JOURNAL OF APPLIED ECONOMETRICS
J. Appl. Econ. **32**: 359–378 (2017)
Published online 18 May 2016 in Wiley Online Library
(wileyonlinelibrary.com) DOI: 10.1002/jae.2525

SPOTTING THE DANGER ZONE: FORECASTING FINANCIAL CRISES WITH CLASSIFICATION TREE ENSEMBLES AND MANY PREDICTORS

FELIX WARD*
Department of Macroeconomics and Econometrics, University of Bonn, Germany

SUMMARY

This paper introduces classification tree ensembles (CTEs) to the banking crisis forecasting literature. I show that CTEs substantially improve out-of-sample forecasting performance over best-practice early-warning systems. CTEs enable policymakers to correctly forecast 80% of crises with a 20% probability of incorrectly forecasting a crisis. These findings are based on a long-run sample (1870–2011), and two broad post-1970 samples which together cover almost all known systemic banking crises. I show that the marked improvement in forecasting performance results from the combination of many classification trees into an ensemble, and the use of many predictors. Copyright © 2016 John Wiley & Sons, Ltd.

Received 16 February 2015; Revised 16 February 2016

Supporting information may be found in the online version of this article.

1. INTRODUCTION

The number of institutions whose explicit goal is to identify and address risks to the financial system has increased in the post-2008 economic policy landscape. In the USA, for example, the Financial Stability Oversight Council (FSOC) has been given a statutory mandate to 'identify risks and respond to emerging threats to financial stability'. [1] A core question policymakers in these institutions face is: where is the economy currently operating relative to the economic danger zones from which banking crises emanate? It is here that formal early-warning systems can make a valuable contribution.

This paper introduces classification tree ensembles (CTEs) (Breiman, 1996b, 2001) to financial crisis forecasting and analyzes their ability in making out-of-sample predictions for binary banking crisis indicators on the basis of several datasets: one long-run annual dataset (1870–2011), covering 17 developed countries, and two post-1970 datasets, the first covering 162 countries annually and the second quarterly. The results suggest that the out-of-sample forecasting performance of CTEs substantially surpasses current best-practice logit specifications. To give a concrete example of the trade-offs involved, the favorite CTE allows policymakers to correctly forecast about 50% of banking crises, at the cost of a 5% chance of wrongly forecasting a crisis when none will actually occur. The best-practice logit specification can achieve the same 50% rate of correct crisis forecasts only at the substantially higher cost of a 25% chance of making a wrong crisis call. If policymakers prefer a higher rate of correct crisis forecasting, both prior models offer one. The CTE can correctly forecast about 90% of banking crises, with a 30% probability of making a false crisis prediction. The best-practice logit specification can achieve the same 90% rate of correctly forecasting a crisis, only at the far higher

* Correspondence to: Felix Ward, Department of Macroeconomics and Econometrics, University of Bonn, Kaiserplatz 7–9, Bonn 53113, Germany. E-mail: s3feward@uni-bonn.de

[1] The citation is taken from the FSOC's website: http://www.treasury.gov/initiatives/fsoc/Pages/home.aspx.

Copyright © 2016 John Wiley & Sons, Ltd.

cost of an 80% chance of erroneously predicting a crisis. In both scenarios, the CTE offers the better trade-off. In light of their superior performance, CTE forecasts should become an important tool of macroprudential policy.

This paper relates to the existent literature in the following ways. It adds to the modern literature on early-warning systems (EWS) for banking crises, which was pioneered by Kaminsky (1998) and Kaminsky and Reinhart (1999) in the wake of the 1997 Southeast Asian crises. More recent contributions that analyze the predictability of banking crises in developed economies in the long run (since 1870) are: Schularick and Taylor (2012) and Jordà (2014), while others rely on post-1970 samples covering more countries (see Drehmann and Juselius, 2012; Drehmann, 2013). This literature has shown that already relatively simple model structures, based on few predictors—most notably credit aggregates—can convey valuable information on the imminence of a banking crisis. This paper will explore whether more complex classification tree structures, based on many predictors can improve upon this.

Thus this paper is related to the literature on economic forecasts based on many predictors (see Stock and Watson, 2002, 2006), which has stressed the possibility of improving forecasts by drawing from a large set of indicators. It will be demonstrated that increasing the number of predictors that are used in best-practice early-warning systems from 7–10 to ∼70–80 will markedly improve forecasts. The literature on economic forecasts based on many predictors has focused largely on factor modeling and prestep-dimensionality reduction techniques. Such approaches do not easily lend themselves to banking crisis forecasting. First, most banking crisis indicators are binary 0–1 dummies that require discrete classification techniques.[2] Furthermore, widely held beliefs on the genesis of banking crises, namely that they are characterized by discontinuous threshold effects and nonlinear interaction effects between several risk factors (see Duttagupta and Cashin, 2011), are more naturally accommodated by methods which dispense with linearity assumptions from the outset. This paper will apply classification tree structures (see Breiman *et al.*, 1984), which naturally accommodate discontinuous threshold effects as well as nonlinear interactions and can harness many predictors in doing so (see also Varian, 2014).

A classification tree can be seen as a recursive version of the more familiar signals approach to crisis forecasting. Similar to the signals approach, classification trees split a sample into two parts by searching for a predictor and a threshold along that predictor which separates the crisis observations in the sample from the non-crisis observations. Credit growth in the 90th percentile, for example, might be indicative of an impending banking crisis. After the sample has been split in two by the first threshold, the procedure is repeated for the two resulting subsamples—containing observations above and below the 90th percentile credit growth-threshold, respectively. In this way a sample can be recursively partitioned into crisis and non-crisis subsamples.[3] Individual classification trees, however, are renowned for being highly unstable, i.e. their high variance in mean-squared-error terms. This instability severely impairs their forecasting ability. To overcome this, Breiman (1996b) has suggested estimating many trees on many bootstrap samples and then aggregating them into a classification tree ensemble—or forest. This so-called bagging (short-hand for bootstrap aggregating) takes high-variance trees and combines them into low-variance forests, which retain the ability of individual trees to deal with many predictors and accommodate nonlinear threshold and interaction effects. Their ability to thus precisely delineate several danger zones, *and* their ability to harness many predictors in doing so, has already made them a mainstay in other research areas, such as genetics, where often thousands of genetic markers are analyzed with respect to their contributions to particular diseases (e.g. Díaz-Uriarte and De Andrés, 2006). Further examples for the wide applicability of tree-based ensemble methods come

[2] Exceptions are continuous crisis indices such as the exchange market pressure index pioneered by Eichengreen *et al.* (1994). Such indices are available for fewer countries and cover shorter time-spans than their binary counterparts.

[3] Recent contributions have already begun to explore the potential of classification trees for the analysis of banking crises (Davis and Karim, 2008; Duttagupta and Cashin, 2011).

Copyright © 2016 John Wiley & Sons, Ltd.

DOI: 10.1002/jae

10991255, 2017, 2, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/jae.2525 by Universitätsbibliothek Mannheim, Wiley Online Library on [03/01/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

10991255, 2017, 2, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/jae.2525 by Universitatsbibliothek Mannheim, Wiley Online Library on [03/01/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

from ecology (e.g. Prasad *et al.*, 2006), bioinformatics (e.g. Chen and Liu, 2005) and high-energy particle physics (Albert *et al.*, 2008). In parallel work, Alessi and Detken (2014) have also recently investigated the potential of CTEs for banking crisis forecasting.[4]

This paper is structured as follows. Section 2 provides an introduction to classification tree ensembles and Section 3 introduces the datasets. These datasets form the basis for the out-of-sample forecasting contest between CTEs and the best-practice logit specifications in Section 4. Section 5 concludes this paper by showing how a particular CTE, random forest, would have fared in forecasting the 2007/2008 financial crisis.

2. METHODOLOGY: CLASSIFICATION TREE ENSEMBLES

This section gives an introduction to classification trees and their ensembles (CTEs). It also contrasts the classification tree approach with the generalized linear models (GLM) framework, in order to clarify how classification trees differ from logit and probit models—the backbone of many current EWSs for banking crises.

2.1. Single Classification Trees

Classification trees separate crisis from non-crisis observations according to a set of discrete threshold rules. For instance, if an economy's private sector indebtedness exceeds a certain threshold, and GDP growth is faltering below another threshold, a classification tree might categorize the observation into the high-risk category. If, on the other hand, indebtedness was lower and GDP growth was higher, the observation might be categorized as low risk.

Figure 1 illustrates this idea graphically. x_1 and x_2 are two predictors conveying information about financial crisis risk. In the two-dimensional predictor space spanned by x_1 and x_2, black dots indicate crisis observations, while white dots stand for non-crisis observations. A classification tree is characterized by a partition of the predictor space into M non-overlapping regions R_m and an associated set of crisis probabilities p_m $(m = 1, \ldots, M)$. The regions are estimated through recursive partitioning—a step-wise algorithm. The algorithm will be described in more detail in the following section, but a short description is given here in order to provide intuition on how the region estimates $\widehat{R}_m$ come about: recursive partitioning searches across predictors for a threshold that separates crisis from non-crisis observations (see upper left panel of Figure 1). Next, the sample splitting continues on the obtained subsamples as indicated by the upper right and lower left panels of Figure 1. Once recursive partitioning stops, a crisis probability $\widehat{p}_m$ for region $\widehat{R}_m$ $(m = 1, \ldots, M)$ is estimated according to the fraction of crisis observations in that region:

$$\widehat{p}_m = \frac{\sum_{i \in \widehat{R}_m} y_i}{\sum_{i \in \widehat{R}_m} 1} \tag{1}$$

where $y_i = 1$ if a crisis occurs within the next 2 years, and $y_i = 0$ in all other cases. In the final partition depicted by the lower left panel of Figure 1, for instance, regions 3 and 5 delineate danger zones of high crisis risk. Regions 1–3, conversely, predict zero crisis risk. The partitioning of the

[4] Alessi and Detken's (2014) approach differs considerably from the approach followed here. They use a tree ensemble to identify important variables, and then estimate a single tree based on these variables. Single-tree forecasts allow for a better interpretation than ensemble-based forecasts. However, single-tree forecasts are plagued by high variance and thus are unlikely to be precise. Furthermore, identifying important variables with a CTE is problematic: several methods exist to determine the importance of a predictor in a CTE—they can produce very different results (see the series of papers published in *BMC Bioinformatics*: Nicodemus *et al.*, 2007; Strobl *et al.*, 2007, 2008; Nicodemus and Malley, 2009; Nicodemus *et al.*, 2010).

Copyright © 2016 John Wiley & Sons, Ltd.

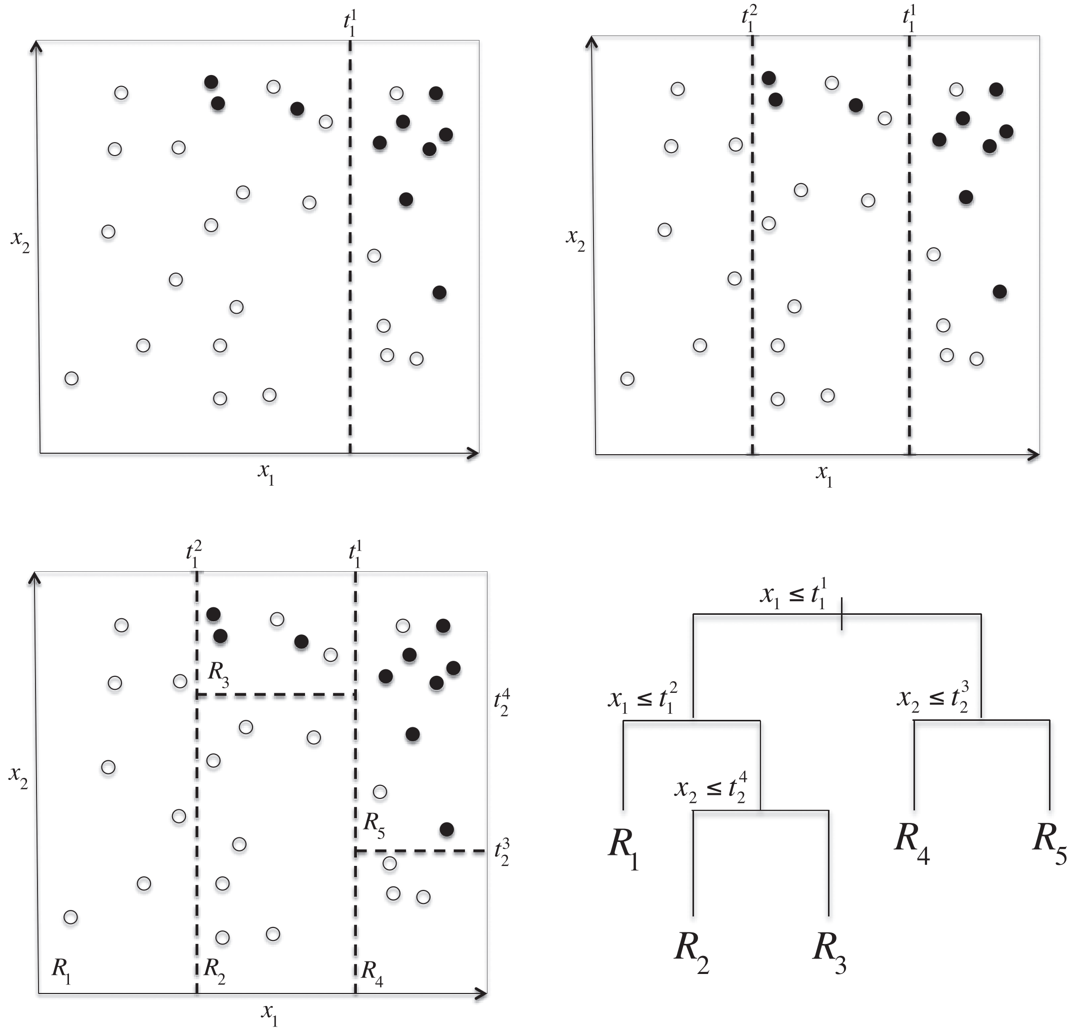

Figure 1. Recursive partitioning: an illustration. The upper left panel shows the first recursive partitioning step, while the upper right panel shows the second recursive partitioning step. The lower left panel shows a third and fourth partitioning step. The lower right panel shows the tree corresponding to the partition in the lower left panel. Filled circles, crisis; empty circles, no crisis; x_j, predictors; t_j^s, splits/thresholds; R_m, terminal regions

predictor space can also be represented as a dendrogram, in which the final nodes correspond to the final region estimates—hence the name classification *tree* (see lower right panel of Figure 1).

Formally, a classification tree predicts crisis probability as

$$\widehat{\mathcal{T}}(X_i) = \sum_{m=1}^{M} \widehat{p}_m I(X_i \in \widehat{R}_m) \tag{2}$$

where $\widehat{p}_m$ is the probability estimate of a crisis occurring within the following 2 years, X_i is a $J \times 1$ vector of predictor values, for observations $i = 1, \ldots, N$ and $I(X_i \in \widehat{R}_m)$ is an indicator function that equals 1 when region $\widehat{R}_m$ contains observation i.

10991255, 2017, 2, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/jae.2525 by Universitatsbibliothek Mannheim, Wiley Online Library on [03/01/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

A comparison to current workhorse specifications for banking crisis EWS, such as logit and probit models, will help to bring the characteristics of classification trees into sharper contrast. In the specification of a generalized linear model (GLM) with a binomially distributed dependent variable

$$P(y_i|X_i) = g^{-1}(X_i'\beta) \tag{3}$$

a dual set of assumption is made: first, in the form of a link function g (e.g. the logit link), whose inverse—the mean function—maps the linear predictor $X_i'\beta$ onto the [0, 1]-range; second, by assuming that the predictor values enter crisis risk only as a linear additive combination $X_i'\beta$.[5]

An advantage of classification trees is that they are more flexible in both regards. First, the nonparametric calculation of crisis probabilities according to equation (1) obviates the need to map an unbounded predictor range to the [0, 1]-crisis probability domain with the help of a particular mean function. Classification trees are thus free to approximate a multitude of functional forms between the dependent variable and the predictors through the combination of discrete thresholding rules.[6] Secondly, contrary to the GLM framework, classification trees are geared towards identifying nonlinear and discontinuous predictor interactions, while maintaining the ability also to approximate smooth and even linear relationships.[7] A downside of this flexibility is that globally optimal estimation of all the parameters that characterize a classification tree $\Theta = \{R_m, p_m\}_{m=1}^{M}$ constitutes an NP-complete problem (Hyafil and Rivest, 1976). Thus classification tree regions are typically estimated through recursive partitioning—a greedy search algorithm that conducts a stepwise locally optimal estimation.

2.2. Recursive Partitioning

At each recursive partitioning step a threshold, or split point, is selected in order to minimize a loss function. This loss function is the (negative) information gain $\mathcal{IG}$, which measures the extent to which a split point is successful in separating crisis from non-crisis observations. Evaluating the homogeneity of a region in terms of the crisis and non-crisis observations contained in it necessitates the definition of a measure of region impurity. Gini impurity—a parabolic function of the proportion of crisis observations p_a in region R_a—is such a measure:

$$\mathcal{GI} = -2p_a^2 + 2p_a$$

$\mathcal{GI}$ reaches minima of 0 in regions that contain only crisis observations ($p_a = 1$) or only non-crisis observations ($p_a = 0$). For $0 < p_a < 1$ $\mathcal{GI}$ exceeds 0 and reaches a maximum of 0.5 for regions that contain an equal amount of crisis and non-crisis observations ($p_a = 0.5$). As its name suggests, Gini impurity is thus a measure of region impurity that penalizes the mixing of crisis and non-crisis observations within a region. On the basis of this measure it is possible to define the loss function according to which split points are selected at each recursive partitioning step—the information gain:

$$\mathcal{IG}(R_a, R_b) = \mathcal{GI}(R_a \cup R_b) - 0.5\,[\mathcal{GI}(R_a) + \mathcal{GI}(R_b)]$$

[5] There exists, however, the possibility to explicitly define some interaction effects and higher-order terms and include them among the other predictors.

[6] The right sort and degree of functional flexibility depends on the problem at hand. Liu *et al.* (2004) present a set of conditions under which classification trees outperform artificial neural networks, although the latter are generally more flexible; the financial crisis forecasting problem seems to fit this set of conditions.

[7] To see this, imagine crisis and non-crisis observations were separated along one of the two linear diagonals in the panels of Figure 1. In this case, a good separation of crisis from non-crisis observations would necessitate the estimation of several more regions, but eventually a satisfying approximation to the diagonal separation could be achieved through a somewhat more finely granulated partitioning of the predictor space. Note, however, that smaller regions tend to contain fewer observations and the corresponding crisis probability estimates would be less precise.

Copyright © 2016 John Wiley & Sons, Ltd.

10991255, 2017, 2, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/jae.2525 by Universitätsbibliothek Mannheim, Wiley Online Library on [03/01/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

$\mathcal{IG}(R_a, R_b)$ compares the Gini impurity of a parent region $\mathcal{GI}(R_a \cup R_b)$ with the average Gini impurity of the two child regions $\mathcal{GI}(R_a)$ and $\mathcal{GI}(R_b)$ that a split point creates. The negative $\mathcal{IG}$ constitutes the loss function that is minimized at each recursive partitioning step $s = 2, \ldots, S$ through the choice of a splitting predictor j and a split point t along the range of that splitting predictor:

$$\hat{t}_j^s = \arg\max_{t_j^s} \mathcal{IG}\left(R_a^s\left(t_j^s|\hat{t}_j^1, \ldots, \hat{t}_j^{s-1}\right), R_b^s\left(t_j^s|\hat{t}_j^1, \ldots, \hat{t}_j^{s-1}\right)\right) \tag{4}$$

The thrust behind equation (4) is to estimate thresholds that separate crisis and non-crisis observations into different regions.[8] Note that only the first split $s = 1$ is an unconditional one; all others depend on all previously estimated splits $\hat{t}_j^1, \ldots, \hat{t}_j^{s-1}$.

Recursive partitioning can end in one of two ways: either running its course until the classification tree has been 'fully grown', i.e. only pure regions are left; or recursive partitioning can be ended through an ad hoc stopping rule. For example, each terminal region can be required to contain a minimum number of observations.[9] The final partition constitutes an estimate of the M terminal regions $\{\widehat{R}_m\}_{m=1}^M$, on the basis of which the classification tree (2) can be completed by estimating crisis probabilities $\{\widehat{p}_m\}_{m=1}^M$ according to equation (1). Algorithm 1 gives an overview of all the steps involved in estimating a classification tree.

Despite their ability to handle many predictors and accommodate nonlinear threshold and interaction effects, classification trees have been associated with poor out-of-sample forecasts of banking crises (see Davis and Karim, 2008). What is the reason for this? The most significant constraint that holds back the forecasting performance of a single classification tree is its high variance—an unwelcome side effect of recursive partitioning. Small changes in the sample under analysis can easily lead to changes in the early partitions and, owing to the dependence of later partitions on earlier ones, this change then reverberates throughout the tree. The results in Section 4 will confirm that this instability deals a severe blow to the forecasting performance of single classification trees. Fortunately, as explained in the next section, combining many classification trees into a CTE can provide a solution to this problem.

Algorithm 1 Classification tree pseudocode

```
1. Estimate regions $\{R_m\}_{m=1}^M$ through recursive partitioning:
repeat                                                    ▷ Recursive partitioning
    Select splitting predictor and split point according to (4).
until Stopping rule applies
⇒ Region estimates $\{\widehat{R}_m\}_{m=1}^M$:

2. Estimate crisis probabilities $\{p_m\}_{m=1}^M$ according to (1).
⇒ Classification tree $\widehat{\mathcal{T}}(X_i) = \sum_{m=1}^M \widehat{p}_m I(X_i \in \widehat{R}_m)$
```

Note: The pseudocode shows the steps involved in generating a single classification tree.

[8] Note that the stepwise estimation through recursive partitioning allows classification trees to make use of many predictors, whereas generalized linear models estimated through maximum likelihood would run into problems associated with the curse of dimensionality.

[9] Usually the application of such an ad hoc stopping rule is necessary to avoid poor out-of-sample predictions due to severe in-sample overfitting. The following analyses impose a lower bound of 10 observations on the terminal region size of single classification trees. The single-tree results are, however, robust to variations in the stopping rule.

Copyright © 2016 John Wiley & Sons, Ltd.

10991255, 2017, 2, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/jae.2525 by Universitatsbibliothek Mannheim, Wiley Online Library on [03/01/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

2.3. Classification Tree Ensembles

As the name suggests, an ensemble of trees—or forest $\mathcal{F}$—consists of many classification trees $\mathcal{T}_b$, $b = 1, \ldots, B$. Each individual tree 'grows' on an i.i.d. bootstrap sample X^b, for which N observations are drawn with replacement from the original data X. Such bootstrapping with subsequent aggregation is referred to as bagging (Breiman, 1996b).[10] If each tree is given the same weight, a forest's crisis probability estimate is

$$\widehat{\mathcal{F}}(X_i) = \frac{1}{B}\sum_{b=1}^{B}\widehat{\mathcal{T}}_b(X_i) \tag{5}$$

Thus the forest's prediction is simply the average prediction of the B single trees.[11]

Why are tree ensembles expected to have better predictive ability than individual classification trees? Consider the variance-bias decomposition of the mean squared error (MSE) of a tree:[12]

$$E\left[(y-\widehat{T}_b)^2\right] = \underbrace{E\left\{\left[\widehat{T}_b - E(\widehat{T}_b)\right]^2\right\}}_{:=\sigma^2_{\text{tree}}} + \underbrace{\left[E(\widehat{T}_b) - y\right]^2}_{:=\text{bias}^2_{\text{tree}}} \tag{6}$$

and a forest:

$$E\left[(y-\widehat{F})^2\right] = \underbrace{E\left\{\left[\widehat{F} - E(\widehat{F})\right]^2\right\}}_{:=\sigma^2_{\text{bag}}} + \underbrace{\left[E(\widehat{F}) - y\right]^2}_{:=\text{bias}^2_{\text{bag}}}$$

The main rationale behind bagging is its variance-reducing effect: the variance of the average of B identically—but not independently—distributed variables (note that trees are grown on overlapping bootstrap samples) is

$$\sigma^2_{\text{bag}} = \rho\sigma^2_{\text{tree}} + \frac{1-\rho}{B}\sigma^2_{\text{tree}}$$

where ρ is the pairwise correlation between any two trees[13] and thus

$$\sigma^2_{\text{bag}} \leq \sigma^2_{\text{tree}}$$

Hence the variance of a tree ensemble can generally be expected to be lower than the variance of an individual tree (see Bühlmann and Yu, 2002; Buja and Stuetzle, 2006), with $\rho\sigma^2_{\text{tree}}$ constituting the lower bound on variance that can be reached through bagging.

[10] Note that the use of the bootstrap methodology in bagging is somewhat unusual, in that it is not used for statistical inference here. Hence the choice of the i.i.d. bootstrap is harmless at the bagging stage. However, temporal and cross-sectional dependencies in the data presumably resurface later in the form of temporally and cross-sectionally dependent crisis probability estimates. Therefore at a later stage, for the evaluation and comparison of the models' predictive ability on the basis of their crisis probability estimates, block-bootstrap procedures become important for robustifying confidence intervals and statistical tests (see the online Appendix, provided as supporting information).

[11] CTEs appear to be rather unaffected by tree growth-stopping rules (Segal, 2004). Fully growing each tree in an ensemble has consequently established itself as a standard and has therefore been applied to the following analysis.

[12] For ease of clarification, the following argument assumes fixed predictors and thus abstracts from population MSE, which is in any case beyond the control of forecasters.

[13] Note the exchangeability assumption needed in the derivation of this expression: $\text{cov}(T_i, T_j) = \text{cov}(T_1, T_2)$ for any $i \neq j$.

Copyright © 2016 John Wiley & Sons, Ltd.

10991255, 2017, 2, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/jae.2525 by Universitätsbibliothek Mannheim, Wiley Online Library on [03/01/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

One of the most prominent ensemble algorithms that will be investigated in the following sections is random forest (Breiman, 2001), which aims at lowering ensemble variance σ^2_{bag} even further, by lowering ρ—or the correlation between trees. This is done by considering only a random subset J_{try} of all J predictors in the maximization problem faced at each recursive partitioning step (4). This subset of predictors is drawn without replacement from the set of all predictors.[14] Individual trees thus no longer vary only with respect to the bootstrap sample on which they are 'grown', but also with respect to the selection of splitting predictors. The trees become less similar (i.e. ρ is lower) and, all else equal, ensemble variance decreases.[15] Besides this, random forest equals bagging:

$$\widehat{\mathcal{RF}}(X_i) = \frac{1}{B}\sum_{b=1}^{B} \widehat{\mathcal{T}_b^{\text{RF}}}(X_i) \tag{7}$$

A concise overview of all the steps involved in obtaining the bagging (5) and random forest crisis probabilities (7) can be found in Algorithm 2.

Note that while the branches and leaves of an individual tree can still be traced and interpreted in economic terms, one drawback of CTEs is that they lose this straightforward interpretability. In this sense CTE-based forecasts might be thought of as a complement to methodologies that are better suited to the evaluation of the relative importance of different risk factors.[16]

Algorithm 2 Random forest pseudocode

```
for $b = 1$ to $B$ do                                                  ▷ Estimate $B$ classification trees
    1. Draw i.i.d. bootstrap sample $X^b$ of size $N$ (with replacement).

    2. Estimate regions $\{R^b_m\}^M_{m=1}$ through recursive partitioning:
    repeat                                                             ▷ Recursive partitioning
        (a) Draw $J_{\text{try}}$ predictors (without replacement).
        (b) Select splitting predictor and split point according to (4).
    until Stopping rule applies (i.e. trees are fully grown)
    $\Rightarrow$ Region estimates $\{\widehat{R}^b_m\}^M_{m=1}$:

    3. Estimate crisis probabilities $\{p^b_m\}^M_{m=1}$ according to (1).
    $\Rightarrow$ classification tree $\widehat{\mathcal{T}_b^{\text{RF}}}(X_i) = \sum_{m=1}^{M} \widehat{p}^b_m I(X_i \in \widehat{R}^b_m)$
end for
$\Rightarrow$ $B$ classification trees $\{\widehat{\mathcal{T}_b^{\text{RF}}}(X_i)\}^B_{b=1}$

$\Rightarrow$ Random forest $\widehat{\mathcal{RF}}(X_i) = \frac{1}{B}\sum_{b=1}^{B} \widehat{\mathcal{T}_b^{\text{RF}}}(X_i)$              ▷ Generate ensemble
```

Note: The pseudocode shows the steps involved in generating the $\mathcal{RF}$ ensemble (7). This pseudocode also covers the generation of the $\mathcal{F}$ ensemble (5) if the number of randomly drawn predictors J_{try} in step 2 (a) is set equal to the total number of predictors J. In this case, the $\mathcal{F}$- and the $\mathcal{RF}$ ensembles become equal.

[14] A widely used default choice for J_{try}, which the following analysis will adhere to, is $\lfloor\sqrt{J}\rfloor$ (see Breiman, 2002). Generally, the paper makes use of default settings sourced from the literature for fine-tuning parameters (J_{try} and tree growth stopping rules). Given the rarity of severe banking crises setting aside part of the data for tuning considerations (i.e. validation data) is an excessive strain on the samples.

[15] Bagging and randomization also have countervailing effects on prediction bias. Averaging across many trees smooths out the discontinuities found in any single tree. This can lower CTE bias (Buja and Stuetzle, 2006). However, each bootstrap sample leaves out ∼37% of observations, which increases finite sample bias compared to larger resamples.

[16] Although several methods exist to determine the importance of a predictor in a CTE, for the applications that follow I find few commonalities between the predictor rankings produced by two of the most common variable importance measures (see online Appendix for results and further discussion).

Copyright © 2016 John Wiley & Sons, Ltd.

10991255, 2017, 2, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/jae.2525 by Universitätsbibliothek Mannheim, Wiley Online Library on [03/01/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

3. DATA

This section introduces three datasets on the basis of which I will evaluate the forecasting performance of logit models, single classification trees and CTEs. Systemic banking crises are rare. Their statistical analysis necessitates datasets that cover large time-spans or many countries—one usually comes at the cost of the other. Therefore I make use of one long-run sample spanning from 1870 to 2011, as well as two post-1970 samples with broader country coverage.

3.1. The Long-Run Sample, 1870–2011

With regard to the long-run sample, this study utilizes the dataset introduced by Schularick and Taylor (2012). After further extensions by Jordà *et al.* (2013), this dataset now ranges from 1870 to 2011 and covers 17 countries. Usually, these countries cumulatively constitute more than half of the world's GDP (according to Maddison GDP estimates).

The dataset features macroeconomic indicators (GDP, consumption, investment, consumer prices, current account and exchange rates) as well as financial indicators (bank loans, total bank assets, stock prices, interest rates, public debt and monetary aggregates). These are the base indicators from which ∼70 predictors are derived (see Table A1 in the online data Appendix). The bare nominal series (n) are utilized when they are deemed to be of interest with respect to crisis risk (e.g. nominal interest rates). CPI-deflated quantities, growth rates (gr), trend deviations (gap), to GDP ratios (/GDP), global (GDP-weighted) averages (glo), real exchange rates and interest rate differentials are also obtained (see Alessi and Detken, 2011, for a similar approach). Furthermore, to obtain an even more detailed snapshot of economic conditions several of these transformations are combined when it makes economic sense, e.g. the gap of the loans-to-GDP ratio (loans/GDP (gap)). Schularick and Taylor (2012) also provide a binary banking crisis indicator, the definition of which follows Laeven and Valencia (2008): the indicator takes a value of 1 for years characterized by bank runs, a jump in default rates and large capital losses associated with public interventions as well as bankruptcies or forced mergers of major financial institutions. Otherwise the indicator takes a value of 0. Overall, the dataset contains 93 systemic banking crises (for the country–year incidence of crises see the crisis map (Figure A1) in the online data Appendix).

3.2. The Broad Post-1970 Samples

For the post-1970 period, this paper makes use of the binary banking crisis indicator provided by Laeven and Valencia (2013). This indicator encompasses 162 countries and 147 systemic banking crises between the years 1970 and 2011 (for the country–year incidence of crises see the crisis map (Figure A2) in the online data Appendix).[17]

Next, annual and quarterly base indicators from the IMF IFS database and Datastream were obtained. When selecting base indicators, it was paramount to consider their availability across a wide range of countries, as a multitude of missing values would further endanger the already small number of financial crises. The annual indicators include consumer prices, net exports, exchange rates, bank loans, stock prices, interest rates and public debt (provided by Abbas *et al.*, 2013). The quarterly indicators include GDP, consumer prices, exchange rates, bank loans, stock prices, house prices, interest rates, foreign liabilities and reserves. For the post-1970 annual sample, further use is made of the GDP, consumption and investment series from the Penn World Tables (Feenstra *et al.*, 2013), as well as the public debt-to-GDP ratios from Abbas *et al.* (2013). A detailed list of all the predictors can be found in the online data Appendix (see Tables A2 and A3). An overview of the characteristics of all datasets is given in Table I.

[17] For the quarterly dataset, the quarterly crisis dummy was set to 1 for all quarters, if the year dummy was 1. This is also the case if a financial crisis began later in the year.

Copyright © 2016 John Wiley & Sons, Ltd.

10991255, 2017, 2, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/jae.2525 by Universitätsbibliothek Mannheim, Wiley Online Library on [03/01/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Table I. Datasets

Dataset	Long-run 1870–2011 sample	Broad post-1970 sample I	Broad post-1970 sample II
Base indicators	Schularick and Taylor (2012)	IFS, PWT, Abbas *et al.* (2013)	IFS, Datastream
Crisis dummy	Schularick and Taylor (2012)	Laeven and Valencia (2013)	Laeven and Valencia (2013)
Frequency	Annual	Annual	Quarterly
Time-span	1870–2011	1970–2011	1970–2011
No. of countries	17	162	162
No. of predictors	77 (Table A1)	70 (Table A2)	73 (Table A3)
No. of crises	93	147	147
N	2414	7081	30,967

Note: N, number of observations; IFS, International Financial Statistics; PWT, Penn World Tables. The Schularick and Taylor (2012) dataset has subsequently been extended and updated (see Jordà *et al.*, 2013). All three datasets are unbalanced. The number of observations and crises will vary across applications.

4. PERFORMANCE COMPARISON

This section stages the competition between logit models, single classification trees and CTEs. The rules are simple: the method whose crisis probability predictions achieve the highest out-of-sample area under the receiver operating characteristic curve (AUC) wins.[18] The following paragraph gives a short introduction to the AUC measure

Each crisis forecasting model faces a true positive rate (TPR)–false positive rate (FPR) trade-off. At one extreme, the model could make a crisis call for each period, thus correctly predicting all crises (100% TPR). However, this comes at the price of never correctly giving the all-clear (100% FPR). At the other extreme, a model could never issue a crisis warning, and thus be correct for all non-crisis periods (0% FPR) at the cost of never correctly predicting a crisis (0% TPR). Crisis probability estimates can be translated into crisis calls or all-clears depending on whether crisis probability passes a certain threshold $\eta \in [0, 1]$. For different thresholds different TPR–FPR combinations are obtained. By slowly shifting the threshold η from 0 to 1 all of the TPR–FPR combinations that a model is capable of can be depicted in the TPF–FPR plane (a unit square). The resulting curve is the ROC curve, which gives a comprehensive description of a model's predictive ability. The area under this curve (AUC) is a slightly more aggregate measure, upon which most of the following model comparisons will be based. The AUC ranges from 0.5 to 1. An AUC of 1 indicates a perfect EWS, which correctly forecasts all crises as crises, and all non-crises as non-crises. An AUC of 0.5 indicates an entirely uninformative EWS. The corresponding ROC curve is a diagonal in the TPR–FPR plane: a higher TPR only comes at the cost of an equally higher FPR. Intuitively, the AUC represents the probability that, for a randomly selected pair of one crisis and one non-crisis observation, the crisis probability estimate for the crisis observation is higher than that for the non-crisis observation. For a comprehensive introduction to the ROC curve and the AUC in the context of financial crisis forecasting see Jordà (2014).

All model evaluations are based on out-of-sample data. For the CTEs, so-called out-of-bag (OOB) data are used (see Breiman, 1996a). A tree's OOB data are those ∼37% of observations that are not contained in the bootstrap sample on which this tree was estimated. Correspondingly, each observation constitutes OOB data to ∼37% of the trees in an ensemble. Out-of-sample crisis probability estimates are obtained by evaluating each observation by only those trees in an ensemble for which it constitutes OOB data. For single classification trees and the logit models, this section conducts Monte Carlo cross-validation (MCCV) evaluations that are comparable to the OOB evaluations. For instance, 100 logit models are estimated based upon 100 bootstrap samples (drawn with replacement).

[18] The receiver operating characteristic (ROC) curve and the AUC are useful for the evaluation of predictive performance in classification problems where one class constitutes a minority class (e.g. banking crises). Under such circumstances, many other criteria tend to inflate the predictive ability of models that blindly predict the majority class.

Copyright © 2016 John Wiley & Sons, Ltd.

10991255, 2017, 2, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/jae.2525 by Universitatsbibliothek Mannheim, Wiley Online Library on [03/01/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

With the observations not contained in the bootstrap samples 100 out-of-sample AUCs are calculated and their average constitutes the MCCV estimate. The MCCV-AUC is comparable to the OOB-AUC in that both are estimates of expected out-of-sample performance ($\text{AUC} = E(\text{AUC}_{\mathcal{T}})$) as opposed to conditional out-of-sample performance—i.e. performance conditional on a particular training dataset $\mathcal{T}(\text{AUC}_{\mathcal{T}})$ (see Hastie *et al.*, 2013, pp. 254–257).

4.1. Logit EWS

To obtain a yardstick against which to measure the performance of CTEs, this subsection first reports logistic regression-based results. Bi- and multivariate logit models were estimated on the basis of a selection of predictors, which are comparable to those found in the literature.

Among the single predictors the largest AUCs come from the private burden (AUC = 0.64) and the loans/GDP gap (AUC = 0.63). They are significantly different from 0.5 at the 1% significance level.[19] The public debt/GDP gap (AUC = 0.59) and the public burden (AUC = 0.58) achieve significance at higher levels. Most of the other AUC estimates hover closely above 0.5—a rather poor result. Generally, these results are similar to those obtained by Jordà (2014), who, based on comparable specifications, reports AUCs ranging from 0.52 to 0.67.[20]

Next are multivariate specifications. The variable selections are displayed on the right-hand side of Table II. They are inspired by similar specifications in Schularick and Taylor (2012) and Jordà *et al.* (2011). AUCs of all three multivariate models are significantly different from 0.5 at the 1% significance level. They range from 0.62 to 0.65.

Compared to the baseline specification, the IA specification with interaction terms is successful in conveying extra information on the imminence of a banking crisis (AUC = 0.65). The AUC remains the same after the additional inclusion of country fixed effects. These results are very close to the out-of-sample results reported by Schularick and Taylor (2012) (AUC = 0.646), which are based on similar logit specifications and data.

4.2. Classification Tree-Based EWS

CTEs are not just an ensemble of trees but also an ensemble of techniques. To obtain an impression of the relative efficacy of bagging, randomization and the use of many predictors, the following analysis will build up to the final $\mathcal{RF}$ model one step at a time. First, a single classification tree, based on the same restricted selection of 10 predictors as the IA logit model, will be presented, before bagging and randomization is added to the recipe. After that, the same three steps—(i) single tree, (ii) bagging, (iii) randomization—will be analyzed on the basis of the broader set of 76 predictors.

4.2.1. Single Tree

The left-hand side of Table II displays results for the restricted predictor selection. Here, a single tree performs badly (AUC = 0.55). This confirms similar findings by Davis and Karim (2008). When put in terms of the MSE equation (6), a likely explanation for this is the high variance of single classification trees. Estimation through recursive partitioning makes them highly susceptible to small changes in finite training samples.

[19] The reported results hold up when confidence bands and tests are robustified against serial and cross-sectional correlation in the crisis probability estimates (see online Appendix).

[20] On the basis of 19 systemic crises (11 of which are associated with the most recent global financial crisis) Drehmann and Juselius (2013) report mean AUC estimates between 0.8 and 0.9 for their logit specifications. These high AUC estimates may hint at important country and time specificities in the development of financial crises. The soon to be introduced CTEs also enter the 0.8–0.9 range of AUC estimates, but on the basis of a more diverse set of banking crises ($\geq$ **70**). This will allow forecasters to predict banking crises, which resemble crises from the more distant past or crises from less similar countries.

Copyright © 2016 John Wiley & Sons, Ltd.

10991255, 2017, 2, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/jae.2525 by Universitätsbibliothek Mannheim, Wiley Online Library on [03/01/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Table II. Logit EWS

	Results				Specification		
	AUC	95% CI	N		Baseline	IA	FE & IA
Bivariate				*Variable*			
Loans/GDP (gap)	0.63**	[0.55, 0.71]	1283	Loans/GDP (gap)	✓	✓	✓
Public debt/GDP (gap)	0.58*	[0.51, 0.66]	1347	Public debt/GDP (gap)	✓	✓	✓
Narrow money/GDP (gap)	0.55	[0.48, 0.63]	1308	Narrow money/GDP (gap)	✓	✓	✓
LT interest rate	0.52	[0.45,0.6]	1425	LT interest rate	✓	✓	✓
GDP (gr)	0.52	[0.44, 0.6]	1402	GDP (gr)	✓	✓	✓
Inflation	0.54	[0.47,0.61]	1492	Inflation	✓	✓	✓
Exchange rate (gap)	0.51	[0.44,0.59]	1502	Exchange rate (gap)	✓	✓	✓
Public burden	0.57†	[0.5, 0.65]	1321				
Private burden	0.64**	[0.56, 0.72]	1223	*Interaction terms*			
Joint burden	0.52	[0.44, 0.61]	1180	Public burden		✓	✓
				Private burden		✓	✓
Multivariate				Joint burden		✓	✓
Baseline	0.62**	[0.55, 0.7]	1144	Loans/GDP × GDP (gr)		✓	✓
IA	0.65**	[0.57, 0.73]	1144	Public debt/GDP × GDP (gr)		✓	✓
FE & IA	0.65**	[0.57, 0.73]	1144	Loans/GDP (gap) × Exchange rate (gap)		✓	✓
				Fixed effects			
				Country-FE			✓

Note: Dependent variable: 2-year horizon before crisis. Out-of-sample AUC and confidence band estimates are based on Monte Carlo cross-validation (see Picard and Cook, 1984; Arlot and Celisse, 2010): 100 MC draws of training (63.2%)—test (36.8%) data partitions. IA, interaction terms; FE, country fixed effects; N, number of training observations ($= 0.632\times$ total number of observations).
$^\dagger p < 0.10$; $^* p < 0.05$; $^{**} p < 0.01$.

10991255, 2017, 2, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/jae.2525 by Universitätsbibliothek Mannheim, Wiley Online Library on [03/01/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

4.2.2. Bagging

Indeed, the dramatic improvement in forecasting performance for an ensemble made up of many trees over an individual tree appears to confirm that tree variance was to blame for an individual tree's bad performance. The second row in the upper left quadrant of Table III displays the effect of bagging in the 10-variable setting. The AUC leaps by more than 0.2 to a value of 0.77. This AUC is significantly higher than that displayed by the FE & IA-logit model.

4.2.3. Random Forest

The third model in the upper left quadrant of Table III is the $\mathcal{RF}$ estimator. The additional randomization, in the form of randomly analyzing only three out of the 10 predictors at each recursive partitioning step, leads to a slightly higher mean AUC estimate of 0.79. CTEs have already left behind their logistic competitors without yet having capitalized on their ability to make forecasts with far more predictors.

4.2.4. Many Predictors

I now turn to the more predictor-intensive contenders. The results are displayed in the upper right quadrant of Table III. The extension of the list of predictors to a total of 76 results in a second significant leap in forecasting performance, by about 0.1 for the $\mathcal{F}$ (AUC = 0.87) and $\mathcal{RF}$ (AUC = 0.88) estimators. Even the single classification tree (AUC = 0.63) now performs similarly to the multivariate logit EWS. In summation, the combination of many classification trees into an ensemble and the making use of many predictors result in marked improvements in banking crisis forecasts.

Table III. $\mathcal{CT}$ EWS: long-run 1870–2012 sample

	Restricted selection			Many predictors		
Results						
Model	AUC	95% CI	N	AUC	95% CI	N
Single tree	0.55	[0.5,0.6]	1816	0.63§	[0.57,0.7]	1742
Bagging	**0.77**	[0.73, 0.81]	1816	**0.87**§	[0.84,0.9]	1742
Random forest	**0.79**	[0.75, 0.83]	1816	**0.88**§	[0.85,0.91]	1742
Specification						
Parameter	Single	Bagging	RF	Single	Bagging	RF
B	1	5000	5000	1	5000	5000
J_{try}	10	10	3	76	76	9
J		10			76	
No. of crises		72			70	

Note: Dependent variable: 2-year horizon before crisis. Restricted selection: loans/GDP (gap), public debt/GDP (gap), narrow money/GDP (gap), LT interest rate, GDP (gr), inflation, exchange rate (gap), loans/GDP, public debt/GDP, LT interest rate (n). Many predictors: see Table A1 in the online data Appendix. For single tree: out-of-sample mean AUC and confidence band estimates are based on Monte Carlo cross-validation (see Picard and Cook, 1984; Arlot and Celisse, 2010); 100 MC draws of training (63.2%)—test (36.8%) data partitions. For ensembles: out-of-sample AUC estimates (and confidence intervals) are based on out-of-bag (OOB) data (see Breiman, 1996a). N, number of observations; J, number of predictors under analysis; J_{try} number of predictors randomly selected and considered as a splitting variable at each recursive partitioning step; B, number of trees. Specification table: if there is only a single entry in the bagging column, this means that all models share the same specification. § H_0: $\text{AUC}_{\text{many}} - \text{AUC}_{\text{restricted}} = 0$. Bold: H_0: $\text{AUC} - \text{AUC}_{\text{logitFE\&IA}} (= 0.65) = 0$.

10991255, 2017, 2, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/jae.2525 by Universitätsbibliothek Mannheim, Wiley Online Library on [03/01/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

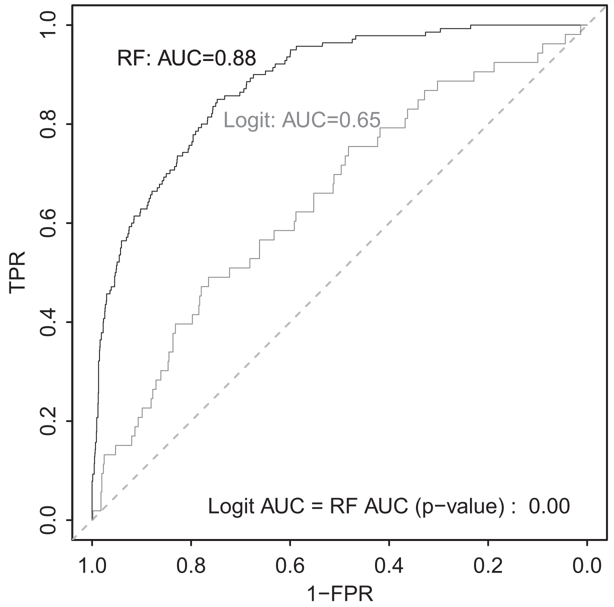

Figure 2. ROC comparison. Receiver operating characteristic curves of the logit model with country FE and interaction terms (grey) and $\mathcal{RF}$ model (black). The p-value corresponds to a test for equality of AUCs according to DeLong *et al.* (1988). TPR, true positive rate; FPR, false positive rate

4.3. ROC Comparison

Depending on how much weight policymakers put on making correct crisis calls as opposed to correct all-clears, the logit EWS might be the preferable EWS after all. To see whether this is the case, Figure 2 displays the ROC curves of both EWSs. The $\mathcal{RF}$ model offers the higher TPR for any given FPR—equivalently, the $\mathcal{RF}$ model offers a lower FPR for any given TPR. Evidently, regardless of policymakers' preferences, the $\mathcal{RF}$ EWS offers the better TPR–FPR trade-off.

In order to get a better indication of $\mathcal{RF}$ EWS's performance, some of the exemplary TPR–FPR combinations from Figure 2 should be studied. The $\mathcal{RF}$ EWS offers a balanced TPR–FPR trade-off at about $\text{TPR} = (1 - \text{FPR}) = 0.80$, i.e. it enables policymakers to correctly forecast 80% of crises and 80% of non-crises. However, if a 20% probability of mistakenly forecasting a crisis is deemed too high by policymakers, the $\mathcal{RF}$ EWS allows for a reduction of the probability of mistakenly forecasting a crisis to 5%, while still correctly forecasting about 50% of banking crises. At the other extreme, policymakers eager not to miss any crisis could use the $\mathcal{RF}$ estimator to correctly forecast 95% of crises. This will, however, result in only correctly predicting about 40% of non-crises. Any of these trade-offs leaves policymakers substantially better off than when using the logit EWS.

4.4. Robustness

In order to check whether the main results generalize, the analysis is repeated for the annual and quarterly post-1970 samples, emerging market and good-quality data subsamples, a different crisis dummy as well as 1-year and 3-year pre-crisis horizons. In an effort to save space and to counteract repetitiveness, Table IV presents an abbreviated analysis that only reports the results for the random forest models.[21]

[21] See the online robustness Appendix for an extended analysis of the annual and quarterly post-1970 samples.

Copyright © 2016 John Wiley & Sons, Ltd.

10991255, 2017, 2, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/jae.2525 by Universitatsbibliothek Mannheim, Wiley Online Library on [03/01/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Table IV. Robustness: various $\mathcal{RF}$ EWS

	Restricted selection			Many predictors		
Model	AUC	95% CI	N [crises]	AUC	95% CI	N [crises]
Long-run 1870–2011 dataset						
1-year horizon	0.63	[0.56, 0.69]	1,816 [71]	**0.79**§	[0.73,0.85]	1742 [70]
2-year horizon	**0.79**	[0.75, 0.83]	1,816 [70]	**0.88**§	[0.86, 0.91]	1,742 [70]
3-year horizon	**0.82**	[0.8, 0.85]	1,816 [69]	**0.89**§	[0.87, 0.91]	1,742 [69]
RR dummy	**0.76**	[0.72, 0.8]	1,800 [84]	**0.86**§	[0.83, 0.89]	1,726 [84]
Post-1970 annual dataset						
1-year horizon	0.64	[0.58, 0.69]	4,465 [103]	**0.74**§	[0.7, 0.79]	4,373 [103]
2-year horizon	**0.78**	[0.75, 0.81]	4,465 [102]	**0.85**§	[0.83, 0.87]	4,373 [102]
3-year horizon	**0.8**	[0.77, 0.82]	4,465 [102]	**0.88**§	[0.87, 0.9]	4,373 [102]
Emerging markets	**0.77**	[0.72, 0.82]	823 [33]	**0.82**	[0.78, 0.87]	804 [33]
Quality data	**0.8**	[0.76, 0.84]	3,325 [77]	**0.86**§	[0.84, 0.89]	3,256 [77]
Post-1970 quarterly dataset						
1-year horizon	**0.84**	[0.82, 0.86]	19,126 [104]	**0.93**§	[0.92, 0.94]	19,061 [104]
2-year horizon	**0.85**	[0.84, 0.86]	19,126 [102]	**0.95**§	[0.95, 0.96]	19,061 [102]
3-year horizon	**0.84**	[0.83, 0.85]	19,126 [101]	**0.96**§	[0.95, 0.96]	19,061 [101]
Q4 only	**0.77**	[0.74, 0.8]	4,820 [103]	**0.83**§	[0.8, 0.85]	4,800 [103]
Emerging markets	**0.88**	[0.86, 0.89]	3,110 [30]	**0.95**§	[0.94, 0.96]	3,110 [30]
Quality data	**0.86**	[0.84, 0.87]	15,033 [84]	**0.96**§	[0.95, 0.96]	15,010 [84]

Note: Dependent variable: 1/2/3-year horizon before crisis. $^{\S}\boldsymbol{H_0}$: $\boldsymbol{AUC_{\text{many}}} - \text{AUC}_{\text{restricted}} = \mathbf{0}$. Bold: $\boldsymbol{H_0}$: AUC $-$ AUC$_{\text{logitFE\&IA}} = \mathbf{0}$. All tests at the 5% significance level. Long-run sample restricted selection: loans/GDP (gap), public debt/GDP (gap), narrow money/GDP (gap), LT interest rate, GDP (gr), inflation, exchange rate (gap), loans/GDP, public debt/GDP, LT interest rate (n). Long-run sample many predictors: see Table A1 in the online data Appendix. Annual post-1970 sample restricted selection: loans/GDP (gap), public debt/GDP (gap), GDP (gap), inflation, real exchange rate (gap), loans/GDP, public debt/GDP, net exports/GDP (gap). Annual post-1970 sample many predictors: see Table A2 in the online data Appendix. Quarterly post-1970 sample restricted selection: loans (gap), loans (gr), foreign liabilities (gap)(glo), LT interest rate (gap)(glo), GDP (gap)(glo), inflation, exchange rate (gap), reserves (gap), GDP (gr)(glo). Quarterly post-1970 sample many predictors: see Table A3 in the online data Appendix.

Concerning the 1-, 2- and 3-year pre-crisis horizons, AUCs generally increase with the length of the horizon. This implies that it is harder to assess whether there will be a crisis next year than to assess whether there will be a crisis within the next few years. This conforms to accounts which picture banking crisis risks as building up slowly over time. At the same time, the actual crisis realization is less determinate—usually triggered by a shock, which may or may not occur in any particular year.

EWSs may provide different results for different banking crisis dummies. In order to investigate whether the high AUCs are specific to the banking crisis dummy by Schularick and Taylor (2012), Table IV displays AUCs obtained for the banking crisis dummy by Reinhart and Rogoff (2010) (RR dummy). The mean AUC estimates for the RR dummy are only marginally lower, otherwise the core results hold: the CTE model outperforms the logit model based on the same set of predictors, and the inclusion of many predictors significantly improves forecasts.

For the post-1970s datasets it is possible to look at emerging markets (EM) subsamples.[22] Despite the fact that the EM subsamples contain only about a third of the crisis events from the full sample, the EM AUC estimates are remarkably similar to the baseline results. Note, however, that in the annual post-1970 EM sample the inclusion of many predictors no longer significantly improves the AUC.

[22] The EM subsample consists of Argentina, Brazil, Bulgaria, Chile, China, Colombia, Costa Rica, Croatia, Ecuador, Egypt, Hungary, India, Indonesia, Lithuania, Malaysia, Mexico, Morocco, Panama, Papua New Guinea, Paraguay, Peru, Philippines, Poland, Romania, Thailand, Turkey and Vietnam.

Copyright © 2016 John Wiley & Sons, Ltd.

10991255, 2017, 2, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/jae.2525 by Universitatsbibliothek Mannheim, Wiley Online Library on [03/01/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Excluding countries of presumably poor data quality[23] reduces the number of crises by about one-fifth (from approximately 100 to around 80). AUCs tend to improve marginally by about 0.01. The fact that the exclusion of noisy data and several outliers does not significantly alter the results highlights the robustness of the recursive partitioning algorithm.[24]

Finally, AUCs for the quarterly dataset are systematically higher than for the two annual samples, particularly in the case with many predictors. This is not due to the slightly different set of predictors contained in the quarterly sample. This can be seen in the 'Q4 only' row of Table IV, which reports the AUCs for a random forest model estimated only with the 4th quarter observations from the quarterly dataset. In this case, AUC estimates converge with those for the two annual samples. CTEs seem to thrive on the larger number of observations contained in the quarterly sample.

In summary, the core results hold up very well: the CTE-based EWSs yield significantly higher AUCs than the logit alternative, and the inclusion of many predictors further improves the accuracy of banking crisis predictions.

5. CASE STUDY: 2007/2008

To round out this study, the performance of the $\mathcal{RF}$ EWS based on many predictors is compared with the IA-logit EWS in forecasting the 2007/2008 global financial crisis. Both EWSs are estimated using the long-run sample, where only the data up to 1997 are incorporated and the rest of the data are used as test data. The resulting crisis probability estimates for the test data (1998–2011) are reported in Figure 3.

It is immediately clear that the $\mathcal{RF}$ crisis risk evaluation exhibits considerably more variation than the logit model. For most countries it would have signaled a build-up in crisis risk in the mid 2000s. Thus the $\mathcal{RF}$ model would have signaled rather clearly that the developed world as a whole was embarking upon a path that historically has often ended in crisis. The evidence for the logit model is less flattering. While for some countries it signals a (slightly) higher crisis risk, for others it signals no big changes or even shows an increasing resilience during the 2000s.

The $\mathcal{RF}$ model produces mixed results with respect to the country-specific incidence of the 2007/2008 crisis. For the following countries crisis risk went up and a crisis did indeed occur: Belgium, Switzerland, Denmark, Spain, France, UK, Italy, Netherlands, Portugal, Sweden and the USA. Although the $\mathcal{RF}$ crisis risk is upward trending for all of these countries, its level is relatively low for some, namely Switzerland and the USA. Germany, for which crisis risk does not even trend upwards, also exhibits a very low risk level. How can these cases be explained? What brought down German and Swiss banks was their exposure to foreign assets. For the USA, non-bank intermediation was at the heart of its banking crisis. Neither exposure to foreign assets nor non-bank intermediation is well reflected by any of the base indicators in the long-run sample. Extending the list of base indicators may help improve forecasts.

Several countries show clear signs of being in a danger zone prior to 2007/2008, but did not experience a systemic banking crisis according to the binary indicator: Australia, Canada, Finland, Norway. There is a notable concentration of Scandinavian countries and primary good exporters in this group. Hardy and Pazarbasioglu (1998) show that primary-product exporting countries possess a distinct set of early-warning indicators, which might explain the poor performance of the $\mathcal{RF}$ EWS in these cases.

[23] The sample excludes Afghanistan, Angola, Benin, Burkina Faso, Burundi, Cambodia, Cape Verde, Central African Republic, Chad, Comoros, Democratic Republic of Congo, Côte d'Ivoire, Djibouti, El Salvador, Eritrea, Ethiopia, Fiji, The Gambia, Grenada, Guinea-Bissau, Haiti, Lao People's Democratic Republic, Liberia, Libya, Mali, Mauritania, Mozambique, Myanmar, Niger, Nigeria, Rwanda, Sierra Leone, Swaziland, Syria, Timor-Leste, Togo, Uganda, Yemen and Zambia.

[24] Recall that recursive partitioning is robust to outliers since extreme values do not influence the internally optimal split points. Noise resilience also appears to make CTEs outperform one of their most prominent competitors—boosting—whose out-of-sample AUC estimates appear to be held back by the level of noise in macroeconomic data (see also Long and Servedio, 2010) (see the online Appendix for results and further discussion of boosting-based EWS).

Copyright © 2016 John Wiley & Sons, Ltd.

10991255, 2017, 2, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/jae.2525 by Universitatsbibliothek Mannheim, Wiley Online Library on [03/01/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

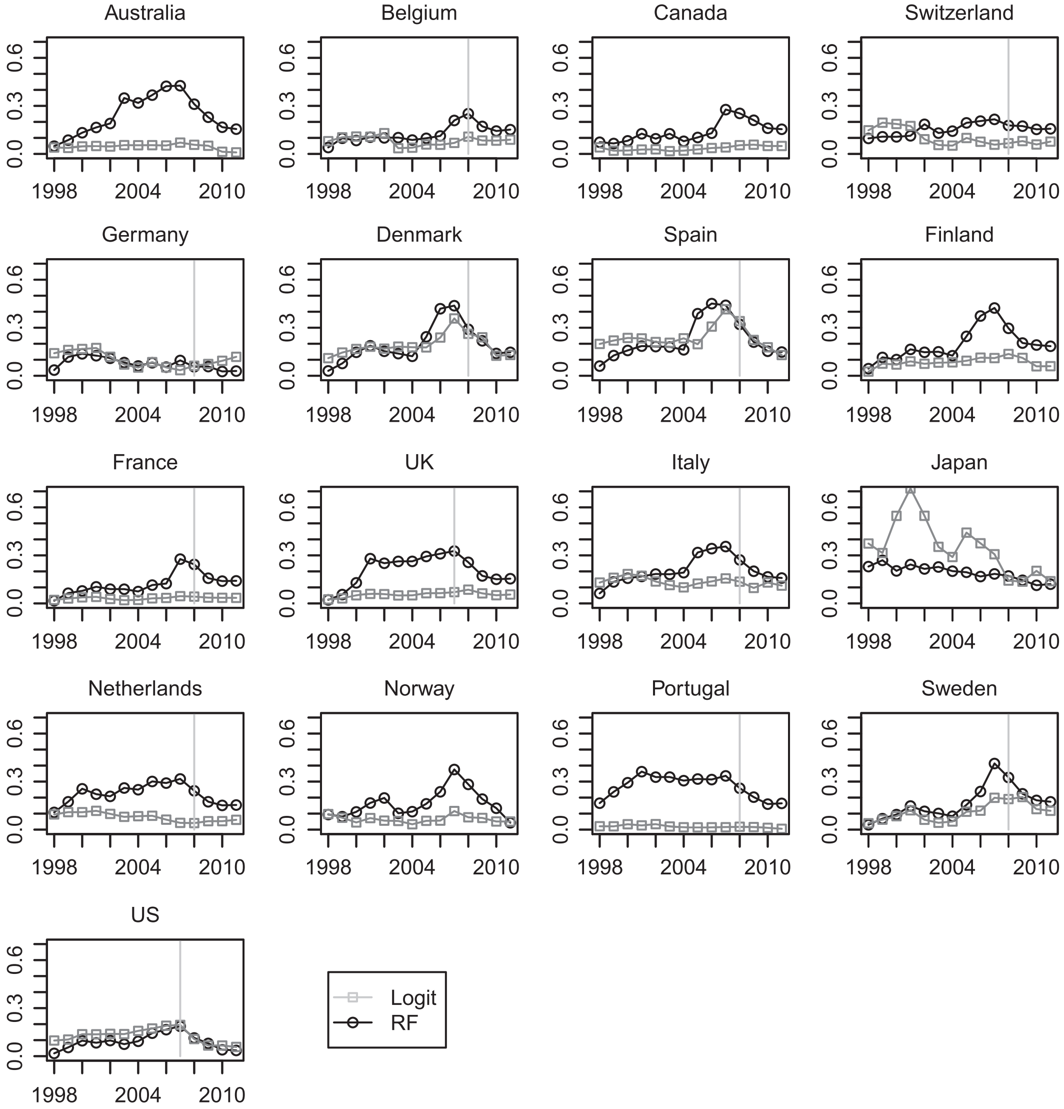

Figure 3. The 2007/2008 global financial crisis. 1998–2011 out-of-sample probability estimates of being in the 2-year horizon before a banking crisis for 17 countries. Ten country–year observations between 1998 and 2011 exhibit missing values and were replaced by the respective variable’s mean to obtain a probability estimate. Vertical gray bars indicate years with a systemic banking crisis

What is also interesting is that, excluding Canada, all of these countries experienced a banking crisis in the late 1980s or early 1990s. The ensuing institutional changes might have rendered their banking systems more resilient 20 years later. Giannone *et al.* (2010) present related evidence for the importance of regulatory quality in credit markets for explaining cross-country differences in weathering the global recession. Also note that in Australia and Norway banking systems did in fact come under considerable stress during the relevant period—they are knife-edge cases with respect to the dummy categorization that was applied.

The last group consists of countries that did not see their risk profiles rise and, indeed, did not experience a systemic event: Japan is the only country in this category.

Copyright © 2016 John Wiley & Sons, Ltd.

10991255, 2017, 2, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/jae.2525 by Universitatsbibliothek Mannheim, Wiley Online Library on [03/01/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Formalizing these observations into conditional[25] out-of-sample AUC measures yields AUCs of 0.53 and 0.76 for the logit and $\mathcal{RF}$ model, respectively. The difference is statistically significant (p-value = 0.00),[26] while the logit AUC does not significantly differ from the uninformative 0.5 at the 5% significance level.

In summary, the results of the $\mathcal{RF}$ model on the most recent crisis is mixed. While the model would not have performed as convincingly with respect to the country-specific incidence of the crisis, it would have clearly signaled that the developed world as a whole was on a dangerous path from the early 2000s on. The first part of this conclusion nicely mirrors results reported by Claessens *et al.* (2010), Rose and Spiegel (2010a,b, 2012), who find that prior to the global financial crisis hardly any predictor conveyed reliable information about the crisis' subsequent cross-country severity. While Rose and Spiegel (2012) continue to argue that their results warrant skepticism towards the potential of EWS to accurately predict a crisis, the analysis provided above suggests a different conclusion. Although even CTE EWS would have found predicting the 2007/2008 crises somewhat more difficult than their historical track record suggests, their use is still generally very promising. Also note that the evaluation of the $\mathcal{RF}$ EWS's performance in 2007/2008 depends on the categorization of two knife-edge cases. Given a more lenient evaluation of these cases (Australia and Norway), Figure 3 shows that even in terms of cross-country incidence for 2007/2008 the $\mathcal{RF}$ predictor did not perform badly. Especially if combined with country-specific knowledge, as exemplified above, the proposed $\mathcal{RF}$ EWS would have given policymakers a valuable warning as to the vulnerability of the world financial system prior to the crisis.

6. CONCLUSION

This paper explored the potential of classification tree ensembles (CTEs) for forecasting binary banking crisis indicators. Their out-of-sample performance surpasses current best-practice early-warning systems that are based on logit models, by a substantial margin. The good forecasting performance of CTEs contrasts with the poor performance of single classification trees. However, the combination of many classification trees into an ensemble on the one hand, and the making use of many predictors on the other, result in an EWS that has the potential to provide policymakers with a substantially more accurate assessment of banking crisis risk than current alternatives.

ACKNOWLEDGEMENTS

The author wishes to thank without implicating Daniel Becker, Jörg Breitung, Yao Chen, Mathias Drehmann, Òscar Jòrda, Alois Kneipp, Keith Küster, Gernot Müller, Moritz Schularick, Heiko Wagner and Liu Yang. The manuscript also benefited from helpful comments by four anonymous referees. All remaining errors are the sole responsibility of the author.

REFERENCES

Abbas SM, Belhocine N, El-Ganainy A, Horton M. 2013. A historical public debt database. IMF working paper.

Albert J, Aliu E, Anderhub H, Antoranz P, Armada A, Asensio M, Baixeras C, Barrio JA, Bartko H, Bastieri D, Becker J, Bednarek W, Berger K, Bigongiari C, Biland A, Bock RK, Bordas P, Bosch-Ramon V, Bretz T, Britvitch I, Camara M, Carmona E, Chilingarian A, Ciprini S, Coarasa JA, Commichau S, Contreras JL, Cortina J, Costado MT, Curtef V, Danielyan V, Dazzi F, De Angelis A, Delgado C, de los Reyes R, De Lotto B, Domingo-Santamaría E. 2008. Implementation of the random forest method for the imaging atmospheric cherenkov telescope MAGIC. *Nuclear Instruments and Methods in Physics Research, Section A* **588**: 424–432.

[25] 'Conditional', as it is used here, refers to the fact that results are conditional on this particular 1998–2011 test dataset.

[26] The p-value corresponds to the test for equality of AUCs by DeLong *et al.* (1988).

Copyright © 2016 John Wiley & Sons, Ltd.

10991255, 2017, 2, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/jae.2525 by Universitatsbibliothek Mannheim, Wiley Online Library on [03/01/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Alessi L, Detken C. 2011. Quasi real time early warning indicators for costly asset price boom/bust cycles: a role for global liquidity. *European Journal of Political Economy* **27**(3): 520–533.

Alessi L, Detken C. 2014. Identifying excessive credit growth and leverage. Working paper. European Central Bank.

Arlot S, Celisse A. 2010. A survey of cross-validation procedures for model selection. *Statistics Surveys* **4**: 40–79.

Breiman L. 1996a. Out-of-bag estimation. Statistics Department, University of California, Berkeley, CA.

Breiman L. 1996b. Bagging predictors. *Machine Learning* **24**(2): 123–140.

Breiman L. 2001. Random forests. *Machine Learning* **45**(1): 5–32.

Breiman L. 2002. Manual on setting up, using, and understanding random forests v3. 1. Statistics Department, University of California, Berkeley, CA.

Breiman L, Friedman J, Stone CJ, Olshen RA. 1984. *Classification and Regression Trees*. Chapman & Hall: London.

Bühlmann P, Yu B. 2002. Analyzing bagging. *Annals of Statistics* **30**(4): 927–961.

Buja A, Stuetzle W. 2006. Observations on bagging. *Statistica Sinica* **16**: 323–351.

Chen XW, Liu M. 2005. Prediction of protein-?protein interactions using random decision forest framework. *Bioinformatics* **21**(24): 4394–4400.

Claessens S, Dell'Ariccia G, Igan D, Laeven L. 2010. Cross-country experiences and policy implications from the global financial crisis. *Economic Policy* **25**(62): 267–293.

Davis EP, Karim D. 2008. Comparing early warning systems for banking crises. *Journal of Financial Stability* **4**(2): 89–120.

DeLong ER, DeLong DM, Clarke-Pearson DL. 1988. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics* **44**(3): 837–845.

Díaz-Uriarte R, De Andrés SA. 2006. Gene selection and classification of microarray data using random forest. *BMC Bioinformatics* **7**: 3.

Drehmann M. 2013. Total credit as an early warning indicator for systemic banking crises. *BIS Quarterly Review* June.

Drehmann M, Juselius M. 2012. Do debt service costs affect macroeconomic and financial stability? *BIS Quarterly Review* September.

Drehmann M, Juselius M. 2013. Evaluating early warning indicators of banking crises: satisfying policy requirements. BIS Working Paper No. 421.

Duttagupta R, Cashin P. 2011. Anatomy of banking crises in developing and emerging market countries. *Journal of International Money and Finance* **30**(2): 354–376.

Eichengreen B, Rose AK, Wyplosz C. 1994. Speculative attacks on pegged exchange rates: an empirical exploration with special reference to the European monetary system. Working paper, National Bureau of Economic Research.

Feenstra RC, Inklaar R, Timmer M. 2013. The next generation of the Penn World Table. *American Economic Review* **105**(10): 3150–3182.

Giannone D, Lenza M, Reichlin L. 2010. Market freedom and the global recession. *IMF Economic Review* **59**: 111–135.

Hardy DCL, Pazarbasioglu C. 1998. Leading indicators of banking crises: was Asia different? Working paper. International Monetary Fund.

Hastie T, Tibshirani R, Friedman J. 2013. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer: Berlin.

Hyafil L, Rivest RL. 1976. Constructing optimal binary decision trees is NP- complete. *Information Processing Letters* **5**(1): 15–17.

Jordà Ò. 2014. Assessing the historical role of credit: business cycles, financial crises and the legacy of Charles S. Peirce. *International Journal of Forecasting* **30**(3): 729–740.

Jordà Ò, Schularick M, Taylor AM. 2011. Financial crises, credit booms, and external imbalances: 140 years of lessons. *IMF Economic Review* **59**(2): 340–378.

Jordà Ò, Schularick M, Taylor AM. 2013. Sovereigns versus banks: credit, crises and consequences. Working paper. NBER.

Kaminsky GL. 1998. Currency and banking crises: the early warnings of distress. Working paper. International Monetary Fund.

Kaminsky GL, Reinhart CM. 1999. The twin crises: the causes of banking and balance-of-payments problems. *American Economic Review* **89**(3): 473–500.

Laeven L, Valencia F. 2008. Systemic banking crises: a new database. Working paper. International Monetary Fund.

Laeven L, Valencia F. 2013. Systemic banking crises database. *IMF Economic Review* **61**: 225–270.

Copyright © 2016 John Wiley & Sons, Ltd.

10991255, 2017, 2, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/jae.2525 by Universitatsbibliothek Mannheim, Wiley Online Library on [03/01/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Liu X, Bowyer KW, Hall LO. 2004. Decision trees work better than feed- forward back-prop neural nets for a specific class of problems. *IEEE International Conference on Systems, Man and Cybernetics*: The Hauge.

Long PM, Servedio RA. 2010. Random classification noise defeats all convex potential boosters. *Machine Learning* **78**(3): 287–304.

Nicodemus KK, Malley JD. 2009. Predictor correlation impacts machine learning algorithms: implications for genomic studies. *Bioinformatics* **25**(15): 1884–1890.

Nicodemus K, Wang W, Shugart Y. 2007. Stability of variable importance scores and rankings using statistical learning tools on single-nucleotide polymorphisms and risk factors involved in gene? gene and gene? environment interactions. *BMC Proceedings* **1**(Suppl 1): S58.

Nicodemus K, Malley J, Strobl C, Ziegler A. 2010. The behaviour of random forest permutation-based variable importance measures under predictor correlation. *BMC Bioinformatics* **11**: 110.

Picard RR, Cook RD. 1984. Cross-validation of regression models. *Journal of the American Statistical Association* **79**(387): 575–583.

Prasad AM, Iverson LR, Liaw A. 2006. Newer classification and regression tree techniques: bagging and random forests for ecological prediction. *Ecosystems* **9**(2): 181–199.

Reinhart CM, Rogoff KS. 2010. From financial crash to debt crisis. Working paper. National Bureau of Economic Research.

Rose AK, Spiegel MM. 2010a. Cross-country causes and consequences of the crisis: an update. *European Economic Review* **55**: 309–324.

Rose AK, Spiegel MM. 2010b. Cross-country causes and consequences of the 2008 crisis: international linkages and American exposure. *Pacific Economic Review* **15**(3): 340–363.

Rose AK, Spiegel MM. 2012. Cross-country causes and consequences of the 2008 crisis: early warning. *Japan and the World Economy* **24**(1): 1–16.

Schularick M, Taylor AM. 2012. Credit booms gone bust: monetary policy, leverage cycles, and financial crises, 1870–2008. *American Economic Review* **102**(2): 1029–1061.

Segal MR. 2004. Machine learning benchmarks and random forest regression. Working paper. Center for Bioinformatics and Molecular Biostatistics.

Stock JH, Watson MW. 2002. Forecasting using principal components from a large number of predictors. *Journal of the American Statistical Association* **97**(460): 1167–1179.

Stock JH, Watson MW. 2006. Forecasting with many predictors. In *Handbook of Economic Forecasting*, Elliott G, Granger C W J, Timmermann A (eds). Elsevier: Amsterdam; 515–554.

Strobl C, Boulesteix AL, Zeileis A, Hothorn T. 2007. Bias in random forest variable importance measures: illustrations, sources and a solution. *BMC Bioinformatics* **8**: 25.

Strobl C, Boulesteix A-L, Kneib T, Augustin T, Zeileis A. 2008. Conditional variable importance for random forests. *BMC Bioinformatics* **9**: 307.

Varian HR. 2014. Big data: new tricks for econometrics. *Journal of Economic Perspectives* **28**(2): 3–28.

Copyright © 2016 John Wiley & Sons, Ltd.

10991255, 2017, 2, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/jae.2525 by Universitätsbibliothek Mannheim, Wiley Online Library on [03/01/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License