CESifo WORKING PAPERS

10449
2023

May 2023

ROC and PRC Approaches to Evaluate Recession Forecasts

Kajal Lahiri, Cheng Yang

CESifo

Impressum:

CESifo Working Papers
ISSN 2364-1428 (electronic version)
Publisher and distributor: Munich Society for the Promotion of Economic Research - CESifo GmbH
The international platform of Ludwigs-Maximilians University's Center for Economic Studies and the ifo Institute
Poschingerstr. 5, 81679 Munich, Germany
Telephone +49 (0)89 2180-2740, Telefax +49 (0)89 2180-17845, email office@cesifo.de
Editor: Clemens Fuest
https://www.cesifo.org/en/wp
An electronic version of the paper may be downloaded
· from the SSRN website: www.SSRN.com
· from the RePEc website: www.RePEc.org
· from the CESifo website: https://www.cesifo.org/en/wp

CESifo *Working Paper No. 10449*

ROC and PRC Approaches to Evaluate Recession Forecasts

Abstract

We have studied the relationship between Receiver Operating Characteristics (ROC) and Precision-Recall Curve (PRC) both analytically and using a real-life empirical example of yield spread as a predictor of recessions. We show that false alarm rate in ROC and inverted precision in PRC are analogous concepts, and their difference is determined by the interaction of sample imbalance and forecast bias. We found that in cases of severe class imbalance, the forecasts need to be adequately biased to mitigate the effect of imbalancedness. The mix of values of precision and recall over six sub-samples show that the predictive power of the spread has not deteriorated in recent decades, provided the optimum values of threshold are used. Using PRC, we quantify the extent to which ROC could be exaggerating the true predictive value of the yield curve in predicting recessions.

JEL-Codes: C180, C220, C250, C530, E170, E370, E470.

Keywords: ROC, PRC, recessions, yield spread, rare events forecasting.

Kajal Lahiri
Department of Economics
University of Albany: SUNY
Albany / NY / USA
klahiri@albany.edu

Cheng Yang
Li Anmin Institute of Economic Research
Faculty of Economics
Liaoning University / Shenyang / China
yangcheng8906@gmail.com

May 18, 2023

1 Introduction

The Receiver Operating Characteristic (ROC) curve, originally used by military radar receivers during the second World War, was further developed in medical diagnostics in the 1970's and since then has become the most common binary assessment tool worldwide in almost any scientific field. In economic forecasting, even though a late comer, ROC has now become a standard evaluation metric over the last few years. Given a continuous predictive index, a decision threshold (or cut off) determines whether a particular value of the index belongs to a positive (recession) or a negative (non-recession) state, and by varying the cut off value over the complete range of the predictive index the ROC defines the Pareto front of attainable true and false positive predictions as a convex hull. The area under the ROC curve (AUROC) is commonly reported as a scalar measure of the overall predictive skill and is related to the Mann-Whitney U statistic. AUROC gives the probability of a randomly chosen positive sample observation to be associated with a higher predictive index than a randomly chosen negative sample. Compared to conventional accuracy measures like the concordance index or Brier's score, Fawcett (2006) attributes to a number of well understood properties and inherent flexibility of the ROC approach for its popularity.[1] Facing the trade-off in multi-object optimization, ROC enables the user to delay the trade off as long as possible in order to satisfy the relative net utilities associate with alternative outcomes.

However, in recent years many authors from diverse fields including ecology, bioinformatics, information retrieval and subfields of machine learning have found that ROC analysis tends to overstate the true predictive value of a classifier in predicting rare or uncommon events in which the data sets are highly imbalanced, see for example, Davis and Goadrich (2006), Ozenn et al. (2015), Saito and Rehmsmeier (2015) and Sofaer et al. (2019). In economics the typical binary events like recessions, bank failures, and unemployment are almost always relatively rare. The intuitive reason for ROC's failure is because AUROC can be interpreted as a measure of overlap between the two conditional distributions of the predictive index in the positive vs the negative state. If the two conditional distributions have a lot of overlap, it will be very difficult to discriminate between the two classes. If the object of interest has a low prevalence rate of say 5%, it is possible that the right tail of the distribution of the dominant event can overlap most of the distribution of the rare event. The performance metrics under ROC get overwhelmed by the success in predicting the dominant event, even when the rare object interest is not predicted well. Additionally, Yang et al. (2023) have shown that the weight distribution over

[1] Currently, Google Scholar lists over a million scientific articles under "ROC curve".

thresholds implicitly used in the calculation of AUROC is hard to justify from a decision theoretic prospective.

Under the rare event scenarios, an alternative approach of Precision-Recall Curve (PRC) is to ignore the true negatives altogether and focus on true positive rate (i.e., recall) and the success rate of positive predictions (i.e. precision). PRC only concentrates on the correct predictions of the minority class. Motivated by AUROC, the area under PRC (AUPRC) has also been suggested that supposedly presents a more nuanced view of the overall predictive power of a predictive index in identifying the event of interest in imbalanced data sets.

Most of the studies that considered ROC or PRC used them independently as distinct alternatives without trying to connect them. The main purpose of our paper is to compare the ROC and PRC approaches to understand under what circumstances ROC approach tends to overstate the predictive power of an index. We try to synthesize the two approaches and show the conditions on the trade off between biasedness of the forecasts and the balancedness of the sample under which the two approaches would converge. Since the latter is given for a sample, the biasedness of forecasts becomes the key choice parameter, which in turn is determined by the choice of the decision threshold. We show that it is the forecast user's relative net utility of making correct predictions of the rare event compared with that of the dominant event that determines the final choice of the threshold, and how much the forecasts need to be biased. Certain threshold-dependent measures are widely used in the PRC framework, whereas Kuipers score is commonly used in the ROC framework as a predictive score. Applicable in both frameworks, Matthews (1975) correlation coefficient (MCC), is another measure that has been recommended in recent years. We have established certain relations between these diverse measures.

In order to illustrate the differences in the ROC and PRC approaches, we use daily interest rate spread as the predictive index to forecast U.S. recessions over 1962-2021. We could characterize the overoptimism in ROC evaluation compared with PRC. Most of the research on the predictive power of yield spread is based on probabilities estimated from econometric models.[2] By using the raw data directly, we could avoid the problems of structural breaks, instability and model mis-specifications, and focus only on the differences in the two approaches in predicting recessions. In addition, Lieli and Hsu (2019) have shown that the tests based on the estimated probabilities would induce severe size distortions. Using our empirical strategy and both ROC and PRC

[2]See Choi et al. (2023) for a recent reference.

approaches, we find no evidence that yield spread has deteriorated in its capacity to predict recessions in the U.S. provided the appropriate threshold values are chosen. However, for a reasonably high hit rate (recall) of around 90%, the precision associated with PRC is so low that the use of yield spread as a predictor of recessions becomes rather unattractive. This explains the unresolved puzzle first pointed out by Rudebusch and Williams (2009) regarding the apparently irrational behavior of the professional forecasters in not using the information embedded in yield spread despite the overwhelming evidence of its predictive power provided by usual probit regressions, cf. Estrella and Mishkin (1996) and Lahiri et al. (2013). These regressions do not adequately evaluate the predictive power for recessions without being affected by the superior predictive success during non-recessions. Also, most of the thresholds that give reasonable recession forecasts are usually associated with biased forecasts, which contributes to the wedge between ROC and PRC and the discordance between different performance measures used in ROC and PRC approaches.

Rest of the paper is organized as follows: In section 2, we introduce the forecast evaluation methods and discuss the relationship between different measures. In section 3, we use daily yield spread as the predictor to forecast recessions, and evaluate the forecasts with ROC, PRC and different measures. Section 4 checks the robustness with other data frequencies and forecast horizons. Finally, conclusions are summarized in section 5.

2 Forecast Evaluation Methods: ROC vs PRC

Let y_t be the object to be forecasted at time t. Given a particular threshold δ and corresponding forecasts $\hat{y}_t$, there are four possibilities for the joint distribution of the binary forecasts and outcomes: (a) forecast is positive (recession) and outcome is positive ($\hat{y}_t = 1, y_t = 1$); (b) forecast is negative (not recession) and outcome is positive ($\hat{y}_t = 0, y_t = 1$); (c) forecast is positive and outcome is negative ($\hat{y}_t = 1, y_t = 0$); and (d) forecast is negative and outcome is negative ($\hat{y}_t = 0, y_t = 0$). These four forecast outcomes are called true positive (TP), false negative (FN), false positive (FP) and true negative (TN). The four cases are summarized in a 2×2 contingency Table 1, often called a Confusion matrix.

We define recall R as $R = TP/(TP+FN)$, which is the proportion of actual recession periods ($y_t = 1$) that are correctly predicted. Recall is called hit rate or true positive rate (TPR) in ROC parlance. False alarm rate is denoted as FA, which is the proportion of non-recession periods ($y_t = 0$) that are mistakenly predicted to be recessions ($\hat{y}_t = 1$), that is $FA = FP/(FP + TN)$. FA

is also called false positive rate, or one minus specificity: $1 - TN/(FP + TN)$. Precision P, a key concept in PRC, is defined as $P = TP/(TP + FP)$, which is the proportion of correct hits out of all the positive (recessionary) forecasts. The conditionality is interchanged in the definitions of R and P. The Confusion matrix for the four possible outcomes and the definitions of R, P and FA are listed in Table 1. Note that since δ can change, R, FA and P are functions of threshold δ. Thus we denote them as $R(\delta)$, $FA(\delta)$ and $P(\delta)$ from now on.

Table 1: Confusion Matrix with Definitions of Precision, Recall, and False Alarm Rate

	recession ($y_t = 1$)	no recession ($y_t = 0$)	
predict recession ($\hat{y}_t = 1$)	TP	FP	$P = \frac{TP}{TP+FP}$
predict no recession ($\hat{y}_t = 0$)	FN	TN	
	$R = \frac{TP}{TP+FN}$	$FA = \frac{FP}{FP+TN}$	

By plotting $FA(\delta)$ against $R(\delta)$ for all possible spread thresholds δ, we have the ROC curve. The ROC curve is monotonically increasing from left to right, and gives a complete summary of the trade-offs between hit rate and false alarm rate. Better forecasts have ROC curves closer to the upper-left corner since upper-left points represent low levels of false alarm and high levels of recall. Points on the 45° line represent thresholds whose forecasts are same as random guesses. Points below the 45° line represent thresholds whose forecasts are worse than random guesses. In case of recession forecasts, ROC curve evaluates how well a forecasting system performs during both recession and expansion periods, and has been used extensively in existing literature. The most commonly used measure of forecast performance in terms of ROC is AUROC, which can be computed as

$$AUROC = \frac{1}{2}\sum_{j=1}^{M}(R(\delta_j) + R(\delta_{j-1}))(FA(\delta_j) + FA(\delta_{j-1})), \tag{1}$$

where M is the number of thresholds. AUROC evaluates the overall performance across all thresholds. It gives the probability that a randomly selected observation from the recession periods will have a higher score than a randomly selected observation from the non-recession periods, and is independent of the prevalence rate. Kuipers score (KS), originally proposed by Peirce (1884), is a threshold-dependent measure related to ROC. KS is defined as

$$KS(\delta) = R(\delta) - FA(\delta). \tag{2}$$

KS puts equal weights on recall and false alarm rate. Larger KS implies better forecast performance. Further, we also explore the maximized values of KS

$$\max_{\delta} KS(\delta), \tag{3}$$

and the optimal thresholds δ that maximize KS. The reason is practitioners are not necessarily interested in all of points on the ROC curve. For instance, people may not be interested in points with extremely low recalls, where almost no recessions are hit. Additionally, we also compute average Kuipers score (AKS) suggested by Yang et al. (2023), who proposed global measures of the degree of conformity based on expected utility of a forecast user. AKS is defined as

$$AKS = \frac{1}{M}\sum_{j=1}^{M} KS(\delta_j), \tag{4}$$

where $a = \delta_0 < \delta_1 < \cdots < \delta_M = b$, a and b are lower and upper bounds of the spread, and M is the number of thresholds. The measure is average over δ.

By plotting $R(\delta)$ on the horizontal axis and $P(\delta)$ on the vertical axis for all possible spread thresholds δ, we have the PRC. Each point on the PRC provides an analyst important information about a threshold value: the fraction of observations predicted to be in recessions and the fraction of this positive predictions that are actually in recessions. PRC is not necessarily monotonic or uniformly convex. Since higher $R(\delta)$ and $P(\delta)$ are preferred, better forecasts have PRC closer to the upper-right corner. Rather than a 45° line, the baseline for PRC is a horizontal line with the value of the fraction of positives. Similar to AUROC, the area under PRC (AUPRC) can be computed to evaluate the overall performance. The computation of AUPRC needs nonlinear interpolation if there are points far away from each other. Davis and Goadrich (2006) discussed how to deal with this issue: A local skew as a function of TP and FP of point A and point B (potentially far away from each other) is defined first; Then new points of TP between A and B are created; Finally, precision, as a function of TP and FP, is computed by linearly increasing FP for each new point. AUPRC can be interpreted as expected precision obtained over uniformly varying recall. Many programs such as the R package PRROC (Keilwagen et al. (2014) and Grau et al. (2015)) can perform this task. As we are using daily data in our empirical application, the data points are very close to each other. As a result, the PRC and AUPRC remain almost the same when different interpolation and area-computation techniques are applied.

Unlike ROC, PRC focuses only on periods in which recessions are coming or periods in which recessions are signaled by a predictor, which are directly

relevant for practitioners. Further, PRC is more informative for imbalanced datasets (Saito and Rehmsmeier (2015)), which is true for us since recession is a rare event.

Since the distinguishing feature of PRC is its focus on precision in place of false alarm rate in ROC, we derive the relationship between FA and P below. With the forecasting rule being $\hat{y}_t(x_t;\delta) = \mathbf{1}(x_t \leq \delta)$,[3] precision, false alarm rate and recall (i.e., hit rate) are given by

$$P(\delta) = Prob(y_t = 1|x_t \leq \delta),$$

$$FA(\delta) = Prob(x_t \leq \delta|y_t = 0),$$

and

$$R(\delta) = Prob(x_t \leq \delta|y_t = 1).$$

False alarm rate can be written as

$$\begin{aligned} FA(\delta) &= \frac{Prob(x_t \leq \delta, y_t = 0)}{Prob(y_t = 0)} \\ &= \frac{Prob(y_t = 0|x_t \leq \delta)Prob(x_t \leq \delta)}{Prob(y_t = 0)} \\ &= \frac{(1 - Prob(y_t = 1|x_t \leq \delta))Prob(x_t \leq \delta)}{Prob(y_t = 0)} \\ &= \frac{(1 - P(\delta))Prob(\hat{y}_t = 1)}{1 - Prob(y_t = 1)} \\ &= \frac{(1 - P(\delta))\mu_{\hat{y}(\delta)}}{1 - \mu_y}, \end{aligned} \tag{5}$$

where $\mu_{\hat{y}(\delta)}$ is the proportion of recessionary forecasts, which is a function of threshold, and μ_y is the proportion of periods with $y_t = 1$. It can be seen that $FA(\delta)$ and $1 - P(\delta)$ are positively related through $\mu_{\hat{y}(\delta)}/(1 - \mu_y)$, which is the ratio of the proportion of positive forecasts to the proportion of actual negative cases on ground.

Since $\mu_{\hat{y}(\delta)} = (TP(\delta) + FP(\delta))/(TP(\delta) + FP(\delta) + FN(\delta) + TN(\delta))$ and $1/(1-\mu_y) = (TP(\delta)+FP(\delta)+FN(\delta)+TN(\delta))/(FP(\delta)+TN(\delta))$, $\mu_{\hat{y}(\delta)}/(1-\mu_y)$

[3]The forecasting rule can be $\hat{y}_t(x_t;\delta) = \mathbf{1}(x_t \geq \delta)$ if x_t is probability forecast. We define the rule as $\mathbf{1}(x_t \leq \delta)$ here since we will directly use the value of the yield spread as the threshold later in our empirical application.

in Equation (5) can be written as

$$\begin{aligned}\frac{\mu_{\hat{y}(\delta)}}{1-\mu_y} &= \frac{TP(\delta)+FP(\delta)}{FP(\delta)+TN(\delta)} \\ &= \frac{TP(\delta)+FP(\delta)}{FP(\delta)+TN(\delta)} \times \frac{TP(\delta)+FN(\delta)}{TP(\delta)+FN(\delta)} \times \frac{TP(\delta)}{TP(\delta)} \\ &= \frac{rR(\delta)}{P(\delta)},\end{aligned} \tag{6}$$

where $r = (TP(\delta) + FN(\delta))/(FP(\delta) + TN(\delta))$ is the ratio of the number of positive cases to the number of negative cases, which measures how imbalanced the data is. Thus Equation (5) is equivalent to

$$FA(\delta) = \frac{rR(\delta)}{P(\delta)}(1 - P(\delta)), \tag{7}$$

which is equivalent to Equation (2) in Williams (2021), who showed the role of sample imbalance on PRC. Again, $FA(\delta)$ and $1 - P(\delta)$ are linked through the ratio $rR(\delta)/P(\delta)$. Note that $\mu_{\hat{y}(\delta)}/(1 - \mu_y)$ can also be written as

$$\begin{aligned}\frac{\mu_{\hat{y}(\delta)}}{1-\mu_y} &= \frac{TP(\delta)+FP(\delta)}{FP(\delta)+TN(\delta)} \times \frac{TP(\delta)+FN(\delta)}{TP(\delta)+FN(\delta)} \\ &= r\frac{TP(\delta)+FP(\delta)}{TP(\delta)+FN(\delta)} \\ &= r\frac{\mu_{\hat{y}(\delta)}}{\mu_y}.\end{aligned} \tag{8}$$

$\mu_{\hat{y}(\delta)}/\mu_y$ is a measure of the biasedness of the forecasts. Thus the link between $FA(\delta)$ and $1 - P(\delta)$ is determined jointly by the balancedness of the sample and the biasedness of forecasts. With imbalanced data when r is sufficiently small, $FA(\delta)$ tends to be lower than $1 - P(\delta)$ as long as the forecasts are not extremely upward biased. Further, combining with Table 1 we can see $FA(\delta)$ and $1 - P(\delta)$ are equal if $TP = TN$ (there are same number of true positives and true negatives). Also, $1 - P(\delta)$ can be thought of as the false alarm rate $\tilde{FA}(\delta)$ with TN replaced by TP in Table 1.

Since KS is extensively used in ROC analysis and R and P in PRC, it will be interesting to see their relation. Plugging Equation (7) in Equation (2), $KS(\delta)$ can be written as a function of $P(\delta)$, $R(\delta)$ and r:

$$KS(\delta) = R(\delta) - \frac{rR(\delta)}{P(\delta)}(1 - P(\delta)). \tag{9}$$

The second term in Equation (9) is the false alarm rate $FA(\delta)$ as defined before.

For evaluating PRC, F statistics are commonly used threshold-dependent measures, which are harmonic means of $P(\delta)$ and $R(\delta)$ and are defined as

$$F_\beta(\delta) = \frac{(1+\beta^2)P(\delta)R(\delta)}{\beta^2 P(\delta) + R(\delta)}, \tag{10}$$

where β defines the relative importance of recall over precision; see van Rijsbergen (1979) and Chinchor and Sundheim (1993). Using Equation (7), $F_\beta(\delta)$ can be written as

$$F_\beta(\delta) = \frac{(1+\beta^2)R(\delta)}{R(\delta) + \frac{1}{r}FA(\delta) + \beta^2}, \tag{11}$$

which depends positively on class imbalance r. Thus everything else unchanged, data with lower fraction of positive cases tends to give worse forecast performance evaluation in terms of F measures. Following the literature, we will consider $F_{0.5}$, F_1 and F_2 in Equation (10):

$$F_{0.5}(\delta) = \frac{1.5P(\delta)R(\delta)}{0.25P(\delta) + R(\delta)}, \tag{12}$$

$$F_1(\delta) = \frac{2P(\delta)R(\delta)}{P(\delta) + R(\delta)}, \tag{13}$$

$$F_2(\delta) = \frac{5P(\delta)R(\delta)}{4P(\delta) + R(\delta)}. \tag{14}$$

Higher values of $F_{0.5}$, F_1 and F_2 would suggest better forecast performance. The F measures originated in statistical ecology, cf. Dice (1945) and Sorensen (1948). Chinchor (1992) and Chinchor and Sundheim (1993) gave the final notations that have been used in recent literature. Note that Chicco and Jurman (2020) criticized F_1 measure as it may provide misleading information about the overall forecast performance when a prediction has many true positives but not enough true negatives (or reverse), since F measures like PRC are independent from true negatives (TN).[4] We still choose to report F measures since 1) they are integral part of PRC methodology and use exclusively $P(\delta)$ and $R(\delta)$; and 2) in our particular application to recession forecasting, correctly predicting recessions is usually more important than correctly predicting expansions. Similar to maximizing KS, F measure can also be maximized ($\max_\delta F_{0.5}(\delta)$, $\max_\delta F_1(\delta)$ and $\max_\delta F_2(\delta)$).

Finally, a less well-known Matthews correlation coefficient (MCC), which uses all four elements of the Confusion matrix including TN, is also considered:

$$MCC(\delta) = \frac{TP(\delta)TN(\delta) - FP(\delta)FN(\delta)}{\sqrt{(TP(\delta)+FP(\delta))(TP(\delta)+FN(\delta))(TN(\delta)+FP(\delta))(TN(\delta)+FN(\delta))}}. \tag{15}$$

[4] See also Flach and Kull (2015) and Yedidia (2016).

The value of MCC ranges from -1 to 1, the same range as the range of KS. Larger MCC represents better forecast performance, and the MCC of no-skill forecasts is 0. Chicco and Jurman (2020) argued that MCC is a better measure than F_1 as it considers all the four forecast outcomes in Table 1 and invariant to data swapping. It generates high score only if a classifier scores high values in all four basic rates - hit rate, false alarm rate, precision and negative predictive value defined as $TN(\delta)/(TN(\delta)+FN(\delta))$. Chicco and Jurman (2023) also argued for MCC in place of ROC. However, if forecast user wants to put more weight on positive outcomes compared to the negative ones, then MCC may not be the most appropriate evaluation statistic.

Using the definitions of μ_y and $\mu_{\hat{y}}$, $TP+FP=\mu_{\hat{y}}(TP+TN+FP+FN)$, $TP+FN=\mu_y(TP+TN+FP+FN)$, $TN+FP=(1-\mu_y)(TP+TN+FP+FN)$ and $TN+FN=(1-\mu_{\hat{y}})(TP+TN+FP+FN)$, the relationship between MCC and KS can be derived in the following steps (dropping δ temporarily):

$$\begin{aligned}
MCC &= \frac{(TP\times TN - FP\times FN)}{\sqrt{\mu_y\mu_{\hat{y}}(1-\mu_y)(1-\mu_{\hat{y}})}(TP+TN+FP+FN)^2} \\
&= \frac{(\frac{TP}{TP+FN}TN(TP+FN) - \frac{FP}{FP+TN}FN(FP+TN))}{\sqrt{\mu_y\mu_{\hat{y}}(1-\mu_y)(1-\mu_{\hat{y}})}(TP+TN+FP+FN)^2} \\
&= \frac{(R\frac{TP}{FP+TN} - FA\frac{FP}{TP+FN})\frac{(TP+FN)(FP+TN)}{(TP+TN+FP+FN)^2}}{\sqrt{\mu_y\mu_{\hat{y}}(1-\mu_y)(1-\mu_{\hat{y}})}} \\
&= \frac{(R(1-FA)-FA(1-R))\mu_y(1-\mu_y)}{\sqrt{\mu_y\mu_{\hat{y}}(1-\mu_y)(1-\mu_{\hat{y}})}} \\
&= \Bigg(R(1-FA)-FA(1-R)\Bigg)\sqrt{\frac{\mu_y(1-\mu_y)}{\mu_{\hat{y}}(1-\mu_{\hat{y}})}} \\
&= \Bigg(R-FA\Bigg)\sqrt{\frac{\mu_y(1-\mu_y)}{\mu_{\hat{y}}(1-\mu_{\hat{y}})}} \\
&= KS\sqrt{\frac{\mu_y(1-\mu_y)}{\mu_{\hat{y}}(1-\mu_{\hat{y}})}}.
\end{aligned} \tag{16}$$

Putting δ back,

$$MCC(\delta) = KS(\delta)\sqrt{\frac{\mu_y(1-\mu_y)}{\mu_{\hat{y}(\delta)}(1-\mu_{\hat{y}(\delta)})}}. \tag{17}$$

Thus, MCC can be written in terms of KS, scaled by the square root of ratio of the standard deviation of the actual values y to that of the binary predictions. This ratio can also be treated as a function of forecast biasedness. MCC and KS are equal if the forecasts are unbiased ($\mu_{\hat{y}(\delta)}=\mu_y$). They can also be equal

under a less natural condition if $\mu_{\hat{y}(\delta)} = 1 - \mu_y$. In case of rare event forecasts, μ_y is small and the variance of y also tends to be small. If correctly forecasting positive cases is important, people may tend to make more positive forecasts and get a relatively larger $\mu_{\hat{y}(\delta)}$ and larger $\mu_{\hat{y}(\delta)}(1-\mu_{\hat{y}(\delta)})$. As a result, it is more likely to get a ratio smaller than 1, hence an MCC smaller than KS.

3 The Empirical Set Up with yield spread

In this section we evaluate recession forecasts using ROC, PRC and different evaluation measures at the optimal thresholds. Saito and Rehmsmeier (2015) identified a large number of bio-informatics studies on rare biological events that used almost always ROC analysis, and found that these studies reached deceptive conclusions about the reliability of their classification methods. Lahiri and Yang (2022) examined the predictability of yield spread in recession forecasting using the ROC approach, but without any reference to the PRC approach. Following Saito and Rehmsmeier (2015) this section will compare and contrast the two approaches in the context of exactly the same empirical illustration and data for precise comparison.

3.1 The background

We use the term spread, which is the difference between the yields on the ten-year Treasury note and the three-month Treasury bill rate, to forecast U.S. recessions. The spread predicts recessions since it not only reflects current stance of monetary policy, but also its complex interactions with expected future monetary policies. These, in turn, are linked to expectations of future economy. Cooper et al. (2020) have provided a summary about how the spread and future recessions are related.

Regardless of the types of spread used, there has also been a parallel body of research on the hypothesis that the yield spread has lost its predictive power after the 1980s. The failure of the experimental recession index by Stock and Watson (1993) was attributed to their reliance on these spread variables at the 6-month horizon, see Dotsey (1998) and Jardet (2004). Many factors, including unconventional monetary policies, financial innovations, deregulation, deepening of the commercial paper market, increasing globalization, structural breaks, and inflation targeting, have been proposed as reasons for the loss in forecasting power, see Giacomini and Rossi (2006) and Pažický (2021). Chauvet and Potter (2005) used a number of models to accommodate some of the documented instabilities in yield curve prediction models. The yield spread is one of the

ten leading indicators of The Conference Board (TCB) since 1996, see Levanon et al. (2015). It has the highest AUROC among all other leading indicators and the value is close to that of the composite LEI, see Lahiri and Yang (2021).

During last few years there have been more than 100 articles published in two major forecasting journals that used ROC. However, since recession is a relatively rare event covering a little over 10% of the sample, the data is highly imbalanced. Saito and Rehmsmeier (2015) showed that with imbalanced datasets, PRC curve can be a more appropriate and informative tool. Since PRC curve does not count true negative forecasts, it is less likely to exaggerate or inflate the forecast performance when the data is imbalanced (Sofaer et al. (2019), Cook and Ramadas (2020)). Pinker (2018) also challenged the reliability of AUROC when relatively rare events are being predicted. Davis and Goadrich (2006) showed that when making comparison, best AUROC does not necessarily imply best AUPRC. Analyses based on PRC have been applied extensively to many different topical areas such as image detection (Liang et al. (2017), prediction of opioid overdose (Lo-Ciganic et al. (2019), crime prediction (Rummens and Hardyns (2021)), solar forecasting (Lin et al. (2023)) and many more. PRC is usually used with machine learning techniques. As of now, Google Scholar lists over half a million citations under it. However, not many studies on recession forecasts have yet adopted PRC approach. One exception is Pigini (2021), who emphasized the importance of PRC in the context of early warning systems using panel logit-based models. Vrontos et al. (2021) also used machine learning techniques to forecast recession, and used F_1 as one of the evaluation measures, but did not focus on PRC. Puglia and Tucker (2021) also recognized the usefulness of PRC but did not pursue the approach due to lack of positive cases in their sample.

The standard workhorse in this sphere has been the probit model, e.g., Estrella and Hardouvelis (1991) and Estrella and Mishkin (1996). The probit link function is symmetric and may not be consistent with the loss function of the forecast user. Similar to Berge and Jordà (2011), who used ROC, we additionally use PRC approach that directly focuses on the identification of a binary event, and evaluates the forecast performance in terms of precision and recall. The commonly used goodness-of-fit criteria such as mean squared forecast errors and Brier's score tend to get overwhelmed by the dominant event of the sample, see Stephenson (2000) and Lahiri and Yang (2013).

Like Berge and Jordà (2011) and Bauer and Mertens (2018), we translate the yields directly into binary predictions without going through an intermediate

step of estimating the probabilities from a probit or other models, which may cause problems. For example, Lieli and Hsu (2019) showed that traditional tests of AUROC can give misleading results if a regression model is used and parameters are estimated in-sample. Also, because the AUROC or AUPRC statistic is obtained by evaluating over the whole probability space, it may fail to identify the goodness of a predictor in the region where it matters, cf. Elliott and Lieli (2013) and Zhou et al. (2011). Note that AUROC and AUPRC are not very useful for issuing forecasts in real time, they are more like R-square statistic to gauge the over all fit of a linear regression. Thus we focus on threshold-specific composite measures under ROC and PRC, which can help determine whether the forecasting power of the yield curve has been falling over time subject to appropriate choice of a threshold. Additionally, since financial and market practitioners directly monitor the yield spreads and watch when the inversion of the yield curve occurs, our approach of generating PRC curves directly from the spread values will enable them to look at alternative threshold spread values, and directly interpret them without any intermediary assumption on the link function.

Finally, and importantly, we use daily data that has important advantages in our context. (1) Since the market analysts and policy makers monitor indicators like yield spreads on a continuous basis, daily data makes the forecasts more timely in real time. (2) Daily data provides more observations. This is important for us, since we focus on recessions that occur infrequently in the sample. (3) Since daily data is less discrete, the points on the Precision-Recall space are relatively dense and as a result we will suffer less from interpolation problem (Davis and Goadrich (2006)).

3.2 Forecasting Rule, Data and Results

Our simple forecasting rule makes a recessionary forecast if the spread value in day t falls to or below some threshold δ:

$$\hat{y}_t(x_t;\delta) = \mathbf{1}(x_t \leq \delta). \tag{18}$$

In our application, the left-hand-side variable y_t is an indicator that is equal to 1 if there is an onset of recession anytime during the next h months and 0 otherwise.[5] $\hat{y}_t(x_t;\delta)$ is a binary forecast for y_t using threshold δ. We take $h = 12$ months for our main results because this is the horizon at which the

[5]This forecast object does not pre-specify a fixed horizon and has been used in a number of recent papers, cf. Wright (2006), Ergungor (2016), Johansson and Meldrum (2018), Bauer and Mertens (2018), Ajello et al. (2022) and others. Also see Lahiri and Yang (2023) for a comparison of different forecast objects in recession forecasting.

spread has the maximum predictive power. Later we will also report results with respect to horizons at 18 and 6 months as robustness checks.

We use monthly NBER recession indicator as the basis for the left-hand-side outturn variable. It is turned into daily to match the frequency of the predictor. The data of the spread is daily 10-Year Treasury Constant Maturity Minus 3-Month Treasury Constant Maturity, spanning from January 2 1962 to November 30 2021. Bauer and Mertens (2018) show that the difference between the 10-year and 3-month rates is the most effective term spread to forecast recessions without any adjustments for term premium or quantitative easing.[6]

In order to simplify our analysis, we make the number of daily observations of the spread in each month fixed at 22. If there are fewer than 22 actual observations or there are holidays on weekdays in a month, we linearly interpolate them, treating them as missing values. If there are more than 22 values in a month, following Ghysels et al. (2020), we replace the last several observations in that month with their average such that every month has 22 values. In the end, the data is adjusted so that there are 22 observations in each month.[7]

The causes and mechanisms of recessions are different historically. In addition, optimal threshold for the spread variable can change due to trends in financial innovations in the economy. Therefore we work with different subsamples. We start every sample from January 1962 and add several years each time in such a way that one more recession is included in the sample that ends 12 months before the next recession begins. By doing this cumulatively and recursively, we are able to explore how the forecast performance and the optimal threshold changed over time before each recession individually. Specifically, we have six sub-samples: sample 1 (1/2/1962 - 7/31/1980), sample 2 (1/2/1962 - 7/31/1989), sample 3 (1/2/1962 to 3/31/2000), sample 4 (1/2/1962 - 12/29/2006, sample 5 (1/2/1962 - 2/28/2019), and finally sample 6 (1/2/1962 - 11/30/2021) which is full sample. Each of the first 5 sub-sample ends exactly 12 months before the beginning of the 1982, 1990, 2001, 2008 and the 2020 recessions.

We plot ROC curves for all six samples in Figure 1. All of the ROC curves look good, and the values of AUROC range from 0.866 to 0.913. Chicco and Jurman (2023) recommend an AUROC value of at least 0.785 to be considered

[6]Pažickỳ (2021) and Choi et al. (2023) reached the same conclusion on the predictive value of yield spread as we defined it.

[7]Our experiments show that the adjustments only account for a negligible fraction of the data, so the results are not affected.

as high. Figure 2 shows the PRC for all the corresponding samples together with AUPRC. It is found that the value of AUPRC is relatively low for sample 1 at 0.502. Starting from sample 2, the AUPRC increases and stays between 0.55 and 0.60. The PRC curves are not strictly monotone in extreme cases when recall is below 0.2, but gradually become monotonically decreasing afterward. Also, no matter how the AUPRC changes, it is always significantly higher than the area under the baseline, which is at most 0.161 over six subsamples, meaning that the predictive power of the spread stayed strong relative to baseline. AUROC tends to give much better-looking results in terms areas under the curves. The reason is ROC curves involve false alarm rates, whose computation is based on the forecast performance during expansions covering the vast majority of the sample. The huge number of true negatives in the numerator of the false alarm rate makes even large changes in false positives to have little effect on FA, see Tables 7-11 in the appendix. Thus the forecast performance using ROC will not be bad as long as the true-negative forecasts are tallied and counted in. Since identifying recessions is usually what most practitioners care about, they may not want to be over-optimistic about the good-looking ROC results because the forecasts were good during expansions. PRC approach can be a reasonable alternative since it does not use the true negatives as part of the evaluation. As Sofaer et al. (2019) quipped, "After all even a poor model can predict that a desert shrub will not occur in a rain forest"!

In order to directly compare PRC and ROC, we change the way PRC is conventionally plotted. In Figure 3 the inverted precision (i.e.,1-precision) is plotted on the horizontal axis against recall (hit rate) on the vertical axis.[8] This way we are able to compare ROC curve and the inverted PRC in one coordinate system. It is obvious to see that the PRC is uniformly below the ROC curve. Note that $1 - P(\delta)$ is conceptually an alternative measure of false alarm rate with the true negatives in $FA(\delta)$ replaced by true positives in $1 - P(\delta)$. Thus $1 - P(\delta)$ is defined over forecasts rather than actual values. The wedge between $1 - P(\delta)$ and $FA(\delta)$ in this diagram brings out directly the essential difference between the ROC and PRC approaches. Algebraically, the explanation can be found in our derivation in section 2, where smaller r tends to make $FA(\delta)$ smaller than $1 - P(\delta)$. The gap between the two curves is driven by the ratio in Equation (8), where the imbalancedness of data overwhelms the biasedness of forecasts most of the time. Figure 4 plots the ratio of $FA(\delta)$ over $1 - P(\delta)$ across different thresholds. The ratio is equal to 1 only when $\delta = 2.91\%$, which is a threshold where the false alarm rate is truly "alarming" at 84.7%. Most of the time the ratio is smaller 1, meaning that PRC would suggest a substantially

[8]The baseline of the inverted PRC becomes a vertical line on the right.

worse forecast performance compared to what ROC would suggest.

Table 2 shows AKS, the optimal threshold that maximizes KS, the maximized KS, and corresponding FA and R for all sub-samples. Note the standard errors of KS were computed as $\sqrt{R(\delta)(1-R(\delta))/n_H + F(\delta)(1-F(\delta))/n_F}$, where $n_H = TP + FN$ and $n_F = FP + TN$; see Stephenson (2000). The optimal threshold has increased to 0.91% with maximized KS of 0.67. The forecast performance in terms of AKS has been stable around 0.20. In the final sample the recall (hit rate) is high at 0.92 with a false alarm rate of 0.25. If the forecast user could stomach a smaller recall of 0.80, the false alarm rate could be reduced further to 0.18, see Figure 3.

Tables 3, 4 and 5 show the optimal thresholds that maximize $F_{0.5}$, F_1 and F_2 respectively. The corresponding precision, recall and maximized $F_{0.5}$, F_1 and F_2 are also reported. It is found the thresholds that maximize $F_{0.5}$ are all around zero for all sub-samples, meaning that looking at the inversion of the yield spread is still a good strategy to forecast recessions if $F_{0.5}$ is believed to be the appropriate measure. The threshold that maximizes F_1 increases from about 0% to 0.21% after reaching sample 4. The threshold that maximizes F_2 increases to 0.91% after reaching sample 3. The values of P/R at maximized thresholds for $F_{0.5}$ are higher/lower compared with values of P/R at maximized thresholds for F_1 and F_2 since the weights in the nonlinear function of $F_{0.5}$ favor precision more relative to recall. Further, it is also found that all three maximized measures drop a bit after sample 2, but creep back gradually afterward. This implies that the predictive power of the spread has not deteriorated in recent few decades. Instead, it has increased slightly.

We plot δ against F measures and KS in Figure 5 to help explain Tables 2-5. $F_{0.5}$ is maximized around zero threshold. As β in Equation (10) increases, recall (hit rate) becomes more and more weighed relative to precision, and higher thresholds become more preferred. Tables 7-10 in Appendix A show, as threshold increases from 0% to 0.91%, the ratio of biasedness $\mu_{\hat{y}(\delta)}/\mu_y$ increases from 0.79 to 2.55, and forecasts get more and more biased because of forecasting more recessions; see Tables 7-10 in Appendix A. Under the global accuracy measure like the Brier's score or RMSE, the results are very similar to those under $F_{0.5}$ with almost unbiased forecasts because the bias term is the dominant part of the RMSE criterion. Interestingly, it is also found that the thresholds that maximize F_2, which prefers recall more compared with $F_{0.5}$ and F_1, coincide with the thresholds that maximize $KS(\delta)$ in all sub-samples; see Tables 2 and 5. The column values under FA and P can be reconciled

exactly following Equation (5) or (8). At the optimum value of F_2, a high hit rate of 0.92 is attained. The accompanying FA and P would be 0.25 (Table 2) and 0.36 (Table 5) respectively. The PRC approach however would suggest that the respectable FA of 0.25 can be considered unduly optimistic, because of its use of TN in its definition. In PRC approach this definition is replaced by $1 - P(\delta) = FP(\delta)/(FP(\delta) + TP(\delta))$, the PRC analogue of false alarm rate based on forecasts, which has a much higher value (0.64). Thus, under PRC the same forecasts would look much less certain in terms of the extra possibility of getting a positive forecast not materialize. The inverted precision is particularly intuitive and useful in rare event forecasting because it measures the fraction of incorrect predictions among the positive predictions. We will focus on F_2 further later in the section on robustness check. Yang et al. (2023) have sown that KS implies a net utility gain from correctly identifying recessions $(1 - \mu_y)/\mu_y$ times greater than that from correctly identifying an expansion. In the current context the ratio turns out to be 6.52. Similarity of KS to F_2 would suggest that the latter formula incorporates a similar relative preference for identifying recessions. In this sense KS and F_2 measures are better suited to evaluate rare-event forecasts.

Note that given a sample of forecasts and actual outcomes, the 2×2 confusion matrix has 3 degrees of freedom. Two of these are used to define (R, FA) in ROC and (P, R) in PRC analysis. The third degree of freedom can be used to calibrate the forecasts. As the confusion matrices in Tables 7-11 show, with changes in thresholds the elements of ROC and PRC together with the biasedness of the forecasts [defined as $\mu_{\hat{y}(\delta)}/\mu_y = (TP(\delta) + FP(\delta))/(TP(\delta) + FN(\delta))$] change too. The latter ratio changes from 0.79 ($\delta = 0$) to 6.52 ($\delta = 2.91\%$). Forecasts were found to be exactly unbiased only when $\delta = 0.21\%$. As we mentioned before the flexibility of ROC and PRC approaches comes from the feature that the choice of the threshold comes at the end to comply with the user's preferred trade offs between R and FA or P and R. So the most fundamental determining criterion in choosing a cut off point is the relative costs and benefits to the forecast user to identify recessions compared with expansions. Fundamentally ROC and PRC contain the same amount of information, see Davis and Goadrich (2006).

Table 6 reports the results for MCC. It is found that the thresholds that maximize MCC are close to zero initially and rose to 0.21 for the last two sub samples, giving forecasts with smaller R (hit rate) compared with F_2. The reason can be that MCC takes TN into account, which prefers lower thresholds as they are likely to generate more negative forecasts. Although F_1 does not

take TN into account, the maximizing thresholds for MCC and F_1 coincide exactly for all sub samples in our case; see Tables 4 and 6 and also Figure 6. Thus even though F_1 can give misleading information in some extreme cases as discussed by Chicco and Jurman (2020), it may not necessarily do so at the optimal threshold. Figure 7 plots the ratio of the standard deviation of the actual values of y to the standard deviation of the binary forecasts as a function of threshold δ.[9] The curve is U-shaped as expected due to the nature of the variance of binary variable, but slightly less than one over meaningful threshold values, making MCC less than KS. This prediction is borne out well in Tables 2 and 6.

4 Robustness Checks

We regenerated the results for the six separate sub-samples reported in Table 5 using both monthly and quarterly observations. These results are reported in Tables 12 (monthly) and 13 (quarterly) respectively. In these tables we report the optimal threshold values that maximize F_2 and also the corresponding values of precision and recall. We see that the optimal threshold has increased away from zero towards one.[10] Although there are some differences in terms of precision and recall, the maximized values of F_2 are very close to the F_2 values in our main findings with daily data. Thus the results are consistent with results using daily data. However, daily data is still preferred due to its additional number of observations, which benefits us numerically in the calculation of different rates in skewed samples.

Figures 8 and 9 plot the PRC of 18-month-ahead and 6-month-ahead forecasts. It is found that the AUPRC values of the 18-month-ahead forecasts tend to be higher, with a maximum value of 0.682 in sample 6. The AUPRC values of the 6-month-ahead forecasts tend to be lower, which are no more than 0.5. The predictive power for 18-month-ahead forecasts gradually increased since a drop in sample 3, which is consistent with our main findings with 12-month-ahead forecasts. The AUPRC values at horizon of 6 months stayed around 0.4 after sample 3. Tables 14 and 15 report the optimal thresholds in terms of F_2 for horizons of 18 and 6 months. As expected, forecast performance in terms of maximized F_2 improves as the horizon increases from 6 to 18 months in most

[9] Some extreme thresholds at both ends are excluded for better demonstration.

[10] With quarterly data in Table 13, although the optimal threshold is 0.27% for sample 4, the F_2 is only slightly lower (0.68 rather 0.69) when a threshold of 0.93% is chosen as found in adjoining sub-samples 3, 5 and 6. This is because with quarterly data F_2 is rather flat between 0.2 and 1.0. At threshold of 0.93%, the recall is higher ($R = 0.92$) than that with threshold 0.27% ($R = 0.76$), but the precision is lower ($P = 0.34$).

of the sub-samples, since a recession that starts anytime in the next 18 months is a wider (thus easier) target to hit compared with a recession that starts anytime in the next 12 months or next 6 months. What is more important to notice is that the optimal thresholds have increased to around 1.0% at longer horizons. Interestingly, at 6-month horizon, the optimal threshold was indeed very close to zero till post-COVID period. At this horizon, the deterioration in the forecasting power of yield spread is very clearly seen after the end of sample 2, i.e., after 1990 the maximized F_2 and the recall fell substantially. Thus, a deterioration of predictive power is only found at horizon of 6 months, which is no longer the best forecast horizon for spreads in recent decades. At longer horizons the predictive power has even been increasing gradually in terms of both AUPRC and maximized F_2, which is consistent with our main findings.

5 Conclusions

In the context of machine learning, Provost et al. (1998) first pointed out the inadequacy of conventional accuracy measures that evaluate estimated probabilities and recommended Receiver Operating Characteristic (ROC) analysis for binary decision making. It is now the most common binary assessment tool in almost any scientific field. Economic forecasting is no exception. However, it is now well recognized that ROC presents an overly optimistic picture of a predictor's true discriminating power when the sample data is highly imbalanced, and Precision-Recall Curve (PRC) has emerged as a more appropriate evaluation tool for rare event forecasts. Due to ROC's failure to handle rare events and many other problems, Chicco and Jurman (2023) have suggested for its retirement after 80 years of honorable service!

In this paper we have studied the relationship between ROC and PRC both analytically and using a real-life empirical example of yield spread as a predictor of recessions. In the forecasting literature events are taken to be rare if it occurs less that 2.5% of the times or less. Recessions in the U.S. are not that uncommon - it occupies 10-15% of the sample in recent decades. So one empirical question we addressed is whether the inadequacy of ROC that has been demonstrated with truly rare events carries over to cases occurring at higher rates too.

The essential difference between ROC and PRC is the way each defines false alarm rates. We show that false alarm rate in ROC and inverted precision in PRC are analogous and their difference is determined by the interaction of sample imbalance and forecast bias. By inverting the PRC and replacing precision

with one minus precision, we are able to plot it on ROC space for direct comparisons. We found that in cases of severe imbalance in the sample, the forecasts need to be adequately biased to mitigate the effect of imbalancedness. This relationship is confirmed in our empirical example and shows that for plausible values of preference parameters, even in recession forecasting where it occurs little over 10% of the sample, ROC severely overstates the true predictive performance.

To aid real-time forecasting, our forecasts are evaluated more in terms of maximized $F_{0.5}$, F_1, F_2, MCC and KS measures than on global measures like AUROC or AUPRC. We derived few relationships between these statistics and focused on their thresholds that can be used optimally to make forecasts. By extending the sample gradually and recursively, we find that the optimal threshold varies depending on the used measures, that reflect their relative preference for making correct recession forecasts compared to non-recessions. The optimal threshold stays around 0 if $F_{0.5}$, F_1 or MCC are used. But at these values, the hit rate or recall will be unacceptably low to many. The optimal threshold has increased to about 0.91% in recent decades if KS or F_2 is used, which are more recall-weighted. Importantly, the mix of values of precision and recall over six sub-samples show that the predictive power of the spread has not deteriorated in recent decades, provided the optimum values of threshold are used.

Our analysis also finds that most of the meaningful thresholds for recession forecasts are upward biased, generating much smaller FA than $1 - P$ for comparable hit rates or recalls. We underscore the importance of deliberate forecast bias in order to attain acceptable levels of hit rate, false alarm rate and precision by choosing appropriate thresholds. For example, at hit rate 92%, the associated false alarm rate with ROC will be 25%. But at the same hit rate, [1-Precision], which is the analogous false alarm rate with PRC, will be 64%. This quantifies the extent to which ROC could be exaggerating the true predictive value of the yield curve. Our results are robust even when we worked with monthly or quarterly data and with alternative forecast horizons.

Figure 1: ROC Curves with Daily Spread as Threshold: 12-Month-Ahead Forecasts

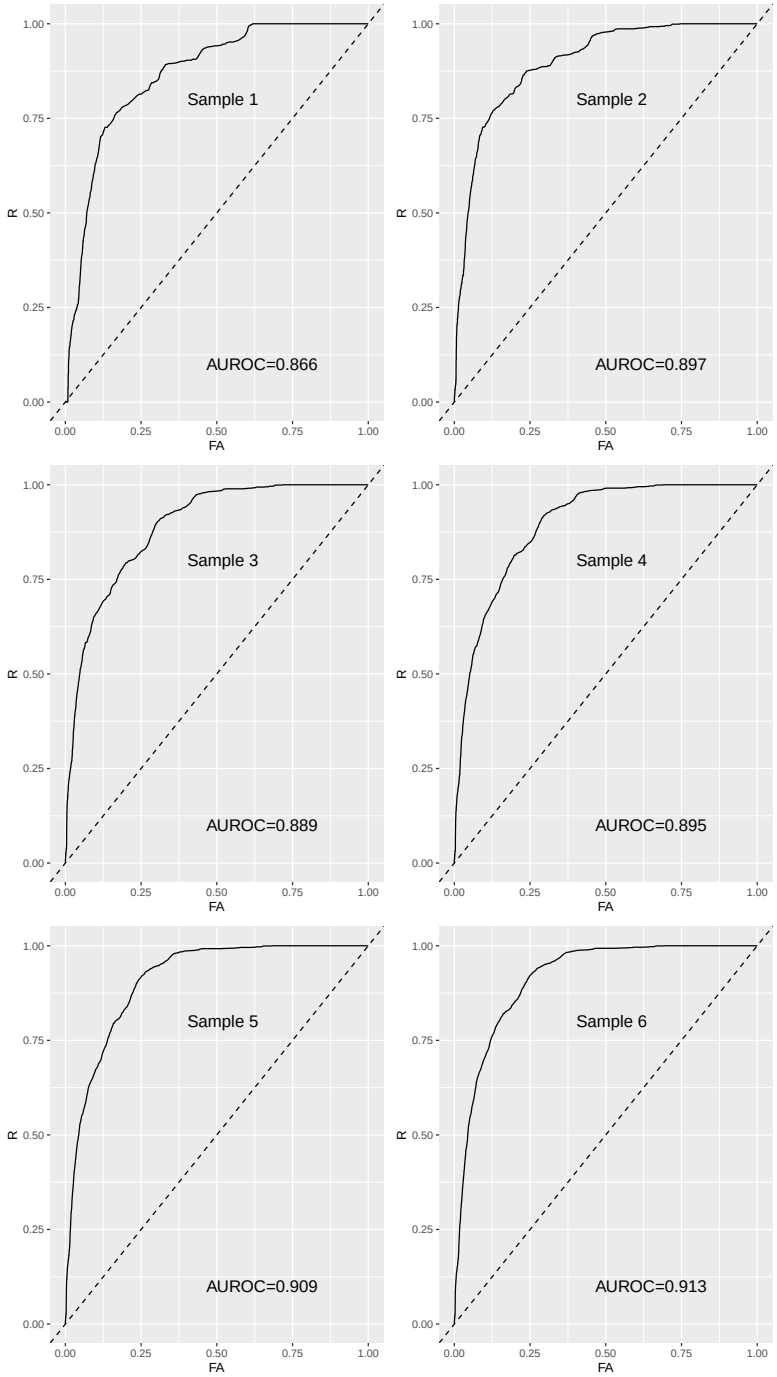
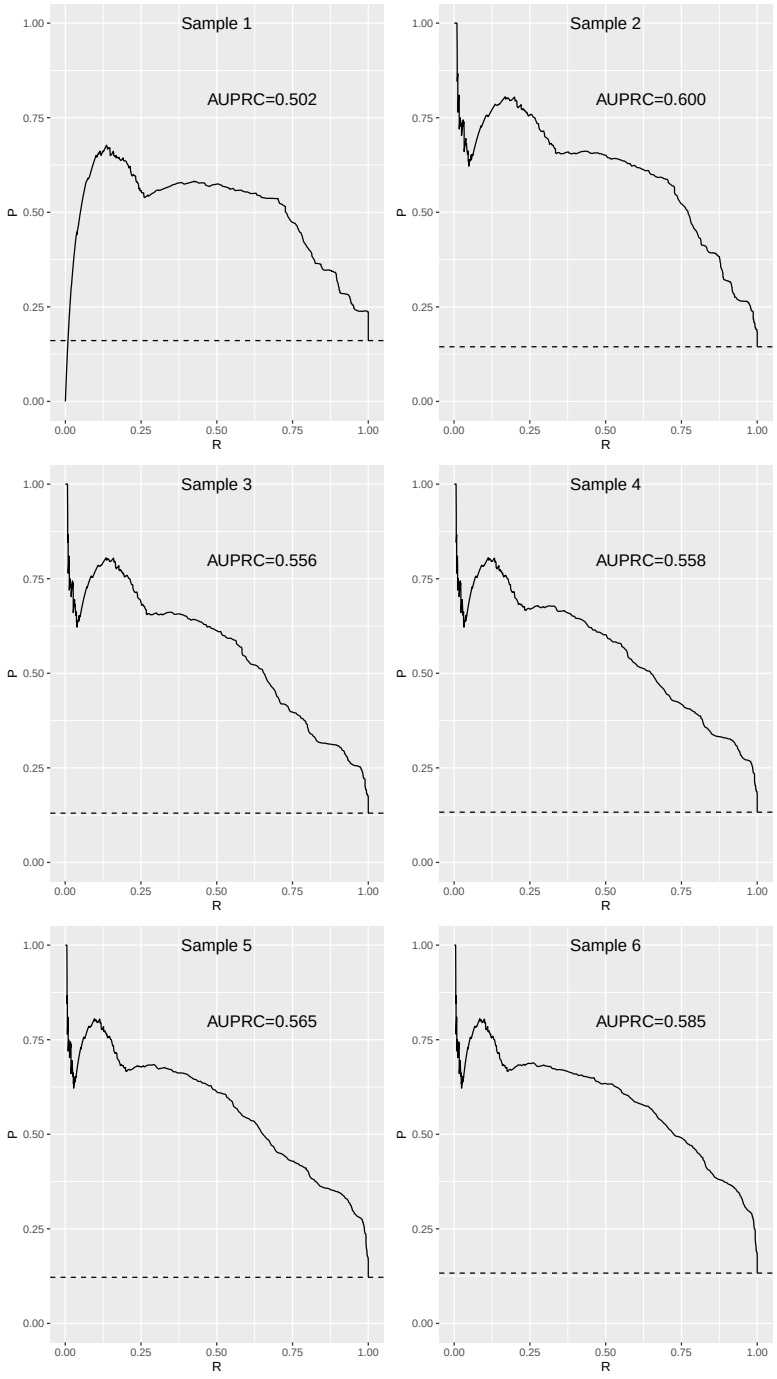


Figure 2: PRC with Daily Spread as Threshold: 12-Month-Ahead Forecasts



The dashed line in each graph is for the baseline no-skill forecasts, which is a horizontal line at the level of the fraction of recession periods. The proportions are 0.161, 0.144, 0.130, 0.133, 0.122, and 0.133 from sample 1 to sample 6.

Figure 3: ROC Curve and Inverted PRC: 12-Month-Ahead Forecasts (1/2/1962 - 11/30/2021)

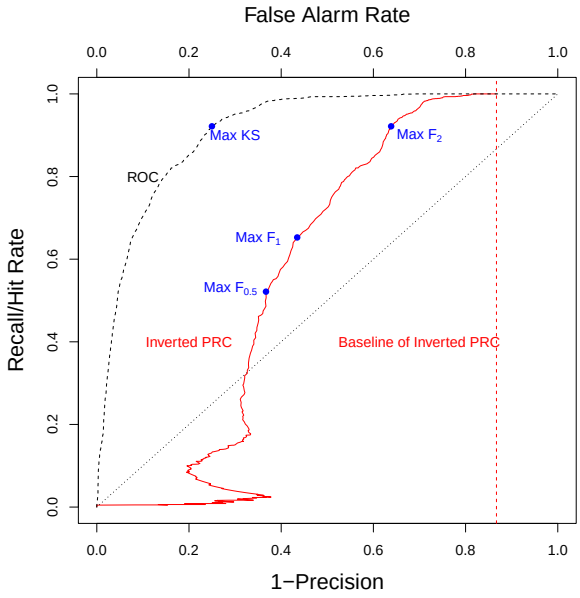


Figure 4: $FA(\delta)$ over $1 - P(\delta)$ (1/2/1962 - 11/30/2021)

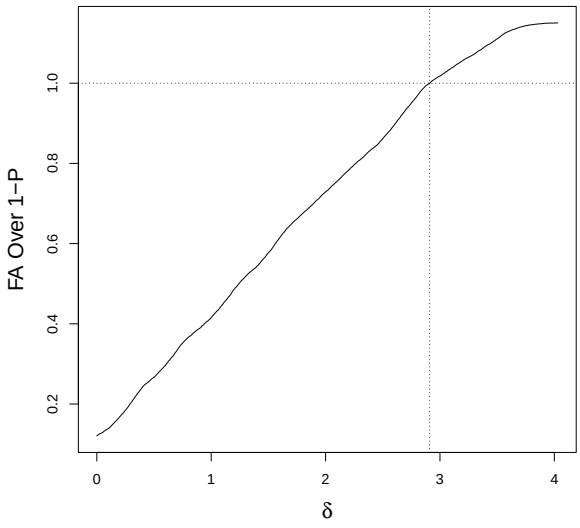


Figure 5: KS and F measures as functions of δ (1/2/1962 - 11/30/2021)

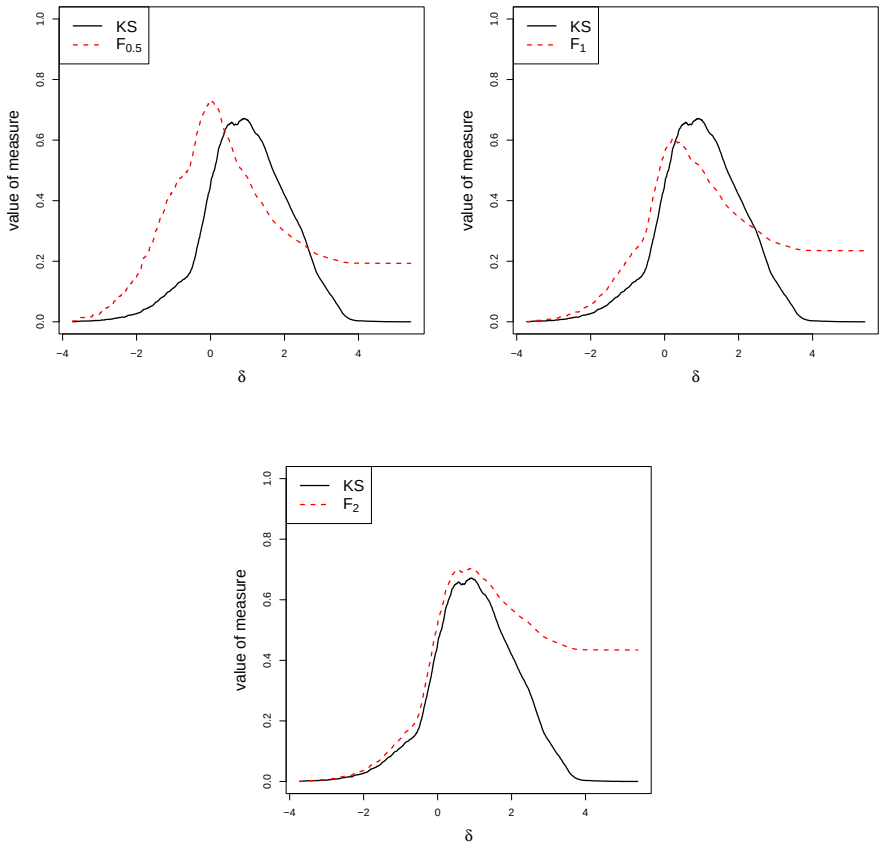


Figure 6: MCC and F_1 as functions of δ (1/2/1962 - 11/30/2021)

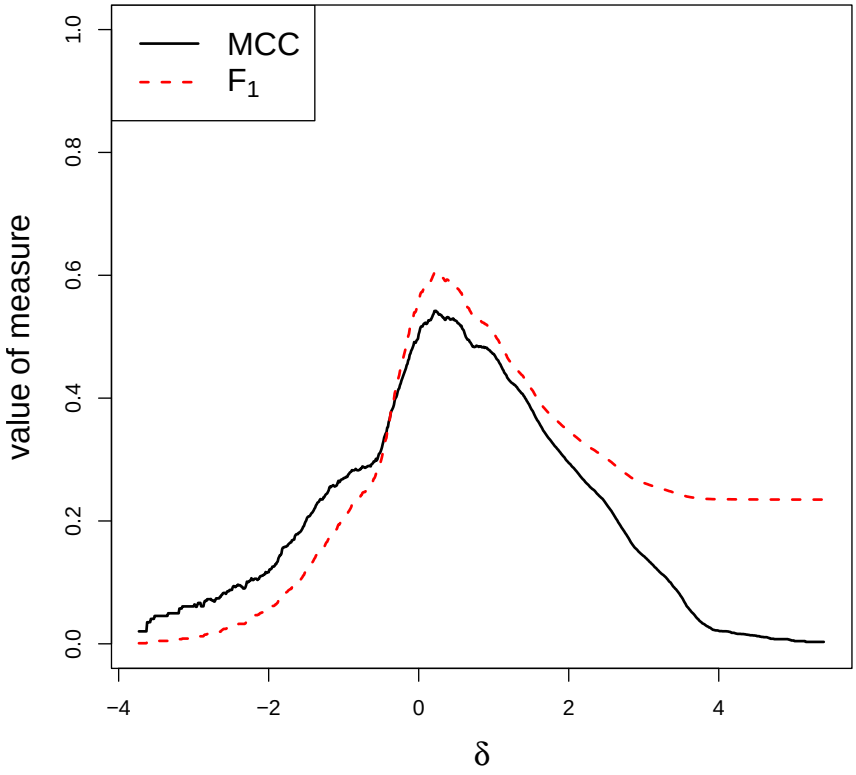
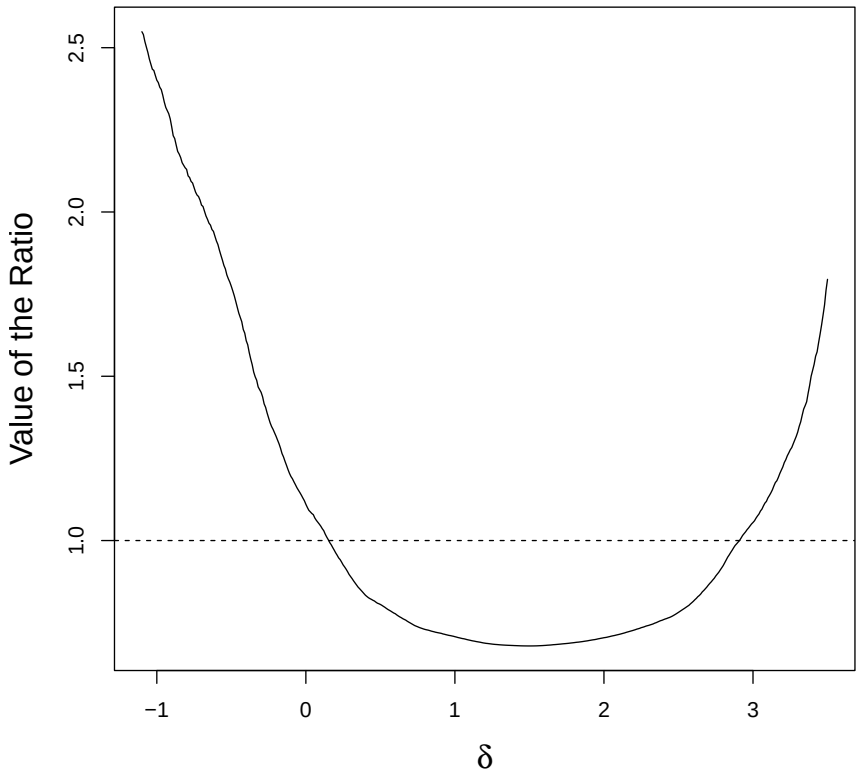


Figure 7: $\sqrt{(\mu_y(1-\mu_y))/(\mu_{\hat{y}(\delta)}(1-\mu_{\hat{y}(\delta)}))}$ as a function of δ (1/2/1962 - 11/30/2021)



The horizontal dashed line at 1 shows that the ratio of the standard deviation of actuals and the standard deviation of binary forecasts are equal at delta $= 0.15\%$ *and* $\delta = 2.91\%$*. By Equation* (16)*,* MCC *is smaller than* KS *below the dashed horizontal line. Note that the forecasts are unbiased at* $\delta = 0.15\%$ *but not at* $\delta = 2.91\%$*. The ratio is equal to 1 at* $\delta = 2.91\%$ *since* $\mu_{\hat{y}(2.91)} = 0.867 = 1 - \mu_y$*.*

Table 2: Optimal Interest Rate Spread Thresholds KS: 12-Month-Ahead Forecasts with Daily Data

Sample	FA	R	Maximized KS	Maximizing Threshold	AKS
1	0.17	0.77	0.60	0.20	0.21
2	0.13	0.77	0.64	0.21	0.20
3	0.30	0.90	0.60	0.91	0.19
4	0.30	0.92	0.62	0.91	0.20
5	0.24	0.91	0.67	0.91	0.20
6	0.25	0.92	0.67	0.91	0.20

Sample 1: 1/2/1962 to 7/31/1980; Sample 2: 1/2/1962 to 7/31/1989; Sample 3: 1/2/1962 to 3/31/2000; Sample 4: 1/2/1962 to 12/29/2006; Sample 5: 1/2/1962 to 2/28/2019; Each sub-sample ended 12 months before the beginning of the recessions of 1981, 1990, 2001, 2008 and 2020, respectively. Sample 6 is our full sample. The standard errors of KS are all smaller than 0.018.

Table 3: Optimal Interest Rate Spread Thresholds for $F_{0.5}$: 12-Month-Ahead Forecasts with Daily Data

Sample	P	R	Maximized $F_{0.5}$	Maximizing Threshold
1	0.55	0.63	0.68	-0.05
2	0.64	0.56	0.74	-0.17
3	0.61	0.51	0.71	-0.06
4	0.64	0.43	0.70	-0.19
5	0.63	0.48	0.71	-0.06
6	0.63	0.52	0.73	0.03

See Table 2 for definition of sub-samples.

Table 4: Optimal Interest Rate Spread Thresholds for F_1: 12-Month-Ahead Forecasts with Daily Data

Sample	P	R	Maximized F_1	Maximizing Threshold
1	0.54	0.70	0.61	0.03
2	0.59	0.70	0.64	0.03
3	0.57	0.58	0.58	0.08
4	0.51	0.65	0.57	0.21
5	0.53	0.62	0.58	0.21
6	0.57	0.65	0.61	0.23

See Table 3 for definition of sub-samples.

Table 5: Optimal Interest Rate Spread Thresholds for F_2: 12-Month-Ahead Forecasts with Daily Data

Sample	P	R	Maximized F_2	Maximizing Threshold
1	0.46	0.77	0.68	0.20
2	0.50	0.77	0.70	0.21
3	0.31	0.90	0.65	0.91
4	0.32	0.92	0.67	0.91
5	0.34	0.91	0.68	0.91
6	0.36	0.92	0.70	0.91

See Table 3 for definition of sub-samples.

Table 6: Optimal Interest Rate Spread Thresholds for MCC: 12-Month-Ahead Forecasts with Daily Data

Sample	P	R	FA	Maximized MCC	Maximizing Threshold
1	0.54	0.70	0.12	0.53	0.03
2	0.59	0.70	0.08	0.58	0.03
3	0.59	0.56	0.06	0.51	0.03
4	0.58	0.55	0.06	0.50	0.03
5	0.53	0.62	0.08	0.51	0.21
6	0.57	0.65	0.08	0.54	0.23

See Table 3 for definition of sub-samples.

Acknowledgments

An earlier version of this paper was presented as plenary lecture at the 36th CIRET Conference September 14 – 17, 2022. We thank the anonymous referee for making many useful comments.

Statements and Declarations

No funding was received for conducting this study. The authors have no competing interests to declare that are relevant to the content of this article.

Appendix

A Confusion Matrices for Different Thresholds

Table 7: Contingency Matrix When $\delta = 0\%$ (1/2/1962 - 11/30/2021)

	recession ($y_t = 1$)	no recession ($y_t = 0$)	
predict recession ($\hat{y}_t = 1$)	TP=1048	FP=604	$P = 0.63$
predict no recession ($\hat{y}_t = 0$)	FN=1056	TN =13110	
	$R = 0.50$	$FA = 0.04$	

$\mu_{\hat{y}}/\mu_y = (TP + FP)/(TP + FN) = 0.79$. *The forecasts are downward biased.*

Table 8: Confusion Matrix When Forecasts are Unbiased at $\delta = 0.15\%$ (1/2/1962 - 11/30/2021)

	recession ($y_t = 1$)	no recession ($y_t = 0$)	
predict recession ($\hat{y}_t = 1$)	TP=1237	FP=853	$P = 0.59$
predict no recession ($\hat{y}_t = 0$)	FN=867	TN =12861	
	$R = 0.59$	$FA = 0.06$	

$\mu_{\hat{y}}/\mu_y = (TP + FP)/(TP + FN) = 0.99$. *The forecasts are unbiased with the ratio close to 1.*

Table 9: Confusion Matrix When $\delta = 0.50\%$ (1/2/1962 - 11/30/2021)

	recession ($y_t = 1$)	no recession ($y_t = 0$)	
predict recession ($\hat{y}_t = 1$)	TP=1677	FP=1980	$P = 0.46$
predict no recession ($\hat{y}_t = 0$)	FN=427	TN =11734	
	$R = 0.80$	$FA = 0.14$	

$\mu_{\hat{y}}/\mu_y = (TP + FP)/(TP + FN) = 1.74$. *The forecasts are upward biased.*

Table 10: Confusion Matrix When $\delta = 0.91\%$ (1/2/1962 - 11/30/2021)

	recession ($y_t = 1$)	no recession ($y_t = 0$)	
predict recession ($\hat{y}_t = 1$)	TP=1939	FP=3425	$P = 0.36$
predict no recession ($\hat{y}_t = 0$)	FN=165	TN =10289	
	$R = 0.92$	$FA = 0.25$	

$\mu_{\hat{y}}/\mu_y = (TP + FP)/(TP + FN) = 2.55$. *The forecasts are upward biased.*

Table 11: Confusion Matrix When $\delta = 2.91\%$ (1/2/1962 - 11/30/2021)

	recession ($y_t = 1$)	no recession ($y_t = 0$)	
predict recession ($\hat{y}_t = 1$)	TP=2104	FP=11612	$P = 0.15$
predict no recession ($\hat{y}_t = 0$)	FN=0	TN =2102	
	$R = 1.00$	$FA = 0.85$	

$\mu_{\hat{y}}/\mu_y = (TP + FP)/(TP + FN) = 6.52$. *The forecasts are upward biased.*

B Tables and Figures for Robustness Checks

Table 12: Optimal Interest Rate Spread Thresholds for F_2: 12-Month-Ahead Forecasts with Monthly Data

Sample	P	R	Maximized F_2	Maximizing Threshold
1	0.53	0.78	0.71	0.14
2	0.58	0.79	0.74	0.14
3	0.41	0.78	0.66	0.49
4	0.42	0.81	0.68	0.49
5	0.39	0.85	0.69	0.65
6	0.42	0.86	0.71	0.65

See Table 3 for definition of sub-samples.

Table 13: Optimal Interest Rate Spread Thresholds for F_2: 12-Month-Ahead Forecasts with Quarterly Data

Sample	P	R	Maximized F_2	Maximizing Threshold
1	0.45	0.77	0.68	0.21
2	0.47	0.82	0.71	0.27
3	0.33	0.90	0.67	0.93
4	0.50	0.76	0.69	0.27
5	0.36	0.93	0.71	0.93
6	0.37	0.94	0.72	0.93

See Table 3 for definition of sub-samples.

Table 14: Optimal Interest Rate Spread Thresholds for F_2: 18-Month-Ahead Forecasts with Daily Data

Sample	P	R	Maximized F_2	Maximizing Threshold
1	0.30	0.98	0.68	2.38
2	0.46	0.82	0.71	0.89
3	0.36	0.88	0.68	1.31
4	0.44	0.84	0.71	1.00
5	0.48	0.85	0.74	1.00
6	0.48	0.87	0.75	1.00

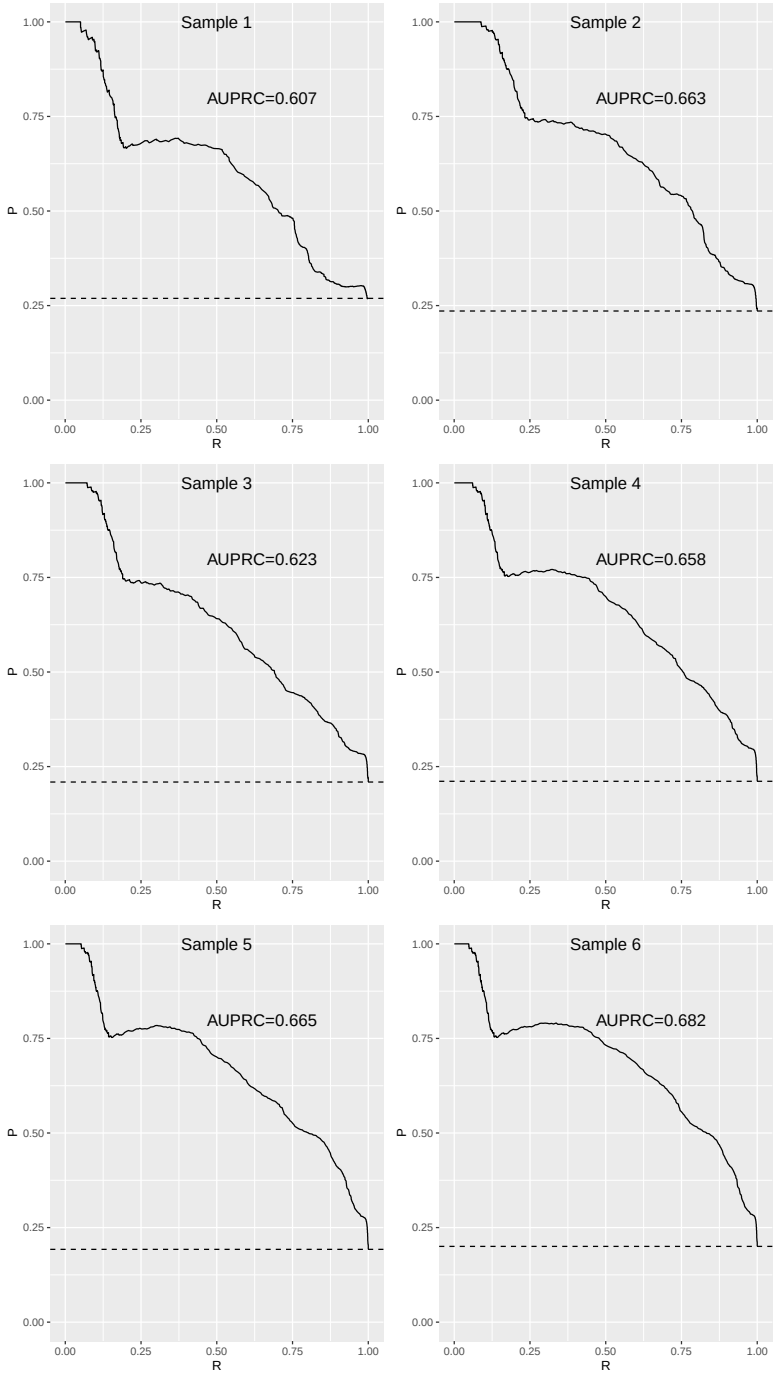
See Table 3 for definition of sub-samples.

Table 15: Optimal Interest Rate Spread Thresholds for F_2: 6-Month-Ahead Forecasts with Daily Data

Sample	P	R	Maximized F_2	Maximizing Threshold
1	0.36	0.95	0.72	0.03
2	0.39	0.93	0.72	0.03
3	0.39	0.74	0.63	0.03
4	0.37	0.72	0.61	0.04
5	0.35	0.63	0.54	0.06
6	0.25	0.76	0.54	0.37

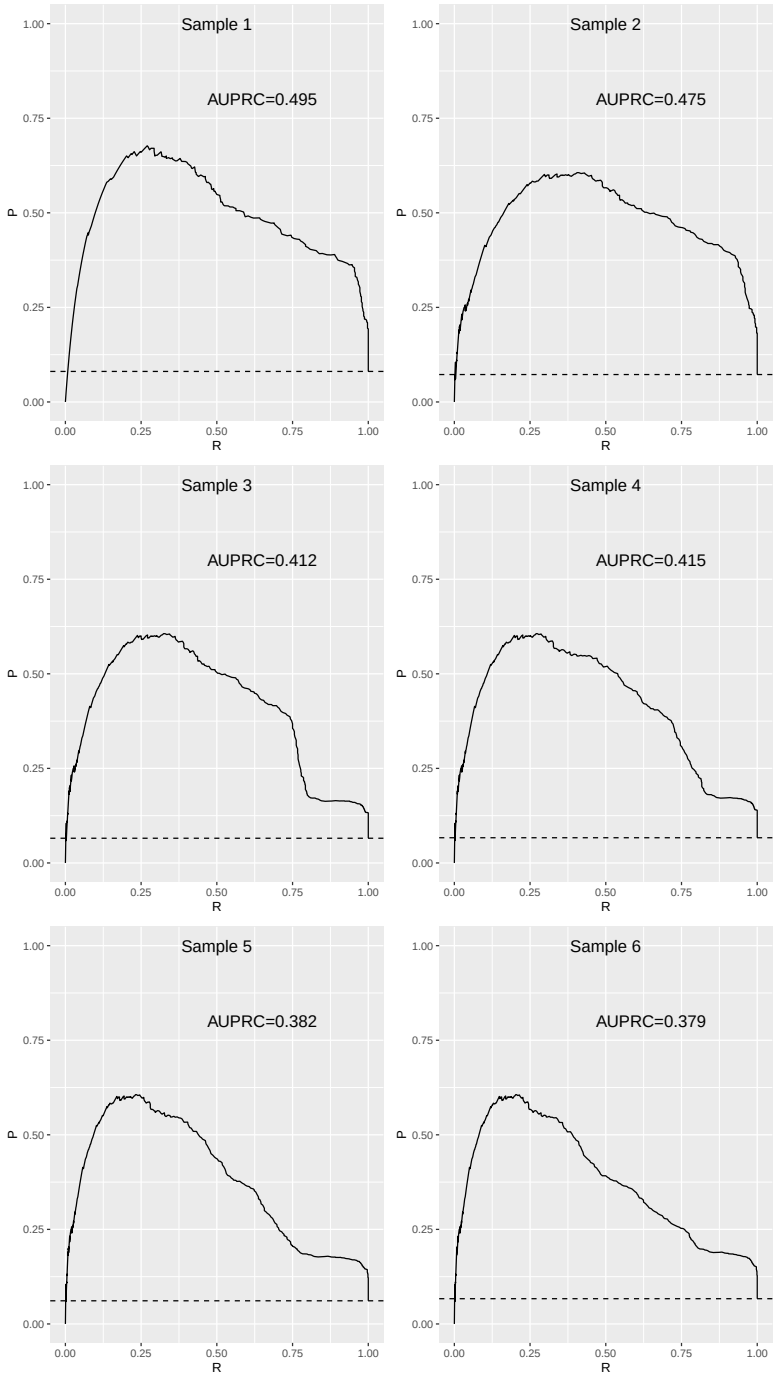
See Table 3 for definition of sub-samples.

Figure 8: PRC with Daily Spread as Threshold: 18-Month-Ahead Forecasts



The dashed line in each graph is for the baseline no-skill forecasts, which is a horizontal line at the level of the fraction of recession periods. The baselines here differ from the baselines for 12-month-ahead forecasts since the definition of the left-hand-side variable depends on forecast horizon, and the fractions of observations with $y_t = 1$ differ.

Figure 9: PRC with Daily Spread as Threshold: 6-Month-Ahead Forecasts



See Figure 8 for explanations of baseline forecasts.

References

Ajello, A., L. Benzoni, M. Schwinn, Y. Timmer, and F. Vazquez-Grande (2022). Monetary policy, inflation outlook, and recession probabilities. FEDS Notes. Washington: Board of Governors of the Federal Reserve System, July 12, 2022, https://doi.org/10.17016/2380-7172.3175.

Bauer, M. D. and T. M. Mertens (2018). Information in the yield curve about future recessions. *FRBSF Economic Letter 20*, 1–5.

Berge, T. J. and Ò. Jordà (2011). Evaluating the classification of economic activity into recessions and expansions. *American Economic Journal: Macroeconomics 3*(2), 246–77.

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

Chicco, D. and G. Jurman (2020). The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC genomics 21*(1), 1–13.

Chicco, D. and G. Jurman (2023). The Matthews correlation coefficient (MCC) should replace the ROC AUC as the standard metric for assessing binary classification. *BioData Mining 16*(4).

Chinchor, N. (1992). MUC-4 Evaluation Metrics. In *Fourth Message Uunderstanding Conference (MUC-4): Proceedings of a Conference Held in McLean, Virginia, June 16-18, 1992*.

Chinchor, N. and B. M. Sundheim (1993). MUC-5 evaluation metrics. In *Fifth Message Understanding Conference (MUC-5): Proceedings of a Conference Held in Baltimore, Maryland, August 25-27, 1993*.

Choi, J., D. Ge, K. H. Kang, and S. Sohn (2023). Yield spread selection in predicting recession probabilities. *Journal of Forecasting*.

Cook, J. and V. Ramadas (2020). When to consult precision-recall curves. *The Stata Journal 20*, 131–148.

Cooper, D., J. C. Fuhrer, and G. Olivei (2020). Predicting recessions using the yield curve: The role of the stance of monetary policy. Available at SSRN 3587629.

Davis, J. and M. Goadrich (2006). The relationship between precision-recall and roc curves. In *Proceedings of the 23rd international conference on Machine learning*, pp. 233–240.

Dice, L. R. (1945). Measures of the amount of ecologic association between species. *Ecology 26*(3), 297–302.

Dotsey, M. (1998). The predictive content of the interest rate term spread for future economic growth. *FRB Richmond Economic Quarterly 84*(3), 31–51.

Elliott, G. and R. P. Lieli (2013). Predicting binary outcomes. *Journal of Econometrics 174*(1), 15–26.

Ergungor, O. E. (2016). Recession Probabilities. Federal Reserve Bank of Cleveland. Federal Reserve Bank of Cleveland, *Economic Commentary* 2016-09. https://doi.org/10.26509/frbc-ec-201609.

Estrella, A. and G. A. Hardouvelis (1991). The term structure as a predictor of real economic activity. *The Journal of Finance 46*(2), 555–576.

Estrella, A. and F. S. Mishkin (1996). The yield curve as a predictor of U.S. recessions. *Current issues in economics and finance 2*(7), 1–6.

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters 27*(8), 861–874.

Flach, P. A. and M. Kull (2015). Precision-Recall-Gain curves: PR analysis done right. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett (Eds.), *Advances in neural information processing systems*, Volume 1, pp. 838–846.

Ghysels, E., J. B. Hill, and K. Motegi (2020). Testing a large set of zero restrictions in regression models, with an application to mixed frequency granger causality. *Journal of Econometrics 218*(2), 633–654.

Giacomini, R. and B. Rossi (2006). How stable is the forecasting performance of the yield curve for output growth? *Oxford Bulletin of Economics and Statistics 68*, 783–795.

Grau, J., I. Grosse, and J. Keilwagen (2015). Prroc: computing and visualizing precision-recall and receiver operating characteristic curves in r. *Bioinformatics 31*(15), 2595–2597.

Jardet, C. (2004). Why did the term structure of interest rates lose its predictive power? *Economic Modelling 21*(3), 509–524.

Johansson, P. and A. C. Meldrum (2018). Predicting recession probabilities using the slope of the yield curve. Board of Governors of the Federal Reserve System (U.S.).

Keilwagen, J., I. Grosse, and J. Grau (2014). Area under Precision-Recall Curves for Weighted and Unweighted Data. *PLOS ONE 9*(3), e92209.

Lahiri, K., G. Monokroussos, and Y. Zhao (2013). The yield spread puzzle and the information content of SPF forecasts. *Economics Letters 118*(1), 219–221.

Lahiri, K. and C. Yang (2022). ROC approach to forecasting recessions using daily yield spreads. *Business Economics 57*(4), 191–203.

Lahiri, K. and C. Yang (2023). A tale of two recession-derivative indicators. *Empirical Economics*.

Lahiri, K. and L. Yang (2013). Forecasting binary outcomes. In *Handbook of economic forecasting*, Volume 2, pp. 1025–1106. Elsevier.

Lahiri, K. and L. Yang (2021). Construction of leading economic index for recession prediction using vine copulas. *Studies in Nonlinear Dynamics & Econometrics 25*(4), 193–212.

Levanon, G., J.-C. Manini, A. Ozyildirim, B. Schaitkin, and J. Tanchua (2015). Using financial indicators to predict turning points in the business cycle: The case of the leading economic index for the united states. *International Journal of Forecasting 31*(2), 426–445.

Liang, S., Y. Li, and R. Srikant (2017). Enhancing the reliability of out-of-distribution image detection in neural networks. arXiv preprint arXiv:1706.02690.

Lieli, R. P. and Y.-C. Hsu (2019). Using the area under an estimated ROC curve to test the adequacy of binary predictors. *Journal of Nonparametric Statistics 31*(1), 100–130.

Lin, F., Y. Zhang, and J. Wang (2023). Recent advances in intra-hour solar forecasting: A review of ground-based sky image methods. *International Journal of Forecasting 39*(1), 244–265.

Lo-Ciganic, W.-H., J. L. Huang, H. H. Zhang, J. C. Weiss, Y. Wu, C. K. Kwoh, J. M. Donohue, G. Cochran, A. J. Gordon, D. C. Malone, et al. (2019). Evaluation of machine-learning algorithms for predicting opioid overdose risk among medicare beneficiaries with opioid prescriptions. *JAMA network open 2*(3), e190968–e190968.

Matthews, B. W. (1975). Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochimica et Biophysica Acta (BBA)-Protein Structure 405*(2), 442–451.

Ozenn, B., F. Subtil, and D. Maucort-Boulch (2015). The precision–recall curve overcame the optimism of the receiver operating characteristic curve in rare diseases. *Journal of Clinical Epidemiology 68*(8), 855–859.

Pažickỳ, M. (2021). Predicting recessions in germany using the german and the us yield curve. *Journal of Business Cycle Research 17*(3), 263–291.

Peirce, C. S. (1884). The numerical measure of the success of predictions. *Science* (93), 453–454.

Pigini, C. (2021). Penalized maximum likelihood estimation of logit-based early warning systems. *International Journal of Forecasting 37*(3), 1156–1172.

Pinker, E. (2018). Reporting accuracy of rare event classifiers. *npj Digital Medicine 1*(56).

Provost, F. J., T. Fawcett, and R. Kohavi (1998). The Case against Accuracy Estimation for Comparing Induction Algorithms. In *International Conference on Machine Learning*, Volume 98, pp. 445–453.

Puglia, M. and A. Tucker (2021). Neural Networks, the Treasury Yield Curve, and Recession Forecasting. *The Journal of Financial Data Science 3*(2), 149–175.

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*(4), 492–503.

Rummens, A. and W. Hardyns (2021). The effect of spatiotemporal resolution on predictive policing model performance. *International Journal of Forecasting 37*(1), 125–133.

Saito, T. and M. Rehmsmeier (2015). The precision-recall plot is more informative than the roc plot when evaluating binary classifiers on imbalanced datasets. *PloS one 10*(3), e0118432.

Sofaer, H. R., J. A. Hoeting, and C. S. Jarnevich (2019). The area under the precision-recall curve as a performance metric for rare binary events. *Methods in Ecology and Evolution 10*(4), 565–577.

Sorensen, T. A. (1948). A method of establishing groups of equal amplitude in plant sociology based on similarity of species content and its application to analyses of the vegetation on Danish commons. *Biol. Skar. 5*, 1–34.

Stephenson, D. B. (2000). Use of the "odds ratio" for diagnosing forecast skill. *Weather and Forecasting 15*(2), 221–232.

Stock, J. H. and M. W. Watson (1993). Introduction to" business cycles, indicators and forecasting". In *Business Cycles, Indicators, and Forecasting*, pp. 1–10. University of Chicago Press.

van Rijsbergen, C. J. (1979). *Information Retrieval* (Second ed.). Butterworth-Heinemann Newton, MA, USA.

Vrontos, S. D., J. Galakis, and I. D. Vrontos (2021). Modeling and predicting US recessions using machine learning techniques. *International Journal of Forecasting 37*(2), 647–671.

Williams, C. K. (2021). The effect of class imbalance on Precision-Recall Curves. *Neural Computation 33*(4), 853–857.

Wright, J. H. (2006). The yield curve and predicting recessions. Finance and Economics Discussion Series 2006-07, Federal Reserve Board.

Yang, L., K. Lahiri, and A. Pagan (2023). Getting the ROC into Sync. *Journal of Business & Economic Statistics*, 1–13.

Yedidia, A. (2016). Against the F-score.

Zhou, X.-H., N. A. Obuchowski, and D. K. McClish (2011). Statistical Methods in Diagnostic Medicine, Second Edition, Chapter 2. Inc: New York: John Wiley & Sons.