Does Machine Learning Help us Predict Banking Crises? [*]

Johannes Beutel
Deutsche Bundesbank

Sophia List
Deutsche Bundesbank

Gregor von Schweinitz
Leipzig University and
Halle Institute for Economic Research,
Member of the Leibniz Association

Abstract

This paper compares the out-of-sample predictive performance of different early warning models for systemic banking crises using a sample of advanced economies covering the past 45 years. We compare a benchmark logit approach to several machine learning approaches recently proposed in the literature. We find that while machine learning methods often attain a very high in-sample fit, they are outperformed by the logit approach in recursive out-of-sample evaluations. This result is robust to the choice of performance metric, crisis definition, preference parameter, and sample length, as well as to using different sets of variables and data transformations. Thus, our paper suggests that further enhancements to machine learning early warning models are needed before they are able to offer a substantial value-added for predicting systemic banking crises. Conventional logit models appear to use the available information already fairly efficiently, and would for instance have been able to predict the 2007/2008 financial crisis out-of-sample for many countries. In line with economic intuition, these models identify credit expansions, asset price booms and external imbalances as key predictors of systemic banking crises.

Keywords: Early warning system; logit; machine learning; systemic banking crises

JEL classification: C35; C53; G01

[*] Contact address: Gregor von Schweinitz, Halle Institute for Economic Research and Leipzig University. E-mail: gregorvon.schweinitz@iwh-halle.de. We thank Marisa Basten, Tobias Berg, Mathias Drehmann, Daniel Fricke, Christoph Memmel, Norbert Metiu, Peter Sarlin, Christian Schumacher, Makram El Shagi, Felix Thierfelder, Benjamin Weigert as well as the anonymous referees and seminar participants at the Bundesbank for many helpful comments. Research of Gregor von Schweinitz was supported by DFG grant GZ:SCHW 1919/1-1. This paper represents the authors' personal opinions and does not necessarily reflect the views of the Deutsche Bundesbank or its staff.

1 Introduction

The global financial crisis has spurred a new wave of research on the importance of a stable financial system for macroeconomic stability. New early warning models for financial crises have been developed and are being employed by central banks to monitor the stability of the financial system and to guide macroprudential policy (see, for example European Central Bank, 2010, 2017; Drehmann and Juselius, 2014). Given the high costs associated with financial crises, it is important to understand the circumstances under which countries are likely to experience them and to provide accurate early warning signals of these events. Failing to activate macroprudential policy tools in time might lead to large costs for taxpayers, policymakers, and society as a whole, while issuing false alarms might lead to costly over-regulation of the financial system.[1]

Recently, early warning models that rely on machine learning methods have been proposed as an alternative to the traditionally employed methods in this field, such as the signaling approach (e.g. Kaminsky and Reinhart, 1999; Knedlik and von Schweinitz, 2012) and discrete choice (probit or logit) models (e.g. Frankel and Rose, 1996; Lo Duca and Peltonen, 2013). For instance, Alessi and Detken (2018) as well as Tanaka, Kinkyo, and Hamori (2016) have argued that random forests may improve early warning predictions in comparison to the logit model and the signaling approach. Holopainen and Sarlin (2017) have extended this argument to at least four other machine learning methods, namely artificial neural networks, support vector machines, k-nearest-neighbors, and decision trees.

Using a comprehensive dataset encompassing systemic banking crises for 15 advanced economies over the past 45 years, we compare the out-of-sample prediction accuracies of the logit model to four machine learning methods employed in the existing literature (random forest, support vector machines, k-nearest neighbors, and decision trees). We come to an interesting and perhaps surprising conclusion: simple logit models systematically outperform all machine learning methods considered under a large variety of circumstances. In particular, we show that, while machine learning methods are able to achieve near perfect in-sample fit, they perform worse than the logit model in recursive out-of-sample prediction, and often even worse than a naïve benchmark. This result is remarkable, as it cautions against the use of machine learning methods whose impressive in-sample performance may backfire in the context of actual out-of-sample forecasting situations.

We subject our key result to a variety of tests. First, we document the superiority of logit models for different combinations of leading indicator variables as well as for different measures of prediction accuracy. Second, we perform standard robustness checks, such as different data transformations, crisis databases, estimation periods, and parameterizations. Finally, we propose a bootstrap as a uniform approach to account for estimation uncertainty, allowing us to establish statistically significant differences in performance between methods. Moreover, we seek to determine ex ante optimal hyperparameters for machine learning methods using a computationally intensive re-sampling procedure (a specific variant of cross-validation). Even with this considerable effort, machine learning methods still generate out-of-sample predictions which are inferior to those of the logit

[1] The costs of financial crises are documented, for instance, in Jordà, Schularick, and Taylor (2011), and Laeven and Valencia (2013). An overview of internationally employed macroprudential policy tools can be found in Lim, Costa, Columba, Kongsamut, Otani, Saiyid, Wezel, and Wu (2011), Cerutti, Claessens, and Laeven (2017) or Claessens (2015).

model.

We suggest an explanation for this result and compare our findings to other studies that use machine learning methods for predicting financial crises. Machine learning methods typically contain a much larger number of parameters than the logit model and are able to flexibly approximate a large space of functions. This allows them to fit in-sample data quite closely, but, at the same time, entails the risk of an overfit, and as a consequence weak out-of-sample performance. We provide empirical and theoretical arguments to show that this risk appears to materialize in the early warning context.

Our paper is related to the new wave of research on early warning models spurred by the global financial crisis of 2008, as, for instance, in Alessi and Detken (2011); Rose and Spiegel (2012); Gourinchas and Obstfeld (2012); Lo Duca and Peltonen (2013); Drehmann and Juselius (2014). These papers construct different early warning models but do not consider machine learning methods or horse races between different methods. Random forests are introduced by Alessi and Detken (2018) in early warning models of systemic banking crises at the country level, and by Tanaka et al. (2016) and Tanaka, Kinkyo, and Hamori (2018) to predict failures at the level of individual banks. These papers evaluate the predictive ability of their models using cross-validation. While cross-validation has appealing features relative to more traditional out-of-sample evaluation, it also suffers from serious drawbacks. As recognized by Holopainen and Sarlin (2017), cross-validation estimates of performance can be inflated and biased towards more complex machine learning methods if (as in the early warning context) cross-sectional and serial correlation are strong features of the data. Related to this, Neunhoeffer and Sternberg (2018) show that performance of machine learning methods has been seriously over-estimated in published articles of the political science literature as a result of using cross-validation for both hyperparameter selection and model evaluation.

By contrast, Holopainen and Sarlin (2017) run out-of-sample comparisons of several methods. Yet, they do so on a dataset containing a relatively small number of crisis episodes. We build on their pioneering work, but refine it in several important ways, namely regarding our careful construction of datasets and robustness checks, as well as our bootstrap and hyperparameter selection schemes, taking into account cross-sectional and serial dependence structures. We show that our out-of-sample results differ with respect to their paper and provide an explanation for this difference.

Our paper complements recent assessments of machine learning methods in other fields, for instance Neunhoeffer and Sternberg (2018) in the context of civil war prediction in the political science literature. We thereby seek to contribute to a realistic assessment of the strengths and limitations of the various methods, and to stimulate further research in this area.

2 Methodology

2.1 Estimation

In line with the recent early warning literature (see Drehmann and Juselius, 2014; Alessi and Detken, 2018; Holopainen and Sarlin, 2017) we estimate the probability of a financial crisis starting between the next 5 to 12 quarters (conditional on not already being in an acute crisis period) based on a set of potential early warning indicators. Details on this window forecasting approach and the resulting definition of the dependent variable for the estimations are given in Appendix A.1.

We employ the following methods for estimating crisis probabilities: Logistic regression, k-nearest neighbors, decision trees, random forests, support vector machines and neural networks. Appendix A.3 provides an overview of these methods as well as details on their implementation. The selection of methods follows the previous literature (Berg and Pattillo, 1999; Bussière and Fratzscher, 2006; Alessi and Detken, 2018; Holopainen and Sarlin, 2017).

While binary choice models such as the logit are standard tools in the early warning literature, machine learning methods are sometimes thought to allow for stronger non-linearities and more flexible distributional assumptions, which might be beneficial when forecasting extreme events such as systemic banking crises.[2] We have a panel dataset with observations for several countries at different points in time. In order to treat methods uniformly and keep the setup parsimonious, we estimate each method on the same pooled sample of observations. That is, observations are pooled in the cross-section and time dimension and we do not include fixed effects in any of the models.

The machine learning methods come with hyperparameters that have to be set exogenously prior to estimation. For example, the k-nearest neighbor method has a single exogenous hyperparameter, k, determining the number of neighbors to consider. Following the standard in the literature (James, Witten, Hastie, and Tibshirani, 2013; Murphy, 2012), hyperparameters for all methods are chosen such that they optimize a performance criterion (relative usefulness as described in Section 2.2) in a cross-validation exercise. We have implemented a fairly sophisticated cross-validation algorithm taking into account the cross-sectional and serial correlation present in our dataset (see Appendix A.4 for details). As a consequence, cases where the performance of machine learning methods falls short of the logit approach cannot be easily attributed to sub-optimal hyperparameters, but appear to be more deeply rooted in the given model. To ensure a strict separation between in-sample and out-of-sample data, our cross-validation routine uses information before the start of the out-of-sample window only (i.e. data before 2005Q3). A list of optimized hyperparameters can be found in Table A.3 in the Appendix.

2.2 Evaluation of Predictions

For every observation, the early warning models estimate the probability of a crisis starting in the following five to twelve quarters. These probabilities can be transformed into binary early-warning signals using an (optimized) threshold. The performance of an early warning model can therefore be evaluated either with respect to signals or probabilities. We employ four different performance measures that are standard in the literature. Relative usefulness (U_r) uses a preference parameter μ to balance type-1 (missed crises) and type-2 (false alarms) errors of binary signals. It is zero for a naïve forecast, and increases

[2] Other commonly mentioned advantages of machine learning methods relate to potential benefits on large datasets ("big data"), and their ability to deal with a large number of potentially relevant variables. In the context of early warning models however, datasets typically contain only a limited number of observations. Moreover, the amount of variables can generally still be challenging for some machine learning methods as can be seen in our discussion of methods' properties in Appendix A.3 and in our empirical results 4.2).

for better forecasts. The F-measure (F_1), as an alternative, relates the number of correctly predicted crises to the number of erroneous predictions, ignoring correctly predicted tranquil periods (Powers, 2011). It is zero for forecasts where no crisis if correctly predicted, and increases for better forecasts. The area under the curve (AUC) measures the relationship between type-1 and type-2 errors at all possible signalling thresholds. It is 0.5 for a naïve forecast, and also increases for better forecasts. The Brier probability score is comparable to the RMSE of a linear regression, and thus lower values are preferable. The AUC and BPS only depend on predicted probabilities, and are therefore independent of additional (preference) parameters used to derive the threshold for relative usefulness and the F-measure. Details on the transformation of crisis probabilities into signals, and the four performance measures are given in Appendix A.2.

We deliberately do not use cross-validation to evaluate models. As Neunhoeffer and Sternberg (2018) show, using cross-validation both for hyperparameter selection and model evaluation may lead to serious over-estimation of model performance. Thus, we use cross-validation only for hyperparameter selection and perform a classic out-of-sample prediction experiment to evaluate models. To this end, we split the panel dataset into two distinct parts: estimations (and hyperparameter selection) are performed on an in-sample part (the *training sample*), while predictions and performance evaluations are derived on an out-of-sample part (the *test sample*). For comparability with previous findings, we follow Holopainen and Sarlin (2017) in setting our out-of-sample window to the period between 2005Q3 and 2016Q4. This leads to a good balance between observations available for estimation and for evaluating predictions, with approximately half of the pre-crisis observations contained in the in-sample part and half of the pre-crisis observations contained in the out-of-sample part.

In most of the paper, we focus on recursive out-of-sample estimations where we predict the crisis probability quarter-by-quarter between 2005Q3 and 2016Q4 based on the information that was available in each respective quarter.[3] The performance measures are then based on the recursive predictions for the out-of-sample part of the dataset. If, instead, we are interested in in-sample performance, we use the same dataset for estimation and performance evaluation, i.e. we set the test and training sample equal to the full sample.

2.3 Bootstrap

Several of our estimation methods do not readily come with measures of estimation uncertainty. Moreover, even if such measures can be derived, they are conditional on very different (distributional) assumptions for different methods, making a comparison difficult. We solve this problem by bootstrapping, which provides a straightforward approach for calculating measures of estimation uncertainty under identical assumptions for all estimation methods. This allows us to test whether differences between model performances are statistically significant.

[3] The definition of the early warning window is forward-looking. In order to account for that, all observations where the dependent variable is yet unknown given information at time t have to be excluded from the training sample. That is, for a forecast made in 2006Q1 we can only estimate the model on observations until 2003Q1 (unless a crisis occurs between 2003Q1 and 2006Q1, in which case the realization of the binary early-warning variable is known for some additional periods).

Bootstrapped measures of estimation uncertainty can be derived from the dispersion of estimates across random variations of the original dataset. These measures of estimation uncertainty are conditional on the statistical properties of the bootstrap datasets. Therefore, it is important to construct bootstrap datasets that preserve those statistical properties of the original dataset that are likely to affect the precision of estimates. In our case, autocorrelation and cross-sectional correlation are strong features of the data. Based on El-Shagi, Knedlik, and von Schweinitz (2013) and Holopainen and Sarlin (2017), we use a panel-block-bootstrap to account for these properties, as described in Appendix A.5 in more detail.

3 Data

3.1 Crisis Variable

We use the database for systemic banking crises established by the European System of Central Banks (ESCB) and the European Systemic Risk Board (ESRB) covering European countries from 1970 to 2016 (Lo Duca, Koban, Basten, Bengtsson, Klaus, Kusmierczyk, Lang, Detken, and Peltonen, 2017). This latest database refines previous crisis databases, both with respect to the identification of events and their timing. Crises are identified by the following two-step procedure. In a first step, "systemic financial stress events" are identified using the quantitative methodology of Duprey, Klaus, and Peltonen (2017). These financial stress events together with additional crises identified in previous databases (Laeven and Valencia, 2013; Babecký, Havránek, Matějů, Rusnák, Šmídková, and Vašíček, 2014; Detken, Weeken, Alessi, Bonfim, Boucinha, Frontczak, Giordana, Giese, Jahn, Kakes, Klaus, Lang, Puzanova, and Welz, 2014) form a list of potential crisis events. In the second step, this list of potential crises is checked against a set of qualitative criteria defining systemic financial crises (see Lo Duca et al., 2017, for details).

Following Drehmann and Juselius (2014), we focus on systemic banking crises with at least partially domestic origins.[4] Furthermore, we expand the coverage of the crisis database to include two additional (non-European) advanced countries with important crisis experience, namely Japan and the United States.[5] As a result, our dataset covers all of the "big five" crises identified by Reinhart and Rogoff (2008). The full list of crisis episodes used in our analysis after taking into account the availability of the explanatory variables may be found in Table B.1 in the Appendix. It includes 19 crises for European countries (of which 11 take place before 2008) as well as three crisis events in the United States and Japan (of which two take place before 2008). The majority of countries are included for a time period starting in the early to mid-1970s until the beginning of 2016.

As a robustness check we also run an estimation with crisis dates taken from the most recent version of the Laeven and Valencia (2018) database. Their database has the

[4] Focusing on crises with at least partially domestic origins makes sense, as our modeling framework (where domestic variables determine the crisis probability of each country) does not allow for cross-country spillover effects. That is, we know a priori that these events are largely unforeseeable given the present modeling framework.

[5] For these countries, we use the crisis episodes identified by Laeven and Valencia (2018) adapting start and end dates such that they are consistent with the definition in our core database.

advantage of being somewhat more agnostic in its definition of crisis events. Yet, the most recent European crisis database, which we use for our core results, provides a more comprehensive and more precise account of the crises in the European countries of our sample.

3.2 Early Warning Indicators

We use a total of ten explanatory variables, capturing key economic channels affecting the likelihood of systemic banking crises as identified in the literature, while balancing data availability.[6] The channels we focus on are (i) asset prices, namely *house prices* and *equity prices*, (ii) credit developments (*total credit to the private non-financial sector relative to GDP*), (iii) the macroeconomic environment, as measured by *GDP*, *gross fixed capital formation relative to GDP*, *inflation* and *three-month interest rates*, and (iv) external and global imbalances given by *real effective exchange rates*, the *current account balance relative to GDP* and *oil prices*. All variables are expressed either in real terms or as a share of GDP.

For our model specifications, we use four different combinations of employed variables (also referred to as datasets). Dataset (iv) uses all of the available variables. In addition, we specify smaller models using subsets of variables in order to illustrate the relative performance of methods across datasets of varying complexity and information content. While reduced information content should generally reduce models' predictive ability, this may in some cases be offset by the gains from estimating less complex models. Datasets (i)-(iii) are mutually exclusive selections of indicators based on the different sources of vulnerabilities: (i) asset prices and credit developments, (ii) macroeconomic environment, and (iii) external and global imbalances. A list of the variables used in each dataset can be found in Table B.6. In order to guarantee comparability across the different datasets, we use the same sample for all datasets.

These a priori specified, economically motivated datasets have been chosen to allow for an economic interpretation of the information contained in each dataset. Moreover, by limiting ourselves to a priori specified, economically motivated variables and transformations, we seek to limit potential problems of data-mining. As Inoue and Kilian (2005) explain, when trying many variables and specifications, one is likely to find "spurious rejections of the no-predictability null and thus overfitting relative to the true model". Thus, avoiding data-mining is important for a realistic assessment of the true out-of-sample performance of early warning models. This is especially critical for early-warning models as policy-relevant analysis tools, as an overestimation of their accuracy might lead to a wrong sense of security.

Several of our potential predictor variables naturally contain a time trend (the exception being inflation, money market rates and current account to GDP), which needs to be removed prior to estimation. We focus on two of the most frequently employed approaches in the early warning literature: A Hodrick-Prescott (HP) filtering approach for our benchmark results and a growth rates approach for robustness.[7]

[6] Appendix B explains the economic intuition motivating the choice of variables for each channel based on previous literature, describes data transformations and reports summary statistics.

[7] To remove extreme outliers, we furthermore winsorize the data at the 1%- and 99%-quantile.

4 Results

4.1 In-sample Predictive Performance across Methods

Table 1 reports the *in-sample* relative usefulness of the six different methods (in rows) for four different sets of explanatory variables (in columns). The best performance on each dataset is indicated in bold. Significance stars (obtained from our bootstrap procedure) indicate whether the respective usefulness is significantly below that of the best-performing method on the same dataset.

In line with the literature (Alessi and Detken, 2018; Holopainen and Sarlin, 2017; Tanaka et al., 2016), machine learning methods such as knn and random forest always attain substantially higher in-sample relative usefulness than the corresponding logit model. Random forests achieve the best in-sample performance on all datasets, with knn close on their tails on dataset 2 (macroeconomic environment) and 4 (all variables). The inferiority of the logit model's in-sample performance relative to the best model on every dataset is statistically and economically significant. As Table C.1 in the Appendix shows, these findings are robust to using alternative measures of prediction performance. Moreover, the fit of knn and random forest is often close to perfect, with relative usefulness, F-measure and area under the curve (AUC) being close to one.[8] The other machine learning methods, trees, support vector machines and neural networks, are in general more in line with the logit results.

4.2 Out-of-sample Predictive Performance across Methods

In line with the standard in the empirical literature, the focus of our evaluation is on recursive out-of-sample performance rather than in-sample performance. For every point in time from 2005Q3 until the end of our sample in 2016Q4, we estimate the model recursively, strictly using only the information available at that time. We thus obtain predictions for approximately 300 out-of-sample observations, including 60 pre-crisis periods. These predictions are then used to calculate out-of-sample performance measures.

Table 2 shows the out-of-sample relative usefulness of six different methods (in rows) for four different sets of explanatory variables (in columns). The table shows that the logit model almost always outperforms the machine learning methods. The only exception is the dataset based on macroeconomic variables. However, in that case the relative usefulness is negative for all methods.[9] That is, a naïve forecast would be better than using any of the considered models. Table C.2 in the Appendix shows that the superiority of the logit model is confirmed by the other three performance measures, namely the area under the curve (AUC) and Brier probability score (BPS) – again, with an exception on the macroeconomic dataset for the F-measure and AUC. The difference of performance measures between the logit model and machine learning methods is statistically significant

[8] Table C.1 also shows that the zero usefulness of trees on dataset 2 and 3 is due to a degenerate tree that always predicts a crisis. This happens because the available variables are not very informative relative to the required tree complexity. The optimal tree (according to the tree's internal cost function) then consists of the unconditional forecast. By definition, relative usefulness and AUC will be zero and 0.5, respectively.

[9] This can happen in out-of-sample predictions where the performance measure uses the ex-post knowledge on actual crisis occurrence – information that is not available at the time the forecast is made.

Table 1: *In-sample* relative usefulness

	(1) Credit/Asset Prices	(2) Macro	(3) External	(4) All
logit	0.347**	0.202***	0.390*	0.511*
	[0.185, 0.440]	[0.018, 0.236]	[0.243, 0.422]	[0.208, 0.561]
trees	0.432**	0.000***	0.000**	0.674*
	[0.158, 0.601]	[-0.031, 0.216]	[0.000, 0.410]	[0.230, 0.711]
knn	0.693*	0.965	0.685	0.955
	[0.353, 0.693]	[0.265, 0.965]	[0.302, 0.685]	[0.370, 0.955]
rf	**0.966**	**0.966**	**1.000**	**0.992**
	[0.411, 0.966]	[0.240, 0.966]	[0.299, 1.000]	[0.347, 0.992]
svm	0.362**	0.481**	0.000**	0.784
	[-0.006, 0.483]	[0.129, 0.489]	[-0.258, 0.320]	[0.334, 0.784]
nen	0.456**	0.266***	0.377*	0.764
	[0.201, 0.503]	[0.042, 0.294]	[0.263, 0.419]	[0.338, 0.793]

Note: Highest usefulness on each dataset in bold. Stars indicate whether the respective usefulness is significantly below the best performance on the same dataset (***/**/* for the 1%/5%/10% level). Numbers in brackets indicate 90% confidence bands.

Table 2: *Out-of-sample* relative usefulness

	(1) Credit/Asset Prices	(2) Macro	(3) External	(4) All
logit	**0.368**	-0.236	**0.438**	**0.605**
	[0.143, 0.431]	[-0.386, -0.036]	[0.309, 0.563]	[0.222, 0.605]
trees	0.201	-0.605	0.390***	0.126***
	[0.084, 0.317]	[-0.605, -0.087]	[0.072, 0.390]	[-0.037, 0.313]
knn	0.247	-0.087	0.293***	-0.062***
	[0.088, 0.384]	[-0.137, 0.034]	[0.133, 0.417]	[-0.166, 0.042]
rf	0.246	-0.245	0.138***	-0.003***
	[0.129, 0.350]	[-0.323, -0.132]	[0.055, 0.283]	[-0.141, 0.146]
svm	0.218*	**-0.015**	0.202**	-0.065***
	[0.018, 0.326]	[-0.257, 0.060]	[0.043, 0.409]	[-0.173, 0.130]
nen	0.251	-0.349	0.434	0.092**
	[0.080, 0.451]	[-0.382, -0.048]	[0.322, 0.551]	[-0.041, 0.250]

Note: Highest usefulness on each dataset in bold. Stars indicate if the respective usefulness is significantly below the best performance on the same dataset (***/**/* for the 1%/5%/10% level). Numbers in brackets indicate 90% confidence bands.

on datasets 3 (external variables) and 4 (all variables). Neural networks are an exception because they are only insignificantly worse in terms of relative usefulness on dataset 3, and in terms of the F-measure and BPS on datasets 3 and 4. On dataset 1 (credit and asset prices), logit models perform significantly better than trees and support vector machines, while the difference to other machine learning methods is insignificant. Taken together, a first impression is that the logit is best suited for recursive out-of-sample predictions, with neural networks being the closest competitor.

4.3 Robustness

In this section, we assess the degree to which our results are robust to four key variations of the modeling setup. These variations concern the choice of the preference parameter for relative usefulness, data transformation, sample length, and the crisis database. We also test an extreme variation of out-of-sample forecasting with one-off-splits.

Preference parameter: First of all, we make sure that our results do not hinge on the choice of loss function preference parameter (see the detailed description of relative usefulness in Appendix A). While AUC and BPS are preference-independent measures, the preference parameter enters into the computation of relative usefulness via the loss function, which is used to evaluate forecasts and to compute optimal thresholds. Our benchmark value for the preference parameter, $\mu = 0.5$, represents a balanced trade-off between type-1 (missed crises) and type-2 errors (false alarms). As we use the same threshold for both the relative usefulness and the F-measure, the preference parameter implicitly also enters the F-measure.

It is easily conceivable that missing a crisis may be more costly than issuing a false alarm. In this case, more weight should be given to type-1 errors. Therefore, we conduct a robustness check for a preference parameter of $\mu = 0.75$ which assigns much more weight to type-1 errors. We re-estimate all models including hyperparameters given this new preference parameter and report results in Table C.3 in the Appendix. We find that relative usefulness drops strongly for all methods, often to negative values. While the logit model still has a positive relative usefulness on all datasets but the second, machine learning methods seem to deteriorate more strongly, for instance on datasets (4) and (1). The F-measure (which does not change systematically) indicates that the logit outperforms the machine learning methods on datasets (2) and (4), while it is only insignificantly worse than the best method on the other two datasets. The change in preference parameters also affects hyperparameters of machine learning methods (because they are based on relative usefulness). As a result, the AUC drops. In many cases, the AUC now takes on values below 0.5. That is, not only the signals derived at "optimal thresholds" but the whole probability predictions for models like knn.4 are de facto misleading predictions. For the AUC and BPS, logit is the best method on all datasets except the second, where it is insignificantly worse than neural networks or support vector machines, respectively. As a result, logit continues to outperform machine learning methods, and the logit with all variables remains the "best" overall model.

Data transformation: As a second robustness test, we check the extent to which our results hinge on the choice of using HP filter gaps for removing the trend in our explanatory variables. To this end, we replace the HP filter gaps by simple growth rates, where lags used to compute growth rates differentiate between business cycle and financial cycle variables. The results of this robustness check are shown in Table C.4. Looking at relative usefulness, we see that some of the machine learning models perform better when using growth rates rather than HP filter gaps. By contrast, all four logit specifications have somewhat lower relative usefulness, F-measure and higher BPS than before. Depending on the performance metric, knn, random forests and even trees may outperform logit on datasets (1) and (2). Across all models considered, the logit model with all variables (logit.4) remains the best model according to relative usefulness and BPS, closely followed by nen.4, which also outperforms logit.4 in terms of AUC.

Sample length: A third important test concerns the robustness of our results relative to variations of our sample. Specifically, we consider a robustness check, where we cut off the first ten years of our dataset, amounting to approximately one-sixth of our observations. Financial repression during the 1970s may have affected the behavior of our explanatory variables and their impact on the probability of future financial crises. More generally, the underlying data-generating process may be time-varying, suggesting a trade-off between sample length and sample homogeneity.

Table C.5 presents results when restricting our sample to exclude all observations prior to 1980Q1. For the logit model, we find that out-of-sample prediction performance based on this smaller set of information is lower than when using the full dataset. Thus, the trade-off between sample length and sample homogeneity appears to be tilted in favor of sample length. The relative ordering of machine learning methods compared with the logit method is unchanged. Thus, we conclude that, while sample length appears to be important, our main finding regarding relative prediction performance between methods is robust to the change in the sample. We also note that this robustness appears to be driven by the robustness of the logit method, while machine learning methods sometimes react quite strongly to this moderate change in the sample.

Crisis database: In our fourth robustness check we replace the ESCB/ESRB crisis database by the well-known database of Laeven and Valencia (2018). Compared to the baseline crisis database, this database has fewer pre-crisis periods in the training sample, and a higher share of pre-crisis periods in the test sample. Results are shown in table C.6. It turns out that changing the dependent variable of our models induces big changes in the results. Looking at the different performance metrics, there are winners and losers across all datasets and methods. However, results for logit.4 and logit.3 remain robust to this change. As a consequence, they continue to outperform their machine learning competitors and logit.4 remains the best overall model according to relative usefulness. However, estimation uncertainty is now much larger: even for logit.4, the confidence band around the point estimate of relative usefulness includes negative values. Therefore, only very few machine learning methods are significantly worse than their corresponding logit model. While the F-measure indicates similar results, performances of BPS and AUC are more mixed. Neural networks have the highest AUC for all datasets but the second, while support vector machines perform well in terms of BPS on datasets 1 and 2. Overall, the highest usefulness is achieved by the logit on datasets 1 and 4, the best AUC by neural networks on dataset 4, and the lowest BPS by logit on dataset 4. Overall, the logit.4 still seems to be preferable.

One-off splits: As an alternative to recursive estimation, we also explore one-off splits, where an estimation on observations until 2005Q2 is used to predict probabilities after that (results reported in Table C.7) or the other way around (see Table C.8). Different to recursive forecasts, these exercises do not take advantage of the most recent developments of data to perform out-of-sample predictions. Instead (at the extreme), crisis probabilities in the 1970's are backcasted based on an estimation on data between 2005Q3 and 2016Q2. This is a very hard task, especially since the pre-crisis probability in the later subsample is considerably larger than in the early parts of the sample. Therefore, it comes at no surprise that the optimal thresholds for the backcasting exercise are mostly much higher than for the forecasting exercise. In the backcasting case, we should also be clear that we are predicting more than 83% of the dataset based on the remaining 17% of observations. Yet, even in these extreme scenarios relative usefulness remains mostly positive. The performance of logit models is mostly at the upper end of all methods. However, we have considerable estimation uncertainty for the reasons mentioned above. Therefore, we find differences between methods to be mostly insignificant.

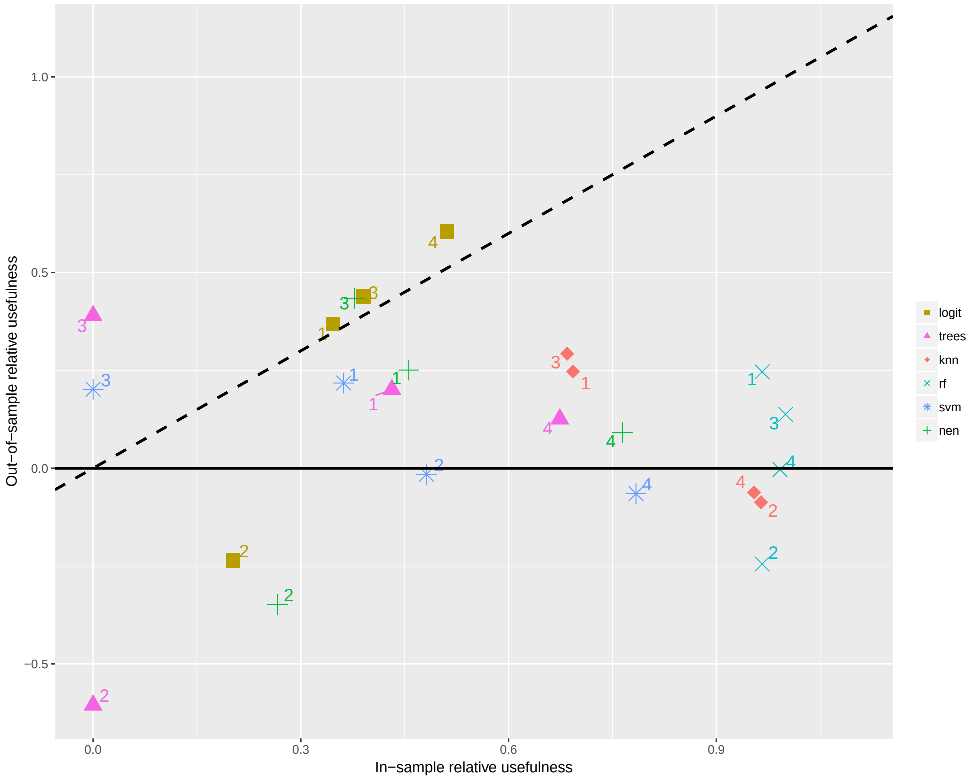
Overall, our robustness checks confirm the finding that a logit model using all variables offers the best predictive performance among all models considered. Moreover, we saw that changes to model specifications, as those considered in this section, can induce substantial changes to some models' performance. Given this, we see the robustness of the logit.4 (and logit.3) model across specifications, as an additional feature of these models. By contrast, performance of machine learning methods is substantially less robust.

4.4 Interpretation and Discussion

A comparison of the performance of in-sample and recursive out-of-sample estimations gives an indication as to why machine learning methods do not outperform the logit approach in this application. Figure 1 displays the relationship between in-sample and recursive out-of-sample performance across models based on our benchmark results shown in Tables 1 and 2. A striking result is that many of the machine learning models (including all random forest models) achieve a near-perfect in-sample fit (relative usefulness close to its theoretical maximum of 1), but, at the same time, show much lower out-of-sample performance. This suggests that overfit may be a major issue for at least some of the models. Moreover, even for those machine learning models where in-sample fit is not perfect, their out-of-sample performance is often markedly below their in-sample performance. By contrast, among logit models this is only the case for logit.2, which we saw is a special case of negative relative usefulness for all methods on this dataset. Consistent with the results in the previous section, neural networks emerge as the closest competitors to logit models also with regards to the consistency between in-sample and out-of-sample prediction performance. Figures C.2 through C.5 in the Appendix show that this pattern holds true more generally across all robustness checks.

The argument made above is not restricted to relative usefulness. We can also look receiver-operator characteristics in Figure 2, divided into four subplots for the four different datasets used above. The AUC is defined as the integral of the area under the ROC

Figure 1: Relative usefulness of in- and out-of-sample estimation, by model.



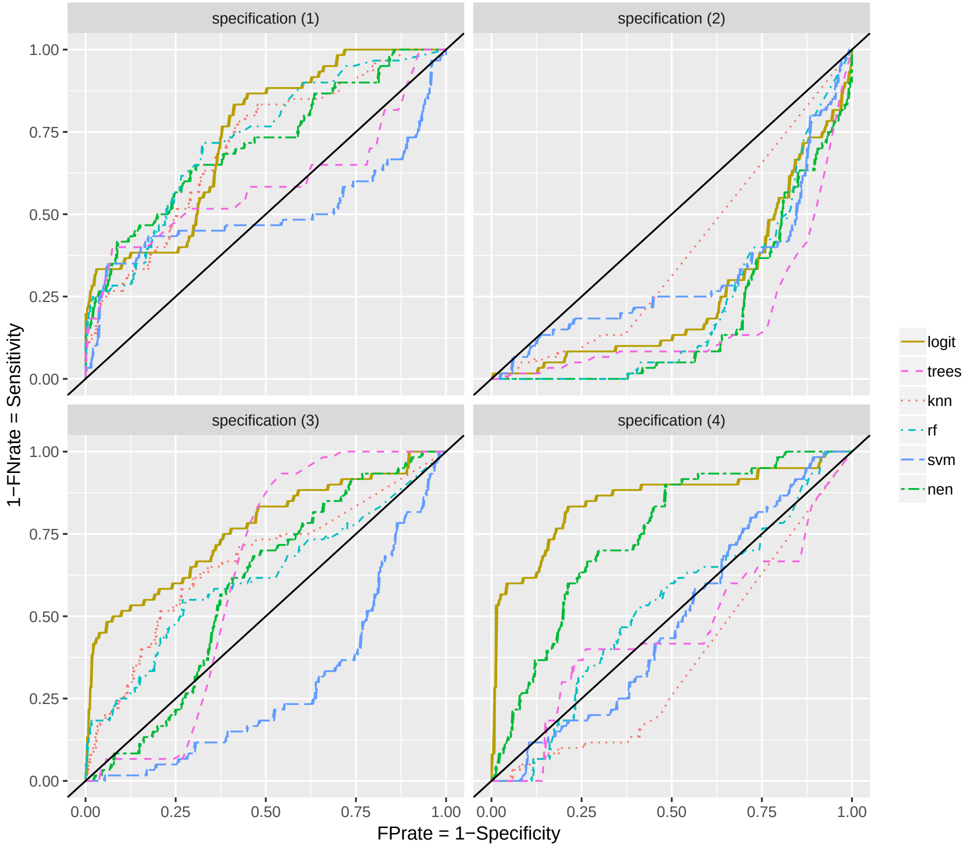
curve.[10] Thus, the figure visualizes the relation of AUC-values presented in Table C.2 in the Appendix. For datasets three and four, the ROC curve of the logit model is (nearly) everywhere above the curves of other models. For dataset one, the logit curve is above the other methods for comparably low thresholds that lead to more (true and false) signals. In total, the AUC of the logit model on the first dataset is higher than for other methods, albeit not always significantly.

A comparison with the ROC curve of the in-sample estimations (shown in Figure C.1 in the Appendix) confirms a finding discussed above for relative usefulness: machine-learning methods with high in-sample fit tend to have a lower out-of-sample fit. Random forests, k-nearest neighbours, support vector machines and trees all loose substantial recursive out-of-sample performance. Only neural networks manage to stay close to their in-sample performance, especially and as comparable to the logit model for datasets 1,3 and 4.

In addition to the empirical evidence in figure 1, a theoretical argument pointing to overfit (relative to the true model) can be made. As a thought experiment, suppose we

[10] If the ROC curve coincides with the diagonal, the AUC would be 0.5. Also, the rate of false positives and false negatives would be identical at every threshold, and it would not be possible to derive an informative signal from the prediction. As ROC curves move closer to the upper left corner (i.e. a point where both the false positive and false negative rates are zero), the prediction becomes more and more informative. This is for example the case in a comparison of logit.4 to other methods. For ROC curves below the diagonal, the signal is counterinformative. This is for example the case for the estimations on the second dataset. If we knew of this in advance, this fact could be exploited: high predicted probabilities would be seen as a sign of no crisis, while policymakers would worry at low predicted probabilities. However, note that we show the ROC curve for recursive out-of-sample predictions, where we do not have this knowledge.

Figure 2: Receiver-operator characteristics for baseline recursive out-of-sample estimations, by model.



Note: Different subplots correspond to the four different datasets. The ROC displays the relation of false positive and false negative rates for different (constant) thresholds that can be applied to the probability predictions in the data. The added points correspond to false positive and false negative rates of the binary predictions derived from time-varying optimal thresholds, see Table C.2.

knew the true data generating process (DGP) and we had a model (and data) at hand that would give us for each observation the true conditional probability of a crisis. Even in this case, the prediction error would still be positive (except in the degenerate case where conditional crisis probabilities could only be either 1 or zero). For that reason, even with a perfect model, we would not expect relative usefulness to approach its maximum value of 1, but rather to converge (both in-sample and out-of-sample) to a DGP specific maximum between 0 and 1 as the sample size increases.[11] This can formally be seen in Table 5 of Boissay, Collard, and Smets (2016). They present a DSGE model generating credit boom driven crises, and run an early warning exercise on simulated data from their model. It turns out that crisis prediction using the true model-implied conditional probability still leads to considerable error rates (around 1/3 missed crises). Their results also show that a logit model estimated on binary crisis realizations is able to converge to a performance similar to that of the true model. The remaining error rates under the true model reflect the fact that crises can only be predicted in probability and not with certainty. In other words, crises are driven by a predictable component, captured by the true model, and a substantial unpredictable component (given the observables), which cannot be forecasted by any model.

To the extent that a logit model is already able to closely approximate the true model (as in Boissay et al. (2016)), it will not be possible to substantially outperform this model. This implies that when machine learning methods fit the data beyond (or below) the predictable component, this comes at the cost of worse performance in the recursive out-of-sample estimation. This may be an issue even for those machine learning models where in-sample fit is not perfect. Theoretically, the hyperparameters of the machine learning methods should provide some safeguard against overfit. However, despite devoting considerable effort to the calibration of these hyperparameters via a sophisticated cross-validation procedure (see Appendix A.4), the overfit still persists for many of the considered models. In sum, it appears that logit models naturally limit the amount of overfit, while being sufficiently flexible in their approximation of the data generating process.

The conclusion from our out-of-sample forecast comparison is different than that of Alessi and Detken (2018) and Tanaka et al. (2016), who argue that a random forest has a better prediction performance than a logit model in an early warning setting. However, Alessi and Detken (2018) do not run an out-of-sample comparison of the two methods. Their argument is rather based on results from k-fold-cross-validation, where they find some differences between the AUC of one random forest specification (AUC = 0.94) and two logit specifications (AUC = 0.84, and 0.93 respectively). Setting aside the question of whether this difference is statistically significant, the high levels of AUC (close to the maximum of 1) suggest that the cross-validation procedure may provide an inflated estimate of the performances of these methods. In fact, cross-validation estimates often appear to be closer to in-sample performance than to out-of-sample performance. The tendency of cross-validation to provide inflated estimates of performance, particularly in the presence of cross-sectional and serial correlation, has also been recognized by Holopainen and Sarlin

[11] As a simple illustration, suppose for example that in 50% of the cases the true crisis probability was 80%, while in 50% of the cases, it was 20%, and that our signaling threshold was 50%. Then, even when knowing the true model, we would still have a false positive rate of 20% and a false negative rate of 20%, leading to a relative usefulness of only 60% (assuming $\mu = 0.5$).

(2017). As a consequence, these estimates are likely to be biased towards (more complex) machine learning methods, given their above-mentioned tendency to overfit in-sample data. Another important point has been noted by Neunhoeffer and Sternberg (2018) based on an example of civil war prediction from the political science literature. They show that performance of machine learning methods has been seriously over-estimated, in studies using cross-validation for both hyperparameter selection and model evaluation. Tanaka et al. (2016) find somewhat more pronounced differences between logit and random forest performance for a bank-level early warning model. However, they also focus on cross-validation estimates of performance. As we do not cover bank-level early warning in our analysis, we cannot make statements about recursive out-of-sample performances in their setup.

The importance of conducting model comparisons via out-of-sample experiments has also been advocated by Holopainen and Sarlin (2017). However, in their out-of-sample forecasting exercise, logit models are outperformed by machine learning methods (except for trees). We conjecture that differences in the employed training sample are a key driver of the contrasting results. In their application, more than 95% of the pre-crisis periods are located in the recursive out-of-sample period and are thus unavailable for the first recursive estimations. This means that their model estimations are driven by a few influential pre-crisis observations in their training sample, such that the selection of these observations is a critical determinant of their out-of-sample results. By contrast, our broader data basis enables us to use nearly 60% of all pre-crisis periods in the first recursive estimation, mimicking more closely actual out-of-sample prediction tasks. The different sample appears to be the most important explanation for our differing results.

As a final word regarding interpretation, we want to make clear that while we think it is important to establish a robust and valid out-of-sample prediction exercise, we do not want to over-interpret its results. After all, our results hold for the given finite dataset at hand. In particular, our out-of-sample window is naturally dominated by the great financial crisis of 2007-2008. Moreover, we cannot fully exclude the possibility that some other (potentially more sophisticated) modeling approach is able to outperform the logit model, or that machine learning methods may be preferable on other (possibly larger) datasets. However, we think that we have established that the logit model is surprisingly hard to beat, in line with findings in the forecasting literature more generally, that simple forecasting models often outperform more complex models. We have provided theoretical and empirical arguments, as well as a discussion of the literature, suggesting systematic issues related to overfit driving this result. Further research is needed, to gain a more complete understanding of the conditions under which machine learning methods can be successfully applied, in general, and in particular to early warning models of financial crises. As neural networks emerge as the most promising machine learning method in our analysis, future research could investigate whether recent advances in deep neural networks (LeCun, Bengio, and Hinton, 2015) can be successfully applied to the prediction of financial crises. Our results suggest that controlling the risk of overfitting may be key for the success of these highly complex neural network models.

5 Conclusion

This paper has presented an analysis of early warning models for systemic banking crises, based on a dataset covering 15 advanced countries over the period 1970-2016. We assess how different methods – a benchmark logit approach and several machine learning methods – perform in a quasi real-time out-of-sample forecasting experiment. It turns out that the logit approach is surprisingly hard to beat, generally leading to lower out-of-sample prediction errors than the machine learning methods. This result holds under different performance measures and different selections of variables, and is robust to alternative choices of crisis variable, variable transformation, sample length or loss function preference parameter.

Our interpretation of this result is that a strong in-sample fit of machine learning methods should not necessarily be taken as an indication of strong out-of-sample prediction performance, since it could alternatively be a sign of overfitting. In addition, the stability of these methods' performances across variations of the setup appears to be less pronounced than that of the logit model.

These results suggest that performance of machine learning methods in real-world out-of-sample prediction situations cannot be taken for granted. Instead, the circumstances under which these methods offer clear advantages as well as potential modifications for improving their stability and performance in early warning applications need further investigation. Neural networks emerge as the most promising machine learning method in our analysis. Future research could therefore investigate whether recent advances in deep neural networks can be successfully applied to the prediction of financial crises.

References

Alessi, L. and C. Detken (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. and C. Detken (2018). Identifying Excessive Credit Growth and Leverage. *Journal of Financial Stability 35*, 215–225.

Allen, F. and D. Gale (2007). *Understanding Financial Crises.* Oxford University Press, Oxford.

Arlot, S., A. Celisse, et al. (2010). A Survey of Cross-Validation Procedures for Model Selection. *Statistics Surveys 4*, 40–79.

Babecký, J., T. Havránek, J. Matějů, M. Rusnák, K. Šmídková, and B. Vašíček (2014). Banking, Debt, and Currency Crises in Developed Countries: Stylized Facts and Early Warning Indicators. *Journal of Financial Stability 15*, 1–17.

Basel Committee on Banking Supervision (2010). Guidance for National Authorities Operating the Countercyclical Capital Buffer. Technical report.

Berg, A. and C. Pattillo (1999). What Caused the Asian Crises: An Early Warning System Approach. *Economic Notes 28*(3), 285–334.

Boissay, F., F. Collard, and F. Smets (2016). Booms and Banking Crises. *Journal of Political Economy 124*(2), 489–538.

Breiman, L. (1996). Bagging Predictors. *Machine Learning 24*(2), 123–140.

Breiman, L. (2001). Random Forests. *Machine Learning 45*(1), 5–32.

Breiman, L., J. Friedman, R. Olshen, and C. Stone (1984). *Classification and Regression Trees.* Wadsworth.

Brier, G. W. (1950). Verification of Forecasts Expressed in Terms of Probability. *Monthly weather review 78*(1), 1–3.

Brunnermeier, M. K. (2009). Deciphering the Liquidity and Credit Crunch 2007-2008. *Journal of Economic Perspectives 23*(1), 77–100.

Brunnermeier, M. K. and M. Oehmke (2013). Bubbles, Financial Crises, and Systemic Risk. In *Handbook of the Economics of Finance*, Chapter 18, pp. 1221–1288.

Bussière, M. and M. Fratzscher (2006). Towards a New Early Warning System of Financial Crises. *Journal of International Money and Finance 25(6)*, 953–973.

Calvo, G. A. (1998). Capital Flows and Capital-Market Crises: The Simple Economics of Sudden Stops. *Journal of Applied Economics 1*(1), 35–54.

Cerutti, E., S. Claessens, and L. Laeven (2017). The Use and Effectiveness of Macroprudential Policies: New Evidence. *Journal of Financial Stability 28*, 203–224.

Claessens, S. (2015). An Overview of Macroprudential Policy Tools. *Annual Review of Financial Economics 7*, 397–422.

Cover, T. and P. Hart (1967). Nearest Neighbor Pattern Classification. *IEEE Transactions on Information Theory 13*(1), 21–27.

Davidson, R. and J. G. MacKinnon (2000). Bootstrap tests: How many bootstraps? *Econometric Reviews 19*(1), 55–68.

DeLong, E. R., D. M. DeLong, and D. L. Clarke-Pearson (1988). Comparing the Areas under Two or More Correlated Receiver Operating Characteristic Curves: a Nonparametric Approach. *Biometrics*, 837–845.

Detken, C., O. Weeken, L. Alessi, D. Bonfim, M. M. Boucinha, S. Frontczak, G. Giordana, J. Giese, N. Jahn, J. Kakes, B. Klaus, J. H. Lang, N. Puzanova, and P. Welz (2014). Operationalising the Countercyclical Capital Buffer: Indicator Selection, Threshold Identification and Calibration Options. ESRB Occasional Paper 5.

Diebold, F. X. and G. D. Rudebusch (1989). Scoring the Leading Indicators. *Journal of Business 62*(3), 369–391.

Drehmann, M., C. Borio, and K. Tsatsaronis (2011). Anchoring Countercyclical Capital Buffers: the Role of Credit Aggregates. *International Journal of Central Banking 7*(4), 189–240.

Drehmann, M., C. Borio, and K. Tsatsaronis (2012). Characterising the Financial Cycle: Don't Lose Sight of the Medium Term! BIS Working Papers 380.

Drehmann, M. and M. Juselius (2014). Evaluating Early Warning Indicators of Banking Crises: Satisfying Policy Requirements. *International Journal of Forecasting 30*(3), 759–780.

Drehmann, M., M. Juselius, and A. Korinek (2017). Accounting for Debt Service: The Painful Legacy of Credit Booms. BIS Working Papers 645.

Duprey, T., B. Klaus, and T. Peltonen (2017). Dating Systemic Financial Stress Episodes in the EU Countries. *Journal of Financial Stability 32*, 30–56.

El-Shagi, M., T. Knedlik, and G. von Schweinitz (2013). Predicting Financial Crises: The (Statistical) Significance of the Signals Approach. *Journal of International Money and Finance 35*, 76–103.

El-Shagi, M., A. Lindner, and G. von Schweinitz (2016). Real Effective Exchange Rate Misalignment in the Euro Area: A Counterfactual Analysis. *Review of International Economics 24*(1), 37–66.

European Central Bank (2010). Financial Stability Review, June 2010.

European Central Bank (2017). Financial Stability Review, May 2017.

Frankel, J. A. and A. K. Rose (1996). Currency Crashes in Emerging Markets: An Empirical Treatment. *Journal of International Economics 41*(3), 351–366.

Gourinchas, P.-O. and M. Obstfeld (2012). Stories of the Twentieth Century for the Twenty-First. *American Economic Journal: Macroeconomics 4*(1), 226–265.

Holopainen, M. and P. Sarlin (2017). Toward Robust Early-Warning Models: A Horse Race, Ensembles and Model Uncertainty. *Quantitative Finance 17*(12), 1–31.

Inoue, A. and L. Kilian (2005). In-Sample or Out-of-Sample Tests of Predictability: Which One Should We Use? *Econometric Reviews 23*(4), 371–402.

James, G., D. Witten, T. Hastie, and R. Tibshirani (2013). *An Introduction to Statistical Learning with Applications in R*. Springer.

Janes, H., G. Longton, and M. Pepe (2009). Accommodating Covariates in ROC Analysis. *The Stata Journal 9*(1), 17.

Jordà, Ò., M. Schularick, and A. M. Taylor (2011). Financial Crises, Credit Booms, and External Imbalances: 140 Years of Lessons. *IMF Economic Review 59*(2), 340–378.

Jordà, O., M. Schularick, and A. M. Taylor (2015). Leveraged Bubbles. *Journal of Monetary Economics 76*, S1–S20.

Kaminsky, G. L. and C. M. Reinhart (1999). The Twin Crises: the Causes of Banking and Balance-of-Payments Problems. *American Economic Review 89*(3), 473–500.

Kaminsky, G. L. and C. M. Reinhart (2000). On Crises, Contagion, and Confusion. *Journal of International Economics 51*, 145–168.

Kindleberger, C. P. and R. Z. Aliber (2005). *Manias, Panics and Crashes – A History of Financial Crises*. Palgrave Macmillan, Hampshire and New York.

Knedlik, T. and G. von Schweinitz (2012). Macroeconomic Imbalances as Indicators for Debt Crises in Europe. *JCMS: Journal of Common Market Studies 50*(5), 726–745.

Laeven, L. and F. Valencia (2013). Systemic Banking Crises Database. *IMF Economic Review 61*(2), 225–270.

Laeven, L. and F. Valencia (2018). Systemic Banking Crises Database Revisited. IMF Working Paper 18/206.

LeCun, Y., Y. Bengio, and G. Hinton (2015). Deep learning. *Nature 521*(7553), 436.

Lim, C. H., A. Costa, F. Columba, P. Kongsamut, A. Otani, M. Saiyid, T. Wezel, and X. Wu (2011). Macroprudential Policy: What Instruments and How to Use Them? Lessons from Country Experiences. IMF Working Paper 11/238.

Lo Duca, M., A. Koban, M. Basten, E. Bengtsson, B. Klaus, P. Kusmierczyk, J. H. Lang, C. Detken, and T. Peltonen (2017). A New Database for Financial Crises in European Countries - ECB/ESRB EU Crises Database. ECB Occasional Paper No 194.

Lo Duca, M. and T. A. Peltonen (2013). Assessing Systemic Risks and Predicting Systemic Events. *Journal of Banking & Finance 37*(7), 2183–2195.

Maddaloni, A. and J.-L. Peydró (2011). Bank Risk-taking, Securitization, Supervision, and Low Interest Rates: Evidence from the Euro-area and the U.S. Lending Standards. *The Review of Financial Studies 24*(6), 2121–2165.

McFadden, D. L. (1984). Econometric Analysis of Qualitative Response Models. *Handbook of econometrics 2*, 1395–1457.

Minsky, H. P. (1982). *Can "It" happen again? – Essays on Instability and Finance.* M.E. Sharpe Inc., Armonk, N.Y.

Murphy, K. (2012). *Machine Learning: A Probabilistic Perspective.* MIT Press.

Neunhoeffer, M. and S. Sternberg (2018). How cross-validation can go wrong and what to do about it. *Political Analysis 27*(1), 101–106.

Powers, D. M. (2011). Evaluation: From precision, recall and f-measure to roc, informedness, markedness and correlation. *Journal of Machine Learning Technologies 2*(1), 37–63.

Rajan, R. G. (2006). Has Finance Made the World Riskier? *European Financial Management 12*(4), 499–533.

Reinhart, C. M. and K. S. Rogoff (2008). Is the 2007 US Sub-Prime Financial Crisis so Different? An International Historical Comparison. *The American Economic Review 98*(2), 339–344.

Reinhart, C. M. and K. S. Rogoff (2009). *This Time is Different: Eight Centuries of Financial Folly.* Princeton University Press, Princeton and Woodstock.

Rose, A. K. and M. M. Spiegel (2012). Cross-Country Causes and Consequences of the 2008 Crisis: Early Warning. *Japan and the World Economy 24*(1), 1–16.

Sarlin, P. (2013). On Policymakers' Loss Functions and the Evaluation of Early Warning Systems. *Economics Letters 119*(1), 1–7.

Sarlin, P. and G. von Schweinitz (2017). Optimizing Policymakers' Loss Functions in Crisis Prediction: Before, Within or After? ECB Working Paper 2025.

Schüler, Y. S., P. Hiebert, and T. A. Peltonen (2015). Characterising the Financial Cycle: a Multivariate and Time-Varying Approach. ECB Working Paper 1846.

Shalev-Shwartz, S. and S. Ben-David (2014). *Understanding Machine Learning: from Theory to Algorithms.* Cambridge University Press.

Tanaka, K., T. Kinkyo, and S. Hamori (2016). Random Forests-Based Early Warning System for Bank Failures. *Economics Letters 148*, 118–121.

Tanaka, K., T. Kinkyo, and S. Hamori (2018). Financial hazard map: Financial vulnerability predicted by a ran-dom forests classification model. *Sustainability 10*(5), 1530.

6 Appendix A: Methodology

A.1 Definition of Dependent Variable

Early warning models (typically) perform window forecasts of crisis probabilities and use thresholds to derive binary signals from these probabilities. The rationale behind this approach has two aspects. First, window forecasts are used since it is hard to predict the exact quarterly start date of a crisis, which may be driven to a large extent by unforecastable shocks. However, recurring patterns before crises may still be informative about their likelihood of occurrence during a given time *interval*. Thus, window forecasts of the probability of a systemic banking crisis can be used to reflect potential buildups of vulnerabilities, which might require for instance the activation of macroprudential policy measures. Second, converting probabilities into clear signals (taking into account the policymaker's preferences) helps to inform policymakers' ultimate decision on whether and when to take action. Moreover, it allows for a straightforward evaluation of predictions in terms of correct or incorrect signals.

To implement these ideas, we follow the literature and define the dependent variable for our estimations as follows. Starting from a crisis database, where $C_{t,n}$ is 1 if a crisis was ongoing in country n at time t and zero otherwise, we define another binary variable $\bar{C}_{t,n}$ as our dependent variable. This dependent variable $\bar{C}_{t,n}$ is set to one during early warning windows between h_1 and h_2 periods before a crisis (pre-crisis periods) and zero for observations that are not followed by a crisis within the next h_2 quarters (tranquil periods).

The resulting gap of length $h_1 - 1$ between early warning windows and crises is excluded from the estimation, as these periods can neither be classified as being in an early warning window, nor as being tranquil periods. Moreover, it is standard to exclude periods where a country is already in a crisis (crisis periods). The reason for excluding crisis periods is that the extreme imbalances during these periods are typically due to being in a crisis (which is assumed to be known), instead of reflecting the buildup of imbalances prior to a crisis.[12]

The timing of the early warning window is chosen to fulfill two criteria. First, the gap $h_1 \geq 0$ between the window and the start of the crisis is chosen to allow for policy action. Second, the window needs to be sufficiently close to the predicted crisis for economic variables to show informative developments. Following the literature, we set the limits of early warning windows to $h_1 = 5$ and $h_2 = 12$ quarters (see Drehmann and Juselius, 2014; Alessi and Detken, 2018; Holopainen and Sarlin, 2017). This allows at least one year for policy measures to become effective and to issue warnings up to three years before a crisis.

In sum, this leads to the following definition of the dependent variable:

$$\bar{C}_{t,n} = \begin{cases} 0 & \text{, if } C_{t+h,n} = 0 \text{, for all } h \in \{0, \dots, h_2\} \\ 1 & \text{, if } C_{t+h,n} = 1 \text{, for some } h \in \{h_1, \dots, h_2\} \text{ and} \\ & C_{t+h,n} = 0 \text{, for all } h \in \{0, \dots, h_1 - 1\} \\ NA & \text{, otherwise.} \end{cases} \quad (1)$$

[12] We do not exclude additional periods after a crisis, as crises in our database are defined such that they already account for the post-crisis bias discussed in Bussière and Fratzscher (2006).

We thus estimate the probability of a crisis between the next $h_1 = 5$ to $h_2 = 12$ quarters, conditional on not already being in an acute crisis period. To do this, the binary dependent variable $\bar{C}$ is linked to a set of early warning indicators X using different modeling choices. Each model is estimated and then used to predict the probability of being in an early warning window at time t in country n conditional on the observables X, $P(\bar{C}_{t,n}|X_{t,n})$.

A.2 Evaluation of Early Warning Models

To inform decision-making, the estimated probabilities may be mapped into binary signals. Using a threshold τ, the signal is set to one if the probability exceeds τ and to zero otherwise. These signals or their absence will ex post turn out as right or wrong, and can be classified into true positives, false positives, true negatives, or false negatives as indicated in Table A.1. False negatives FN (also called type-1 errors) are observations where no signal is given during an early warning window (missed crises), while false positives FP (type-2 errors) result from observations where a signal is given outside of an early warning window (false alarms). A higher classification threshold τ implies fewer signals, reducing both true and false positives. Converting probabilities into signals via a threshold thus entails a trade-off between type-1 errors (missed crises) and type-2 errors (false alarms). Selecting the optimal threshold generally depends on the loss function of the forecast user. The current standard is to choose the classification threshold τ such that it maximizes the relative usefulness function of Alessi and Detken (2011), which weighs errors (as a share of the respective actual class) by a parameter μ representing the forecast user's preferences.[13]

Table A.1: A contingency matrix.

		Actual class C	
		Pre-crisis period	Tranquil period
Predicted class S	Signal	Correct call *True positive (TP)*	False alarm *False positive (FP)*
	No signal	Missed crisis *False negative (FN)*	Correct silence *True negative (TN)*

Note: This contingency matrix follows Holopainen and Sarlin (2017).

The relative usefulness (U_r) as a function of the preference parameter μ sets the loss of misspecification, $L(\mu) = \mu\frac{FN}{FN+TP} + (1-\mu)\frac{FP}{FP+TN}$, in relation to the loss of a naïve decision rule, $\min(\mu, 1-\mu)$, resulting from either always or never signaling a crisis depending on the preference parameter (Alessi and Detken, 2011):[14]

$$U_r(\mu) = 1 - \frac{L(\mu)}{\min(\mu, 1-\mu)}$$

[13] In an out-of-sample experiment, it is important to choose the threshold using in-sample information only and not by maximizing out-of-sample performance, which is a priori unknown to the forecaster.

[14] An alternative usefulness function proposed by Sarlin (2013) was shown to be equivalent under a constant unconditional crisis probability (Sarlin and von Schweinitz, 2017).

This implies a maximum relative usefulness of one, if $L(\mu) = 0$, and a usefulness of zero, if $L(\mu) = \min(\mu, 1-\mu)$. A usefulness above (below) zero therefore means that the model is more (less) informative than the naïve decision rule. We use a standard choice of $\mu = 0.5$ for our baseline results, thus weighting the two types of errors equally.

The F-measure uses the same contingency matrix as the relative usefulness. For a given signaling threshold, it relates true positives to the mean of the number of "Signals" and pre-crisis periods,[15]

$$F_1 = \frac{TP}{TP + \frac{(FP+FN)}{2}}.$$

In other words, the bigger the ratio of correctly predicted crises (TP) relative to erroneous predictions ($FP + FN$) is, the bigger is F_1. A perfect prediction implies a F-measure of 1. A prediction without a single true positive results in $F_1 = 0$. The fact that the number of correctly predicted tranquil periods does not enter the formula, introduces a bias into the F-measure (Powers, 2011). That is, it disregards the benefit an early-warning model generates in correctly predicting periods where no costly crisis-prevention mechanisms need to be enacted.

In contrast to relative usefulness and the F-measure, the two other performance measures do not rely on an additional preference parameter. The Brier probability score (Brier, 1950; Diebold and Rudebusch, 1989; Knedlik and von Schweinitz, 2012) operates directly on probabilities instead of signals. It is simply given by the mean of the squared differences between predicted probabilities and actual outcomes (i.e. a special case of mean squared forecast error for binary dependent variables). By contrast, the area under the (receiver-operator characteristic) curve (AUC or AUROC) does operate on signals, but aggregates type-1 errors and type-2 errors over *all* possible classification thresholds τ (Janes, Longton, and Pepe, 2009; Drehmann and Juselius, 2014). The AUC can take on values between 0 and 1, with 0 being a misleading, 0.5 an uninformative and 1 a perfect set of forecasts.[16]

A.3 Description of Estimation Methods

This section provides a brief overview of each method. Table A.2 summarizes the discussion in this section by highlighting some key benefits and drawbacks of the employed methods. Of course, this is only a snapshot of the more complete description of these methods in the mentioned references.[17]

Logistic regression (logit): Logit models are the workhorse models in the early warning literature (Frankel and Rose, 1996; Bussière and Fratzscher, 2006; Lo Duca and Peltonen, 2013). They are based on two assumptions. First, the dependent binary variable is assumed to be driven by a latent process y^*, which is in turn linearly related to the

[15] This formula follows from the definitions in Powers (2011).

[16] If the AUC takes the value 0, it is a perfect albeit negative signal. That is, an interpretation of low probabilities as signals of upcoming crises and vice-versa would result in a perfect prediction. However, this knowledge does not exist ex-ante. Therefore, misleading predictions and AUC values below 0.5 can occur in out-of-sample predictions. We thank a reviewer for pointing this out.

[17] More detailed introductions to these methods may be found, for instance, in Murphy (2012), James et al. (2013), or Shalev-Shwartz and Ben-David (2014).

Table A.2: Comparison of employed methods: benefits and drawbacks

	Benefits	**Drawbacks**
logit	explicit probabilistic foundations high interpretability	pre-specified functional form
knn	simple approach	strong curse of dimensionality
trees	automatic variable selection intuitive approach	instability across time / samples
rf	more stable than trees improves on tree accuracy	risk of overfitting complex drivers of predictions
svm	flexible nonlinear fitting computationally efficient	risk of overfitting ad hoc in probabilistic setups difficult to communicate
nen	flexible functional form recent advances in the literature	risk of overfitting computationally expensive difficult to communicate

employed explanatory variables: $y^* = X\beta + \varepsilon$. Second, the latent process is assumed to be linked to the binary variable by a logistic transformation (or, equivalently, estimation errors ε follow a logistic distribution). Hence, a key advantage of logit models is that they are based on a clear and straightforward statistical model, which explicitly takes uncertainty into account. Compared to machine learning methods, they are easy to interpret (for instance, in terms of coefficients), but, at the same time, restricted to the specific functional form just described. A key issue in their estimation is to make sure that a sufficient number of observations in each category is available (McFadden, 1984). In the context of early warning models, it is crucial to have a sufficient number of pre-crisis periods (which are much less frequent than tranquil periods) available for estimation. When the number of crisis events contained in the sample is reduced, estimation uncertainty increases, and, in the extreme case, perfect discrimination can prevent a proper estimation of the model's parameters. To put the logit method on equal footing with the machine learning methods, we estimate a non-dynamic logit model, pooling observations both in the cross-section and the time dimension. We do not include fixed effects.

K nearest neighbors (knn): The idea of knn[18] (Cover and Hart, 1967) is to predict the probability of an event (here: pre-crisis or tranquil period) for a given observation $X_{t,n}$, where X is a vector of early-warning indicators at time t in country n. The probability of being in a pre-crisis period conditional on the vector of observables $X_{t,n}$ is estimated by the share of pre-crisis observations among its K closest (nearest) neighbors. Closeness of two observations X and X', is measured by the Euclidean distance, i.e. $||X - X'|| = \sqrt{\sum_{i=1}^{d}(X_i - X_i')^2}$, where d refers to the number of early-warning indicators included in the model. That is, two observations are close if the realizations of the explanatory

[18] We implement knn using the R-package 'kknn'.

variables associated with these observations are similar.

The hyperparameter K is chosen by cross-validation. Moreover, we use a knn algorithm that refines the method by weighting each of the neighboring points by their distance to the given point $X_{t,n}$. A key problem of knn is that it is subject to a strong "curse of dimensionality". Shalev-Shwartz and Ben-David (2014) show that the sample size required to achieve a given error grows exponentially with the number of explanatory variables in the dataset. In our empirical application, we use sets of explanatory variables of different dimension in order to gain insights on the tradeoff between additional information and additional complexity.

Decision trees (trees): Binary decision trees[19] (Breiman, Friedman, Olshen, and Stone, 1984) essentially cluster observations into different groups by successively comparing the values of their early-warning indicators to specific thresholds. The key estimation task performed by the tree is to optimally determine these thresholds, and to decide on the sequence of variables to compare. Taken together, this determines the structure of the tree. Similar to KNN, the pre-crisis probability of an observation in a given final cluster (or node) is estimated as the average share of pre-crisis observations within that cluster. In contrast to KNN, which treats all explanatory variables uniformly, the tree optimally determines the importance of each explanatory variables, allowing to ignore variables deemed to be uninformative. Thus, the tree is, in principle, able to perform variable selection.

More formally, trees consist of a root, interior nodes (branches) and final nodes (leafs). The root and every branch consist of a decision rule based on a single explanatory variable X_i and a threshold τ_i. The decision rules assign observations to the left subtree if $X_i > \tau_i$ and to the right subtree otherwise. Starting at the root, observations are thus passed down the tree until they end up in a final node. For every node, the (predicted) probability of an event is equal to the average occurrence of said event among observations from the training sample assigned to the same final node.

The estimation of the tree entails choosing simultaneously the variables x and thresholds τ to split on. Efficient algorithms have been developed for approximating the optimal solution to this non-trivial task. These proceed by starting at the root and recursively constructing the tree from there, based on a measure of gain from each considered split and several potential stopping criteria for limiting the complexity of the tree. In our case, the number of branches is determined by a "pruning" parameter which balances increasing complexity against the homogeneity in final leaves.[20]

The selection of the pruning parameter (the hyperparameter of this method) thus decides on the complexity of the tree. Lower complexity costs imply additional splits which decrease classification errors on the training sample and thus increase the sharpness of estimated probabilities (pushing them closer to either zero or one). At the same time, the larger number of final nodes implies fewer training observations per final node, which increases estimation uncertainty and the potential for overfit. As the sensitivity of estimated trees to small changes in the underlying dataset can be high, the method of

[19] We implement decision trees using the R package 'rpart'.

[20] This is, of course, only one way to limit tree complexity. Other approaches, for example, set a minimum number of observations per final node (used in our implementation of random forest), or, alternatively, a maximum number of final nodes.

random forests has been developed to mitigate this undesirable feature.

Random forest (rf): Random forests[21] (Breiman, 1996, 2001) generalize decision trees by averaging over the predictions of a large number of different decision trees. This can reduce the variance of estimates and, hence, prediction errors. Random forests generate heterogeneity among its trees by (a) estimating trees on randomly chosen subsets of observations (also called bootstrap aggregating or bagging), and (b) considering only a randomly chosen subset of early warning indicators at each split (also called random subspace method, or attribute bagging). Both components are needed in order to decorrelate individual trees sufficiently, so as to achieve the desired variance reduction, while maintaining a high degree of prediction accuracy.

To put the random forest method into practice, we have to select three different hyperparameters. First, we set the number of trees used in each random forest to 1'000 such that the average prediction of the trees in the forest converges. Cross-validation is used to set the further two hyperparameters of this method. Heterogeneity between trees is driven largely by the number of randomly drawn variables to be considered at each split (the second hyperparameter). Third, complexity of the trees in the forest is limited by setting a minimum number of observations per terminal node, which is the third hyperparameter of this method.

Random forests have so far been the most frequently employed machine learning method in the early warning literature (Alessi and Detken, 2018; Holopainen and Sarlin, 2017; Tanaka et al., 2016). However, their success in reducing variance and improving out-of-sample performance depends on achieving a sufficiently low correlation between the randomly generated trees (Breiman, 2001). We conjecture that achieving such a low degree of correlation could be especially challenging in the presence of serial and cross-sectional correlation of the underlying training data.

Support vector machine (svm): svm[22] constructs a hyperplane in order to separate observations into distinct groups, pre-crisis and tranquil periods in our case. When the data is linearly separable, the main question is which hyperplane to choose from an infinite space of possible separating hyperplanes. svm uniquely determines the hyperplane by maximizing the distance of the two closest observations to the separating hyperplane (this distance is called margin).

For illustration, let us consider a one-dimensional example. Suppose a dataset is univariate, with observations given as points on the real line, namely $x = \{-3, -2, -1, 1, 2, 3\}$ and suppose that observations are linearly separable, namely $y(x = \{-3, -2\}) = 1$ and $y(x = \{-1, 1, 2, 3\}) = 0$. Then, obviously, any rule which assigns $y(x \leq -2) = 1$ and $y(x \geq -1) = 0$ perfectly separates the observations. The svm method would choose the point -1.5 for a separating rule that maximizes the margin. Obviously, we achieve separation by a point in this one-dimensional example, by a line in 2-d, and by a hyperplane in 3-d or higher.

However, in typical applications observations are not linearly separable. Consider a modification of the above example where $y = 1$ for all observations where $\mid x \mid> 2$ and zero

[21] We implement random forests using the R package 'randomForest'.

[22] We implement support vector machines using the R package 'e1071'.

otherwise (this example is inspired by Shalev-Shwartz and Ben-David (2014, p. 179)). This is not linearly separable in the original space, but can be made separable by mapping to a two-dimensional space, for instance by using $\phi : \mathbb{R} \rightarrow \mathbb{R}^2$ where $\phi(x) = (x, x^2)$. Then, any rule that assigns, $y = 1$ whenever $x^2 > 4$ separates the observations. This simple example illustrates the more general idea that is typically used in combination with svm: By mapping non-linear transformations of the original data into a higher dimensional space, linear separability of the dataset can be achieved or, at least, enhanced.

Mapping data into higher dimensional feature spaces enhances the expressiveness of methods (enlarging the space of functions considered for describing the data), but, obviously, higher dimensionality comes at the cost of increased complexity (the number of parameters can rise exponentially in the multivariate case, for example when polynomials using cross-products of variables are considered). To deal with this issue, the machine learning literature has developed the so-called 'kernel trick'. This allows an efficient computation of svm classifiers when such non-linear mappings into high-dimensional spaces are used. Broadly speaking, kernel functions describe similarities between observations and have special properties that allow the svm calculation to be based on these kernel functions without explicitly handling the high-dimensional representation of the data. A formal description of the kernel trick may be found, for instance, in Shalev-Shwartz and Ben-David (2014, p. 181). The complexity of the high-dimensional function space is controlled by a hyperparameter gamma inside the kernel function, with higher values of gamma leading to more complexity (we use the standard choice of a radial basis kernel).

While it is possible to linearly separate observations after mapping them into an arbitrarily complex space (n observations can always be perfectly fitted using a n-1 degree polynomial), this is generally not desirable. Instead, a penalty term for misclassified observations is added to the svm classifier loss function. Allowing for misclassification makes it possible to use the parameter svmg (see Table A.3) to separately control the complexity of the classifier. Moreover, it induces a tradeoff between large margins and misclassification, which is controlled by a second hyperparameter svmc (cost of soft margin constraint violation). The more tolerant we are towards misclassification on the training sample, the larger the margin can be (ceteris paribus). A larger margin then makes the classification more robust towards perturbations of the original data, for example when predicting the label of new data points. Thus, both hyperparameters seek to strike a balance between perfectly fitting the training data (potentially overfitting relative to the true model), and correctly classifying new data.

Support vector machines are among the most frequently used machine learning algorithms. Their ability to flexibly fit complex functions to the data at the same time entails the risk of overfitting. Moreover, the probabilistic foundations for the svm method are rather ad-hoc (Murphy, 2012). In the early warning context, the presence of a substantial unpredictable component as well as of cross-sectional and serial correlation may dampen the method's out-of-sample performance (see results and discussion). The communication of this method may present additional challenges.

Neural networks (nnet): Inspired by the structure of the human brain, artificial neural networks[23] consist of a directional network of simple neurons (i.e., network nodes),

[23] We implement neural networks using the R package 'nnet'.

arranged in sequential layers. Taken together, they are able to flexibly approximate functional forms linking input and output of the neural network. In the early warning context, the observations$X_{t,n}$ of early-warning indicators are the input to the network, while the probability of each observation belonging to a pre-crisis state conditional on $X_{t,n}$ are its output.

The neurons are the basic building blocks of the network. The input layer of our neural network with d early-warning indicators consists of $d+1$ neurons. That is, there is one neuron for each of the explanatory variables, and a neuron for a constant. The output layer of our neural network consists of a single neuron. Between the input layer and the output layer, we add a single hidden layer consisting of $1 \leq m \leq d$ hidden nodes.[24]

The output of each neuron of the input layer is the value of the corresponding explanatory variable. The neuron in the output layer and each neuron in the hidden layer receives as input a weighted sum of the output of all neurons in the previous layer. The output of these neurons is a logistic function of its inputs.

The basic architecture of the network, that is the number of hidden layers and the number of neurons in each hidden layer are exogenously given to the neural network algorithm as hyperparameters. The key task of the algorithm is to estimate the weights connecting the neurons of adjacent layers. As each neuron can be connected to each neuron in the previous layer, the number of weights can quickly become large, even in relatively simple architectures with few hidden layers. This makes neural networks universal function approximators, but also leads to a challenging estimation task. Generally, weights are estimated such that they minimize a (potentially non-convex) loss function. Efficient algorithms for dealing with this problem have been developed in the machine learning literature (see Shalev-Shwartz and Ben-David (2014) for an introduction).

Neural networks subsume simple logit models as a special case where there is no hidden layer, i.e. $J = 0$. In this case, each neuron of the input layer is directly linked to the output layer and the weights assigned to each input of the neuron in the output layer correspond to the coefficients of a logistic regression. This implies that in cases where a simple logit model is able to approximate the (sample) data relatively closely, the neural network is likely to yield predictions similar to that of the logit model. Generally, the logit model has the advantage of having to estimate fewer parameters, while the neural network is able to approximate more complicated functional forms.

Following Holopainen and Sarlin (2017) we focus on single-hidden layer neural networks, which can be substantially more flexible than logit models. Neural networks with more than two hidden layers are considered deep and have recently been successfully applied for instance to speech and image recognition (LeCun et al., 2015).

A.4 Cross-validation

We use cross-validation to select optimal hyperparameters from a predefined grid. The idea of cross-validation is to obtain an estimate of how well a model is able to make predictions on previously unseen data. To this end, the sample is cut repeatedly into an estimation sample and a test sample. In this sense, cross-validation is similar to out-of-sample prediction, but without paying as much attention to the time dimension of the dataset. In particular, we use panel block leave-p-out cross-validation (Arlot, Celisse,

[24] In general, there can be a number of say J so-called hidden layers each consisting of m_J neurons.

et al., 2010). In this variant, a block of twelve consecutive quarters (corresponding to the horizon of the early warning window) across all countries is used as test sample, while all other observations are included in the training sample. This is done repeatedly for all possible blocks until the sample observations are exhausted.

Using whole blocks of observations in the test sample instead of randomly selected observations has two advantages. First, it captures the serial and cross-sectional correlation of the data in a similar way as recursive out-of-sample estimation. Second, the number of possible splits of the dataset is limited, making an exhaustive cross-validation over all possible combinations possible. Using all possible splits into blocks of twelve quarters causes each observation (time t, country n) to be contained in twelve different panel blocks. That is, for each observation (t, n), the panel block leave-p-out cross-validation results in twelve different predictions for every model (with each model being defined by a combination of method, hyperparameters, and explanatory variables). In order to calculate the relative usefulness for a given model, we average performance over all cross-validation predictions from that model. We can then perform a grid search to select for each method the hyperparameters maximizing its cross-validation performance. Table A.3 displays the resulting optimal hyperparameters.

A.5 Panel-block-bootstrap

The aim of our panel-block-bootstrap is to draw random datasets with similar autocorrelation and cross-sectional dependence patterns as in the original dataset. To achieve this, we construct bootstrap datasets from blocks of observations that are jointly sampled from the original dataset. Drawing blocks of consecutive observations retains the autocorrelation structure of the data, while the panel structure of blocks captures cross-sectional correlation.

For every estimation, we sample $R = 1'000$ different bootstrap datasets from the respective training sample, which covers the time-country specific observations $\{(t, n) | t \in \{1, \ldots, T\},\ n \in \{1, \ldots, N\}\}$. A block B_t (with blocklength $b = 8$) starts at time t and contains the following observations of both X and $\bar{C}$:

$$B_t = \begin{bmatrix} (t, 1) & \cdots & (t, N) \\ \vdots & \ddots & \vdots \\ (t + b - 1, 1) & \cdots & (t + b - 1, N) \end{bmatrix} \tag{2}$$

Bootstrap samples $r \in \{1, \ldots, R\}$ are drawn randomly from the original data such that every observation has an equal probability of entering the random sample. Thus, we proceed as follows:

1. Initialize with an empty bootstrap sample $r = \emptyset$.
2. Draw a random starting period $t^* \in \{2 - b, \ldots, T\}$. If we would not allow for early or late starting periods (that effectively generate blocks with missing observations), observations at the beginning or end of the original sample would have a lower probability of entering the bootstrap sample.
3. Obtain B_{t^*}, corresponding to t^*, from the original training dataset. Some observations may be missing due to (a) shorter sample length for an individual country, (b)

an early starting period $t^* < 1$ or (c) a late starting period $t^* > T - b$. In this case, only include nonempty observations in B_{t^*}.

4. Concatenate the bootstrap sample r and B_{t^*}.

5. If the bootstrap sample r has fewer observations than the original in-sample dataset, return to step 2. Otherwise, return the bootstrap sample r.

We estimate every model m on every bootstrap sample $r \in \{1, \ldots, R\}$. Results are used to predict probability estimates $p_{t,n}^{m,r}$ for every observation (t, n) in the test sample. From this, we calculate the different performance measures (relative usefulness, AUC and BPS) for every bootstrap sample r and model m. The bootstrap distribution of performance measures across r yields estimates of confidence bands for each model m.[25] Moreover, it allows us to test whether model m_1's performance is statistically significantly better than model m_2's performance. For example, the probability that the relative usefulness of model m_1 is higher than that of model m_2 is given by $\frac{1}{R}\sum_{r=1}^{R} 1_{U_r^{m_1,r} > U_r^{m_2,r}}$.

[25] Conventional confidence bands from the 5^{th} to the 95^{th} quantile sometimes do not cover the point estimate. In case of relative usefulness and BPS, we therefore report confidence bands that cover both the point estimate and 90% of the probability mass. We still work with a sample of 1'000 boostrap draws even though we would need a higher number of bootstrap samples for confidence bands at extreme probabilities to converge (Davidson and MacKinnon, 2000). The reason is that our main focus is on the point estimates and the probability that one method outperforms another. In case of AUC, we recourse to a non-parametric approach developed specifically for this measure (DeLong, DeLong, and Clarke-Pearson, 1988).

Table A.3: Hyperparameters for machine learning methods (for baseline results)

Method	Hyperparameter name	Opt. value	Hyperparameter name	Opt. value
trees.1	cp (tree complexity parameter controlling cost of adding another split to the tree)	0.021212		
trees.2		0.027273		
trees.3		0.012121		
trees.4		0		
knn.1	k (number of nearest neighbours to use for each prediction)	49		
knn.2		7		
knn.3		29		
knn.4		15		
rf.1	nodesize (minimum number of observations per terminal node of each tree in the forest)	16	rfmtry (number of variables randomly sampled as candidates at each split)	2
rf.2		14		3
rf.3		2		2
rf.4		16		9
svm.1	svmg (parameter in radial basis function)	0.5	svmc (cost of soft margin constraint violation)	0.01
svm.2		0.5		0.026667
svm.3		0.000488		0.026667
svm.4		0.03125		8
nen.1	nendecay (learning rate parameter controlling the rate of convergence of the learning algorithm)	0.1	nensize (# nodes in hidden layer)	3
nen.2		0.1		2
nen.3		10		1
nen.4		0.1		7

7 Appendix B: Data

B.1 Sources of Vulnerabilities and Corresponding Indicators

Asset prices: Historically, banking crises have often been preceded by asset price booms. Banking crises associated with house price booms and busts, could, for example, not only be observed during the global financial crisis of 2008, but also in a number of industrial countries in the late 1970s to early 1990s, such as in Spain, Sweden, Norway, Finland, and Japan (Reinhart and Rogoff, 2008, 2009). We therefore include *house prices* and *equity prices* to capture booms and busts in asset prices.

Credit developments: High private sector indebtedness poses risks to the financial system when asset price booms are debt-financed, asset prices decrease and borrowers are unable to repay their debt (Kindleberger and Aliber, 2005; Jordà, Schularick, and Taylor, 2015). As a consequence of decreasing asset values, banks may be forced to deleverage, in particular when market liquidity is low and banks relying mainly on short-term funding face a liquidity mismatch (Brunnermeier and Oehmke, 2013; Brunnermeier, 2009). Deleveraging may induce a credit crunch and potentially lead to a recession. The effects of losses in asset values may be amplified by fire sales and may spill over to other assets as these are sold to meet regulatory and internal standards, such as capital and liquidity ratios. Moreover, bank runs may occur when the net worth of banks decreases and depositors lose confidence in the affected institutions (Allen and Gale, 2007). To capture risks related to high private sector indebtedness, we use *total credit to the private non-financial sector relative to GDP* as an indicator of how far credit developments are in line with real economic developments.

Macroeconomic environment: Closely related to credit and asset prices are real economic developments. On the one hand, rapid economic growth may increase risk appetite, asset prices and credit growth (Drehmann, Borio, and Tsatsaronis, 2011; Kindleberger and Aliber, 2005; Minsky, 1982). On the other hand, real economic downturns may lead to repayment difficulties on the borrower side inducing asset price declines and financial sector difficulties (Allen and Gale, 2007). To capture real economic developments we include *GDP*, *gross fixed capital formation relative to GDP* and *inflation*. Furthermore, we include *three-month interbank interest rates*, as banks and investors may take on excessive risks when interest rates are low and, hence, low-risk assets are less attractive (Maddaloni and Peydró, 2011; Allen and Gale, 2007; Rajan, 2006). Conversely, an abrupt increase in interest rates may put pressure on banks as well (Minsky, 1982).

External and global imbalances: The external sector played a prominent role in the first seminal contributions to the early warning literature (Frankel and Rose, 1996; Kaminsky and Reinhart, 1999). These papers tended to focus more on balance-of-payment crises than on systemic banking crises. However, both types of crises may occur jointly and often reinforce each other as "twin crises" (Kaminsky and Reinhart, 1999). While classic balance-of-payment crises may be less of a concern for the countries considered in this paper, external imbalances may still add to vulnerabilities. Similarly to the reasoning on credit expansion and asset prices, large capital inflows from abroad may support asset price booms and induce a reversal in asset prices when these inflows decline or stop (Kaminsky and Reinhart, 1999; Calvo, 1998). Hence, we include the *real effective exchange rate* and the *current account balance relative to GDP*. Furthermore, global shocks may affect the domestic banking system through various channels of contagion, such as financial sector interconnectedness and trade links (Kaminsky and Reinhart, 2000). We therefore add *oil prices* as an indicator for global developments.

Any list of potential early warning indicators is naturally incomplete. Yet, for the purpose of comparing predictions across methods, this is not the key point, as long as the same variables are used across all methods. Furthermore, it turns out that data availability is a key issue. While several additional variables would have been plausible predictors on economic grounds, these variables are not available for a long enough time span and/or not available for all countries in our sample. In addition, lack of comparability across countries can be an issue for some variables. For instance, while both theoretical and empirical arguments for the inclusion of a debt service ratio variable can be made (e.g. Drehmann and Juselius, 2014; Drehmann, Juselius, and Korinek, 2017), the extent to which this variable would truncate the sample outweighs its potential benefit in our case.[26] Another important class of variables that we cannot include are those based on bank balance sheet data, where availability in the time series is even much more restricted than for the debt service ratio. Nevertheless, we see in the results that the variables we were able to include do have substantial explanatory power for predicting banking crises. Thus, for the purpose of comparing different prediction methods (as opposed, say, to the question of finding the most important early warning indicator(s) as, for instance, in Drehmann and Juselius, 2014), having a sufficient number of observations in the sample appears to outweigh the benefits of using a complete set of all potentially important early warning indicators. Indeed, a robustness check using a shorter sample length shows that reducing the amount of observations available for estimation substantially reduces out-of-sample prediction performance (see Section 4.3).

B.2 Filtering

In the HP filtering approach, we transform early warning indicators into gaps by calculating deviations from the trend computed by a one-sided HP filter. Using a one-sided filter ensures that the information set at every point in time does not contain future information.[27] For variables such as total-credit-to-GDP ratio, real residential real estate prices and real equity prices we take into account recent evidence on lower frequency financial cycles, as documented, for instance, in Drehmann, Borio, and Tsatsaronis (2012),

[26] Proprietary debt service ratio data from the BIS is available starting at the earliest in 1980 (public debt service ratio data from 1999). We compared the availability of the debt service ratio by country with our sample of crises and early warning indicators described in Table B.1. Including the debt service ratio starting from 1980 would exclude five of the 13 crises episodes prior to the financial crisis of 2007/2008 and one crisis episode starting in 2008 (as BIS debt service ratio data is not available for Ireland). In total, a substantial amount of around 400 out of 1801 total observations would be excluded from our dataset.

[27] For the first k observations in every country, we need to apply a two-sided filter instead of the recursive one-sided version, given that the filter needs a certain minimum number of observations to compute a trend. We set k to ten years in the case of $\lambda = 400'000$ and to four years in the case of $\lambda = 1'600$.

and Schüler, Hiebert, and Peltonen (2015). Thus, for these variables we take the value of $\lambda = 400'000$ often employed for early warning models (Drehmann and Juselius, 2014), corresponding to financial cycles being roughly four times as long as business cycles, which is broadly in line with the findings in the aforementioned literature on financial cycles. Moreover, this ensures that the total credit-to-GDP gap used in our analysis is in line with the definitions of the Basel Committee on Banking Supervision (BCBS) used for Basel III and for setting countercyclical capital buffers (Drehmann and Juselius, 2014; Basel Committee on Banking Supervision, 2010). For typical business cycle variables such as real GDP, gross fixed capital formation-to-GDP, and the real oil price we use a standard HP filter smoothing parameter of $\lambda = 1'600$. In the case of the real effective exchange rate, we also use $\lambda = 400'000$. The reason for this is that real effective exchange rate imbalances have been found to be extremely persistent, especially since the introduction of the Euro made adjustments via nominal exchange rate movements impossible (El-Shagi, Lindner, and von Schweinitz, 2016).[28]

Robustness checks are performed by transforming the variables into growth rates. In line with the reasoning for applying different HP filter smoothing parameters for capturing business cycles and financial cycles, we also use two different growth rate horizons. Our business cycle variables are transformed into four-quarter growth rates, while our financial cycle variables are transformed into 16-quarter growth rates.

A detailed description of all variables and their transformations may be found in Table B.2 in the Appendix. A list of non-transformed original data with the corresponding sources is documented in Table B.3. Summary statistics of the transformed (standardized) predictor variables are displayed in Table B.4 for the HP filter and in Table B.5 for the growth rate transformation. When comparing the means of the selected indicators in the pre-crisis and tranquil periods, we note that the difference is particularly pronounced for the credit-to-GDP gap, the residential real estate price gap, the gross fixed capital formation-to-GDP gap and the current account balance relative to GDP. The first three of these indicators are, on average, higher and the current account balance to GDP is, on average, lower during pre-crisis periods. The volatility of these indicators is similar across pre-crisis and tranquil periods. Similar findings are obtained when using the growth rates transformation.

[28] We also follow Drehmann and Juselius (2014) in calculating relative gaps (i.e. the deviations from trend normalized by the trend) for certain HP filtered variables, which can be useful to improve the comparability of gaps across time and countries. Relative gaps are used for real equity prices, real residential real estate prices, the real oil price, real GDP and the real effective exchange rate.

B.3 Overview of Datasets

Table B.1: Country coverage and crisis dates

	Data availability			Crisis dates					
Country	**Start**	**End**	**No. of quarters**	**Start**	**End**	**Start**	**End**	**Start**	**End**
BE	1975 Q1	2016 Q2	166	no crisis					
DE	1971 Q1	2016 Q2	182	1974 Q2	1974 Q4	2001 Q1	2003 Q4		
DK	1975 Q1	2016 Q1	165	1987 Q1	1995 Q1	2008 Q1	2013 Q4		
ES	1975 Q1	2016 Q1	165	1978 Q1	1985 Q3	2009 Q1	2013 Q4		
FI	1981 Q4	2016 Q2	139	1991 Q3	1996 Q4				
FR	1973 Q1	2016 Q1	173	1991 Q2	1995 Q1	2008 Q2	2009 Q4		
GB	1972 Q1	2016 Q1	177	1973 Q4	1975 Q4	1991 Q3	1994 Q2	2007 Q3	2010 Q1
IE	1990 Q4	2016 Q1	102	2008 Q3	2013 Q4				
IT	1971 Q1	2016 Q1	181	1991 Q3	1997 Q4	2011 Q3	2013 Q4		
JP	1971 Q1	2016 Q2	182	1997 Q4	2001 Q4				
NL	1971 Q1	2016 Q1	181	2008 Q1	2013 Q1				
NO	1975 Q1	2016 Q2	166	1988 Q3	1992 Q4				
PT	1988 Q1	2016 Q2	114	2008 Q4	ongoing				
SE	1975 Q1	2016 Q1	165	1991 Q1	1997 Q2				
US	1971 Q1	2016 Q2	182	1988 Q1	1995 Q4	2007 Q4	2010 Q4		

Table B.2: Data transformation and sources - variables used

Category	**Variable name**	**Source**	**Transformation / Description (for non-transformed variables)**	**Inputs**	**Combinations**
Asset prices	Real equity price gap	OECD, Eurostat and own calculations	Relative gap using HP filter with lambda of 400,000	Equity prices, consumer price index	
Asset prices	Real residential real estate price gap	OECD, Eurostat and own calculations	Relative gap using HP filter with lambda of 400,000 of real residential real estate price	Residential real estate prices, consumer price index	
Credit	Total credit-to-GDP gap	BIS, OECD, Eurostat and own calculations	Absolute gap using HP filter with lambda of 400,000 of total credit-to-GDP ratio	Total credit, nominal GDP	
External	Current account-to-GDP ratio	OECD, Eurostat and own calculations	OECD data: Current account to GDP without transformations, Eurostat data: Current account to GDP calculated as ratio of current account balance to GDP (both summed up over four quarters)	Current account balance, Nominal GDP	Take longest time series available of OECD data or Eurostat data
External	Real oil price gap	OECD and own calculations	Relative gap using HP filter with lambda of 1,600 of real oil price	Oil price, consumer price index (US)	
External	Real effective exchange rate gap	OECD, IMF and own calculations	Relative gap using HP filter with lambda of 400,000 of real effective exchange rate	Real effective exchange rate	
Macro	3-month real money market rate	OECD, ECB, Eurostat and own calculations	Real interbank lending rate	Nominal 3-month money market rate, consumer price index	
Macro	Inflation rate	Eurostat, OECD	Annual rate of inflation (y-o-y growth rate of quarterly data)	Consumer price index	
Macro	Real GDP gap	OECD, Eurostat and own calculations	Relative gap using HP filter with lambda of 1,600 of real GDP	Nominal GDP, consumer price index	
Macro	Gross fixed capital formation-to-GDP gap	OECD, Eurostat, Bundesbank and own calculations	Absolute gap using HP filter with lambda of 1,600 of gross fixed capital formation-to-GDP ratio	Gross fixed capital formation, nominal GDP	

Table B.3: Data transformation and sources - input data

Variable name	**Source**	**Transformation / Description (for non-transformed variables)**	**Combinations**
Consumer price index	Eurostat, OECD	Consumer price index, end of quarter values, rebased to 2015=100	Take longest time series available of OECD data or Eurostat data
Total credit	BIS	Total credit to the private non-financial sector, domestic currency, billions	
Oil price	OECD	Brent crude oil price, USD per barrel	
Real effective exchange rate	OECD, IMF	Real effective exchange rate, CPI based index, base year: 2010	Take longest time series available of OECD data or IMF data
Current account balance	OECD, Eurostat	OECD: Current acount balance as percentage of GDP Eurostat: Current account balance (own calculations: sum of last four quarters, as percentage of GDP)	Take longest time series available of OECD data or Eurostat data
Nominal GDP (national currency)	OECD, Eurostat	Gross domestic product at market prices, seasonally adjusted, domestic currency, billions, sum of last four quarters	Take longest time series available of OECD data or Eurostat data
Nominal GDP (in EUR, for current account-to-GDP calculation)	Eurostat	Gross domestic product at market prices, seasonally adjusted, euro, millions, sum of last four quarters	
Gross fixed capital formation	OECD, Eurostat, Bundesbank	Gross fixed capital formation, domestic currency, millions. For DE: Bundesbank data (including calculations) for long time series of GFCF	Take longest time series available of OECD data or Eurostat data
3-month nominal money market rate	OECD, ECB, Datastream	Interbank interest rate, average through quarter	Take longest time series available of OECD, ECB and Datastream data
Equity prices	OECD, Bloomberg, Datastream	Equity price index, 2010=100, average through quarter	Take longest time series available of OECD, Bloomberg and Datastream data
Residential real estate prices	OECD	Index of residential real estate price, based in 2010, seasonally adjusted.	

Table B.4: Summary statistics: Gap dataset

	Pre-crisis periods					**Tranquil periods**				
Variable name	**mean**	**sd**	**min**	**max**	**obs**	**mean**	**sd**	**min**	**max**	**obs**
Total credit-to-GDP gap	0.76	1.03	-1.16	3.19	171	-0.08	0.96	-3.06	3.19	1608
Real residential real estate price gap	0.58	1.06	-1.36	2.48	171	-0.06	0.97	-2.58	2.48	1608
Real equity price gap	0.23	0.74	-1.73	1.68	171	-0.02	1.02	-1.96	3.61	1608
Real GDP gap	0.10	0.94	-3.15	2.20	171	-0.01	1.01	-4.81	3.71	1608
Inflation rate	-0.08	0.87	-1.24	3.64	171	0.01	1.01	-1.38	4.83	1608
Gross fixed capital formation-to-GDP gap	0.32	1.02	-2.46	2.97	171	-0.03	0.99	-4.63	3.19	1608
Real 3-month money market rate	0.06	1.10	-3.29	2.32	171	-0.01	0.99	-3.29	2.61	1608
Current account-to-GDP ratio	-0.73	1.08	-3.50	2.17	171	0.08	0.96	-3.50	2.75	1608
Real effective exchange rate gap	0.20	0.85	-2.08	2.72	171	-0.02	1.01	-2.61	2.72	1608
Real oil price gap	0.26	1.00	-2.56	2.34	171	-0.03	1.00	-2.56	3.13	1608

Note: Data have been standardized using the unconditional mean and standard deviation across all periods. Since data are winsorized at the 1% and 99% level, minimum (maximum) values may be the same in pre-crisis and tranquil periods.

Table B.5: Summary statistics: Growth rate dataset

	Pre-crisis periods					**Tranquil periods**				
Variable name	**mean**	**sd**	**min**	**max**	**obs**	**mean**	**sd**	**min**	**max**	**obs**
4-year growth rate of credit-to-GDP ratio	0.62	1.04	-1.19	3.46	176	-0.07	0.97	-2.34	3.46	1525
4-year growth rate of real residential real estate prices	0.52	1.09	-1.27	2.63	176	-0.06	0.97	-2.47	2.63	1525
4-year growth rate of real equity prices	0.55	1.01	-1.12	3.67	176	-0.06	0.98	-1.35	3.67	1525
1-year growth rate of real GDP	0.09	0.88	-3.38	1.78	176	-0.01	1.01	-3.98	4.65	1525
Inflation rate	-0.07	0.90	-1.22	3.87	176	0.01	1.01	-1.37	5.11	1525
1-year growth rate of gross fixed capital formation-to-GDP ratio	0.27	1.09	-3.96	3.75	176	-0.03	0.98	-3.96	3.75	1525
3-month real money market rate	0.07	1.10	-3.33	2.38	176	-0.01	0.99	-3.33	2.67	1525
Current account-to-GDP ratio	-0.75	1.14	-3.22	2.06	176	0.09	0.95	-3.22	2.60	1525
4-year growth rate of real effective exchange rate	0.24	0.82	-1.45	3.48	176	-0.03	1.02	-2.21	3.48	1525
1-year growth rate of real oil price	0.02	0.88	-1.82	3.10	176	0.00	1.01	-1.82	3.79	1525

Note: Data have been standardized using the unconditional mean and standard deviation across all periods. Since data is winsorized at the 1% and 99% level, minimum (maximum) values may be the same in pre-crisis and tranquil periods.

Table B.6: Variables used in specifications (1) - (4)

Credit and Asset Prices	**Macro**	**External**	**All**
(1)	(2)	(3)	(4)
Total credit-to-GDP gap	Real GDP gap	Current account-to-GDP ratio	Total credit-to-GDP gap
Real residential real estate price gap	Inflation rate	Real effective exchange rate gap	Real residential real estate price gap
Real equity price gap	3-month real money market rate	Real oil price gap	Real equity price gap
	Gross fixed capital formation-to-GDP gap		Real GDP gap
			Inflation rate
			3-month real money market rate
			Gross fixed capital formation-to-GDP gap
			Current account-to-GDP ratio
			Real effective exchange rate gap
			Real oil price gap

8 Appendix C: Results

Table C.1: *In-sample* performance using different performance measures

	Threshold	TP	FP	TN	FN	U_r		F_1		AUC		BPS	
logit.1	0.109	102	402	1207	69	0.347**	[0.185, 0.440]	0.302**	[0.217, 0.352]	0.736***	[0.703, 0.769]	0.077**	[0.076, 0.092]
trees.1	0.097	84	96	1513	87	0.432**	[0.158, 0.601]	0.479***	[0.246, 0.541]	0.776***	[0.742, 0.810]	0.063***	[0.060, 0.101]
knn.1	0.126	152	315	1294	19	0.693*	[0.353, 0.693]	0.476*	[0.327, 0.561]	0.928***	[0.915, 0.942]	0.057*	[0.057, 0.086]
rf.1	0.228	170	45	1564	1	**0.966**	[0.411, 0.966]	**0.881**	[0.432, 0.881]	**0.994**	[0.992, 0.996]	**0.029**	[0.029, 0.078]
svm.1	0.079	157	895	714	14	0.362**	[-0.006, 0.483]	0.257**	[0.092, 0.437]	0.837***	[0.808, 0.866]	0.073**	[0.064, 0.094]
nen.1	0.091	121	405	1204	50	0.456**	[0.201, 0.503]	0.347**	[0.238, 0.467]	0.777***	[0.743, 0.812]	0.070**	[0.069, 0.092]
logit.2	0.098	107	682	927	64	0.202***	[0.018, 0.236]	0.223*	[0.141, 0.232]	0.613***	[0.576, 0.651]	0.086	[0.086, 0.099]
trees.2	0.096	171	1609	0	0	0.000***	[-0.031, 0.216]	0.175**	[0.033, 0.278]	0.500***	[0.500, 0.500]	0.087	[0.078, 0.109]
knn.2	0.315	171	57	1552	0	0.965	[0.265, 0.965]	0.857	[0.350, 0.857]	0.988***	[0.984, 0.991]	**0.036**	[0.036, 0.101]
rf.2	0.189	170	45	1564	1	**0.966**	[0.240, 0.966]	**0.881**	[0.308, 0.881]	**0.995**	[0.994, 0.997]	0.036	[0.036, 0.090]
svm.2	0.097	92	91	1518	79	0.481**	[0.129, 0.489]	0.520*	[0.206, 0.520]	0.867***	[0.842, 0.891]	0.074	[0.074, 0.113]
nen.2	0.130	74	268	1341	97	0.266***	[0.042, 0.294]	0.288*	[0.136, 0.315]	0.678***	[0.640, 0.716]	0.081	[0.081, 0.099]
logit.3	0.097	118	482	1127	53	0.390*	[0.243, 0.422]	0.306**	[0.255, 0.358]	0.745***	[0.712, 0.778]	0.081**	[0.081, 0.090]
trees.3	0.096	171	1609	0	0	0.000**	[0.000, 0.410]	0.175***	[0.155, 0.404]	0.500***	[0.500, 0.500]	0.087***	[0.074, 0.107]
knn.3	0.150	148	291	1318	23	0.685	[0.302, 0.685]	0.485**	[0.321, 0.485]	0.919***	[0.905, 0.932]	0.062**	[0.062, 0.090]
rf.3	0.373	171	0	1609	0	**1.000**	[0.299, 1.000]	**1.000**	[0.411, 1.000]	**1.000**	[1.000, 1.000]	**0.012**	[0.012, 0.077]
svm.3	0.042	171	1609	0	0	0.000**	[-0.258, 0.320]	0.175**	[0.021, 0.313]	0.390***	[0.350, 0.431]	0.089**	[0.084, 0.800]
nen.3	0.156	117	494	1115	54	0.377*	[0.263, 0.419]	0.299**	[0.240, 0.332]	0.744***	[0.711, 0.777]	0.089**	[0.087, 0.122]
logit.4	0.109	119	297	1312	52	0.511*	[0.208, 0.561]	0.405**	[0.250, 0.424]	0.810***	[0.779, 0.840]	0.073**	[0.073, 0.098]
trees.4	0.100	138	214	1395	33	0.674*	[0.230, 0.711]	0.528***	[0.301, 0.689]	0.901***	[0.879, 0.923]	0.041***	[0.041, 0.111]
knn.4	0.339	166	26	1583	5	0.955	[0.370, 0.955]	0.915	[0.419, 0.915]	0.997***	[0.996, 0.999]	0.021	[0.021, 0.083]
rf.4	0.280	171	13	1596	0	**0.992**	[0.347, 0.992]	**0.963**	[0.392, 0.963]	**0.999**	[0.999, 1.000]	**0.017**	[0.017, 0.082]
svm.4	0.110	143	84	1525	28	0.784	[0.334, 0.784]	0.719*	[0.354, 0.719]	0.946***	[0.930, 0.961]	0.045*	[0.044, 0.101]
nen.4	0.230	139	78	1531	32	0.764	[0.338, 0.793]	0.716	[0.400, 0.794]	0.882***	[0.846, 0.917]	0.037	[0.034, 0.083]

Note: Best performance on given dataset in bold. Stars indicate if the respective method's performance is significantly worse than the best model's performance on the same dataset (***/**/* for the 1%/5%/10% significance level). For U_r, F_1 and AUC higher values indicate better prediction performance, while for BPS lower values are preferable. Numbers in brackets indicate 90% confidence bands. Model names reported in format "method.dataset", where datasets consist of the following sets of variables (1): credit and asset prices, (2) macro variables, (3) external imbalances, and (4) all variables (see also Table B.6).

Table C.2: *Out-of-sample* performance using different performance measures

	Threshold	TP	FP	TN	FN	U_r		F_1		AUC		BPS	
logit.1	0.096	46	96	145	14	**0.368**	[0.143, 0.431]	**0.455**	[0.345, 0.481]	**0.737**	[0.682, 0.792]	**0.139**	[0.131, 0.177]
trees.1	0.065	24	48	193	36	0.201	[0.084, 0.317]	0.364*	[0.265, 0.456]	0.591***	[0.510, 0.672]	0.148*	[0.142, 0.202]
knn.1	0.124	30	61	180	30	0.247	[0.088, 0.384]	0.397	[0.279, 0.481]	0.704	[0.644, 0.765]	0.153	[0.141, 0.185]
rf.1	0.197	22	29	212	38	0.246	[0.129, 0.350]	0.396	[0.293, 0.478]	0.725	[0.666, 0.785]	0.150	[0.134, 0.177]
svm.1	0.078	27	56	185	33	0.218*	[0.018, 0.326]	0.378***	[0.267, 0.441]	0.507***	[0.418, 0.595]	0.161***	[0.161, 0.237]
nen.1	0.124	26	44	197	34	0.251	[0.080, 0.451]	0.400	[0.281, 0.515]	0.711	[0.646, 0.776]	0.144	[0.131, 0.181]
logit.2	0.100	7	85	156	53	-0.236	[-0.386, -0.036]	0.092	[0.037, 0.253]	0.261	[0.203, 0.318]	**0.182**	[0.177, 0.207]
trees.2	0.076	8	178	63	52	-0.605	[-0.605, -0.087]	0.065	[0.019, 0.224]	0.176	[0.126, 0.227]	0.204***	[0.193, 0.256]
knn.2	0.309	4	37	204	56	-0.087	[-0.137, 0.034]	0.079	[0.000, 0.186]	**0.385**	[0.339, 0.431]	0.231***	[0.210, 0.259]
rf.2	0.204	0	59	182	60	-0.245	[-0.323, -0.132]	0.000	[0.000, 0.092]	0.231	[0.185, 0.276]	0.227***	[0.214, 0.260]
svm.2	0.087	17	72	169	43	**-0.015**	[-0.257, 0.060]	**0.228**	[0.076, 0.306]	0.312	[0.242, 0.382]	0.182**	[0.182, 0.260]
nen.2	0.098	0	84	157	60	-0.349	[-0.382, -0.048]	0.000	[0.000, 0.209]	0.195	[0.152, 0.238]	0.190	[0.179, 0.212]
logit.3	0.087	40	55	186	20	**0.438**	[0.309, 0.563]	**0.516**	[0.438, 0.600]	**0.762**	[0.699, 0.825]	**0.153**	[0.131, 0.164]
trees.3	0.077	56	131	110	4	0.390***	[0.072, 0.390]	0.453***	[0.237, 0.514]	0.618***	[0.567, 0.670]	0.181***	[0.154, 0.206]
knn.3	0.151	30	50	191	30	0.293***	[0.133, 0.417]	0.429**	[0.305, 0.510]	0.663***	[0.594, 0.732]	0.159**	[0.148, 0.181]
rf.3	0.423	10	7	234	50	0.138***	[0.055, 0.283]	0.260**	[0.140, 0.351]	0.615***	[0.541, 0.689]	0.161**	[0.145, 0.187]
svm.3	0.094	37	100	141	23	0.202**	[0.043, 0.409]	0.376***	[0.277, 0.491]	0.291	[0.233, 0.348]	0.176***	[0.150, 0.219]
nen.3	0.157	41	60	181	19	0.434	[0.322, 0.551]	0.509	[0.448, 0.595]	0.584***	[0.523, 0.644]	0.161	[0.155, 0.164]
logit.4	0.091	45	35	206	15	**0.605**	[0.222, 0.605]	**0.643**	[0.383, 0.643]	**0.852**	[0.797, 0.906]	**0.125**	[0.106, 0.178]
trees.4	0.065	18	42	199	42	0.126***	[-0.037, 0.313]	0.300***	[0.131, 0.455]	0.456***	[0.383, 0.530]	0.255***	[0.215, 0.289]
knn.4	0.287	4	31	210	56	-0.062***	[-0.166, 0.042]	0.084***	[0.000, 0.196]	0.366***	[0.312, 0.419]	0.217***	[0.167, 0.239]
rf.4	0.265	10	41	200	50	-0.003***	[-0.141, 0.146]	0.180***	[0.060, 0.319]	0.525***	[0.458, 0.592]	0.199***	[0.147, 0.222]
svm.4	0.105	19	92	149	41	-0.065***	[-0.173, 0.130]	0.222***	[0.129, 0.322]	0.486***	[0.423, 0.548]	0.217***	[0.170, 0.269]
nen.4	0.292	8	10	231	52	0.092**	[-0.041, 0.250]	0.205	[0.109, 0.400]	0.747***	[0.694, 0.801]	0.170	[0.141, 0.190]

Note: Best performance on given dataset in bold. Stars indicate if the respective method's performance is significantly worse than the best model's performance on the same dataset (***/**/* for the 1%/5%/10% significance level). For U_r, F_1 and AUC higher values indicate better prediction performance, while for BPS lower values are preferable. Numbers in brackets indicate 90% confidence bands. Model names reported in format "method.dataset", where datasets consist of the following sets of variables (1): credit and asset prices, (2) macro variables, (3) external imbalances, and (4) all variables (see also Table B.6).

Table C.3: Robustness: *Out-of-sample* performance for $\mu = 0.75$

	Threshold	TP	FP	TN	FN	U_r		F_1		AUC		BPS	
logit.1	0.026	60	182	59	0	**0.245**	[-0.527, 0.378]	0.397	[0.335, 0.444]	**0.737**	[0.682, 0.792]	**0.139**	[0.128, 0.178]
trees.1	0.086	60	241	0	0	0.000	[-0.886, 0.099]	0.332	[0.271, 0.388]	0.235	[0.186, 0.283]	0.177*	[0.147, 0.199]
knn.1	0.066	42	98	143	18	-0.307***	[-1.078, 0.006]	**0.420**	[0.299, 0.495]	0.680*	[0.610, 0.749]	0.155	[0.134, 0.180]
rf.1	0.177	23	30	211	37	-0.974***	[-1.378, -0.599]	0.407	[0.305, 0.590]	0.736	[0.678, 0.794]	0.150	[0.119, 0.178]
svm.1	0.044	41	167	74	19	-0.643***	[-1.235, -0.139]	0.306*	[0.242, 0.383]	0.529***	[0.449, 0.610]	0.144**	[0.142, 0.224]
nen.1	0.000	60	241	0	0	0.000	[-0.296, 0.058]	0.332***	[0.251, 0.341]	0.500***	[0.500, 0.500]	0.199***	[0.185, 0.204]
logit.2	0.025	60	241	0	0	**0.000**	[-1.051, 0.032]	**0.332**	[0.249, 0.376]	0.261	[0.203, 0.318]	0.182	[0.176, 0.207]
trees.2	0.069	60	241	0	0	0.000*	[-1.741, 0.083]	0.332***	[0.155, 0.365]	0.176	[0.126, 0.227]	0.204	[0.190, 0.254]
knn.2	0.077	22	128	113	38	-1.431***	[-1.998, -0.994]	0.210	[0.098, 0.296]	0.294	[0.238, 0.350]	0.197	[0.182, 0.209]
rf.2	0.146	0	77	164	60	-2.320***	[-2.353, -1.936]	0.000***	[0.000, 0.103]	0.231	[0.186, 0.276]	0.224**	[0.209, 0.263]
svm.2	0.076	57	235	6	3	-0.125	[-0.622, 0.116]	0.324	[0.289, 0.369]	0.272	[0.215, 0.330]	**0.178**	[0.166, 0.228]
nen.2	0.000	60	241	0	0	0.000	[-0.309, 0.100]	0.332	[0.265, 0.339]	**0.500**	[0.500, 0.500]	0.199	[0.194, 0.206]
logit.3	0.050	50	110	131	10	0.044	[-0.394, 0.127]	0.455	[0.385, 0.492]	**0.762**	[0.699, 0.825]	**0.153**	[0.136, 0.165]
trees.3	0.089	60	241	0	0	0.000*	[-0.914, 0.078]	0.332***	[0.275, 0.403]	0.233	[0.180, 0.286]	0.177***	[0.146, 0.195]
knn.3	0.071	41	79	162	19	-0.278**	[-0.874, -0.236]	0.456***	[0.347, 0.494]	0.656***	[0.585, 0.727]	0.165**	[0.146, 0.176]
rf.3	0.133	18	42	199	42	-1.274***	[-1.545, -0.878]	0.300***	[0.155, 0.383]	0.605***	[0.533, 0.677]	0.171***	[0.148, 0.189]
svm.3	0.069	59	211	30	1	**0.074**	[-0.381, 0.212]	0.358***	[0.316, 0.384]	0.328*	[0.263, 0.393]	0.176***	[0.149, 0.222]
nen.3	0.077	50	107	134	10	0.056	[-0.381, 0.152]	**0.461**	[0.388, 0.500]	0.757	[0.700, 0.814]	0.159	[0.149, 0.175]
logit.4	0.046	55	111	130	5	**0.289**	[-0.598, 0.289]	**0.487**	[0.368, 0.533]	**0.852**	[0.797, 0.906]	**0.125**	[0.114, 0.177]
trees.4	0.089	60	241	0	0	0.000	[-0.793, 0.220]	0.332**	[0.281, 0.404]	0.233**	[0.180, 0.286]	0.177**	[0.147, 0.210]
knn.4	0.165	6	47	194	54	-1.895***	[-2.128, -1.478]	0.106***	[0.020, 0.243]	0.381***	[0.319, 0.443]	0.199***	[0.166, 0.216]
rf.4	0.279	9	39	202	51	-1.712***	[-2.049, -1.169]	0.167***	[0.041, 0.323]	0.519***	[0.453, 0.586]	0.199***	[0.161, 0.221]
svm.4	0.011	60	241	0	0	0.000	[-0.921, 0.100]	0.332**	[0.256, 0.361]	0.238**	[0.190, 0.287]	0.179**	[0.171, 0.232]
nen.4	0.000	60	241	0	0	0.000	[-0.913, 0.054]	0.332***	[0.250, 0.347]	0.500***	[0.500, 0.500]	0.199***	[0.183, 0.211]

Note: Best performance on given dataset in bold. Stars indicate if the respective method's performance is significantly worse than the best model's performance on the same dataset (***/**/* for the 1%/5%/10% significance level). For U_r, F_1 and AUC higher values indicate better prediction performance, while for BPS lower values are preferable. Numbers in brackets indicate 90% confidence bands. Model names reported in format "method.dataset", where datasets consist of the following sets of variables (1): credit and asset prices, (2) macro variables, (3) external imbalances, and (4) all variables (see also Table B.6).

Table C.4: Robustness: *Out-of-sample* performance when using growth rates (instead of gaps) as data transformation

	Threshold	TP	FP	TN	FN	U_r		F_1		AUC		BPS	
logit.1	0.092	39	74	179	34	0.242	[0.077, 0.342]	0.419	[0.317, 0.480]	**0.746**	[0.696, 0.796]	0.169	[0.158, 0.190]
trees.1	0.071	31	46	207	42	0.243	[0.063, 0.347]	0.413**	[0.239, 0.490]	0.450***	[0.374, 0.527]	0.189**	[0.174, 0.224]
knn.1	0.111	37	54	199	36	**0.293**	[0.084, 0.329]	**0.451**	[0.293, 0.477]	0.638***	[0.570, 0.707]	0.181	[0.171, 0.200]
rf.1	0.214	25	29	224	48	0.228	[0.083, 0.314]	0.394	[0.243, 0.472]	0.700*	[0.641, 0.759]	**0.166**	[0.162, 0.202]
svm.1	0.100	30	85	168	43	0.075	[-0.108, 0.206]	0.319**	[0.172, 0.404]	0.445***	[0.378, 0.512]	0.195**	[0.174, 0.259]
nen.1	0.119	36	62	191	37	0.248	[0.135, 0.356]	0.421	[0.320, 0.480]	0.711***	[0.656, 0.766]	0.180	[0.164, 0.189]
logit.2	0.095	10	130	123	63	-0.377	[-0.432, -0.109]	0.094	[0.048, 0.217]	**0.236**	[0.185, 0.287]	**0.209**	[0.204, 0.234]
trees.2	0.088	13	181	72	60	-0.537	[-0.537, -0.110]	**0.097**	[0.031, 0.277]	0.193*	[0.154, 0.232]	0.225***	[0.212, 0.269]
knn.2	0.159	1	55	198	72	-0.204	[-0.281, -0.105]	0.016	[0.000, 0.082]	0.216	[0.174, 0.258]	0.230*	[0.221, 0.242]
rf.2	0.199	0	43	210	73	**-0.170**	[-0.245, -0.103]	0.000	[0.000, 0.086]	0.202*	[0.158, 0.245]	0.248***	[0.237, 0.272]
svm.2	0.095	3	69	184	70	-0.232	[-0.315, -0.110]	0.041	[0.000, 0.173]	0.171**	[0.131, 0.211]	0.220***	[0.220, 0.301]
nen.2	0.121	6	86	167	67	-0.258	[-0.332, -0.095]	0.073	[0.014, 0.147]	0.214	[0.168, 0.260]	0.219	[0.212, 0.242]
logit.3	0.079	47	81	172	26	**0.324**	[0.193, 0.448]	**0.468**	[0.388, 0.558]	**0.693**	[0.628, 0.758]	**0.172**	[0.157, 0.190]
trees.3	0.070	49	194	59	24	-0.096**	[-0.156, 0.197]	0.310***	[0.200, 0.519]	0.583***	[0.508, 0.658]	0.196***	[0.190, 0.236]
knn.3	0.096	48	96	157	25	0.278	[0.107, 0.352]	0.442***	[0.316, 0.489]	0.679	[0.625, 0.734]	0.180***	[0.174, 0.206]
rf.3	0.257	11	26	227	62	0.048***	[-0.015, 0.200]	0.200***	[0.128, 0.368]	0.667	[0.614, 0.720]	0.194***	[0.177, 0.219]
svm.3	0.089	28	106	147	45	-0.035***	[-0.193, 0.107]	0.271***	[0.192, 0.350]	0.408**	[0.354, 0.462]	0.201***	[0.186, 0.266]
nen.3	0.105	44	81	172	29	0.283	[0.103, 0.388]	0.444***	[0.317, 0.519]	0.663***	[0.601, 0.724]	0.180***	[0.172, 0.203]
logit.4	0.092	51	54	199	22	**0.485**	[0.153, 0.485]	**0.573**	[0.356, 0.573]	0.712***	[0.648, 0.776]	**0.163**	[0.145, 0.204]
trees.4	0.083	3	7	246	70	0.013***	[-0.133, 0.106]	0.072***	[0.020, 0.279]	0.661***	[0.609, 0.712]	0.224***	[0.207, 0.273]
knn.4	0.309	2	20	233	71	-0.052***	[-0.121, 0.019]	0.042***	[0.000, 0.165]	0.377***	[0.329, 0.426]	0.230***	[0.206, 0.247]
rf.4	0.264	0	14	239	73	-0.055***	[-0.149, 0.100]	0.000**	[0.000, 0.286]	0.495***	[0.432, 0.558]	0.205**	[0.176, 0.221]
svm.4	0.062	25	93	160	48	-0.025***	[-0.203, 0.086]	0.262***	[0.174, 0.400]	0.506***	[0.442, 0.569]	0.197***	[0.197, 0.287]
nen.4	0.140	39	30	223	34	0.416	[0.157, 0.434]	0.549	[0.343, 0.556]	**0.791**	[0.737, 0.845]	0.164	[0.149, 0.185]

Note: Best performance on given dataset in bold. Stars indicate if the respective method's performance is significantly worse than the best model's performance on the same dataset (***/**/* for the 1%/5%/10% significance level). For U_r, F_1 and AUC higher values indicate better prediction performance, while for BPS lower values are preferable. Numbers in brackets indicate 90% confidence bands. Model names reported in format "method.dataset", where datasets consist of the following sets of variables (1): credit and asset prices, (2) macro variables, (3) external imbalances, and (4) all variables (see also Table B.6).

Table C.5: Robustness: *Out-of-sample* performance when starting the dataset in 1980Q1 (instead of 1970Q1)

	Threshold	TP	FP	TN	FN	U_r		F_1		AUC		BPS	
logit.1	0.104	37	81	160	23	0.281	[0.097, 0.435]	0.416	[0.319, 0.490]	**0.732**	[0.676, 0.788]	**0.139**	[0.125, 0.180]
trees.1	0.077	26	52	189	34	0.218	[0.072, 0.368]	0.377**	[0.247, 0.488]	0.441***	[0.355, 0.527]	0.156**	[0.148, 0.205]
knn.1	0.129	26	52	189	34	0.218	[0.063, 0.347]	0.377	[0.262, 0.467]	0.622***	[0.546, 0.699]	0.158	[0.134, 0.185]
rf.1	0.214	19	31	210	41	0.188	[0.071, 0.292]	0.345	[0.228, 0.426]	0.680*	[0.616, 0.744]	0.155	[0.137, 0.185]
svm.1	0.084	28	87	154	32	0.106	[-0.015, 0.276]	0.320***	[0.219, 0.417]	0.526***	[0.441, 0.610]	0.162***	[0.162, 0.249]
nen.1	0.123	27	37	204	33	**0.296**	[0.055, 0.405]	**0.435**	[0.258, 0.500]	0.692**	[0.625, 0.759]	0.144	[0.134, 0.182]
logit.2	0.100	5	83	158	55	-0.261	[-0.369, -0.066]	0.068	[0.039, 0.214]	0.235***	[0.179, 0.291]	0.189	[0.181, 0.221]
trees.2	0.089	0	34	207	60	-0.141	[-0.220, 0.072]	0.000**	[0.000, 0.118]	0.173***	[0.131, 0.215]	0.234**	[0.218, 0.292]
knn.2	0.130	1	85	156	59	-0.336	[-0.369, -0.120]	0.014	[0.000, 0.124]	0.221***	[0.171, 0.270]	0.205	[0.194, 0.217]
rf.2	0.176	0	46	195	60	-0.191	[-0.282, -0.132]	0.000	[0.000, 0.071]	0.212***	[0.169, 0.255]	0.211	[0.201, 0.231]
svm.2	0.110	15	79	162	45	**-0.078**	[-0.253, 0.018]	**0.195**	[0.053, 0.241]	0.283*	[0.224, 0.343]	**0.185**	[0.185, 0.249]
nen.2	0.140	1	97	144	59	-0.386	[-0.423, -0.094]	0.013	[0.000, 0.160]	**0.356**	[0.302, 0.409]	0.191	[0.187, 0.233]
logit.3	0.096	39	59	182	21	**0.405**	[0.264, 0.513]	**0.494**	[0.408, 0.566]	**0.753**	[0.691, 0.815]	**0.153**	[0.124, 0.168]
trees.3	0.067	41	107	134	19	0.239**	[0.025, 0.455]	0.394***	[0.212, 0.534]	0.595***	[0.526, 0.664]	0.175***	[0.159, 0.217]
knn.3	0.147	19	51	190	41	0.105***	[-0.004, 0.209]	0.292***	[0.217, 0.385]	0.555***	[0.486, 0.625]	0.175***	[0.152, 0.198]
rf.3	0.393	7	18	223	53	0.042***	[-0.112, 0.134]	0.165***	[0.000, 0.277]	0.563***	[0.492, 0.635]	0.179***	[0.163, 0.206]
svm.3	0.103	21	131	110	39	-0.194***	[-0.194, 0.080]	0.198***	[0.118, 0.286]	0.234	[0.177, 0.291]	0.183***	[0.168, 0.232]
nen.3	0.121	37	64	177	23	0.351*	[0.101, 0.480]	0.460***	[0.293, 0.563]	0.664***	[0.597, 0.732]	0.164***	[0.143, 0.200]
logit.4	0.090	37	27	214	23	**0.505**	[0.180, 0.559]	**0.597**	[0.351, 0.597]	**0.809**	[0.754, 0.863]	**0.143**	[0.127, 0.192]
trees.4	0.070	1	51	190	59	-0.195***	[-0.207, 0.118]	0.018***	[0.018, 0.290]	0.172	[0.125, 0.219]	0.298***	[0.223, 0.303]
knn.4	0.267	2	36	205	58	-0.116***	[-0.216, 0.009]	0.041***	[0.000, 0.178]	0.355***	[0.302, 0.409]	0.211***	[0.173, 0.231]
rf.4	0.191	2	28	213	58	-0.083***	[-0.211, 0.139]	0.044	[0.044, 0.322]	0.429***	[0.366, 0.491]	0.180	[0.169, 0.196]
svm.4	0.104	14	68	173	46	-0.049***	[-0.170, 0.076]	0.197***	[0.129, 0.324]	0.474***	[0.409, 0.539]	0.210***	[0.181, 0.270]
nen.4	0.054	9	9	232	51	0.113**	[-0.058, 0.226]	0.231**	[0.094, 0.378]	0.736**	[0.677, 0.794]	0.187**	[0.149, 0.224]

Note: Best performance on given dataset in bold. Stars indicate if the respective method's performance is significantly worse than the best model's performance on the same dataset (***/**/* for the 1%/5%/10% significance level). For U_r, F_1 and AUC higher values indicate better prediction performance, while for BPS lower values are preferable. Numbers in brackets indicate 90% confidence bands. Model names reported in format "method.dataset", where datasets consist of the following sets of variables (1): credit and asset prices, (2) macro variables, (3) external imbalances, and (4) all variables (see also Table B.6).

Table C.6: Robustness: *Out-of-sample* performance when using Laeven & Valencia crisis database

	Threshold	TP	FP	TN	FN	U_r		F_1		AUC		BPS	
logit.1	0.048	63	61	131	26	**0.390**	[0.072, 0.443]	**0.592**	[0.357, 0.620]	0.575***	[0.514, 0.637]	0.267	[0.230, 0.294]
trees.1	0.045	17	16	176	72	0.108	[-0.112, 0.436]	0.279*	[0.044, 0.519]	0.217	[0.155, 0.280]	0.263	[0.257, 0.312]
knn.1	0.090	42	20	172	47	0.368	[0.028, 0.402]	0.556**	[0.208, 0.586]	0.587***	[0.522, 0.652]	0.277	[0.240, 0.304]
rf.1	0.183	22	3	189	67	0.232	[0.045, 0.260]	0.386	[0.163, 0.424]	0.534***	[0.468, 0.601]	0.266	[0.234, 0.293]
svm.1	0.061	29	38	154	60	0.128**	[-0.087, 0.234]	0.372***	[0.125, 0.437]	0.351	[0.277, 0.425]	**0.243**	[0.243, 0.372]
nen.1	0.057	49	37	155	40	0.358	[0.076, 0.455]	0.560	[0.297, 0.627]	**0.679**	[0.621, 0.737]	0.267	[0.246, 0.296]
logit.2	0.076	0	144	48	89	-0.750*	[-0.750, -0.221]	0.000	[0.000, 0.263]	0.086***	[0.056, 0.117]	0.321	[0.313, 0.385]
trees.2	0.056	77	190	2	12	-0.124	[-0.468, 0.184]	**0.433**	[0.000, 0.462]	0.073***	[0.047, 0.099]	0.305	[0.300, 0.372]
knn.2	0.099	3	93	99	86	-0.451	[-0.451, -0.190]	0.032	[0.000, 0.098]	0.161***	[0.125, 0.197]	0.340	[0.316, 0.362]
rf.2	0.176	0	62	130	89	-0.323	[-0.417, -0.176]	0.000	[0.000, 0.043]	0.075***	[0.050, 0.100]	0.348	[0.327, 0.385]
svm.2	0.058	41	107	85	48	**-0.097**	[-0.312, 0.323]	0.346	[0.221, 0.528]	0.087***	[0.059, 0.114]	**0.302**	[0.283, 0.371]
nen.2	0.055	18	115	77	71	-0.397	[-0.555, -0.013]	0.162	[0.000, 0.340]	**0.381**	[0.322, 0.439]	0.337	[0.300, 0.374]
logit.3	0.060	37	16	176	52	**0.332**	[0.105, 0.429]	**0.521**	[0.344, 0.607]	**0.401**	[0.337, 0.465]	0.295***	[0.281, 0.304]
trees.3	0.055	89	192	0	0	0.000	[-0.256, 0.300]	0.481**	[0.214, 0.705]	0.081***	[0.053, 0.109]	0.302***	[0.278, 0.333]
knn.3	0.060	28	91	101	61	-0.159***	[-0.205, 0.085]	0.269**	[0.163, 0.342]	0.336***	[0.282, 0.390]	0.305***	[0.289, 0.321]
rf.3	0.138	7	18	174	82	-0.015**	[-0.137, 0.060]	0.123*	[0.018, 0.211]	0.396	[0.338, 0.454]	0.305***	[0.285, 0.321]
svm.3	0.065	27	89	103	62	-0.160	[-0.374, 0.204]	0.263**	[0.086, 0.475]	0.088***	[0.057, 0.118]	0.305***	[0.284, 0.361]
nen.3	0.086	36	15	177	53	0.326	[0.137, 0.446]	0.514	[0.364, 0.619]	0.387	[0.332, 0.441]	**0.285**	[0.269, 0.292]
logit.4	0.076	49	32	160	40	**0.384**	[-0.143, 0.384]	**0.576**	[0.215, 0.576]	0.456***	[0.390, 0.522]	**0.275**	[0.239, 0.345]
trees.4	0.056	6	168	24	83	-0.808	[-0.808, 0.155]	0.046	[0.000, 0.400]	0.081	[0.048, 0.113]	0.302	[0.269, 0.339]
knn.4	0.228	6	42	150	83	-0.151	[-0.187, -0.023]	0.088*	[0.000, 0.154]	0.355***	[0.312, 0.399]	0.350*	[0.329, 0.381]
rf.4	0.173	12	17	175	77	0.046	[-0.085, 0.165]	0.203	[0.061, 0.292]	0.259	[0.198, 0.320]	0.299	[0.279, 0.311]
svm.4	0.056	29	83	109	60	-0.106*	[-0.270, 0.082]	0.289*	[0.136, 0.370]	0.348*	[0.284, 0.413]	0.295*	[0.289, 0.386]
nen.4	0.160	12	10	182	77	0.083	[-0.122, 0.199]	0.216*	[0.123, 0.400]	**0.736**	[0.687, 0.785]	0.304*	[0.283, 0.378]

Note: Best performance on given dataset in bold. Stars indicate if the respective method's performance is significantly worse than the best model's performance on the same dataset (***/**/* for the 1%/5%/10% significance level). For U_r, F_1 and AUC higher values indicate better prediction performance, while for BPS lower values are preferable. Numbers in brackets indicate 90% confidence bands. Model names reported in format "method.dataset", where datasets consist of the following sets of variables (1): credit and asset prices, (2) macro variables, (3) external imbalances, and (4) all variables (see also Table B.6).

Table C.7: Robustness: *Out-of-sample* forecast performance with a data sample split in 2005Q2

	Threshold	TP	FP	TN	FN	U_r		F_1		AUC		BPS	
logit.1	0.098	37	71	170	23	0.322	[-0.149, 0.410]	0.440	[0.291, 0.530]	0.768*	[0.717, 0.820]	0.139	[0.126, 0.187]
trees.1	0.067	21	17	224	39	0.279	[0.000, 0.521]	0.429	[0.097, 0.606]	0.697***	[0.636, 0.759]	0.149*	[0.113, 0.205]
knn.1	0.124	40	40	201	20	**0.501**	[0.001, 0.626]	**0.571**	[0.169, 0.695]	**0.815**	[0.766, 0.864]	0.144	[0.103, 0.193]
rf.1	0.179	23	19	222	37	0.304	[-0.021, 0.459]	0.451	[0.090, 0.581]	0.797	[0.740, 0.854]	**0.136**	[0.101, 0.181]
svm.1	0.071	18	19	222	42	0.221	[-0.053, 0.547]	0.371*	[0.131, 0.542]	0.636***	[0.565, 0.708]	0.178**	[0.137, 0.236]
nen.1	0.089	35	47	194	25	0.388	[0.000, 0.509]	0.493	[0.151, 0.580]	0.786	[0.735, 0.837]	0.142	[0.126, 0.184]
logit.2	0.103	4	3	238	56	**0.054**	[-0.147, 0.360]	0.119	[0.000, 0.334]	**0.640**	[0.573, 0.708]	0.178	[0.164, 0.193]
trees.2	0.075	60	241	0	0	0.000	[-0.179, 0.126]	**0.332**	[0.000, 0.332]	0.500***	[0.500, 0.500]	**0.175**	[0.157, 0.305]
knn.2	0.304	3	18	223	57	-0.025	[-0.144, 0.071]	0.074	[0.000, 0.200]	0.506***	[0.452, 0.560]	0.197	[0.172, 0.244]
rf.2	0.163	2	30	211	58	-0.091	[-0.211, 0.126]	0.043	[0.000, 0.159]	0.372	[0.306, 0.437]	0.194	[0.168, 0.289]
svm.2	0.074	8	48	193	52	-0.066	[-0.298, 0.218]	0.138	[0.000, 0.457]	0.460***	[0.391, 0.530]	0.179	[0.164, 0.244]
nen.2	0.105	0	0	241	60	0.000	[-0.394, 0.058]	0.000	[0.000, 0.316]	0.566***	[0.493, 0.640]	0.180	[0.159, 0.197]
logit.3	0.078	37	51	190	23	0.405	[-0.409, 0.467]	0.500	[0.086, 0.568]	0.803	[0.746, 0.860]	**0.154**	[0.109, 0.179]
trees.3	0.075	60	241	0	0	0.000*	[-0.083, 0.396]	0.332*	[0.113, 0.641]	0.500***	[0.500, 0.500]	0.175*	[0.135, 0.207]
knn.3	0.154	23	30	211	37	0.259*	[0.001, 0.389]	0.407	[0.140, 0.504]	0.730**	[0.667, 0.792]	0.154	[0.122, 0.184]
rf.3	0.433	4	4	237	56	0.050***	[-0.129, 0.234]	0.118	[0.000, 0.385]	0.659***	[0.590, 0.727]	0.159	[0.129, 0.190]
svm.3	0.075	34	109	132	26	0.114*	[-0.281, 0.438]	0.335	[0.000, 0.542]	0.572***	[0.510, 0.634]	0.175*	[0.131, 0.190]
nen.3	0.148	39	53	188	21	**0.430**	[0.241, 0.496]	**0.513**	[0.000, 0.583]	**0.804**	[0.748, 0.860]	0.160	[0.151, 0.165]
logit.4	0.081	47	22	219	13	**0.692**	[0.029, 0.763]	**0.729**	[0.242, 0.786]	**0.880**	[0.832, 0.928]	**0.132**	[0.087, 0.211]
trees.4	0.072	16	15	226	44	0.204	[-0.108, 0.700]	0.352*	[0.035, 0.786]	0.542***	[0.469, 0.615]	0.172*	[0.091, 0.289]
knn.4	0.266	22	19	222	38	0.288	[-0.116, 0.446]	0.436	[0.000, 0.580]	0.600***	[0.531, 0.668]	0.167	[0.114, 0.235]
rf.4	0.262	9	36	205	51	0.001	[-0.373, 0.363]	0.171	[0.000, 0.491]	0.677***	[0.619, 0.734]	0.177	[0.071, 0.239]
svm.4	0.100	22	75	166	38	0.055	[-0.386, 0.393]	0.280	[0.000, 0.511]	0.576***	[0.514, 0.638]	0.197	[0.145, 0.752]
nen.4	0.274	24	3	238	36	0.388	[-0.012, 0.726]	0.552	[0.105, 0.727]	0.850**	[0.794, 0.905]	0.132	[0.088, 0.205]

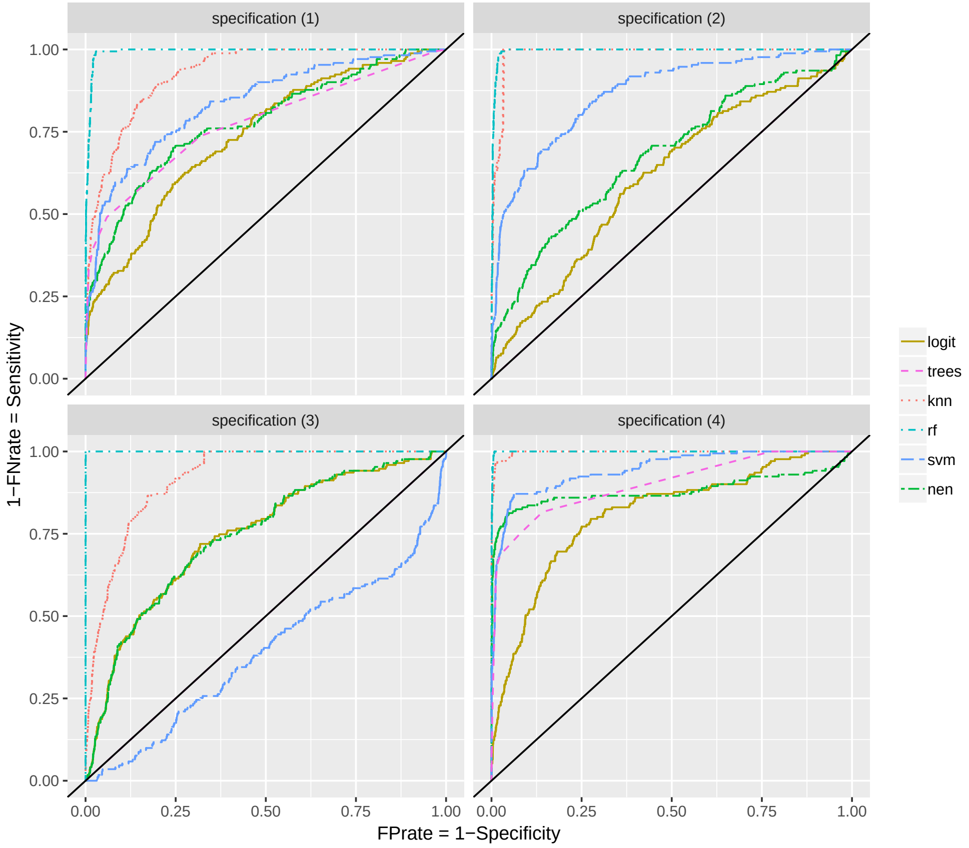
Note: Best performance on given dataset in bold. Stars indicate if the respective method's performance is significantly worse than the best model's performance on the same dataset (***/**/* for the 1%/5%/10% significance level). For U_r, F_1 and AUC higher values indicate better prediction performance, while for BPS lower values are preferable. Numbers in brackets indicate 90% confidence bands. Model names reported in format "method.dataset", where datasets consist of the following sets of variables (1): credit and asset prices, (2) macro variables, (3) external imbalances, and (4) all variables (see also Table B.6).

Table C.8: Robustness: *Out-of-sample* backcast performance with a data sample split in 2005Q2

	Threshold	TP	FP	TN	FN	U_r		F_1		AUC		BPS	
logit.1	0.336	33	97	1271	78	0.226	[-0.060, 0.344]	**0.274**	[0.055, 0.292]	**0.687**	[0.643, 0.732]	**0.075**	[0.066, 0.143]
trees.1	0.100	27	131	1237	84	0.147	[-0.020, 0.315]	0.201	[0.014, 0.288]	0.561***	[0.514, 0.608]	0.120	[0.070, 0.350]
knn.1	0.179	78	605	763	33	**0.260**	[-0.013, 0.373]	0.196	[0.014, 0.294]	0.667	[0.620, 0.715]	0.086	[0.070, 0.215]
rf.1	0.195	50	372	996	61	0.179	[-0.003, 0.334]	0.188	[0.026, 0.240]	0.659	[0.615, 0.702]	0.105	[0.069, 0.192]
svm.1	0.146	82	780	588	29	0.169	[-0.003, 0.298]	0.169	[0.010, 0.329]	0.663	[0.609, 0.716]	0.101	[0.064, 0.202]
nen.1	0.360	27	117	1251	84	0.158	[-0.033, 0.334]	0.212	[0.000, 0.281]	0.606***	[0.560, 0.653]	0.085	[0.066, 0.153]
logit.2	0.226	94	1165	203	17	-0.005	[-0.219, 0.081]	0.137	[0.066, 0.181]	0.572	[0.524, 0.620]	0.494	[0.075, 0.819]
trees.2	0.109	74	878	490	37	0.025	[-0.119, 0.236]	0.139	[0.064, 0.205]	0.473***	[0.436, 0.511]	0.140	[0.069, 0.513]
knn.2	0.320	29	429	939	82	-0.052	[-0.125, 0.182]	0.102	[0.000, 0.179]	0.443	[0.394, 0.491]	0.140	[0.075, 0.450]
rf.2	0.362	70	653	715	41	**0.153**	[-0.068, 0.180]	**0.168**	[0.061, 0.201]	**0.574**	[0.522, 0.626]	0.175	[0.072, 0.457]
svm.2	0.183	85	1071	297	26	-0.017	[-0.089, 0.032]	0.134	[0.000, 0.142]	0.490***	[0.445, 0.535]	**0.106**	[0.070, 0.350]
nen.2	0.219	32	681	687	79	-0.210	[-0.218, 0.031]	0.078	[0.000, 0.142]	0.404	[0.357, 0.451]	0.146	[0.072, 0.567]
logit.3	0.223	71	473	895	40	0.294	[0.107, 0.407]	0.217	[0.155, 0.267]	0.711**	[0.674, 0.747]	0.120	[0.069, 0.202]
trees.3	0.162	33	324	1044	78	0.060	[-0.134, 0.276]	0.141	[0.023, 0.233]	0.508***	[0.453, 0.563]	0.163*	[0.099, 0.550]
knn.3	0.225	59	395	973	52	0.243	[0.031, 0.361]	0.209	[0.130, 0.241]	0.665***	[0.620, 0.709]	**0.098**	[0.069, 0.237]
rf.3	0.367	30	176	1192	81	0.142	[-0.003, 0.239]	0.189	[0.131, 0.225]	0.646***	[0.601, 0.691]	0.109	[0.074, 0.334]
svm.3	0.237	79	475	893	32	**0.364**	[0.142, 0.409]	**0.238**	[0.175, 0.265]	**0.723**	[0.685, 0.760]	0.110	[0.067, 0.249]
nen.3	0.303	73	445	923	38	0.332	[0.090, 0.409]	0.232	[0.154, 0.273]	0.721	[0.684, 0.758]	0.121	[0.093, 0.221]
logit.4	0.392	72	514	854	39	0.273	[0.003, 0.451]	0.207	[0.134, 0.290]	0.681	[0.631, 0.731]	0.237*	[0.085, 0.640]
trees.4	0.040	60	340	1028	51	**0.292**	[0.068, 0.366]	**0.235**	[0.134, 0.286]	0.662	[0.620, 0.703]	0.171	[0.074, 0.237]
knn.4	0.318	56	483	885	55	0.151	[-0.044, 0.315]	0.172	[0.014, 0.256]	0.581***	[0.530, 0.632]	0.155	[0.080, 0.429]
rf.4	0.190	68	450	918	43	0.284	[0.059, 0.368]	0.216	[0.134, 0.280]	0.649**	[0.599, 0.699]	**0.141**	[0.074, 0.204]
svm.4	0.189	81	916	452	30	0.060*	[-0.035, 0.372]	0.146	[0.033, 0.313]	0.593***	[0.542, 0.644]	0.320	[0.081, 0.617]
nen.4	0.290	62	429	939	49	0.245	[0.015, 0.415]	0.206	[0.126, 0.290]	**0.688**	[0.643, 0.732]	0.148	[0.074, 0.385]

Note: Best performance on given dataset in bold. Stars indicate if the respective method's performance is significantly worse than the best model's performance on the same dataset (***/**/* for the 1%/5%/10% significance level). For U_r, F_1 and AUC higher values indicate better prediction performance, while for BPS lower values are preferable. Numbers in brackets indicate 90% confidence bands. Model names reported in format "method.dataset", where datasets consist of the following sets of variables (1): credit and asset prices, (2) macro variables, (3) external imbalances, and (4) all variables (see also Table B.6).

Figure C.1: Receiver-operator characteristics for baseline in-sample estimations, by model.



Note: Different subplots correspond to the four different datasets. The ROC displays the relation of false positive and false negative rates for different (constant) thresholds that can be applied to the probability predictions in the data. The added points correspond to false positive and false negative rates of the binary predictions derived from time-varying optimal thresholds, see Table C.1.

Figure C.2: Robustness (preference parameter): Relative usefulness of in- and out-of-sample estimation by model.

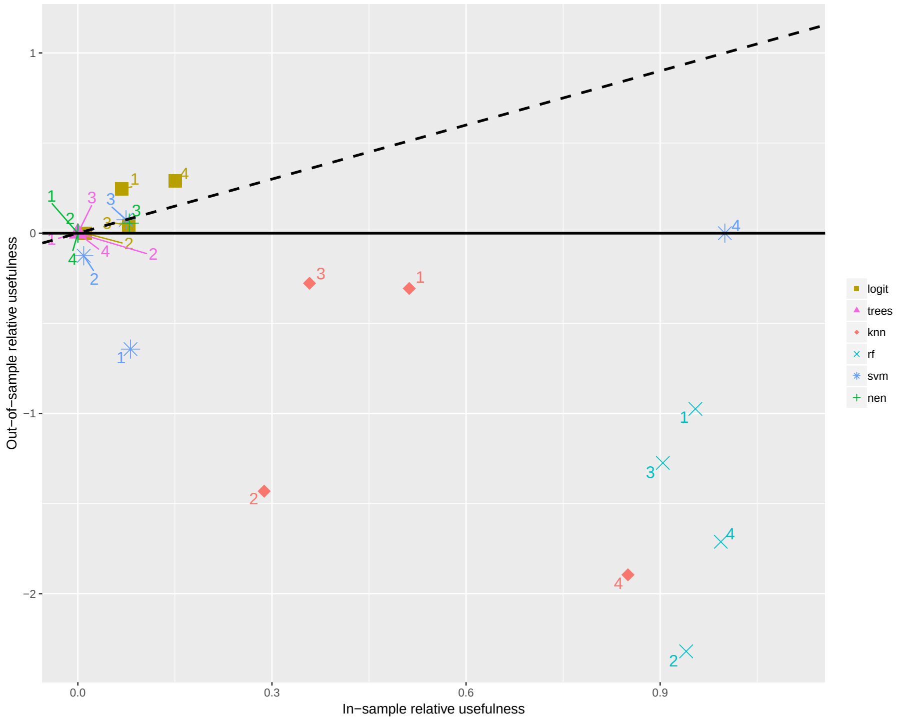


Figure C.3: Robustness (data transformation): Relative usefulness of in- and out-of-sample estimation by model.

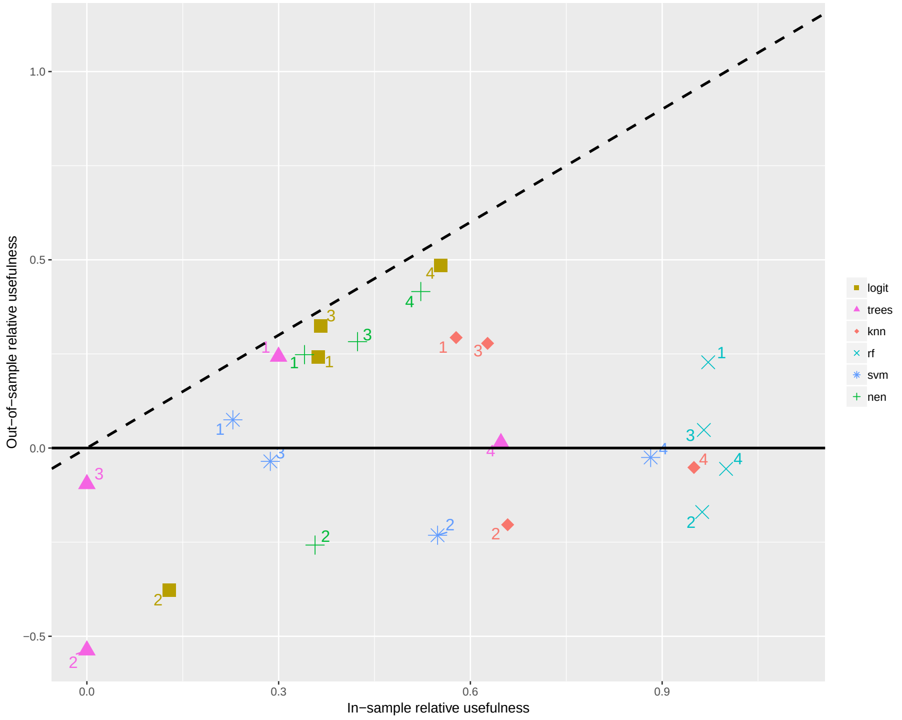


Figure C.4: Robustness (sample length): Relative usefulness of in- and out-of-sample estimation by model.

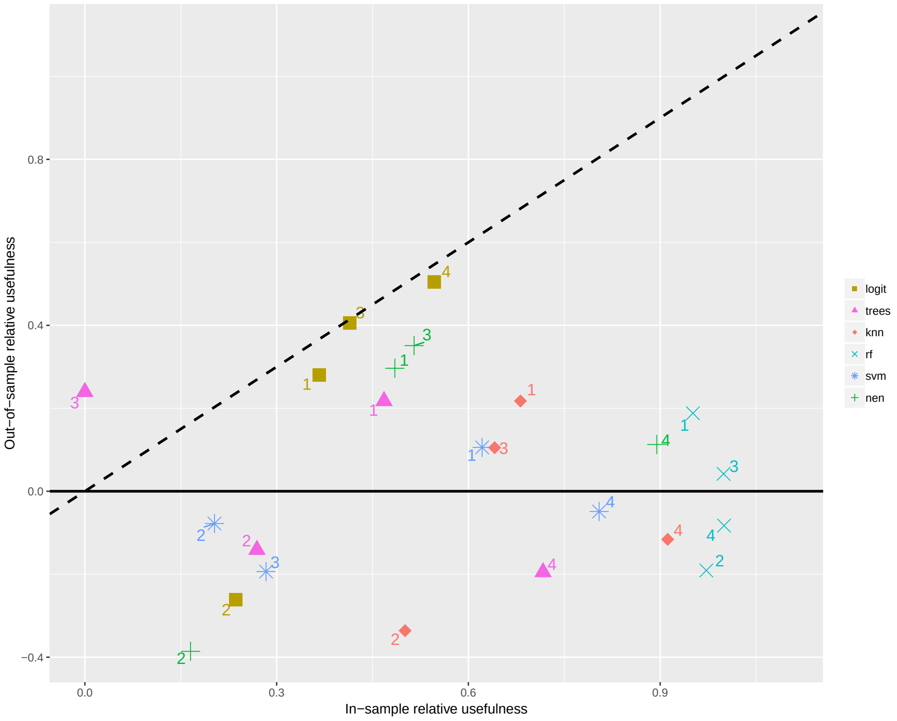
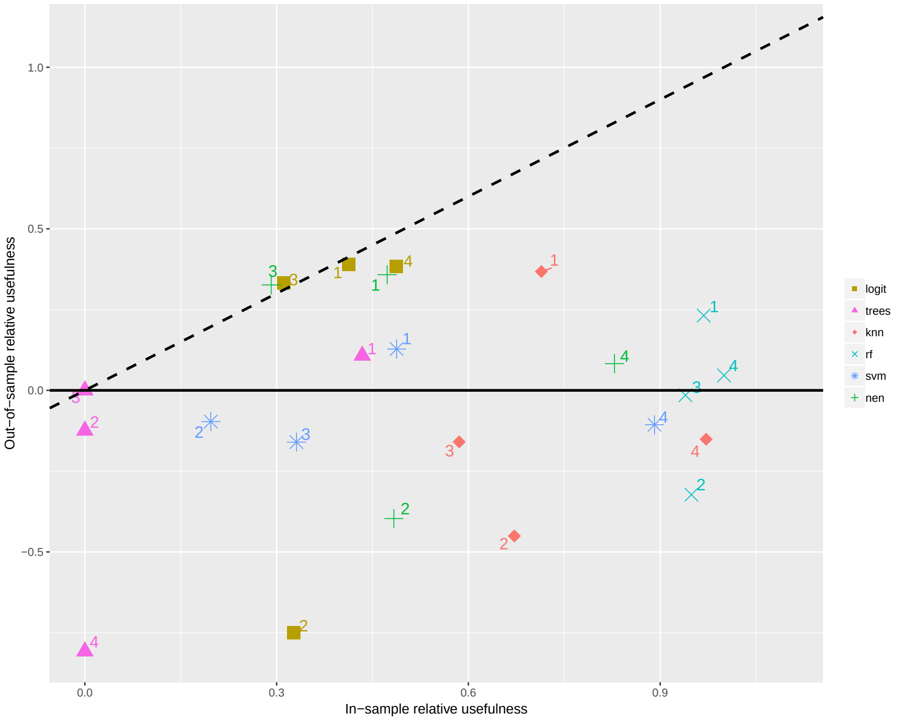


Figure C.5: Robustness (crisis database): Relative usefulness of in- and out-of-sample estimation by model.



9 Appendix D: Economic Interpretation

In this section, we would like to provide interested readers with the economic intuition behind the predictions for the example of the logit.4 model. This model had the best performance in our horse race, and encompasses all other models in terms of variables used. Table D.1 contains coefficients of the in-sample estimation. The most important variables in our model are the current account balance, credit-to-GDP, residential real-estate prices and gross fixed capital formation. This is consistent with existing theoretical and empirical evidence, which documents the vulnerability of the banking sector to many different channels, as described in Section 3. We find that the vast majority of coefficients have the expected sign. Credit growth above long-run trend increases the probability of being in an early warning window, as do high residential real estate prices and high equity prices. Thus, debt-financed asset price booms are found to be major drivers of crises (Kindleberger and Aliber, 2005; Jordà et al., 2015). The positive coefficient on the gap of gross fixed capital formation can be interpreted in a similar way: high levels of investment in fixed capital may be driven by overly optimistic expectations, leading to problems when future returns are lower than expected. Economic downturns, indicated by lower growth and lower inflation rates, also increase the crisis probability (albeit insignificantly). Last but not least, current account deficits and overvaluation of the real effective exchange rate are signs of an uncompetitive and (external) debt-financed economy.

Table D.1: Logit coefficients (full sample)

	Coefficient	**Std. Error**	**Marg. Effects**	**Stdev. recursive**
Constant	-2.771***	0.113	-0.199	0.151
Total credit-to-GDP gap	0.531***	0.102	0.038	0.033
Real residential real estate price gap	0.425***	0.096	0.030	0.073
Real equity price gap	0.200*	0.108	0.014	0.029
Real GDP gap	-0.141	0.123	-0.010	0.049
Inflation rate	-0.104	0.119	-0.007	0.102
Gross fixed capital formation-to-GDP gap	0.499***	0.106	0.036	0.018
Real 3-month money market rate	0.007	0.100	0.001	0.091
Current account-to-GDP ratio	-0.63***	0.093	-0.045	0.077
Real effective exchange rate gap	0.015	0.101	0.001	0.031
Real oil price gap	0.002	0.090	0.000	0.031

Note: This table reports coefficients, standard errors and marginal effects for the logit model using all variables, estimated on all available observations. The model is estimated with standardized data to make coefficients comparable. The last column shows the standard deviation of coefficient estimates across recursive estimations.

Table D.1 also allows us to look at the average marginal effects, which are of particular interest for the significant variables. Their average marginal effects, approximating the effect of a one standard deviation change in the respective indicator on the predicted crisis probability, are 1.4 percentage points for equity price gap, 3.0 percentage points for residential real estate price gap, 3.6 percentage points for gross-fixed capital formation-to-GDP gap, 3.8 percentage points for credit-to-GDP gap, and -4.5 percentage points for the current account balance. Compared to the unconditional probability of being in an early warning window, which is just above 9.5%, these effects are substantial. In comparison to that, the marginal effects of the insignificant variables are mostly negligible. Even though the effects of the significant variables are sizeable, it has to be noted that the model never implies a probability above 90% of being in an early warning window. For such a probability, all important variables need to be around two standard deviations away from their mean at the same time, which is an extremely rare event. This is in line with the view that, while our observables may signal the buildup of vulnerabilities in probability, there remains a substantial unpredictable component driving the ultimate realization or non-realization of crises.

In addition to their economic and statistical significance, coefficients are quite stable over time. In our recursive out-of-sample forecasting exercise, we can observe how (re-)estimated coefficients change across time, as more and more information becomes available. To summarize this, the last column of Table D.1 (Stdev. recursive) reports the standard deviation of coefficients across recursive estimations. As we can see, the magnitude of changes in coefficients during the out-of-sample window is relatively small for the significant coefficients, which is even more remarkable given the occurrence of the great financial crisis during this time period. This suggests a degree of robustness of the estimated model with respect to the addition of new information, which is promising regarding the potential use of such models for future (true) out-of-sample predictions.