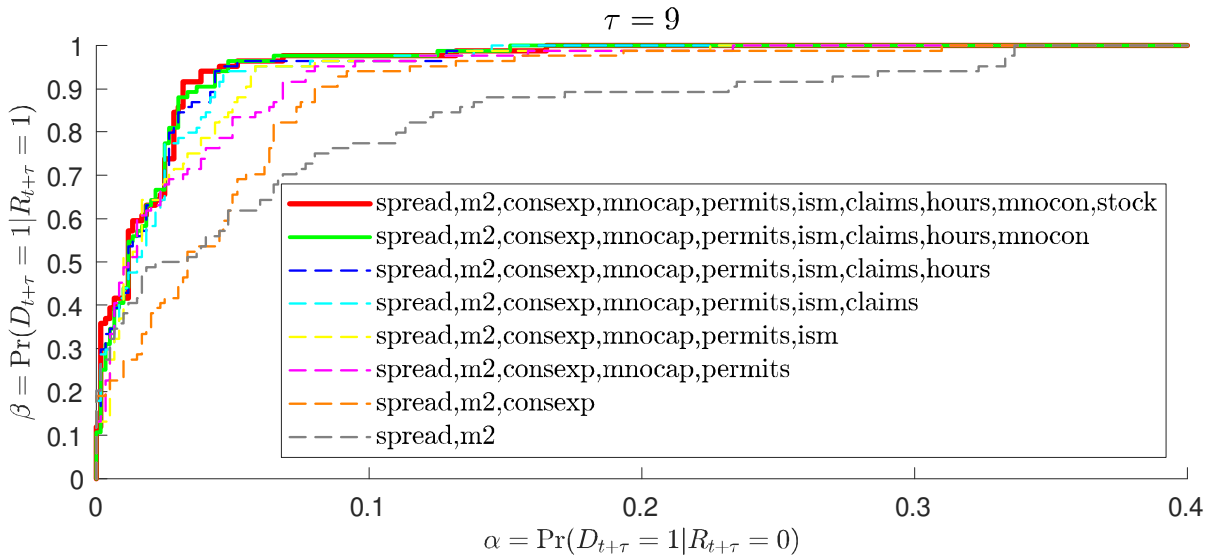
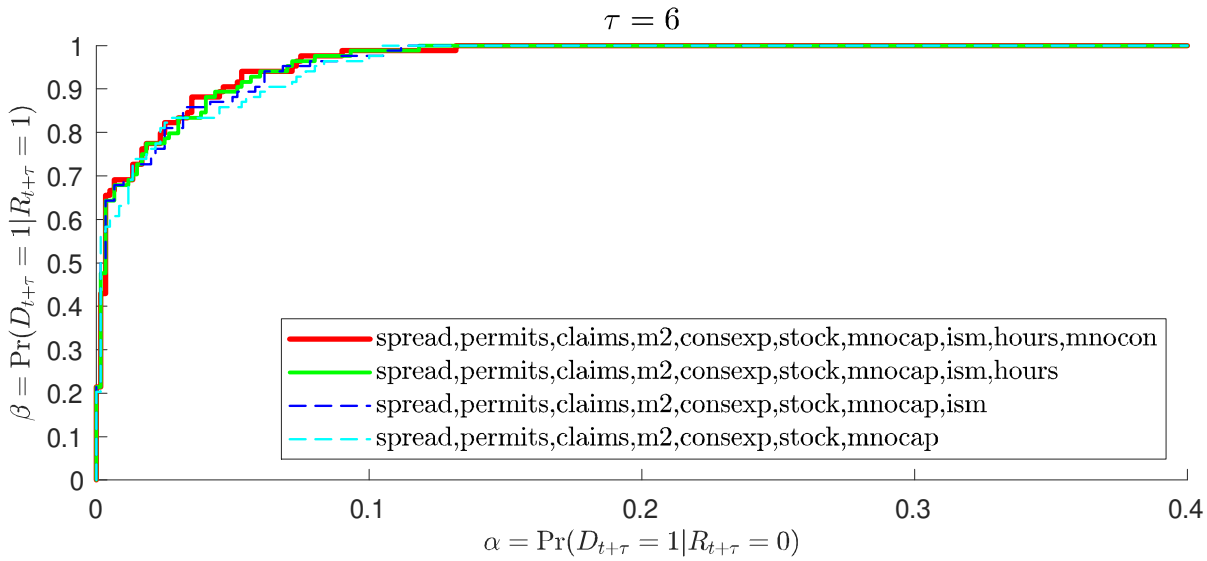


Figure 8: ROC Curves for 12 Month Ahead Models with Alternative Predictor Sets





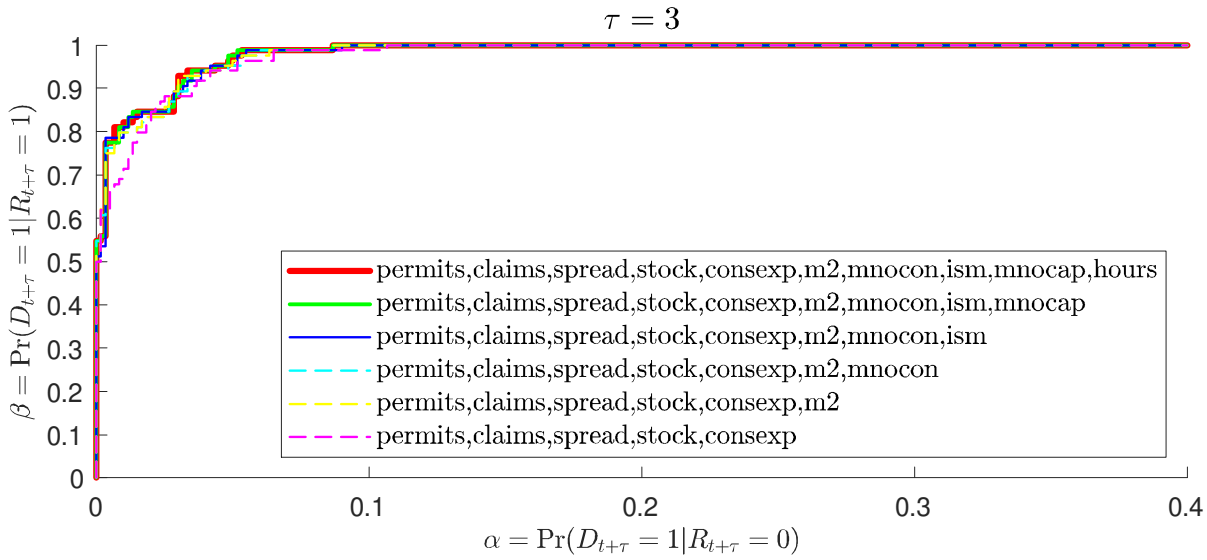


Figure 9: ROC Curves for 9, 6 and 3 Month Ahead Models

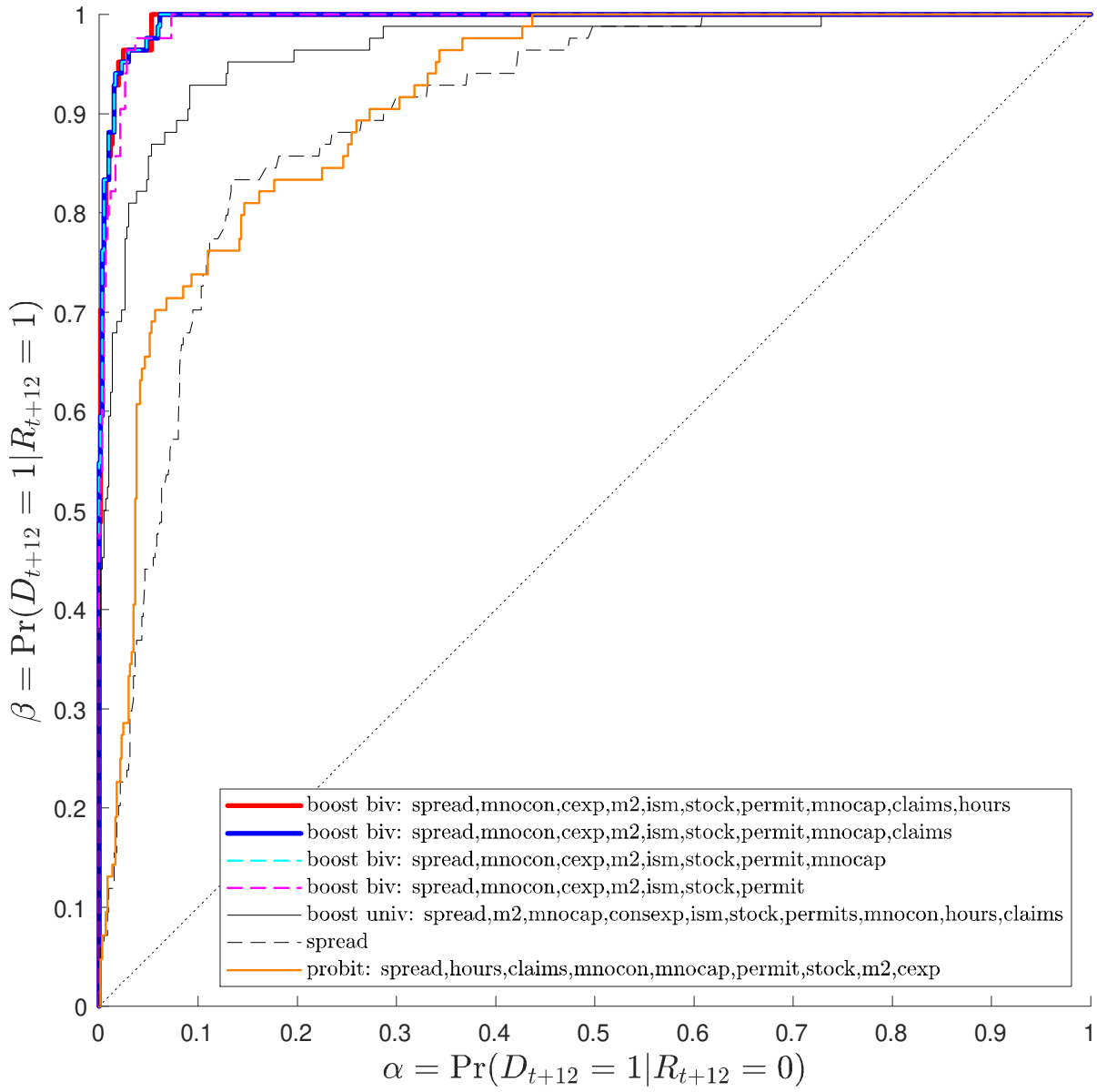


Figure 10: ROC Curves for 12 Month Ahead Models

The **Aboa Centre for Economics (ACE)** is a joint initiative of the economics departments of the Turku School of Economics at the University of Turku and the School of Business and Economics at Åbo Akademi University. ACE was founded in 1998. The aim of the Centre is to coordinate research and education related to economics.

Contact information: Aboa Centre for Economics, Department of Economics, Rehtorinpellonkatu 3, FI-20500 Turku, Finland.

www.ace-economics.fi

ISSN 1796-3133