WP/21/150

IMF Working Paper

Predicting Fiscal Crises: A Machine Learning Approach

by Klaus-Peter Hellwig

IMF Working Papers **describe research in progress by the author(s) and are published to elicit comments and to encourage debate.** The views expressed in IMF Working Papers are those of the author(s) and do not necessarily represent the views of the IMF, its Executive Board, or IMF management.

INTERNATIONAL MONETARY FUND

© 2021 International Monetary Fund WP/21/150

IMF Working Paper

Asia Pacific Department

Predicting Fiscal Crises: A Machine Learning Approach

Prepared by Klaus-Peter Hellwig

Authorized for distribution by Rahul Anand

May 2021

***IMF Working Papers* describe research in progress by the author(s) and are published to elicit comments and to encourage debate.** The views expressed in IMF Working Papers are those of the author(s) and do not necessarily represent the views of the IMF, its Executive Board, or IMF management.

Abstract

In this paper I assess the ability of econometric and machine learning techniques to predict fiscal crises out of sample. I show that the econometric approaches used in many policy applications cannot outperform a simple heuristic rule of thumb. Machine learning techniques (elastic net, random forest, gradient boosted trees) deliver significant improvements in accuracy. Performance of machine learning techniques improves further, particularly for developing countries, when I expand the set of potential predictors and make use of algorithmic selection techniques instead of relying on a small set of variables deemed important by the literature. There is considerable agreement across learning algorithms in the set of selected predictors: Results confirm the importance of external sector stock and flow variables found in the literature but also point to demographics and the quality of governance as important predictors of fiscal crises. Fiscal variables appear to have less predictive value, and public debt matters only to the extent that it is owed to external creditors.

JEL Classification Numbers: C53, F34, F37, H63

Keywords: Early warning systems, sovereign default, random forest

Author’s E-Mail Address: khellwig@imf.org

©International Monetary Fund. Not for Redistribution

I. Introduction[1]

Fiscal crises are extremely disruptive events that remain seared into the memories of entire generations. They are typically accompanied by a significant loss of annual output that is often permanent (Medas et al. 2018). Therefore, in the hope of preventing future crises, economists have long been working to develop early warning systems that help detecting crises before they occur, a quest that has been ongoing since at least the 1970s.[2] Such early warning systems have become tools that directly inform policy decisions, so that limitations in their accuracy can have far-reaching consequences. In this paper, I assess the out-of-sample predictive performance of econometric methods and ask whether our ability to predict fiscal crises can be enhanced through machine learning methods. I then explore whether the predictors identified by these relatively novel algorithms offer new insights into economic variables that can serve as robust early warning indicators.

The appeal of machine learning methods is that they allow us to tackle several well-known challenges in the literature on macroeconomic early warning systems: First, the dynamics preceding fiscal crises are most likely very complex, and simple linear or threshold models, while providing an intuitive narrative, may have a hard time capturing such complexities adequately. Second, given the small sample sizes in macroeconomic panels, with at best a few thousand (often correlated) observations, it is easy to identify patterns – narratives to explain past crises – that are simply spurious and will not be relevant for predicting future crises. This latter concern, the risk of overfitting (i.e., of explaining patterns that are specific to the estimation sample and don't generalize to other samples), is particularly pressing for the methods most popular among applied economists, maximum-likelihood and least-squares which, by definition, try to fit the data as much as possible as they are "tuned to generating

[1] I am grateful to Philip Barrett, Andy Berg, Fabrice Collard, Patrick Feve, Philippe Goulet Coulombe, Andreas Joseph, Christian Hellwig, Martin Hellwig, Jonathan Hersh, Andrew Hodge, Catherine Pattillo, Mehdi Raissi, Chris Redl, Kevin Wiseman, Yunhui Zhao, and seminar participants at the IMF, the 2020 World Congress of the Econometric Society, the 2020 conference on "Modelling with Big Data & Machine Learning: Measuring Economic Instability" at the Bank of England, and the 2020 European Winter Meeting of the Econometric Society for their comments and suggestions. The dataset used in this paper was developed jointly with Marialuz Moreno Badia and Paulo Medas. Juliana Gamboa Arbelaez, Chuqiao Bi, and Yuan Xiang provided excellent research assistance. All errors are my own.

[2] See Moreno Badia et al. (2020) for an extensive survey.

©International Monetary Fund. Not for Redistribution

unbiased estimates of coefficients rather than minimizing prediction error" (Kleinberg et al., 2015). The two concerns about underfitting and overfitting create a well-known trade-off for any predictive modeling exercise: predicting fiscal crises potentially requires complex models; but by adding complexity we may increase the risk of overfitting. Machine learning algorithms try to optimally solve the trade-off between underfitting and overfitting, which has made them the method of choice for a range of prediction problems. [3]

When predicting fiscal crises, the risk of picking up spurious patterns is particularly high, as there are reasons to think that such crises may not be predictable at all: For one, the possibility of multiple equilibria in the context of sovereign debt (Cole and Kehoe 1996, 2000; Detragiache 1996) means that, even though it may be possible to identify fundamentals that make a country vulnerable to self-fulfilling shifts in expectations, the timing of these shifts would be inherently unpredictable. Moreover, as highlighted by Berg and Pattillo (1999), crisis risk is endogenous to policy choices: if crises are predictable, then government and creditors can take action to prevent crises before they occur.[4] Finally, crises are rare events, which compounds the problem of small sample size. Berg and Pattillo (1999) and Christofides et al. (2016) show empirically that the early warning systems in use at the time were not able to predict the Asian currency crises of the 1990s or the Global Financial Crisis, respectively. In a similar vein, my first result in this paper is that the *out-of-sample* predictions for fiscal crises obtained from common econometric approaches, on average, cannot outperform an uninformed heuristic rule of thumb and are considerably less accurate than what their in-sample performance suggests.

By using some of the most popular machine learning approaches, I obtain improvements in predictive performance relative to established econometric methods. These methods, *elastic net* (Chou and Hastie, 2005), *random forest* (Breiman, 2001), and *gradient boosted trees* (Chen et al., 2015) build on models that are familiar to many economists (*logit* regressions and *classification trees*), and they can be implemented with standard statistical software. I first

[3] See Mullainathan and Spiess (2017) for an introduction to machine learning for applied economists. For a technical treatment of specific methods, see e.g., Hastie et al. (2012).

[4] Berg and Pattillo (1999) point out that the notion of predictable crises is difficult to reconcile with the Lucas critique. A similar argument can be found in Taleb (2007). Political scientists have also discussed the predictability of rare events using models (e.g., Ulfelder, 2012) or expert judgement (e.g., Tetlock, 2017).

©International Monetary Fund. Not for Redistribution

show that, even when using only a small pre-selected set of predictors, machine learning can yield sizable improvements in performance. The gains in performance become larger when I expand the set of candidate predictors and delegate variable selection to algorithms. I find that these gains are statistically significant and not driven by just a few individual countries or years in the evaluation sample.

How are these gains in accuracy achieved? First, unlike most econometric estimators, machine learning techniques impose limits on the degree to which models are allowed to fit the data in the estimation sample. Relative to an unconstrained maximum likelihood estimator (e.g., logit), this results in a *poorer in-sample accuracy* and in estimated marginal effects that are biased towards zero.[5] At the margin, however, reducing the in-sample fit also reduces the risk of overfitting and can therefore lead to better *out-of-sample* predictive performance. Second, these constraints on model fit make it safe for algorithms to explore a rich set of interactions between predictor variables and uncover non-linear relationships without the increased risk of overfitting that such explorations would carry in unconstrained settings.[6] Third, ensemble approaches (e.g. random forest) achieve additional gains in accuracy by averaging the predictions of several models instead of relying on a single model. Using the “crowd wisdom” of a large number of weak models can lead to strong average predictions if the prediction errors of individual models tend to cancel each other out.

In approaching the crisis prediction problem, I am equipped not just with an expanded statistical toolkit but also with richer data than previous studies. Overfitting, if left unaddressed, is first and foremost the consequence of small sample size, and I explore several ways to enhance or at least preserve the sample size. These methods are common in predictive modeling but less so in the economics literature: First, unlike previous authors, I explore the pooling of countries with very heterogeneous characteristics into a single large sample instead of estimating separate models for countries with different characteristics (e.g., by income level). This pooling may abstract from important structural differences between advanced,

[5] Machine learning methods share this feature with Bayesian techniques which also limit the degree to which a model is allowed to learn from a small sample.

[6] Goulet Coulombe et al. (2020a) show that this ability to address non-linearities is the main source of performance gains from machine learning methods in macroeconomic forecasting applications.

©International Monetary Fund. Not for Redistribution

emerging, and low-income countries. But I find that any reduction in in-sample fit from this abstraction is offset (and in some cases outweighed) by the reduced risk of overfitting achieved through a larger estimation sample.[7] Second, I investigate whether, by relying on imputation techniques to avoid the loss of observations due to missing values, I can obtain additional gains in accuracy. Of course, imputation adds noise to my predictors, which leads to biased coefficients and reduces the in-sample fit. But, again, the larger estimation sample reduces the risk of overfitting. Hence, imputation is another, complementary, way to navigate the trade-off between underfitting and overfitting. I show that pooling of observations and imputation can help improve model accuracy. Moreover, imputation allows me to make predictions for all observations, not just those with better data coverage. My imputation approach is extremely simple: missing values are replaced by their sample median. Future research could explore more elaborate imputation techniques to obtain additional gains in accuracy.

Machine learning algorithms and larger estimation samples reduce but don’t eliminate the risk of overfitting. Therefore, in addition to exploring novel algorithms and data imputation, the paper emphasizes the need for a machine learning approach to model *evaluation*: throughout the paper, the focus is on out-of-sample prediction; and great care is applied to avoid any spillover of information from the evaluation sample to the estimation sample.

The risk of overfitting is present not just in parameter estimation but also in variable selection. Variable selection based on economic theory or on in-sample fit is inevitably driven by hindsight. As already mentioned, machine learning algorithms allow me to consider a large number of predictor variables. The more candidate predictors we add, the lower the risk that variable selection is biased by our own judgement. I take this idea to the extreme by including a large number of series on economic, demographic, and political conditions, and taking into account sources of contagion as well as global variables and cross-sectional averages of country-specific variables. I am also agnostic about the exact form in which a series should enter the model and hence include current levels, lags, and changes over time at various frequencies. All told, I arrive at 748 individual series.

[7] Intuitively, estimating a single model on the full sample requires half as many parameters as estimating two separate models on two sub-samples. The smaller the number of parameters, the lower the risk of overfitting. Bolhuis and Rayner (2020) explore the question of optimal sample pooling more systematically.

©International Monetary Fund. Not for Redistribution

The set of variables selected from this large number of candidate predictors varies somewhat across modeling techniques. But when it comes to the most important predictors, there is considerable agreement across models. To be sure, predictor importance reflects *correlations*, not *causation*: my gains in predictive performance are achieved at the cost of biased model parameters, so that an economic interpretation of model parameters – establishing a narrative of fiscal crises – is difficult if not impossible. Machine learning algorithms in this paper identify *predictors*, not *causes* of fiscal crises.

I find that the algorithms confirm, as highlighted in the literature, the strong predictive power of the current account balance and of public external debt stocks. On the other hand, unlike many previous authors, I do not find a strong role for the real exchange rate, the GDP growth rate, or trade openness. Moreover, my results indicate that demographics and the quality of governance may be stronger indicators of crisis risk than recognized in previous work. Perhaps surprisingly, fiscal variables appear to have less predictive value, and public debt matters only to the extent that it is owed to external creditors.

The paper is organized as follows: the next section revisits the literature on fiscal crisis prediction. Section III specifies the definition of fiscal crises. Section IV describes the methodological approach and data. Section V assesses the predictive performance for a limited set of variables. Section VI discusses performance with algorithmic variable selection, and Section VII discusses the ranking of predictors by importance.

II. Past Attempts at Predicting Crises

Although crisis prediction has a long history (see, for example, Frank and Cline 1971), it was only in the wake of the Asian crisis of the late 1990s that the literature on EWS experienced a renaissance.[8] Many of the earlier studies only look at currency and/or financial crises. EWS for fiscal crises are far less common, and they tend to cover a relatively small sample of mostly

[8] See Moreno Badia et al (2020) for an in-depth review of the literature on fiscal crises, including of predictors commonly identified as informative. For a review of the literature on early warning systems in macroeconomics, see Kaminsky, Lizondo and Reinhart (1998), Hawkins and Klau (2000), Abiad (2003), and Frankel and Saravelos (2012)

©International Monetary Fund. Not for Redistribution

emerging market economies over different time periods.[9] Only a few studies include LIDCs, notably Cerovic et al. (2018). The literature has generally focused on external debt crises although a wave of recent studies encompasses a broader dimension of fiscal stress. While crisis definitions, sample coverage, forecast horizon, and evaluation methodology differ widely across studies, the empirical research can be classified into two broad categories, in terms of the methodology used:

(1) *Multivariate regressions.* This approach relies on the Generalized Linear Model (GLM – typically the probit or logit version). Several studies use GLM to identify the determinants of fiscal crises, including Marashaden (1997), Detragiache and Spilimbergo (2001), Peter (2002), Manasse et al. (2003), Ciarlone and Trebeschi (2005), Kraay and Nehru (2006), Gourinchas and Obstfeld (2012), Berg et al. (2014), Dawood, Horsewood, and Strobel (2017), and Pamies Sumner and Berti (2017).[10] Probit models also have an important role in policy practice, notably in the IMF/World Bank debt sustainability framework for low-income countries (see IMF, 2015). In most papers, predictor variables are manually pre-selected based on the authors' judgement. Only a handful of papers choose more agnostic approaches such as Extreme Bound Analysis (e.g. Chakrabarti and Zeaiter, 2014; and Bruns and Poghosyan, 2018) or selection algorithms based on bivariate correlations with the outcome (e.g., Cerovic et al., 2018).

(2) *Tree-based approaches*. An alternative strand of the literature has used models based on classification trees to predict crises. This approach was developed by Breiman et al. (1984) and is more flexible in capturing non-linear structures and complex variable interactions. The basic idea is to look for the characteristics that are most closely associated to a class membership (crisis versus non-crisis) and to iteratively sort the sample using binary splits according to those characteristics (see Section IV.C below for more details). Examples of this approach for sovereign debt crises can be found in

[9] Most of these early studies focused on debt crises of the 1980s and 1990s. In contrast, the empirical literature looking at fiscal crises in advanced economies has only taken off in the aftermath of the European sovereign debt crises of 2010 and thereafter.

[10] In some instances, the focus is on the duration of crises, using survival analysis (see, for example, Ghulam and Derber 2018).

©International Monetary Fund. Not for Redistribution

Manasse et al. (2003), Manasse and Roubini (2009), van Rijkjeghem and Weder (2009), Savona and Vezzoli (2015), and Savona et al. (2015).[11] In some cases, predictions from several such trees are aggregated. A special case of very short classification trees is the so-called "signaling" approach.[12] A few papers looking at sovereign defaults and/or fiscal stress have followed this methodology (see, Reinhart, 2002; Baldacci et al., 2011; Berti et al., 2012; De Cos et al. 2014; and Cerovic et al., 2018).

The use of machine learning techniques in the EWS literature is a more recent development in the EWS literature. Examples include Celiku and Kraay (2017) for predicting conflicts, Weisfeld et al. (2020) and Basu et al. (2019) for balance of payments crises, and Bluwstein et al. (2020) for financial crises. Similar to my present paper, Savona et al. (2015) and Jarmulska (2020) use random forest to predict fiscal crises, though with a narrower scope in terms of country and predictor coverage. Moreno Badia et al. (2020) explore a range of alternative variable selection techniques for the random forest model. A few papers have used artificial neural networks to predict sovereign debt crises (Rodriguez and Rodriguez, 2006; Fioramanti, 2008).

III. Defining Fiscal Crises

I use the term fiscal crises to describe a period of heightened budgetary distress. Although a sovereign default is the canonical example, not all fiscal crises are associated with debt defaults or pre-emptive restructuring. In some instances, they entail other forms of expropriation—such as domestic arrears or inflation—that erodes the value of debt (Reinhart and Rogoff 2011a), or a default is avoided altogether thanks to assistance from official creditors (Manasse et al.

[11] Outside economics, tree-based methods have become popular for many predictive tasks, including email spam filters, fraud detection, and image recognition. The use of trees in empirical macroeconomics reaches back at least to the 1990s (e.g. Durlauf and Johnson, 1995). Duttagupta and Cashin (2011) use classification trees to model the risk of financial crises.

[12] Pioneered by Kaminsky, Lizondo, and Reinhart (1998) for currency crises, this approach selects several leading indicators and derives threshold values beyond which the indicator signals a predicted crisis. Like in a very short tree, the thresholds are endogenously derived (within sample) to maximize the predictive power of the indicator. In an extension, Kaminsky (1998) aggregates the information from several signals models, effectively creating a tree ensemble model.

©International Monetary Fund. Not for Redistribution

2003). Thus, to capture a broader notion of fiscal stress, I rely on the IMF's fiscal crises database (Medas et al., 2018), which uses the following four criteria:[13]

1. *Credit events*. I include all sovereign defaults to private or official creditors as well as debt restructurings. To exclude small technical defaults and avoid the perpetuation of a crisis being classified as a string of new events, I impose some minimum requirements in terms of the size and accumulation of defaulted amounts (for more details on the definitions and data sources, see Annex Table 1).

2. *Exceptionally large official financing*. Under this criterion, any IMF financial arrangement with a fiscal adjustment objective and access above 100 percent of quota is counted as a crisis episode. I also consider financial assistance programs by the European Union.

3. *Implicit domestic public debt default*. Two types of events are included: (1) high inflation (thresholds vary by income group to reflect the different degree of monetary deficit financing); and (2) accumulation of domestic arrears proxied by other accounts payable.

4. *Loss of market confidence*. To account for both the volume and price dimensions, I consider two criteria: (1) loss of market access, capturing bond issuance coming to a halt; and (2) large spikes in sovereign yields.

A country is classified as being in a fiscal crisis in any given year if at least one of the four criteria is met. Consecutive crisis years count as a single crisis episode. To separate between crisis episode, I require at least two years of no crisis between two distinct events. Based on this definition, there are 418 crisis episodes for a sample of 188 countries over the period 1980–2016.[14] On average, countries have undergone two fiscal crises since 1980, but there is large heterogeneity. At one end, low-income developing countries (LIDCs) have experienced more than three crises on average while advanced economies (AEs) have less than one. The duration

[13] In addition to some changes in the definition of each criterion (see Annex Table 1), the set of crises may depart from Medas et al. (2018) due to data revisions.

[14] For a description of the data and country groupings, see Annex Table 2.

©International Monetary Fund. Not for Redistribution

of crises also varies significantly, with emerging market economies (EMEs) and LIDCs showing the longest episodes—on average five years. There are also a few cases of serial default, resulting in some countries being in crisis virtually throughout the entire sample.

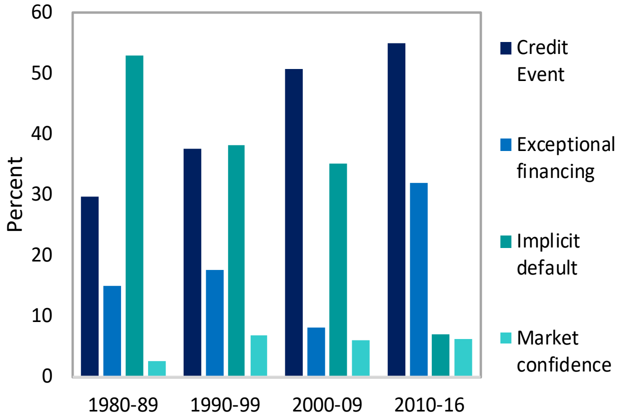
Historically, fiscal crises tend to come in waves—a pattern consistent with the sovereign debt cycles discussed in Reinhart and Rogoff (2011b). The largest concentration of episodes took place in the 1990s with 93 countries (more than 50 percent of which were emerging market economies) undergoing a crisis at its peak. But there has also been bunching in the early 1980s—reflecting the collapse of commodity prices and a surge in global interest rates—and in 2010, in the aftermath of the global financial crisis. LIDCs are the group with the largest incidence of events—about two-thirds of them are in a fiscal crisis at any point in time—but EMEs also have a relatively high frequency—on average, 40 percent. By contrast, less than 15 percent of AEs experience a fiscal crisis in any given year.

Credit events are the most frequent trigger of crises accounting for over 40 percent of episodes (Figure 3.1). This means that countries commonly resort to either default or restructuring as a way out of their fiscal troubles. An exception, however, was the 1980s when inflation became the most prominent form of implicit default on domestic currency debt, surpassing the incidence of external credit events. Only in the last decade has exceptionally large official financing become the second-most common criterion, accounting for almost a third of crisis episodes, underlying the importance of IMF programs during the global financial crisis.

©International Monetary Fund. Not for Redistribution

FIGURE 3.1. CRISIS CRITERIA TRIGGERED

(*PERCENT OF TOTAL CRISIS EPISODES*)



Sources: Bloomberg; Datastream; Eurostat; Gelos, Sahay, and Sandleris (2004); Guscina, Sheheryar, and Papaioannou (2017); IMF, International Financial Statistics; OECD; Reuters; and authors' calculations.

IV. Empirical Strategy and Data

A. The prediction problem and sample

The objective of crisis prediction models is to map a vector of k observables at time t into the probability of a crisis start occurring sometime between the end of year t and the end of year $t+h$.[15] In line with the literature, I choose a prediction window h =2, giving policy makers some time to take corrective policy actions when observing a crisis warning.[16]

The analysis covers 188 countries (see Annex Table 2), and the sample spans from 1979 to 2015. Since much of the literature has focused on advanced and emerging market economies – i.e., countries that regularly borrow from markets – I analyze model performance for these economies separately from performance for low-income developing countries.

It is important to note that, while I separate the sample by income group for the *evaluation* of the models, I don't do so for the *estimation* process. A priori, as pointed out in Weisfeld et al

[15] At any point in time a country is either in a state of crisis or in a non-crisis state. Since we are interested in transitions from non-crisis to crisis state, I only consider observations in which a country is *not* in a crisis in year t and drop from the analysis all observations in which a country is in a crisis in year t.

[16] See Berg and Pattillo (1999) and Bluwstein et al. (2020) for further discussions.

©International Monetary Fund. Not for Redistribution

(2020), there are many similarities between some of the countries classified as emerging market economies and some countries in the low-income group, so that much could potentially be gained by pooling countries from different income groups in the estimation sample, allowing us to learn about the former from the experience of the latter and vice versa.[17] By contrast, van den Berg et al. (2008) question whether there is sufficient homogeneity across countries and argue for a separation of countries into clusters in which crisis prediction models are then estimated separately.[18] I follow an approach of pooling all countries together and letting the algorithms decide how to split the data. For logit models, I check whether this approach leads to losses in performance.

B. Model evaluation

Sample splitting

As mentioned earlier, model evaluation is done out of sample. That is, I split the data into two disjoint sub-samples: The *training sample* is used to estimate the model parameters. The held-out *test sample* is then used to generate predicted crisis probabilities as fitted values from the model. These predicted probabilities are then assessed against the actual crisis outcomes.

Out-of-sample testing requires that no information from the test sample influences the process of variable selection and model estimation. Test sample information spilling into the training sample can occur…

- when variable selection is done using all observations (including on the test sample); the selected variables are then used to estimate the model on the training sample and make predictions for the test sample. Strictly speaking, variable selection based on judgement (or based on the recent literature) compromises the principle of out-of-sample prediction, since our economic judgement is influenced by recent events that could form part of the test sample.

[17] E.g., Angola and Gabon are classified as emerging market economies, whereas the neighboring Republic of Congo and Cameroon, both fellow oil producers, are counted as low-income countries.

[18] Their recommendation is based on in-sample results. Moreover, their finding is limited to linear models which assume homogeneity, whereas tree-based models, by design, can endogenously sort countries into the relevant clusters.

©International Monetary Fund. Not for Redistribution

- when the training and test samples are not separated by year – for example, Spain 2009 could be in the training sample, while Portugal 2009 is in the test sample. [19] Bluwstein et al. (2020) demonstrate that this can lead to biased performance assessments.

- when predictor variables are constructed so that they incorporate test sample information. Examples include output gaps or credit gaps that are based on two-sided (rather than purely backward looking) filtering approaches.

To avoid any use of test sample information in the model estimation process, I use a cutoff year to separate training and test sample. In choosing the cutoff year, we face a trade-off: a late cutoff year yields a larger training sample, so that the quality of the model is likely to be close to the quality we would obtain when estimating the model on the full sample to make predictions about the future – the ultimate objective of any early warning system. By contrast, an earlier cutoff year would yield a larger test sample, allowing for a more robust assessment of the model's accuracy.

To ease this trade-off, I use an iterative procedure with a rolling threshold: I estimate the model with data up to year $t - h$ to then make predictions using data observed at time t. With each new year $t+j$ in the test sample, the cutoff year for the end of the training sample is then moved by one year to $t+j-h$. I choose a test sample that spans 2001-2015 and that includes 1878 country-year observations, 301 of which include a crisis start in year t+1 or t+2. Hence, for a test sample spanning 15 years, I estimate 15 models, each one based on a larger training sample than the previous one.[20]

Performance measures

I assess performance according to several measures. It is important to note upfront that the choice of performance measure should depend on the application. For example, some

[19] This would occur if test and training sample were split by country or randomly. In the case of many prediction tasks, such as e-mail spam filters, observations could be assigned randomly the test and training samples. In the case of cross-country panel data, however, the sample splitting should take into account that, within any given year, global variables (e.g., oil prices) are common across countries, so that it can be unwise to assign observations from the same year to both the training and the test sample.

[20] Hyperparameters (see below) are retuned accordingly.

©International Monetary Fund. Not for Redistribution

applications require us to provide a ranking of crisis risk across countries, whereas others demand a crisis probability for just one specific country. In the first case, one would put more weight on a performance measure that reward an accurate ranking of probabilities. By contrast, in the second case, one would put more weight on whether the predicted probabilities accurately reflect the crisis risk in absolute terms. Since different performance measures capture different aspects of predictive performance, any measure of accuracy reflects the (somewhat subjective) weights one assigns to different types and magnitudes of prediction errors.

In terms of accuracy of predicted probabilities, I choose three measures, each of which penalizes prediction errors differently:

- log-likelihood, defined as

$$\frac{1}{N_{crisis}+N_{non-crisis}}\left[\sum_{i\in I_{crisis}}\log(p_i)+\sum_{i\in I_{non-crisis}}\log(1-p_i)\right]$$

- weighted log-likelihood, defined as

$$\frac{0.5}{N_{crisis}}\sum_{i\in I_{crisis}}\log(p_i)+\frac{0.5}{N_{non-crisis}}\sum_{i\in I_{non-crisis}}\log(1-p_i)$$

- mean squared error (MSE), also known as the Brier score,

$$\frac{1}{N_{crisis}+N_{non-crisis}}\left[\sum_{i\in I_{crisis}}(1-p_i)^2+\sum_{i\in I_{non-crisis}}{p_i}^2\right]$$

where p_i denotes the prediction for observation *i,* I_{crisis} and $I_{non-crisis}$ denote the sets of crisis start and non-crisis observations in the test sample, and N_{crisis} and $N_{non-crisis}$ are the respective numbers of crisis and non-crisis observations in the test sample.[21]

Predictions are not binary, so that individual prediction errors are somewhere between zero and one.[22] In practice, one of the main differences between the performance measures is how

[21] Note that the log-likelihood is the objective function of logit or probit regressions.

[22] Many authors transform the continuous predicted probabilities into binary predictions, by applying a probability threshold, and then count the number of missed crises and false alarms. While this approach is intuitive, it requires the estimation of an additional parameter – the threshold – making it impossible to separate the accuracy of the learning algorithm from that of the threshold.

©International Monetary Fund. Not for Redistribution

they penalize crises that are missed by a large margin. Whenever the predicted probability is close to zero for an observation that is within two years of a crisis start, the log-likelihood assigns a score near -∞, while the MSE assigns at most a loss of 1. The main difference between the unweighted and weighted log-likelihood is the relative importance of missing crises versus issuing false alarms. Since the majority of observations are non-crisis observations which are easier to predict than crisis observations, the unweighted log-likelihood is typically larger than the weighted likelihood.[23]

Accuracy can be visualized with the help of calibration curves (see Tetlock, 2017, for an extensive discussion). These graphical devices plot a model's predicted probabilities against observed crisis frequencies. The calibration curve thereby answers the question how many times a prediction of, say, 40 percent is actually followed by a crisis. If two out of five 40-percent predictions are followed by a crisis, then the model is very well calibrated. Calibration curves are also informative about a model's discrimination, the extent to which the model captures variation in crisis risk. A model that assigns the same probability to each observation has zero discrimination, whereas a model that is willing to assign probabilities close to 0 and 1 exhibits stronger discrimination. Accuracy is assessed as a combination of calibration and discrimination. Models with perfect calibration but little discrimination may not be very useful. Neither are models with strong discrimination but poor calibration.

I also display receiver-operator characteristic curves which, for each possible probability cutoff, plot true positives against false positives. And I report the area under the receiver-operator characteristics curve (AUROC), a popular measure of how well the *rankings* of probabilities correspond to observed outcomes (see, e.g., Jorda et al. (2011) or Berg et al. (2014)). Intuitively, the AUROC captures the probability that an algorithm selects the crisis out of a randomly selected crisis and non-crisis case. The ranking of countries by their risk level is important, for example, for creditors trying to identify a group of at-risk countries that require more regular monitoring. However, a model's AUROC does not always provide

[23] This issue becomes important when comparing predictive performance of a model in advanced vs developing countries where the frequency of crises is considerably higher.

©International Monetary Fund. Not for Redistribution

guidance on the absolute accuracy of individual probabilities: Models with poor calibration and poor discrimination can have an AUROC close to 1.

Significance

In principle, measures of model performance are specific to the test sample. Given the fact that outcomes are correlated within countries and within years across countries, it is important to know whether differences in performance are driven by just a few countries or years in the test sample. To quantify the degree to which results are sensitive to variations in the test sample, I report standard errors around all performance measures. I obtain these standard errors by bootstrapping on the test sample. Since crisis starts are correlated across countries and within countries across time, I adjust standard errors for two-way clustering, following the procedure described in Cameron et al. (2011).

Similarly, I compute standard errors for the differences in performance between methods (e.g., between the AUROC of logit and the AUROC of elastic net), which allows us to perform t-tests of differences in performance. For log-likelihoods and MSE, these tests correspond to Diebold-Mariano (2002) tests, common in the forecasting literature but less so in the literature on early warning systems.

Heuristic benchmarks

Papers on early warning systems typically rank models in terms of one or several of the performance measures presented above. However, while such rankings can give a sense of a model's usefulness relative to other models, interpreting measures of predictive accuracy in terms of their absolute usefulness is not straightforward. For example, a log-likelihood score of, say, -.2 is better than a log-likelihood score of -.5, but it is not clear a priori whether a score of -.2 is good enough for a model to be deemed informative for policy purposes. To put these measures into perspective, I let model predictions compete against the predictions obtained from simple rules of thumb. I consider three such rules:

1. Pooled global averages: this rule of thumb takes the empirical frequency at which all countries have entered a crisis between 1980 and year t-1 as the predicted probability of any

©International Monetary Fund. Not for Redistribution

country entering a crisis in year t+1.[24] According to this very naïve rule, all countries are assigned the same probability at any point in time, as in Hellwig (2018).[25]

2. Country specific averages: this rule of thumb takes the empirical frequency at which a specific country has entered a crisis between 1980 and year t-1 as the predicted probability of that country entering a crisis in year t+1.

3. A combination (simple average) of the first two rules.

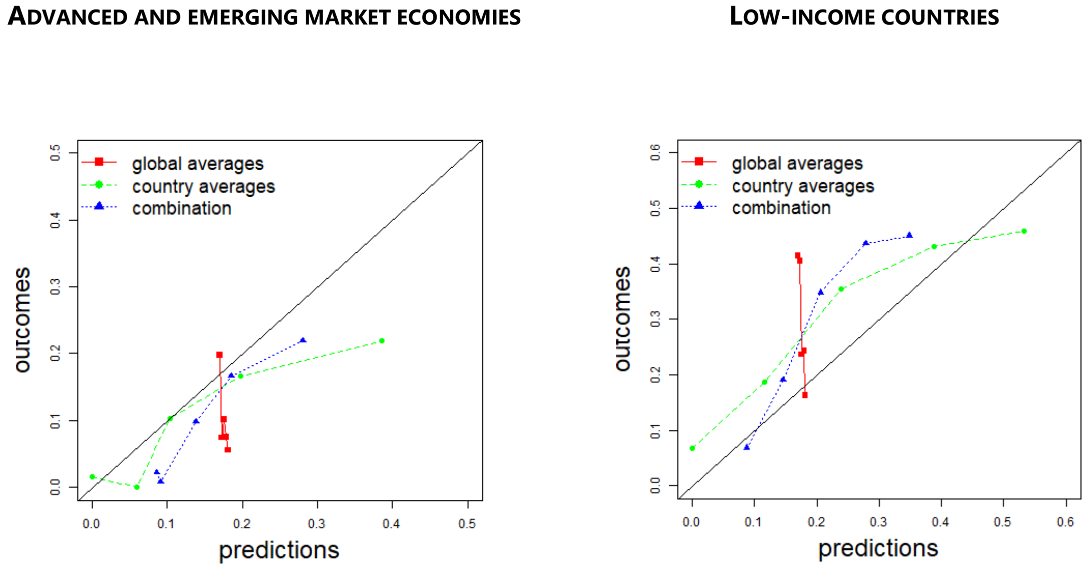
Figure 4.1 shows the calibration plots for the out-of-sample predictions generated using the three historical benchmarks, by income group. The global averages show, unsurprisingly, very little variation. And for advanced and emerging markets, global averages overpredict crises while they underpredict crises for low-income countries. By contrast, the country-specific averages show a relatively large dispersion and are relatively close to the 45-degree line. This suggests that, by simply looking at a country's history, a lot can already be learned about that country's crisis risk – a finding reminiscent of Reinhart et al. (2003) who find that countries have idiosyncratic levels of "debt tolerance" that move very little over time.

[24] For a two-year prediction window, the predicted probability is computed as $p_t + (1 - p_t)p_t$.

[25] Predictions vary from year to year as the training sample grows.

©International Monetary Fund. Not for Redistribution

FIGURE 4.1. CALIBRATION PLOTS FOR HEURISTIC BENCHMARKS USING HISTORICAL AVERAGES



ADVANCED AND EMERGING MARKET ECONOMIES

LOW-INCOME COUNTRIES

Note: Plot shows average predictions and outcomes (observed crisis frequencies) by quintile of predictions.

Table 4.1 illustrates how different aspects of predictions affect the performance assessment metrics. Since the pooled averages do not discriminate between high and low risk countries, their AUROC is very low. By contrast, using country-specific averages leads to better rankings into high risk and low risk countries. However, country-specific averages are prone to costly mistakes: whenever a country with no crisis history experiences a crisis, the likelihood score assigns a heavy penalty which, on average offsets the gains from better discrimination. The large mistakes incurred from country-specific averages motivate the choice of the third rule of thumb which makes only partial use of country-specific information.

It is noteworthy that predictive performance in terms of log-likelihood, MSE, and AUROC, based on historical experience is higher for market access countries than for developing countries. This is purely because there are a lot fewer crises in market access countries, so that the underlying uncertainty is significantly smaller.

It is also interesting to compare the AUROC values in Table 4.1 with some of the values reported in the literature and thereby put into perspective the informativeness of those models. The combination of historical averages delivers an AUROC of .744 for market access countries

©International Monetary Fund. Not for Redistribution

and .695 for low-income developing countries. By contrast, the largest in-sample AUROC for model-based predictions of financial crises in Jorda et al. (2011) and Jorda et al. (2016) is 0.719. And for fiscal crises, Cerovic et al. (2018) report an out-of-sample AUROC of 0.69 for market access countries and 0.68 for developing countries.[26]

C. Estimation methods

I now turn to the prediction algorithms used in this paper and their implementation. I limit my attention to methods that are fairly close to those commonly used in the applied economics literature (see Section II). The descriptions provided in this section are kept relatively non-technical. A more detailed introduction to machine learning techniques can be found in Hastie et al. (2012). The cross-validation procedure for selecting tuning parameters and the tuning grids are described in Annex B.

Penalized logit: elastic net

To address the potential overfitting problem within the logit framework, I us the *elastic net* (Zou and Hastie, 2005).[27] While maximum likelihood estimators like the logit attempt to fit the data in the estimation sample as tightly as possible, elastic net limits the model's ability to fit the data by adding a penalty term, punishing larger slope coefficients, to the likelihood maximization problem. I follow the conventional notation,

$$\max_{\beta} \left\{ L(X, y; \beta) - \lambda \left[\alpha \sum_{i=1}^{k} ||\beta_i|| + (1-\alpha) \sum_i \ {\beta_i}^2 \right] \right\},$$

where L is the likelihood function, X is the matrix of predictors, y is a vector of outcomes, β is a vector of linear model coefficients, the parameter λ is the marginal size of the penalty and the parameter α is the relative importance of the LASSO penalty relative to the Ridge penalty.[28] The values of hyperparameters λ and α are chosen by searching over a grid of values and evaluating model-performance through cross-validation. When α=1, the above maximization

[26] For the subsample of commodity exporting developing countries, they report an AUROC as high as 0.78.

[27] Elastic net generalizes the Ridge (Hoerl and Kennard, 1970) and LASSO techniques (Tibshirani, 1996).

[28] When λ is zero, then the estimated β corresponds to the logit estimator. When λ → ∞ , then the estimated values of β converge to zero. When $\alpha = 1$ (LASSO) and λ is sufficiently large, some elements of β are set to exactly zero. Whenever $\alpha = 0$ (Ridge), all elements of β are non-zero.

©International Monetary Fund. Not for Redistribution

problem can be interpreted as a constrained optimization problem where λ is the Lagrange multiplier on a fixed budget constraint: $\sum_{i=1}^{k} ||\beta_i|| < c$.

Classification trees

Classification trees, formally introduced by Breiman et al. (1984), are sequential decision rules by which a sample is recursively divided into bins with different levels of crisis risk. The sample splitting is done by applying thresholds to the variable that is (locally) the most informative (in terms of Gini index) about crisis risk within a given subsample.[29]

Trees also have a range of hyperparameters: To reduce the risk of overfitting, I impose a maximum tree depth of 4 levels. In addition, I impose that each node must have at least 7 observations, that any split must improve the Gini by at least 0.01, and that additional splits are attempted only if a node has at least 20 observations. While this (ad-hoc) parametrization is more restrictive than in Manasse et al. (2003), I show that it is already extremely prone to overfitting, despite allowing for only up to 16 leaves.

While classification trees have been used in various economic applications, they also have some well-known weaknesses: unlike for most linear estimators, estimating the globally optimal tree (i.e., the maximum likelihood fit for a given tree depth) would require evaluating an infinite number of combinations. Instead, trees are estimated to find a locally optimal tree, so that finding a good fit may require deep trees (i.e., a large number of parameters) which come with an increased risk of overfitting. Moreover, trees are inherently unstable: the tree structure is path dependent in that the first split (i.e., at the top node) determines the importance of variables in downstream nodes. Small variations in the underlying estimation sample can lead to changes in the first split which then propagate through the entire tree.

Stand-alone classification trees are found to be unreliable in many applications, but their performance is often greatly enhanced when predictions made by several such trees are combined to generate "crowd wisdom". Gains from aggregating predictions across tree models are achieved to the extent that these trees are sufficiently diverse so that they offset each other's prediction errors, thereby complementing each other. I use two of the most popular machine

[29] The Gini measures the class imbalance within a sample and is defined as $\sum_{i\,\in\{crisis,non-crisis\}}(1-p_i{}^2)$.

©International Monetary Fund. Not for Redistribution

learning methods to showcase the gains from aggregating the predictions made by weak individual models: *gradient boosted trees* and *random forest*.

Gradient boosting

Trees can be grown in a targeted way so that they complement each other. If a model consists of N trees, the additional N+1th tree can be grown to explain the unexplained residuals from the first N trees, thereby "boosting" the fit of the ensemble. I use the "XG boost" implementation by Chen et al (2015) which relies on gradient descent. I tune over (i) the number of trees; (ii) the maximum permissible depth of any single tree; (iii) the minimum gain required for an additional split; and (iv) the minimum size of any leaf.

Random Forest

In the *random forest* algorithm (Breiman, 2001), a large number of trees is aggregated, where diversity is created through double randomization: (i) each tree is estimated on a synthetic sample that is drawn at random from the original estimation sample; and (ii) at each node of each tree, the tree growing algorithm is limited in its choice of splitting variables. That is, while the standard classification trees can choose among all k predictors, the trees in a random forest can only choose among a subset of $m_{try} < k$ randomly selected predictors. The tuning parameter m_{try} is chosen from a grid of candidate values through cross-validation. As suggested by Breiman, I do not place any other restrictions on the tree growing process, so that each tree is grown exhaustively and can perfectly separate crisis starts from non-crisis observations in the random sample it is grown on.[30] The number of trees is set to 3000.

D. Data

The analysis relies on a large database of possible predictors of fiscal crises (see Annex Table 3 for a detailed variable list). The predictors include a large array of country-specific economic and institutional variables and a number of global variables, such as interest rates, commodity prices, and information on financial market conditions.

[30] Each tree is overfitted to perfectly explain some (randomly selected) observations. But different trees are overfitted to different samples, and their prediction errors due to overfitting tend to cancel each other out as the number of trees becomes large. See Goulet Coulombe (2020) for an instructive interpretation of Random Forest.

©International Monetary Fund. Not for Redistribution

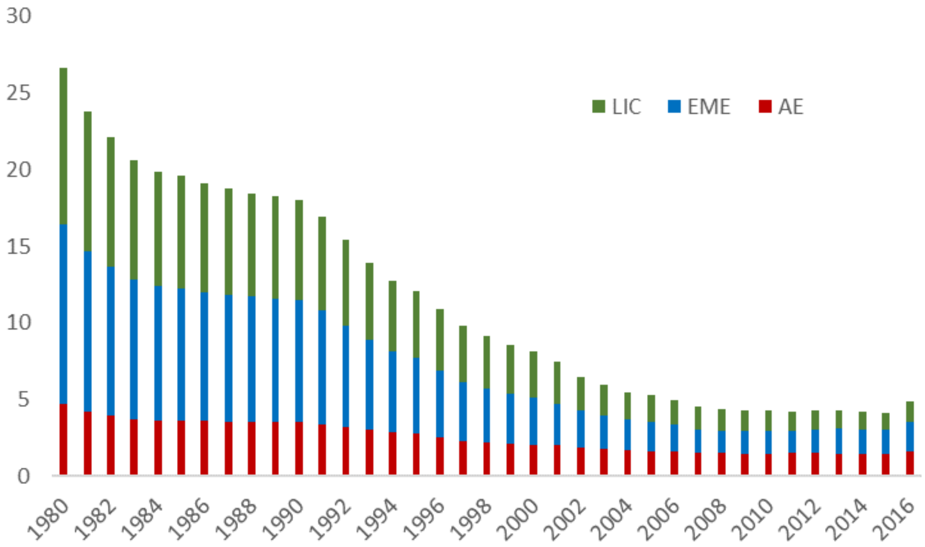
Since I am agnostic about the form in which a variable would have its strongest predictive power, I make heavy use of feature engineering by using various permutations (e.g., levels, first differences, 3-year differences) and lags. In addition, I include global averages for a considerable number of series, to reduce the risk of confounding country-specific with global developments. As a result of these operations, the number of individual series used in the analysis reaches 748.

While many of these predictors are likely to be collinear and thus potentially redundant, a notable feature of the machine learning methods used here is that (unlike OLS or logit models) they are able to use collinear variables. This will allow us to investigate questions such as whether the models pay more attention to revenues, expenditure, or the overall fiscal balance. Or whether they put more weight on public external debt, private external debt, or total external debt. Similarly, we can include the current level, the lag, and the first difference of a variable.

To avoid any undue influence of outliers in linear models, I winsorize each variable at the 1st and 99th percentile.

FIGURE 4.2. TOTAL IMPUTATION BY INCOME GROUP AND YEAR

(PERCENT)



©International Monetary Fund. Not for Redistribution

E. Missing values

A potentially problematic issue for any empirical work is the treatment of missing values. My aim is for the models to provide predictions for *all* countries, regardless of data availability. From a policy perspective, a model that refrains from making predictions for certain countries is of limited use, so that we should favor keeping all observations in the sample regardless of data coverage. From a methodological point of view, crisis risk is likely correlated with data availability and quality, so that any performance assessment would be affected by dropping observations. As will become clear below, losing observations from the training sample due to missing values can also be extremely costly, because smaller samples exacerbate the overfitting problem.

Hence, I favor keeping all observations in the sample, which requires some form of imputation. Whenever the value of a variable is missing, I replace it with the median over the entire training sample of the non-missing values for that variable. While this is arguably the most unsophisticated form of imputation (see Hastie et al. 2012), I choose it to demonstrate that even a very simple approach can lead to sizable gains in performance. I expect that future research into alternative and more involved imputation methods will yield additional gains. As Figure 4.2 shows, most of the missing data is in the 1980s, while since 2000 it is significantly less of a problem and similar across country groups.

Imputation is particularly useful when using a large number of variables since, without imputation, we would lose additional observations for any variable added to the model. By imputing variables, I avoid variable selection based on data availability. Variables with smaller coverage are more noisy, but I leave the decision whether the noise outweighs the information content to the algorithm.

V. Predictive Performance: Baseline Predictors

I now compare the out-of-sample performance of the various modeling approaches. In this section, I rely only on a small set of predictor variables that have been identified as relevant by the literature. Keeping the set of variables fixed for now allows me to analyze the choice of prediction models separately from the issue of variable selection. Hence, the question in this

©International Monetary Fund. Not for Redistribution

first part is whether some models are better at aggregating the information contained in a given set of variables.

I use ten variables that are closely related to those frequently found to be statistically significant in the literature. These baseline predictors are: log(GDP per capita), real GDP growth, foreign exchange reserves (in months of imports), the current account balance (in percent of GDP), trade openness (measured as the 10-year average of the sum of exports and imports in percent of GDP), the real exchange rate (measured using PPP price levels, as in Rodrik, 2008), total external debt (in percent of GDP), general government interest expense (in percent of GDP), Public debt (in percent of GDP), and the Polity IV score.[31] These predictors are already relevant in the literature preceding my test sample. Thus, in the spirit of Berg and Pattillo (1999), I ask how well one would have been served in the recent past by the predictions of models developed more than 15 years ago.

The section starts by demonstrating the overfitting problem of econometric approaches without pooling or imputation. It then compares the econometric approach with machine learning algorithms before discussing imputation and sample pooling.

A. Econometric approach

As a first step, I revisit the logit model and classification trees, the most commonly used approaches in the literature on fiscal crises. To be closer to the literature, I also refrain from imputing missing values for now, so that any observations with missing values are dropped from the sample. And, in line with previous authors, I split the training samples by income group.

The first three columns in panel (i) of Table 5.1. report the performance of the logit model for advanced and emerging market economies. The in-sample performance (column (i)) reflects the model estimated on market access country data from 2001-2015. By construction, the in-sample likelihood fit is the best a logit model can attain. When looking at out-of-sample

[31] The choice of variables is guided by the survey in Moreno Badia et al. (2020); I use interest expense instead of debt service, due to the wider data coverage.

©International Monetary Fund. Not for Redistribution

performance (column (ii)), the performance deteriorates dramatically across all measures. Note that the AUROC of 0.697 is nearly identical to that reported in Cerovic et al (2018).

In principle, the poor out-of-sample performance could be due to structural breaks: a model estimated on data from the 1979-2000 may be of limited use for the post-2000 era. However, performance deteriorates further when I drop the 1980s from the training data (column (iii)), which suggests that the smaller sample leads to even more overfitting.

Columns (iv)-(vi) of Table 5.1, panel (i), report the results from using a classification tree instead of the logit. In sample, the tree easily outperforms the logit model. This suggests that non-linearities could be a promising way to improve predictive performance. Out-of-sample, however, the classification tree model proves to be even less useful than the logit model. This comparison of the tree-based approach highlights the dilemma of predictive modeling: a model that allows for non-linearities appears to provide a better narrative for the historical experience. But the better in-sample fit comes at the cost of worse out-of-sample performance.

To put the performance measures further into perspective, the significance stars indicate the extent to which performance is statistically different from that of my third heuristic benchmark introduced in Section IV.B above (i.e., the combination of country-specific and pooled averages), reported in column (vii).[32] While in-sample fit of the logit model is better than the heuristic benchmark for all but one measure, the differences in AUROC and MSE are not statistically significant. By contrast, the in-sample fit of the tree model is significantly better than the historical benchmark along all dimensions. Out of sample, neither model can robustly outperform the historical benchmark. And the weighted log-likelihood is significantly worse. These findings should give the reader pause: modeling approaches that are widely used in policy applications are not any more useful than an uninformed rule of thumb.

When looking at low-income countries (panel (ii) of Table 5.1), a similar picture emerges. One key difference in performance relative to advanced and emerging markets is that LIDCs experience more crises, and crisis years are more difficult to predict. Hence, the unweighted

[32] Note that the numbers in column (vii) of Figure 5.1 are somewhat different from those in Table 4.1, as the test sample now covers only those observations with non-missing values for the ten baseline predictors.

©International Monetary Fund. Not for Redistribution

likelihood is worse for low-income countries than for more developed countries. The *weighted* likelihood, by contrast is higher for low-income countries, because models assign higher crisis probabilities to low-income country observations, so that crises are "missed" by a smaller margin. Overall, the mixed performance for LIDCs suggests that models with the most frequently used predictors are not sufficiently useful for predicting fiscal crises in developing countries. Of course, the variable selection is based on a literature that has mainly focused on countries with market access, so that the worse performance in a LIDC context is not entirely surprising.

B. Machine learning algorithms

Table 5.2.A compares the logit model (column (i)) with the three machine learning algorithms used in this paper. For advanced and economies, the logit model is outperformed by all three other methods along nearly every performance measure. And random forest dominates the other three methods, not just in terms of the level of performance measures, but also in terms of stability, as can be seen from the smaller standard deviations. Tree ensemble methods (i.e., random forest and XG boost) outperform linear methods (logit / elastic net), and – except for the MSE – the improvements in performance are statistically significant, as can be seen from Table 5.2.B. Random forest also outperforms all three heuristic benchmarks, and with high probability, though the probability is somewhat lower for the weighted log-likelihood.

For low-income developing countries, the picture is more mixed. Random forest still outperforms the logit model along all criteria, but otherwise the ranking is more ambiguous. And not all of the differences between random forest and logit are statistically significant (see Table 5.2.B). More concerning is that none of the four models can systematically outperform the heuristic benchmarks. Again, this suggests that the variable selection for this baseline specification is not well suited for low-income countries.

C. Imputation and pooling

As discussed above, one way to potentially improve predictive performance is by expanding the training sample, so that models become more robust to overfitting. I now explore two ways to expand the training sample: (i) pooling all countries into a single sample instead of splitting

©International Monetary Fund. Not for Redistribution

the estimation sample by income group; and (ii) replacing missing values with an imputed value (the sample median).

To what extent can imputation or pooling improve the performance of a logit model? Table 5.3.A shows the effect of adding variable imputation and sample pooling to the logit estimator. When looking at countries with market access (panel (i)), expanding the training sample to include LIDCs (column(ii)) leads to marginally better performance, except for the weighted log-likelihood. Imputation of missing values without sample pooling (column (iii)) leads to a doubling of the training sample size relative to column (i) and to pronounced gains in performance that are statistically significant (see Table 5.3.B). Once missing values are imputed, there appear to be no benefits from pooling, as can be seen by comparing columns (iii) and (iv). These results suggest that imputation of missing values reduces the overfitting problem to an extent that pooling the training sample to include LIDCs adds more noise than valuable information.

For LIDCs (panel (ii of Table 5.3.A), pooling and imputation combined (column (iv)) improves the performance relative to column (i) across all criteria. Table 5.3.B shows that, for the weighted log-likelihood and AUROC, we can say with relatively a high degree of confidence (78 percent and 84 percent, respectively) that these improvements are not just driven by a few countries or years in the test sample. When applied individually (columns (ii) and (iii)), the benefits from pooling or imputation are not as clear. The doubling of the sample through imputation (column (iii)) appears to add a large amount of noise, so that the log-likelihoods, both weighted and unweighted, deteriorate relative to column (i). By contrast, when using information from advanced and emerging market economies (column (ii)), the benefits from tripling the sample size marginally outweigh the costs of a more heterogeneous training sample.

Table 5.4. reports the performance measures when pooling and imputing missing values for all learning algorithms. Again, the *test* sample consists only of complete observations, so that it is identical to the test sample in Table 5.2.A and results are comparable. Most changes relative to Table 5.2.A are relatively small. For advanced and emerging market economies, the unweighted likelihood and AUROC systematically improves, and random forest does

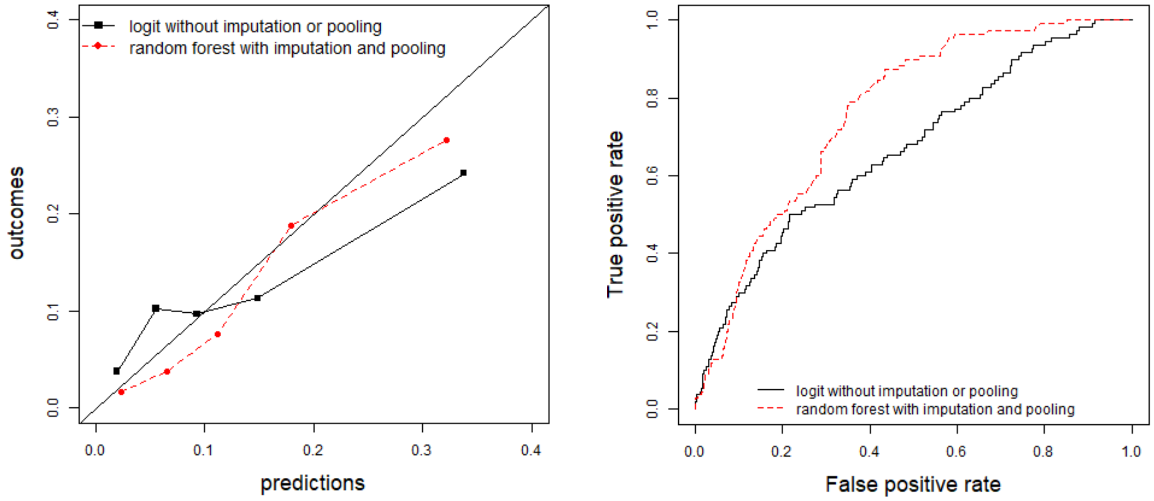
©International Monetary Fund. Not for Redistribution

marginally better across all performance measures – and standard errors are smaller than in Table 2.5.A. It is worth noting that logit exhibits the largest gains in performance relative to Table 5.2.A. Unlike the other methods, logit has no built-in way of addressing the overfitting problem, so that it benefits the most from the expanded training sample.

D. Graphical Comparison

Figures 5.1 and Figure 5.2 visualize the results so far. For advanced and emerging market economies, the logit models are too confident about the ability to assign high and low probabilities. And observations in the middle quintile of predicted probabilities have a lower actual crisis frequency than those in the 2nd lowest quintile – a poor ranking of crisis risk. By contrast, random forest predictions are relatively well calibrated, with a monotonically increasing curve: on average, higher predictions imply higher observed crisis frequencies, and the calibration curve is close to the 45-degree line. And random forest is particularly reliable in identifying observations with low crisis risk. The ability to better rank crisis risk is also expressed in the shape of the ROC curve (Figure 5.1.B).

FIGURE 5.1. OUT-OF-SAMPLE PERFORMANCE WITH 10 PRE-SELECTED VARIABLES, ADVANCED AND EMERGING MARKET ECONOMIES

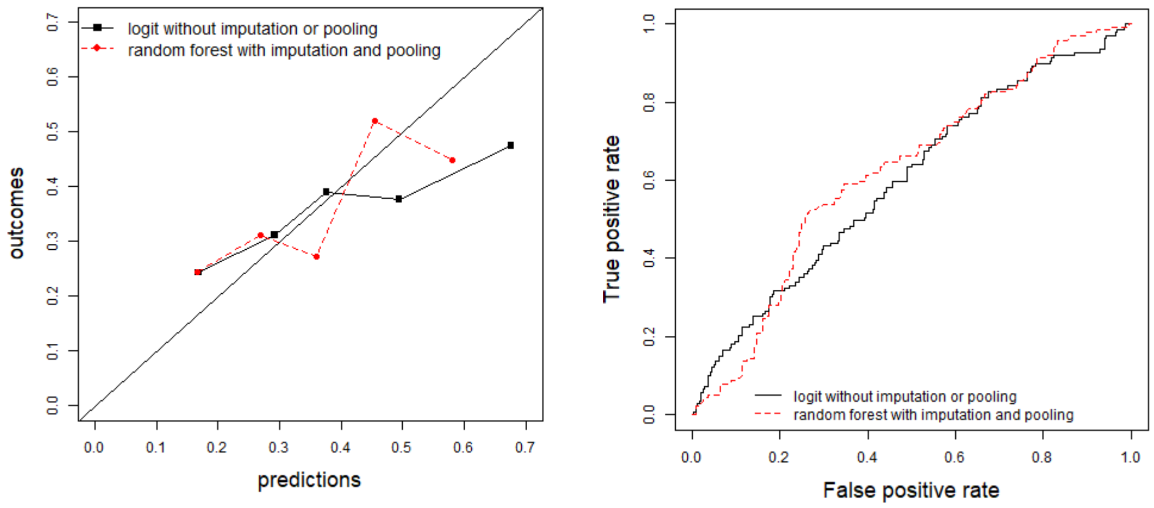


A. CALIBRATION CURVES

B. ROC CURVES

©International Monetary Fund. Not for Redistribution

FIGURE 5.2. OUT-OF-SAMPLE PERFORMANCE WITH 10 PRE-SELECTED VARIABLES, LOW-INCOME COUNTRIES



A. CALIBRATION CURVES

B. ROC CURVES

For low-income developing countries, the differences in performance in Figure 5.2 are at best marginal. The random forest model is less prone to assigning high probabilities and hence overpredicts less than the logit model. But both calibration curves are non-monotonic.

VI. Predictive Performance: Model-Based Variable Selection

How do the various statistical methods fare if we don't impose my literature-based priors on what the relevant predictors should be? And can we improve the predictive performance for low-income countries if we allow for additional predictors? I now expand the set of variables to the full set of all 748 variables, as described in Section IV.D above. And, to make use of the largest possible training samples, I impute missing values (see Section IV.E) and pool the sample across income groups (see Section IV.A). The results are reported in Tables 6.1.A and 6.1.B. Note that I now impute missing values not just in the training but also in the test sample, so that the performance measures reflect performance for all observations, not just for those with complete observations.

©International Monetary Fund. Not for Redistribution

Linear models

For the logit model, variable selection is done by applying a stepwise forward selection algorithm. That is, starting with the intercept, I expand the model by iteratively adding the variable that maximizes the model fit, evaluated using the Bayesian Information Criterion (BIC), until there is no more scope for improvement. The BIC offers a parametric approach to address the risk of overfitting from too many variables (see, e.g., Berg et al., 2014). It selects only 25 variables on average (over the 15 rolling regressions). Even so, Table 6.1.A shows that the resulting model (column (ii)) performs worse than the model with ten pre-selected variables (shown in column (i)) – only the AUROC shows marginal improvements.

The elastic net, by contrast, is able to consider a large number of candidate predictors without a deterioration in performance (columns (iii) and (iv)). And for low-income developing countries, the additional predictors lead to improved performance measures. Interestingly, the elastic net models are less sparse than the logit models with BIC selection: On average over the 15 rolling regressions, elastic net makes use of more than 116 variables out of a possible 748, suggesting that some redundancy inherent in a large set of predictors enhances the robustness. As a result, the performance improves relative to the logit/BIC approach along all dimensions. Table 6.2.B shows that the differences in performances between logit and elastic net are unlikely to be spurious. Still, when compared against the historical benchmarks, the performance of the elastic net algorithm remains mixed.

The finding that BIC leads to poor out-of-sample fit and a smaller set of predictors is in line with the findings from a more systematic comparison of methods conducted by Goulet Coulombe et al. (2020a). To understand the differences between logit and elastic net, it is important to keep in mind that the BIC approach limits the number of variables but maintains the maximum likelihood principle. That is, once the variables are selected, their coefficients are selected to maximize the in-sample fit which, in small samples, mechanically leads to overfitting (see, e.g., Copas, 1983). By contrast, the elastic net restricts the estimator's ability to fit the data and thereby reduces the risk of overfitting. A second difference is that the BIC determines the number of variables based on in-sample fit, whereas the penalty parameter for elastic net is cross-validated, based on out-of-sample accuracy.

©International Monetary Fund. Not for Redistribution

Tree ensembles

Gradient boosted trees (XG boost) deal with the large number of variables by using only those variables that are deemed informative and ignoring the rest. For market access countries, performance improves marginally relative to the elastic net, though not always in a statistically significant way. For LIDCs, performance improves relative to the logit approaches, but not relative to elastic net. And performance is not always superior to that of the heuristic benchmarks.

In the case of random forest, the large number of predictors is potentially problematic. It is likely that many of the predictors are noise and – unlike logit with BIC, elastic net, or gradient boosted trees – a stand-alone random forest has no way of entirely ignoring those noisy variables. Indeed, the variable importance chart (Figure 7.1) below shows that more than half of all variables provide no valuable information. Hence, there are potential benefits from pre-selecting variables. I do so by using *recursive feature elimination* (RFE).[33] The algorithm iteratively estimates a random forest, drops the least important variables, and re-estimates the model until performance (cross-validated, within the training sample) stops improving.

In columns (viii) and (ix) of Table 6.1.A, I report both the stand-alone random forest and the random forest with RFE. Despite having to use many noisy predictors, the stand-alone random forest outperforms the logit with BIC along every dimension. Note also that performance is more robust than for logit across the test sample, as can be seen from the smaller standard deviations. And, finally, random forest outperforms the heuristic benchmarks across all measures, and with relatively high degree of confidence (see Table 6.1.B).

For random forest, the AUROCs and weighted log-likelihoods are particularly large. The fact that these measures put more weight on crisis than non-crisis observations suggests that random forest with its many built-in redundancies is particularly well-suited to avoiding "missed" crises. And for low-income countries, the performance of the random forest with 748

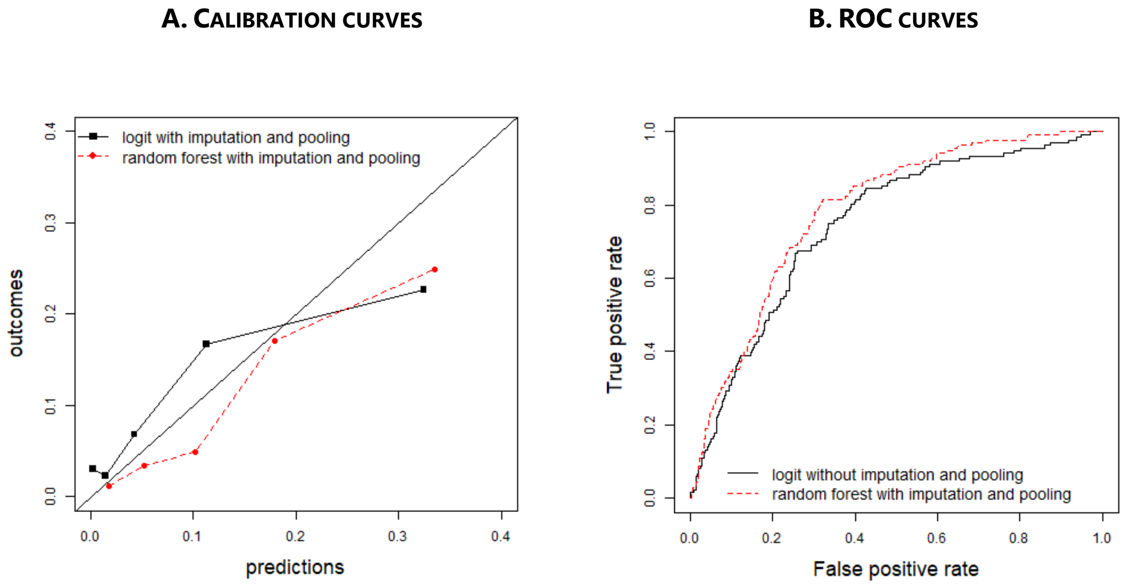
[33] Degenhardt et al. (2011) identify RFE as the most popular variables selection method for random forest. For a comparison of alternative selection algorithms in the context of fiscal crises, see Moreno Badia et al. (2020).

©International Monetary Fund. Not for Redistribution

variables is substantially better than the performance with pre-selected variables, suggesting that the set of predictors used in the baseline specification above was too narrow.

When I combine the random forest with RFE, the results are qualitatively similar but more significant, statistically speaking.[34] While the ranking between random forest with and without RFE is ambiguous, performance with RFE improves relative to all other methods. That said, for advanced and emerging markets, the AUROC and the weighted log-likelihood decline when using RFE.[35]

Figure 6.1. Out-of-sample performance with Variable Selection, Advanced and Emerging Market Economies


A. Calibration curves

B. ROC curves

Figure 6.1.A illustrates the calibration of logit and random forest for advanced and emerging economies. As in the specification with preselected variables, the logit model is too confident when assigning low crisis probabilities. The bottom quintile of predictions is very close to zero,

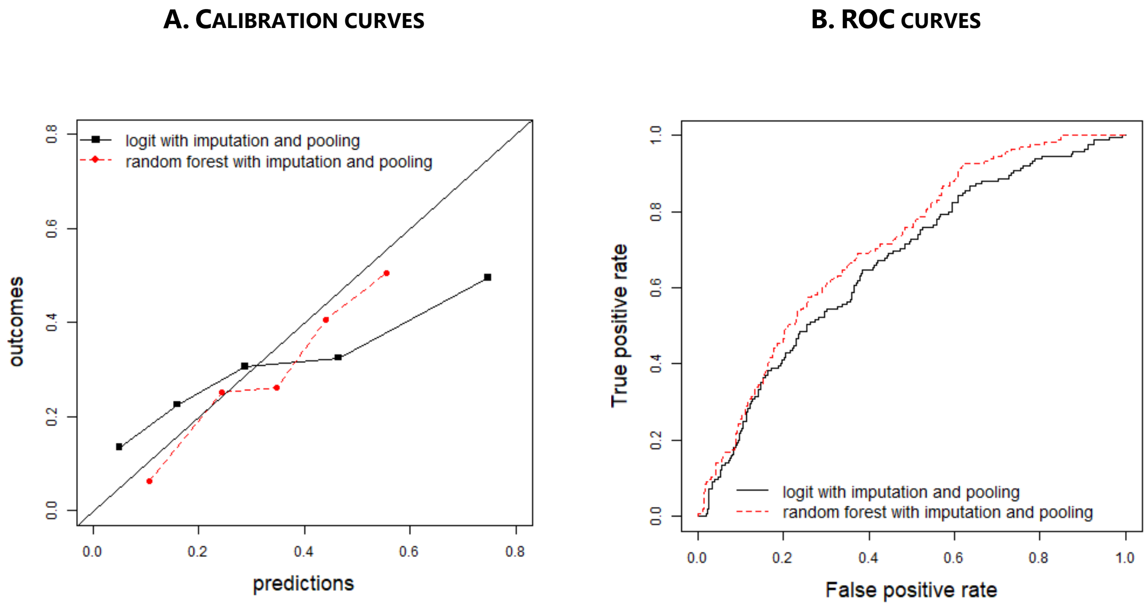
[34] RFE selects 93 out of the 748 variables.

[35] These declines could be, to some extent, mechanical and owed to the variable selection procedure: selection is based on a variable importance measure that is calculated by placing equal weight on each observation. For policy applications where the weighted likelihood or AUROC is the relevant performance measure, a RFE with variable selection based on a weighted variable importance measure would be more suitable.

©International Monetary Fund. Not for Redistribution

but this group includes more crisis observations than the second quintile. By contrast, the calibration curve of the random forest model is increasing and remains fairly close to the 45-degree line. For the top quintile of predictions, the random forest model is exaggerating crisis risk, but less than the logit model.

FIGURE 6.2. OUT-OF-SAMPLE PERFORMANCE WITH VARIABLE SELECTION, LOW-INCOME DEVELOPING COUNTRIES


A. CALIBRATION CURVES

B. ROC CURVES

For low-income countries (Figure 6.2), the differences in calibration are more pronounced. While the random forest's calibration curve follows the 45-degree line closely, the logit model with BIC obtains a flatter calibration curve, overpredicting crises in the upper quintiles and underpredicting in the lower quintiles.

VII. Predictors of Fiscal Crises

I now turn to the issue of predictor importance. That is, I ask which variables are identified as important predictors by the various statistical algorithms. It should once more be emphasized that the purpose of this exercise is to identify early warning indicators, not causes of fiscal crises.[36]

[36] A canary in the coalmine is a useful early warning indicator if it prompts miners to evacuate their shaft. It is less useful if miners confound correlation and causation and decide to resuscitate the canary.

©International Monetary Fund. Not for Redistribution

To assess predictor importance, I re-run the algorithms used above, this time on the full 1979-2015 sample (i.e., without retaining a test sample). Annex Table 4 lists the tuning parameter values for each model.

The variable importance measures are as follows:

- For linear models (logit and elastic net), I use each predictor variable's slope coefficient and scale it by multiplying with the variable's standard deviation.
- For gradient boosted trees, I compute the improvement in fit (measured by the reduction in Gini) attributed to each predictor.
- For random forest, I compute the out-of-bag permuted predictor importance (see e.g., Hastie et al., 2012).

Dense and sparse models

When interpreting these importance measures, it is useful to keep in mind that there are substantial differences in *shrinkage* – i.e., in how the individual methods treat predictors that are collinear and hence potentially redundant:[37] At the one extreme, *sparse* estimators try as much as possible to reduce the number of predictors used, so that the weight of two collinear variables is allocated entirely to one variable while the other variable is dropped. This is exactly the case for elastic net which in our case corresponds to a LASSO regression, since the optimal value for the tuning parameter α is found at 1. At the other end of the spectrum, *dense* estimators keep all variables, distributing the weights evenly over collinear predictors. This is the case for random forest, where it is difficult for any single variable to get an outsize share of importance. Giannone et al. (2018) and Goulet Coulombe (2020b) find that, in macroeconomic applications, dense models are typically superior in performance. Gradient boosted trees inhabit a space somewhere in the middle.

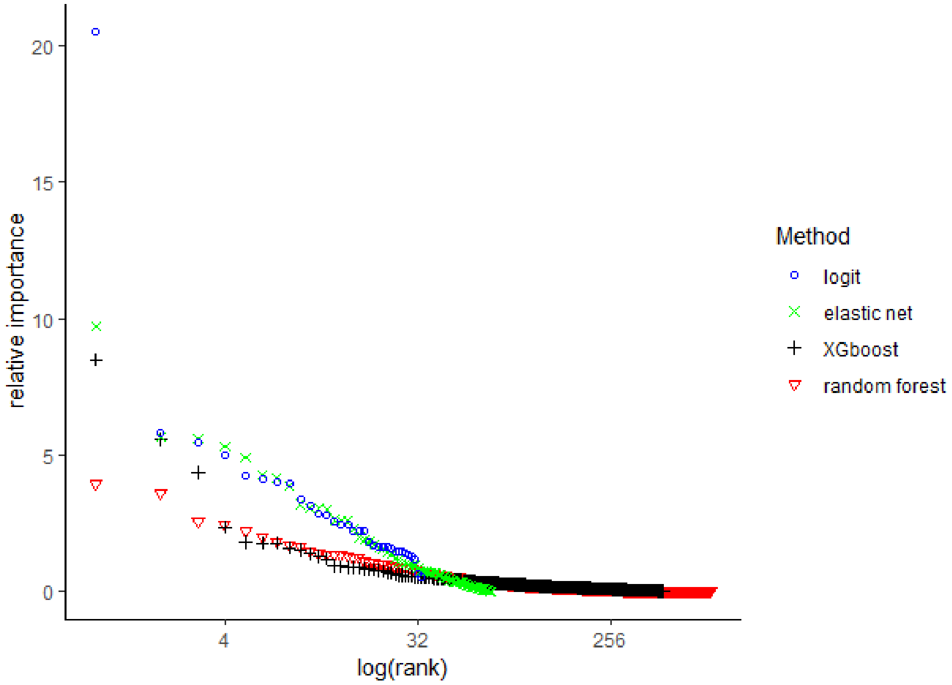
The differences in variable selection are illustrated in Figure 7.1. which ranks the variables according to their importance. While random forest uses all variables, elastic net reduces the set to 120 variables, and gradient boosted trees are somewhere in the middle. I plot the variable

[37] See Belloni and Chernozhukov (2011) and Ng (2013) for a more formal discussion of sparse and dense estimators.

©International Monetary Fund. Not for Redistribution

importance measures relative to the importance of the 10th most important variable. The elastic net model attributes greater importance to the top 10 variables, whereas the weights are distributed more evenly in the random forest model where important variables have to share their role with other variables that are closely correlated.

FIGURE 7.1. VARIABLE IMPORTANCE MEASURES



Notes: importance measures refer to (i) slope coefficient (scaled by predictor's standard deviation) for logit and elastic net; (ii) reduction in Gini for XG boost; and (iii) out-of-bag permuted predictor importance for random forest. Importance measures are expressed relative to the 10th most important variable.

Due to the differences in shrinkage, a predictor could be identified as highly informative by one algorithm and as merely redundant by a different algorithm. Moreover, the informativeness of each variable naturally depends on whether other, correlated, variables are available. Despite these differences and despite using model-specific measures of variable importance, I find a remarkable stability of predictor importance rankings across the various approaches.

The plot for random forest in Figure 7.1. also illustrates the source of the gains obtained from recursive feature elimination in the previous section: most of the variables' importance is close zero or even negative. In other words, some predictors are not just redundant but irrelevant for

©International Monetary Fund. Not for Redistribution

predicting crises. Eliminating such noise factors allows the model to focus on the more important predictors.

Top 30 predictors

In Table 7.1 I document the rankings of variables within the top 30 for each prediction method. A few results stand out: first, there is considerable overlap across machine learning methods in the variables selected. The fact that crisis history and GDP per capita are within the top 3 for elastic net, boosted trees, and random forest is perhaps not surprising. But the agreement across methods goes further than that: boosted trees and random forest agree on 7 variables that are in the top 10.

Results from machine learning algorithms confirm the importance of several predictors used in my literature-based model in Section V: GDP per capita, debt service, external debt, and the current account balance are found among the more important predictors. For random forest, the current account's importance is shared with external gross financing needs, due to the correlation between the two series. Foreign exchange reserves are found to be important, but in random forest this importance again is divided between several indicators, notably ratios of debt amortization to reserves. By contrast, none of the algorithms identify trade openness, the real exchange rate, or the real GDP growth rate as important predictors.

The rankings moreover suggest that both demographics (captured by the size of the working-age population and the age dependency ratio) and governance (measured by the quality of bureaucracy) have a strong role in predicting crises. Other important variables include gross domestic savings, inflation, and volatility measures. The role of volatility is intuitive and consistent with economic theory. Merton (1974) suggests that debt default can be seen as an option. The higher the volatility of the underlying asset (in this case: the economy), the more likely it is that an option is exercised. It is not the materialization of shocks but their volatility and implied likelihood of materializing that helps us predict crises ex ante.

It is also worth taking note of some of the series that are *not* found in the top 30 ranking for random forest and boosted trees: government revenues (apart from foreign development aid),

©International Monetary Fund. Not for Redistribution

expenditures, or deficits are absent, as are interest-growth differentials.[38] And public debt only matters to the extent that it is owed to external creditors. The same is the case for global variables such as oil prices or interest rates. These variables are found to be either relatively redundant or irrelevant for predicting crises.

Overall, the predictors suggest that external sector variables, both stocks and flows, have key role as early warning indicators. This finding is consistent with narratives in which foreign currency borrowing by both the private and public sector creates a vulnerability and exposes countries to external shocks (see, e.g., Eichengreen and Hausman, 1999).

VIII. Conclusion

This paper has explored the benefits of using machine learning tools to assess the risk of fiscal crises. In several ways, the approach taken in this paper is a significant departure from statistical analysis in macroeconomics. The focus is not on parameter identification and statistical significance (the internal validity of narratives) but on external validity – the extent to which a narrative generalizes over time and are helpful for making accurate predictions.

My analysis shows that crisis predictions based on established econometric approaches are of limited use, given that they cannot outperform a relatively naïve prediction rule. It also shows that the performance ranking of algorithms can depend the performance measure used and that differences are not always statistically significant. Even so, random forest approaches systematically outperform the heuristic benchmarks and other statistical models. Relying on a single model for all countries and imputation of missing values does not lead to lower predictive performance. Neither does taking an agnostic approach by allowing for a large set of predictor variables. For low-income developing countries, expanding the set of variables and leaving variable selection to an algorithm leads to considerable gains in accuracy.

[38] The low importance of fiscal balances, while somewhat surprising, is in line with the earlier literature, as pointed out by Van Rijkjeghem and Weder (2009). Van Rijkjeghem and Weder (2009) find that fiscal balances have some predictive value only when reserves are low or international liquidity conditions are tight.

©International Monetary Fund. Not for Redistribution

References

Abbas, S. A., N. Belhocine, A. ElGanainy, and M. Horton. 2011. “A Historical Public Debt Database,” *IMF Economic Review*, vol. 59, issue 4, pp. 717–42.

Abiad, A., 2003. “Early Warning Systems: A Survey and a Regime-Switching Approach,” IMF Working Paper No. 03/32 (International Monetary Fund).

Arslanalp, S. and T. Tsuda, 2012. “Tracking Global Demand for Advanced Economy Sovereign Debt,” IMF Working Papers 12/284.

Baldacci, E., I. Petrova, N. Belhocine, G. Dobrescu, and S. Mazraani. 2011. “Assessing Fiscal Stress,” IMF Working Paper No. 11/100 (International Monetary Fund).

Basu, S.S., R. Perrelli, and W. Xin, 2019. „External Crisis Prediction Using Machine Learning: Evidence from Three Decades of Crises Around the World.”

Belloni, A. and Chernozhukov, V., 2011. “High dimensional sparse econometric models: An introduction,” In *Inverse Problems and High-Dimensional Estimation* (pp. 121-156). Springer, Berlin, Heidelberg.

Berg. A. and C. Pattillo. 1999. “Predicting Currency Crises: The Indicators Approach and An Alternative,” *Journal of International Money and Finance*, 18, pp. 561–86.

Berg, A., E. Borensztein and C. Pattillo, 2005, “Assessing Early Warning Systems: How Have They Worked in Practice?” *IMF Staff Papers*, 52(3), pp. 462–502.

Berg, M.A., Berkes, M.E., Pattillo, M.C.A., Presbitero, A. and Yakhshilikov, M.Y., 2014. “Assessing bias and accuracy in the World Bank-IMF's Debt Sustainability Framework for Low-income Countries,” IMF Working Paper No. 14/48 (International Monetary Fund).

Berti, K., M. Salto, M. Lequien. 2013. “An Early-detection Index of Fiscal Stress for EU Countries,” European Economy. *Economic Papers*. 475. Brussels.

Bolhuis, M., and Brett Rayner, 2020, “The More the Merrier? A Machine Learning Algorithm for Optimal Pooling of Panel Data,” IMF Working Paper No. 20/44 (International Monetary Fund).

Bluwstein, C., M. Buckmann, A. Joseph, M. Kang, S. Kapadia, and O. Simsek, 2020, “Credit Growth, the Yield Curve and Financial Crisis Prediction: Evidence from a Machine Learning Approach.”, Bank of England Staff Working Paper No. 848.

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

Breiman, L., Friedman, J. H., Olshen, R. A., Stone, C. J., 1984. “Classification and Regression Trees.” Chapman & Hall, London.

©International Monetary Fund. Not for Redistribution

Bruns, M. and T. Poghosyan, 2018, "Leading Indicators of Fiscal Distress: Evidence from the Extreme Bound Analysis," *Applied Economics*, 50(13), pp. 1454-78.

Cameron, A., J. Gelbach, J., and D. Miller, 2011. "Robust Inference with Multiway Clustering," *Journal of Business & Economic Statistics*, 29(2), 238-249.

Cerovic, S., K. Gerling, A. Hodge, and P. Medas. 2018. "Predicting Fiscal Crises," IMF Working Paper No. 18/181 (Washington: International Monetary Fund).

Chakrabarti, A., and H. Zeaiter. 2014. "The Determinants of Sovereign Default: A Sensitivity Analysis," *International Review of Economics and Finance*, 33, pp. 300–318.

Chen, T., He, T., M. Benesty, V. Khotilovich, and Y. Tang, 2015. "Xgboost: extreme gradient boosting." *R package version 0.4-2*, pp.1-4. Christofides, C., T. Eicher, and C. Papageorgiou, 2016, "Did Established Early Warning Signals Predict the 2008 Crises?," *European Economic Review*, 81, pp. 103-114.

Ciarlone, A., and G. Trebeschi. 2005. "Designing an early warning system for debt crises." *Emerging Markets Review* 6, pp. 376–395.

Cole, H.L, and T.J. Kehoe, 1996. "A Self-fulfilling Model of Mexico's 1994–1995 Debt Crisis," *Journal of International Economics* 41, pp. 309–330.

Cole, H.L., and T.J. Kehoe, 2000, "Self-fulfilling Debt Crises," *Review of Economic Studies* 67, pp. 91–116.

Copas, J. ,1983, "Regression, Prediction and Shrinkage," *Journal of the Royal Statistical Society. Series B (Methodological)*, 45(3), 311-354.

Cruz, C., P. Keefer, and C. Scartascini, 2016, "Database of Political Institutions." Inter-American Development Bank.

Dawood, M., N. Horsewood, and F. Strobel. 2017. "Predicting sovereign debt crises: An Early Warning System approach," *Journal of Financial Stability*, 28, pp. 16-28

De Cos, Pablo Hernández, G. Koester, E. Moral-Benito, C. Nickel. 2014. "Signaling Fiscal Stress in the Euro Area: A Country-Specific Early Warning System," ECB Working Paper No. 1712 Frankfurt: European Central Bank.

Degenhardt, F., S. Seifert, and S. Szymczak, 2017, „Evaluation of variable selection methods for random forests and omics data sets," *Briefings in bioinformatics*, 20(2), pp.492-503.

Detragiache, E., 1996, "Rational Liquidity Crises in the Sovereign Debt Market: In Search of a Theory," *IMF Staff Papers* 43, pp. 545–570.

Detragiache, E. and A. Spilimbergo, 2001, "Crises and Liquidity: Evidence and Interpretation," IMF Working Paper No. 01/02 (Washington: International Monetary Fund).

©International Monetary Fund. Not for Redistribution

Diebold, F.X. and Mariano, R.S., 2002. Comparing predictive accuracy. *Journal of Business & economic statistics*, *20*(1), pp.134-144.

Durlauf, S.N. and P.A. Johnson, 1995, Multiple regimes and cross-country growth behavior. *Journal of applied econometrics*, *10*(4), pp.365-384.

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

Eichengreen, B. and R. Hausmann, 1999. "Exchange Rates and Financial Fragility," *NBER Working Papers* 7418, National Bureau of Economic Research

Fioramanti, M. 2008. "Predicting Sovereign Debt Crises using Artificial Neural Networks: A Comparative Approach," *Journal of Financial Stability*, Vol. 4, pp. 149–164.

Fischer, S., R. Sahay, and C. Vegh. 2002. "Modern Hyper- and High Inflations," *Journal of Economic Literature*, 40, pp. 837–880.

Frank, C. R., and Cline,W. R. 1971. "Measurement of debt servicing capacity: An application of discriminant analysis," *Journal of International Economics*, 1, pp. 327–344.

Frankel, J., and G. Saravelos. 2012. "Can Leading Indicators Assess Country Vulnerability? Evidence from the 2008–09 Global Financial Crisis," *Journal of International Economics* 87, pp. 216–231.

Friedman, J.H., 2001. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pp.1189-1232.

Fuertes, A., and E. Kalotychou, 2006, "Early Warning Systems for Sovereign Debt Crises: The Role of Heterogeneity," *Computational Statistics & Data Analysis* 51, pp. 1420–1441.

Gelos, R., R. Sahay, and G. Sandleris, 2004, "Sovereign Borrowing by Developing Countries: What Determines Market Access?" IMF Working Paper No. 04/211 (Washington: International Monetary Fund).

Ghulam, Y., and J. Derber. 2018. "Determinants of Sovereign Defaults," *The Quarterly Review of Economics and Finance* 69, pp. 43–55.

Giannone, D., M. Lenza, and G. Primiceri, 2018. "Economic Predictions with Big Data: The Illusion of Sparsity," *Staff Report No. 847*, Federal Reserve Bank of New York.

Gigerenzer, G., and H. Brighton, 2009, "Homo heuristicus: Why biased minds make better inferences." Topics in cognitive science, 1(1):107–143, 2009.

Goulet Coulombe, P., 2020. "To Bag is to Prune".

Goulet Coulombe, P., Leroux, M., Stevanovic, D. and Surprenant, S., 2020a. "How is Machine Learning Useful for Macroeconomic Forecasting?" Goulet Coulombe, P., Leroux,

©International Monetary Fund. Not for Redistribution

M., Stevanovic, D. and Surprenant, S., 2020b. “Macroeconomic Data Transformations Matter.”

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

Guscina, A., M. Sheheryar, and M. Papaioannou. 2017. “Assessing Loss of Market Access: Conceptual and Operational Issues,” IMF Working Paper No. 17/246 (Washington: International Monetary Fund).

Hastie, T., Friedman, J., and Tibshirani, R., 2012. *The elements of statistical learning*, Second Edition. New York: Springer series in statistics.

Hawkins, J., and Klau, M. 2000. “Measuring Potential Vulnerabilities in Emerging Market Economies,” BIS Working Papers 91. Bank for International Settlements.

Hellwig, K.P., 2018. “Overfitting in Judgment-based Economic Forecasts: The Case of IMF Growth Projections.” IMF Working Paper No. 18/260 (Washington: International Monetary Fund).

Hoerl, A.E. and Kennard, R.W., 1970. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, *12*(1), pp.55-67.

IMF. 2015, “The Fund's Lending Framework and Sovereign Debt—Further Considerations,” Board Paper (Washington: International Monetary Fund).

Jarmulska, B., 2020. “Random Forest Versus Logit Models: Which Offers Better Early Warning of Fiscal Stress?” ECB Working Paper No. 20202408

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

Jordà Ò., M. Schularick, and A. M. Taylor. 2016. “Sovereigns versus Banks: Credit, Crises, and Consequences,” *Journal of the European Economic Association* 14 (1), pp. 45–79.

Kaminsky, G.L., S. Lizondo, and C.M. Reinhart. 1998. “Leading Indicators of Currency Crises,” *IMF Staff Papers*, Vol. 45, No. 1, pp. 1–48.

Kleinberg, J., Ludwig, J., Mullainathan, S. and Obermeyer, Z., 2015. Prediction policy problems. *American Economic Review*, 105(5), pp.491-95.

Kraay, A., and V. Nehru. 2006. “When Is External Debt Sustainable?” *The World Bank Economic Review* 20 (3): 341–365.

Laeven, L. and F. Valencia. 2018. “Systemic Banking Crises Revisited,” IMF Working Paper No. 18/206 (Washington: International Monetary Fund).

©International Monetary Fund. Not for Redistribution

Lane, P. 2012. “The European Sovereign Debt Crises,” *Journal of Economic Perspectives*—Volume 26, Number 3—Summer 2012—pp. 49–68.

Lane, P., and G. Milesi-Ferretti, “The External Wealth of Nations Mark II,” Journal of International Economics, November 2007

Manasse, P., N. Roubini, and A. Schimmelpfennig. 2003. “Predicting Sovereign Debt Crises,” IMF Working Paper No. 03/221 (Washington: International Monetary Fund).

Marashaden, O. 1997. “A Logit Model to Predict Debt Rescheduling by Less Developed Countries.” Asian Economies 26, pp. 25–34.

Mauro, P., R. Romeu, A. Binder, and A. Zaman, 2011, “A Modern History of Fiscal Prudence and Profligacy,” IMF Working Paper No. 13/5 (Washington: International Monetary Fund).

Mbaye, S., M. Moreno-Badia, and K. Chae. 2018. “The Global Debt Database: Methodology and Sources.” IMF Working Paper No. 18/111 (Washington: International Monetary Fund).

Medas, P., T. Poghosyan, Y. Xu, J. Farah-Yacoub, and K. Gerling. 2018, “Fiscal Crises,” *Journal of International Money and Finance*, Vol. 88, 191-207.

Merton, Robert C. (1974). “On the Pricing of Corporate Debt: The Risk Structure of Interest Rates”. *Journal of Finance*. 29 (2): 449–470.

Moreno Badia, M., P. Gupta, P. Medas, and Y. Xiang, 2020, “*Debt is not Free*”, IMF Working Paper No. 20/1 (Washington: International Monetary Fund).

Mullainathan S. and J. Spiess. 2017. “Machine Learning: An Applied Econometric Approach,” *Journal of Economic Perspectives*—Volume 31, Number 2, pp. 87–106.

Ng, S., 2013. “Variable Selection in Predictive Regressions”, *Handbook of Economic Forecasting*, vol. 2, 752–789, Elsevier.

Pamies S., S. and K. Berti. 2017. “A Complementary Tool to Monitor Fiscal Stress in European Economies,” EC Discussion Paper, 49 (June).

Peter, M. 2002. “Estimating Default Probabilities of Emerging Market Sovereigns: A New Look at A Not-So-New Literature,” HEI Working Paper No: 06/2002, Geneva: Graduate Institute for International Studies.

Reinhart, C. M. 2002. “Default, Currency Crises, and Sovereign Credit Ratings,” *World Bank Economic Review*, Oxford University Press, vol. 16(2), pages 151-170

Reinhart, C. M, and K. Rogoff. 2009, *This time is different: Eight centuries of financial folly* (Princeton, NJ: Princeton University Press).

©International Monetary Fund. Not for Redistribution

———. 2011a, "The Forgotten History of Domestic Debt," *Economic Journal* 121 (552), pp. 319–350.

———. 2011b, "From Financial Crash to Debt Crisis," *American Economic Review* 101(5), pp. 1676–1706.

Reinhart, C.M., Rogoff, K.S. and Savastano, M.A., 2003. "Debt intolerance", *NBER Working Paper No.9908,* National Bureau of Economic Research.

Rodriguez A., and P. N. Rodriguez. 2006. "Understanding and Predicting Sovereign Debt Rescheduling: A Comparison of the Areas Under Receiver Operating Characteristic Curves," *Journal of Forecasting*, 25, pp. 459-479.

Rodrik, D., 2008. "The real exchange rate and economic growth." Brookings papers on economic activity, 2008(2), pp.365-412.

Savona, R., and M. Vezzoli. 2015. "Fitting and Forecasting Sovereign Defaults using Multiple Risk Signals," *Oxford Bulletin of Economics and Statistics*, 77 (1), pp. 66-91

Savona, R., Vezzoli, M., and E. Ciavolino. 2015. "A Data-Driven Explanation of Country Risk: Emerging Markets vs. Eurozone Debt Crises," SYRTO Working Paper n.17

Taleb, N.N., 2007, "The black swan: The impact of the highly improbable" (Vol. 2). Random house.

Tetlock, P.E., 2017. "Expert Political Judgment: How Good Is It? How Can We Know?"- New Edition. Princeton University Press.

Tibshirani, R., 1996. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, *58*(1), pp.267-288.

Ulfelder, J. 2012. "Why the World Can't Have a Nate Silver". Foreign Policy. https://foreignpolicy.com/2012/11/08/why-the-world-cant-have-a-nate-silver/

Van den Berg, J., Candelon, B. and Urbain, J.P., 2008. "A cautious note on the use of panel models to predict financial crises." *Economics Letters*, *101*(1), pp.80-83.

Van Rijckeghem, C. and Weder, B., 2009. "Political Institutions and Debt Crises." *Public Choice*, *138*(3-4), p.387

Weisfeld, H., I. de Carvalho Filho, F. Comelli, R. Giri, C. Huang, F. Liu, S. Lizarazo Ruiz, A. Meyer Cirkel, and A. Presbitero, 2020, "Predicting Macroeconomic and Macrofinancial Stress in Low-Income Countries," IMF Working Paper No. 20/289 (Washington: International Monetary Fund).

Zou, H. and Hastie, T., 2005. "Regularization and variable selection via the elastic net." *Journal of the royal statistical society: series B (statistical methodology)*, *67*(2), pp.301-320.

©International Monetary Fund. Not for Redistribution

Tables for main text

Table 4.1. Predictive performance of historical averages

	advanced and emerging market economies			low-income developing countries		
	pooled	country-specific	combination	pooled	country-specific	combination
accuracy of predictions:						
log(likelihood)	-0.353	-0.484	-0.314	-0.655	-1.01	-0.578
	(0.032)	(0.126)	(0.031)	(0.056)	(0.231)	(0.049)
log(likelihood), weighted	-0.972	-1.778	-0.886	-0.97	-1.515	-0.82
	(0.003)	(0.416)	(0.025)	(0.003)	(0.424)	(0.023)
MSE	0.098	0.092	0.089	0.224	0.193	0.198
	(0.014)	(0.012)	(0.012)	(0.024)	(0.019)	(0.02)
accuracy of rankings:						
AUROC	0.362	0.75	0.744	0.379	0.701	0.695
	(0.031)	(0.03)	(0.031)	(0.033)	(0.035)	(0.035)
size of test sample	1323	1323	1323	555	555	555
of which: crisis observations	136	136	136	165	165	165

Note: bootstrapped standard deviations in parentheses (with two-way clustering by country and year, following Cameron et al. 2008); backward-looking averages are used as the predicted probability of a crisis start occurring in year t+1 or t+2; test sample covers t = 2001-2015.

©International Monetary Fund. Not for Redistribution

Table 5.1. Baseline predictors: benchmarking in-sample and out-of-sample performance of econometric approaches

(i) Advanced and emerging market economies

	(i)	(ii)	(iii)	(iv)	(v)	(vi)	(vii)
	logit			**classification tree**			**historical averages**
	in-sample	**out-of-sample**		**in-sample**	**out-of-sample**		
impute in training sample	no	no		no	no		combination of country-specific and pooled average
impute in test sample	no	no		no	no		
training sample coverage	AE + EM	AE + EM		AE + EM	AE + EM		
training sample start year	2001	1979	1989	1979	1979	1989	
accuracy of predictions:							
log(likelihood)	-0.28*	-0.302	-0.31	-0.231***	-0.332	-0.33	-0.307
	(0.037)	(0.034)	(0.036)	(0.035)	(0.041)	(0.044)	(0.03)
log(likelihood), weighted	-1.03***	-1.038***	-1.094***	-0.844	-0.968**	-1.06***	-0.878
	(0.058)	(0.062)	(0.078)	(0.071)	(0.07)	(0.093)	(0.026)
MSE	0.081	0.086	0.087	0.064***	0.099**	0.095	0.086
	(0.013)	(0.012)	(0.012)	(0.012)	(0.014)	(0.014)	(0.012)
accuracy of rankings:							
AUROC	0.756	0.697	0.682*	0.822**	0.694	0.718	0.745
	(0.037)	(0.039)	(0.04)	(0.027)	(0.036)	(0.04)	(0.032)
average size of training sample	1148	1309	994	1148	1309	994	
of which: crisis observations	112	146	108	112	146	108	
size of test sample	1148	1148	1148	1148	1148	1148	1148
of which: crisis observations	112	112	112	112	112	112	112

(ii) Low-income countries

	logit			**classification tree**			**historical averages**
	in-sample	**out-of-sample**		**in-sample**	**out-of-sample**		
impute in training sample	no	no		no	no		combination of country-specific and pooled average
impute in test sample	no	no		no	no		
training sample coverage	LIDC	LIDC		LIDC	LIDC		
training sample start year	2001	1979	1989	2001	1979	1989	
accuracy of predictions:							
log(likelihood)	-0.638	-0.692	-0.705	-0.506***	-1.596*	-1.911*	-0.676
	(0.033)	(0.038)	(0.034)	(0.065)	(0.501)	(0.739)	(0.055)
log(likelihood), weighted	-0.697***	-0.741**	-0.733***	-0.548***	-1.67***	-1.86***	-0.805
	(0.023)	(0.034)	(0.033)	(0.121)	(0.51)	(0.567)	(0.022)
MSE	0.223	0.246	0.253	0.167***	0.318***	0.298*	0.24
	(0.015)	(0.017)	(0.015)	(0.018)	(0.024)	(0.033)	(0.024)
accuracy of rankings:							
AUROC	0.627	0.553	0.547*	0.803***	0.483***	0.53*	0.628
	(0.044)	(0.054)	(0.052)	(0.042)	(0.039)	(0.042)	(0.049)
average size of training sample	355	276	218	355	276	218	
of which: crisis observations	134	122	97	134	122	97	
size of test sample	355	355	355	355	355	355	355
of which: crisis observations	134	134	134	134	134	134	134

Note: bootstrapped standard deviations in parentheses (with two-way clustering by country and year, following Cameron et al. 2008). All models use the same 10 variables to predict the probability of a crisis start occurring in year t+1 or t+2; test sample covers t = 2001-2015. Out-of-sample performance obtained from 15 rolling regressions. Stars indicate whether performance is significantly different from the historical benchmark in column (vii), sith significance levels 10%, 5%, and 1% indicated by *,**, and ***, respectively.

©International Monetary Fund. Not for Redistribution

Table 5.2.A. Baseline predictors: predictive performance of different models (without imputation or pooling)

	(i)	(ii)	(iii)	(iv)
	Linear		Tree ensembles	
method	Logit	Elastic net	XG boost	Random forest
accuracy of predictions:				
log(likelihood)	-0.302	-0.3	-0.287	-0.281
	(0.034)	(0.033)	(0.034)	(0.027)
log(likelihood), weighted	-1.038	-1.053	-0.902	-0.906
	(0.062)	(0.035)	(0.055)	(0.045)
MSE	0.086	0.085	0.083	0.081
	(0.012)	(0.012)	(0.012)	(0.011)
accuracy of rankings:				
AUC	0.697	0.697	0.75	0.772
	(0.039)	(0.033)	(0.033)	(0.024)
average size of training sample	1309	1309	1309	1309
of which: crisis observations	146	146	146	146
size of test sample	1148	1148	1148	1148
of which: crisis observations	112	112	112	112
(ii) Low-income developing countries				
impute in training sample	no	no	no	no
impute in test sample	no	no	no	no
training sample coverage	LIDC	LIDC	LIDC	LIDC
method	Logit	Elastic net	XG boost	Random forest
accuracy of predictions:				
log(likelihood)	-0.692	-0.666	-0.68	-0.681
	(0.038)	(0.012)	(0.024)	(0.026)
log(likelihood), weighted	-0.741	-0.7	-0.718	-0.702
	(0.034)	(0.004)	(0.016)	(0.022)
MSE	0.246	0.237	0.243	0.244
	(0.017)	(0.006)	(0.011)	(0.012)
accuracy of rankings:				
AUROC	0.553	0.546	0.53	0.577
	(0.054)	(0.028)	(0.041)	(0.043)
average size of training sample	276	276	276	276
of which: crisis observations	122	122	122	122
size of test sample	355	355	355	355
of which: crisis observations	134	134	134	134

Note: bootstrapped standard deviations in parentheses (with two-way clustering by country and year, following Cameron et al. 2008). All models use the same 10 variables to predict the probability of crisis start occurring in year t+1 or t+2; test sample covers t = 2001-2015. Out-of-sample performance obtained from 15 rolling regressions. Training samples start in 1979.

©International Monetary Fund. Not for Redistribution

Table 5.2.B. Baseline predictors: significance of differences in performance of different models (without imputation or pooling)

log(likelihood), AE + EM

	global average	country average	mix of averages	logit	elastic net	XG boost	random forest
logit	0.83	0.97	0.24	-	0.38	0.13	0.02
elastic net	0.92	0.98	0.23	0.63	-	0.09	0.02
XG boost	0.98	0.98	0.7	0.87	0.91	-	0.22
random forest	0.98	0.99	0.87	0.98	0.98	0.78	-

log(likelihood), LIDC

	global average	country average	mix of averages	logit	elastic net	XG boost	random forest
logit	1	0.91	0.61	-	0.16	0.33	0.34
elastic net	1	0.92	0.78	0.84	-	0.83	0.76
XG boost	1	0.91	0.69	0.67	0.17	-	0.52
random forest	1	0.91	0.68	0.66	0.24	0.48	-

weighted log(likelihood), AE + EM

	global average	country average	mix of averages	logit	elastic net	XG boost	random forest
logit	0.99	0.97	0.08	-	0.77	0.00	0
elastic net	1	0.97	0.01	0.23	-	0.00	0
XG boost	1	0.99	0.87	1	1	-	0.55
random forest	1	0.99	0.89	1	1	0.45	-

weighted log(likelihood), LIDC

	global average	country average	mix of averages	logit	elastic net	XG boost	random forest
logit	1	0.9	1.0	-	0.08	0.22	0.09
elastic net	1	0.92	1.0	0.92	-	0.91	0.54
XG boost	1	0.91	1.0	0.78	0.09	-	0.18
random forest	1	0.92	1.0	0.91	0.46	0.82	-

MSE, AE + EM

	global average	country average	mix of averages	logit	elastic net	XG boost	random forest
logit	0.62	0.81	0.28	-	0.42	0.22	0.06
elastic net	0.73	0.84	0.22	0.58	-	0.16	0.05
XG boost	0.86	0.95	0.56	0.78	0.84	-	0.31
random forest	0.87	0.99	0.69	0.94	0.95	0.69	-

MSE, LIDC

	global average	country average	mix of averages	logit	elastic net	XG boost	random forest
logit	0.99	0.08	0.57	-	0.18	0.4	0.41
elastic net	0.99	0.31	0.74	0.82	-	0.84	0.78
XG boost	0.99	0.16	0.62	0.6	0.16	-	0.52
random forest	0.99	0.17	0.61	0.59	0.22	0.48	-

AUC, AE + EM

	global average	country average	mix of averages	logit	elastic net	XG boost	random forest
logit	1	0.07	0.09	-	0.53	0.0	0
elastic net	1	0.06	0.09	0.47	-	0.0	0
XG boost	1	0.5	0.59	0.98	0.98	-	0.11
random forest	1	0.83	0.89	1	1	0.9	-

AUC, LIDC

	global average	country average	mix of averages	logit	elastic net	XG boost	random forest
logit	1	0.05	0.05	-	0.58	0.68	0.29
elastic net	0.99	0.06	0.07	0.42	-	0.63	0.23
XG boost	0.99	0.02	0.02	0.32	0.37	-	0.12
random forest	1	0.11	0.12	0.71	0.77	0.88	-

Note: tables indicate degree of confidence (1 - p-value) that the model in row i outperforms the model (or historical benchmark) in column j, based on a one-sided t-test using bootstrapped standard errors, adjusted for two-way clustering by country and year (following Cameron et al., 2008).

©International Monetary Fund. Not for Redistribution

Table 5.3.A. Baseline predictors: predictive performance of logit model (with and without imputation and pooling)

(i) Advanced and emerging market economies

	(i)	(ii)	(iii)	(iv)
impute in training sample	no		yes	
impute in test sample	no		no	
training sample coverage	AE + EM	all countries	AE + EM	all countries
accuracy of predictions:				
log(likelihood)	-0.302	-0.298	-0.297	-0.295
	(0.034)	(0.038)	(0.033)	(0.039)
log(likelihood), weighted	-1.038	-1.093	-0.978	-1.087
	(0.062)	(0.06)	(0.057)	(0.059)
MSE	0.086	0.084	0.085	0.083
	(0.012)	(0.013)	(0.012)	(0.013)
accuracy of rankings:				
AUROC	0.697	0.705	0.714	0.712
	(0.039)	(0.037)	(0.038)	(0.036)
average size of training sample	1309	1585	1888	2679
of which: crisis observations	146	268	238	471
size of test sample	1148	1148	1148	1148
of which: crisis observations	112	112	112	112

(ii) Low-income countries

impute in training sample	no		yes	
impute in test sample	no		no	
training sample coverage	LIDC	all countries	LIDC	all countries
accuracy of predictions:				
log(likelihood)	-0.692	-0.695	-0.709	-0.672
	(0.038)	(0.04)	(0.063)	(0.045)
log(likelihood), weighted	-0.741	-0.724	-0.817	-0.733
	(0.034)	(0.051)	(0.059)	(0.049)
MSE	0.246	0.25	0.246	0.238
	(0.017)	(0.016)	(0.024)	(0.018)
accuracy of rankings:				
AUROC	0.553	0.536	0.572	0.556
	(0.054)	(0.051)	(0.048)	(0.048)
average size of training sample	276	1585	791	2679
of which: crisis observations	122	268	233	471
size of test sample	355	355	355	355
of which: crisis observations	134	134	134	134

Note: bootstrapped standard deviations in parentheses (with two-way clustering by country and year, following Cameron et al. 2008). All models use the same 10 variables to predict the probability of crisis start occurring in year t+1 or t+2; test sample covers t = 2001-2015. Out-of-sample performance obtained from 15 rolling regressions. Training samples start in 1979.

©International Monetary Fund. Not for Redistribution

Table 5.3.B. Baseline predictors: significance of differences in performance of logit, with and without imputation and pooling

	global average	country average	mix of averages	no pooling, no imputation	pooling, no imputation	no pooling, imputation	pooling, imputation	global average	country average	mix of averages	no pooling, no imputation	pooling, no imputation	no pooling, imputation	pooling, imputation
	log(likelihood), AE + EM							log(likelihood), LIC						
no pooling, no imputation	0.97	0.92	0.6	-	0.34	0.02	0.17	0.98	0.96	0.23	-	0.45	0.51	0.28
pooling, no imputation	1	0.93	0.76	0.66	-	0.17	0.07	0.96	0.96	0.33	0.55	-	0.56	0.18
no pooling, imputation	0.99	0.94	0.88	0.98	0.83	-	0.58	0.99	0.96	0.18	0.49	0.45	-	0.27
pooling, imputation	1	0.95	0.92	0.83	0.93	0.42	-	0.98	0.97	0.43	0.72	0.82	0.74	-
	weighted log(likelihood), AE + EM							weighted log(likelihood), LIC						
no pooling, no imputation	0.57	0.95	0.05	-	0.57	0.00	0.84	1	0.95	0.99	-	0.12	1.00	0.16
pooling, no imputation	0.56	0.95	0.01	0.44	-	0.01	0.93	1	0.96	1	0.88	-	1.00	0.57
no pooling, imputation	0.91	0.97	0.19	1.00	1	-	1.00	1	0.94	0.85	0.00	0.00	-	0.00
pooling, imputation	0.35	0.95	0	0.17	0.07	0	-	1	0.96	1	0.84	0.43	1.00	-
	MSE, AE + EM							MSE, LIC						
no pooling, no imputation	0.92	0.96	0.72	-	0.57	0.19	0.48	0.96	0.05	0.29	-	0.55	0.44	0.37
pooling, no imputation	0.97	0.98	0.78	0.43	-	0.27	0.31	0.93	0.11	0.32	0.45	-	0.41	0.18
no pooling, imputation	0.94	0.99	0.8	0.82	0.73	-	0.62	0.99	0.08	0.28	0.56	0.59	-	0.4
pooling, imputation	0.99	0.98	0.87	0.52	0.69	0.38	-	0.96	0.19	0.42	0.64	0.82	0.6	-
	AUC, AE + EM							AUC, LIC						
no pooling, no imputation	1	0.05	0.07	-	0.11	0	0.02	0.64	0.01	0.01	-	0.35	0.1	0.22
pooling, no imputation	1	0.14	0.17	0.89	-	0.06	0.03	0.71	0.03	0.02	0.65	-	0.27	0.20
no pooling, imputation	1	0.32	0.38	1.00	0.94	-	0.52	0.81	0.06	0.04	0.90	0.73	-	0.50
pooling, imputation	1	0.31	0.36	0.98	0.97	0.48	-	0.8	0.09	0.07	0.78	0.81	0.5	-

Note: tables indicate degree of confidence (1 - p-value) that the model in row i outperforms the model (or historical benchmark) in column j, based on a one-sided t-test using bootstrapped standard errors, adjusted for two-way clustering by country and year (following Cameron et al., 2008). Results refer to logit models evaluated without imputation in the test sample (i.e., "imputation" refers to imputation in the training samples).

©International Monetary Fund. Not for Redistribution

Table 5.4. Baseline predictors: predictive performance of different models (with imputation and pooling)

(i) Advanced and emerging market economies				
	(i)	(ii)	(iii)	(iv)
impute in training sample	yes	yes	yes	yes
impute in test sample	no	no	no	no
training sample coverage	all countries	all countries	all countries	all countries
method	Logit	Elastic net	XG boost	Random forest
accuracy of predictions:				
log(likelihood)	-0.295	-0.294	-0.281	-0.283
	(0.039)	(0.035)	(0.032)	(0.024)
log(likelihood), weighted	-1.087	-1.057	-0.924	-0.882
	(0.059)	(0.043)	(0.044)	(0.036)
MSE	0.083	0.083	0.082	0.082
	(0.013)	(0.013)	(0.012)	(0.01)
accuracy of rankings:				
AUROC	0.712	0.716	0.761	0.777
	(0.036)	(0.037)	(0.033)	(0.022)
average size of training sample	2679	2679	2679	2679
of which: crisis observations	471	471	471	471
size of test sample	1148	1148	1148	1148
of which: crisis observations	112	112	112	112

(ii) Low-income developing countries				
impute in training sample	yes	yes	yes	yes
impute in test sample	no	no	no	no
training sample coverage	all countries	all countries	all countries	all countries
method	Logit	Elastic net	XG boost	Random forest
accuracy of predictions:				
log(likelihood)	-0.672	-0.669	-0.688	-0.669
	(0.045)	(0.046)	(0.049)	(0.03)
log(likelihood), weighted	-0.733	-0.737	-0.778	-0.734
	(0.049)	(0.044)	(0.041)	(0.031)
MSE	0.238	0.237	0.244	0.237
	(0.018)	(0.018)	(0.021)	(0.013)
accuracy of rankings:				
AUROC	0.556	0.555	0.555	0.581
	(0.048)	(0.048)	(0.047)	(0.038)
average size of training sample	2679	2679	2679	2679
of which: crisis observations	471	471	471	471
size of test sample	355	355	355	355
of which: crisis observations	134	134	134	134

Note: bootstrapped standard deviations in parentheses (with two-way clustering by country and year, following Cameron et al. 2008). All models use the same 10 variables to predict the probability of crisis start occurring in year t+1 or t+2; test sample covers t = 2001-2015. Out-of-sample performance obtained from 15 rolling regressions. Training samples start in 1979.

©International Monetary Fund. Not for Redistribution

Table 6.1.A. Model-based variable selection: predictive performance (with imputation and pooling)

Table 6.1.A. Predictive performance for baseline specification

	(i)	(ii)	(iii)	(iv)	(v)	(vi)	(vii)	(viii)	(ix)	(x)
method	logit	logit with BIC	elastic net		XG boost		random forest		random forest with rfe	combination of country-specific and pooled average
number of predictors	10	748	10	748	10	748	10	748	748	
(i) Advanced and emerging market economies										
accuracy of predictions:										
log(likelihood)	-0.305	-0.322	-0.304	-0.294	-0.293	-0.289	-0.289	-0.3	-0.293	-0.314
	(0.038)	(0.047)	(0.047)	(0.037)	(0.032)	(0.036)	(0.039)	(0.027)	(0.029)	(0.031)
log(likelihood), weighted	-1.081	-1.158	-1.051	-0.978	-0.94	-0.945	-0.875	-0.747	-0.799	-0.886
	(0.052)	(0.101)	(0.059)	(0.051)	(0.041)	(0.061)	(0.044)	(0.029)	(0.041)	(0.025)
MSE	0.087	0.09	0.087	0.087	0.085	0.085	0.085	0.088	0.087	0.089
	(0.013)	(0.013)	(0.017)	(0.013)	(0.012)	(0.013)	(0.015)	(0.01)	(0.011)	(0.012)
accuracy of rankings:										
AUROC	0.714	0.748	0.717	0.766	0.75	0.771	0.78	0.802	0.787	0.744
	(0.033)	(0.032)	(0.038)	(0.032)	(0.03)	(0.031)	(0.031)	(0.026)	(0.029)	(0.031)
(ii) Low-income developing countries										
accuracy of predictions:										
log(likelihood)	-0.593	-0.645	-0.59	-0.576	-0.591	-0.575	-0.58	-0.557	-0.558	-0.578
	(0.042)	(0.062)	(0.027)	(0.046)	(0.046)	(0.049)	(0.022)	(0.036)	(0.038)	(0.049)
log(likelihood), weighted	-0.73	-0.771	-0.735	-0.715	-0.767	-0.74	-0.725	-0.665	-0.661	-0.82
	(0.046)	(0.08)	(0.044)	(0.047)	(0.038)	(0.055)	(0.039)	(0.028)	(0.038)	(0.023)
MSE	0.203	0.216	0.202	0.198	0.202	0.195	0.199	0.189	0.19	0.198
	(0.017)	(0.021)	(0.011)	(0.019)	(0.019)	(0.019)	(0.009)	(0.015)	(0.016)	(0.02)
accuracy of rankings:										
AUROC	0.62	0.662	0.623	0.676	0.634	0.683	0.654	0.705	0.7	0.695
	(0.039)	(0.041)	(0.039)	(0.042)	(0.044)	(0.042)	(0.034)	(0.04)	(0.041)	(0.035)

Note: bootstrapped standard deviations in parentheses (with two-way clustering by country and year, following Cameron et al. 2008). Models predict probability of crisis start occurring in year t+1 or t+2; test sample covers t = 2001-2015. Out-of-sample performance obtained from 15 rolling regressions. Training samples start in 1979 and pool all countries. Missing values in the test and training samples are imputed using the training sample median.

©International Monetary Fund. Not for Redistribution

Table 6.1.B. Model-based variable selection: significance of differences in performance (with imputation and pooling)

	global average	country average	mix of averages	logit with BIC	elastic net	XG boost	random forest	random forest with rfe	global average	country average	mix of averages	logit with BIC	elastic net	XG boost	random forest	random f
	log(likelihood), AE + EM								**log(likelihood), LIDC**							
logit with BIC	0.88	0.95	0.37	-	0.049	0.04	0.19	0.1	0.59	0.95	0.04	-	0.00	0.02	0.01	0.01
elastic net	1	0.97	0.96	0.95	-	0.25	0.65	0.45	0.99	0.97	0.53	1.00	-	0.47	0.14	0.13
XG boost	1	0.97	0.99	0.96	0.746	-	0.78	0.66	1.00	0.97	0.56	0.98	0.53	-	0.17	0.16
random forest	1	0.95	0.93	0.82	0.349	0.22	-	0.09	1.00	0.98	0.82	0.99	0.86	0.83	-	0.57
random forest with rfe	1	0.96	0.99	0.9	0.547	0.34	0.91	-	1.00	0.98	0.81	0.99	0.87	0.84	0.43	-
	weighted log(likelihood), AE + EM								**weighted log(likelihood), LIDC**							
logit with BIC	0.03	0.95	0	-	0.00	0.00	0.00	0.00	0.99	0.97	0.75	-	0.10	0.28	0.06	0.05
elastic net	0.45	0.98	0.01	1.00	-	0.15	0.00	0.00	1.00	0.97	1.00	0.90	-	0.83	0.04	0.03
XG boost	0.67	0.99	0.11	1.00	0.85	-	0.00	0.00	1.00	0.97	0.96	0.72	0.17	-	0.01	0.00
random forest	1	1	1	1.00	1.00	1.00	-	1.00	1.00	0.98	1.00	0.94	0.96	0.99	-	0.40
random forest with rfe	1	0.99	1	1.00	1.00	1.00	0.001	-	1.00	0.98	1.00	0.96	0.97	1.00	0.60	-
	MSE, AE + EM								**MSE, LIDC**							
logit with BIC	0.91	0.68	0.44	-	0.203	0.144	0.323	0.25	0.69	0.03	0.09	-	0.00	0.03	0.01	0.01
elastic net	0.99	0.9	0.69	0.8	-	0.21	0.546	0.47	0.98	0.24	0.50	1.00	-	0.31	0.06	0.08
XG boost	1	0.98	0.91	0.86	0.79	-	0.758	0.75	1.00	0.37	0.66	0.98	0.70	-	0.18	0.21
random forest	0.97	0.91	0.64	0.68	0.454	0.242	-	0.34	1.00	0.68	0.84	0.99	0.94	0.82	-	0.70
random forest with rfe	0.99	0.95	0.72	0.75	0.534	0.252	0.658	-	0.99	0.62	0.80	0.99	0.93	0.79	0.31	-
	AUC, AE + EM								**AUC, LIDC**							
logit with BIC	1	0.47	0.56	-	0.13	0.12	0.01	0.03	1.00	0.12	0.15	-	0.21	0.24	0.08	0.10
elastic net	1	0.75	0.82	0.87	-	0.39	0.02	0.11	1.00	0.20	0.25	0.79	-	0.37	0.09	0.12
XG boost	1	0.81	0.87	0.88	0.62	-	0.02	0.1	1.00	0.27	0.33	0.77	0.64	-	0.09	0.13
random forest	1	1	1	1	0.98	0.99	-	0.95	1.00	0.56	0.64	0.92	0.91	0.91	-	0.65
random forest with rfe	1	0.96	0.98	0.97	0.90	0.90	0.05	-	1.00	0.50	0.58	0.90	0.88	0.87	0.35	-

Note: tables indicate degree of confidence (1 - p-value) that the model in row i outperforms the model (or historical benchmark) in column j, based on a one-sided t-test using bootstrapped standard errors, adjusted for two-way clustering by country and year (following Cameron et al., 2008)

©International Monetary Fund. Not for Redistribution

Table 7.1. Ranking of predictors by importance

Table lists all predictors that are ranked 30th or higher for at least one prediction method. Numbers indicate ranks in terms of variable importance (see Section 7)

Color coding: 🟩 **= Top 10** 🟨 **= Top 20**

		logit	elastic net	XG boost	random forest
Crisis history					
Historical crisis frequency			31 (+)	7	3
Years since last crisis		8 (-)	1 (-)	2	4
Number of countries with banking crisis starts, current year		21 (+)			486
Output, demand, prices					
log(GDP per capita), PPP	level	2 (-)	2 (-)	3	1
log(GDP per capita), PPP	lag			1	2
log(GDP per capita), PPP	10-year change		66 (-)	27	89
log(GDP), US dollars	level		8 (-)	22	23
Real GDP per capita	3-year growth rate	19 (-)	42 (-)	90	112
Gross domestic savings / GDP	level		16 (-)	23	56
Gross domestic savings / GDP	lag			19	62
Gross domestic savings / GDP	10-year average			15	40
Trading partner real GDP	3-year growth rate			30	163
Natural resource rents / GDP	level	12 (-)		54	121
Agricultural GDP (share of total)	level			45	13
CPI	percentage change	20 (+)	7 (+)	16	28
CPI	3-year pct change			107	27
Real exchange rate	3-year depreciation		28 (+)	62	173
Volatility					
CPI inflation	10-year std. dev.		30 (+)	12	14
Terms of trade growth	10-year std. dev.	26 (+)	56 (+)	20	55
Nominal exchange rate depreciation rate	10-year std. dev.			26	132
External sector					
Current account balance / GDP	level	10 (-)	9 (-)	21	25
External gross financing needs / GDP	level			161	22
External gross financing needs / GDP	lag	23 (-)		130	46
Official developent assistance / GDP	level			129	26
Official developent assistance / GDP	lag	22 (-)		415	41
Official reserves (in months of imports)	level			6	50
Official reserves (in US dollars)	percentage change			18	128
Population					
Working-age population, share of total	level		12 (-)		6
Working-age population	growth rate			24	80
Age dependency ratio	level			8	5
Age dependency ratio	10-year change			28	146
Population density	level	6 (-)	10 (-)	14	71
Urban population (share of total)	level			280	10

©International Monetary Fund. Not for Redistribution

Table 7.1. Ranking of predictors by importance (continued)

Table lists all predictors that are ranked 30th or higher for at least one prediction method. Numbers indicate ranks in terms of variable importance (see Section 7)

Color coding: 🟩 **= Top 10** 🟨 **= Top 20**

		logit	elastic net	XG boost	random forest
Fiscal					
Government revenues / GDP	lag		6 (-)	134	96
Debt stocks					
Public external debt / GDP	level	7 (+)	11 (+)	4	8
Public external debt / GDP	lag		47 (+)	25	15
Public external debt / exports	level	15 (-)		9	7
Public external debt / exports	lag			92	12
Short-term public external debt / GDP	lag	30 (+)		389	293
Total external debt / exports	lag			10	60
Private debt / GDP	level			290	24
Private debt / GDP	lag			178	29
Private debt / GDP	5-year change	11 (+)	22 (+)	291	151
Debt service					
Government interest expenditure / GDP	3-year change	28 (+)	33 (+)		245
Public external debt amortization / reserves	level			53	9
Public external debt amortization / reserves	lag			31	20
Public external debt amortization / GDP	level		14 (+)	322	47
Total external debt amortization / exports	level		26 (+)	135	65
Total external debt amortization / reserves	level		25 (+)		19
Total external debt amortization / reserves	lag	13 (+)	18 (+)	233	30
Total external debt amortization / reserves	5-year change			5	37
Total external debt amortization / reserves	10-year change	29 (-)	21 (-)	13	33
Total external debt service / reserves	level		64 (-)	11	34
Total external debt service / reserves	lag	5 (-)	13 (-)	122	53
Total external debt service / GDP	level	18 (+)		79	131
Money					
Broad money / GDP	level			240	21
Broad money / GDP	lag			116	18
Broad money / GDP	5-year change	27 (-)		123	298
External assets					
External assets / GDP	level	1 (-)		32	11
External assets / GDP	lag			68	17
External assets / GDP	10-year change	3 (+)		320	190
Quality of governance					
Bureaucracy, quality	level	4 (-)	4 (-)	17	16
Corruption control	level			29	32
Corruption control	10-year change	17 (+)	19 (+)	153	239
Regulatory quality	5-year change	24 (-)	17 (-)	88	233
Absence of violence	5-year change		20 (-)	34	267

©International Monetary Fund. Not for Redistribution

Table 7.1. Ranking of predictors by importance (continued)

Table lists all predictors that are ranked 30th or higher for at least one prediction method. Numbers indicate ranks in terms of variable importance (see Section 7)

Color coding: 🟩 **= Top 10** 🟨 **= Top 20**

		logit	**elastic net**	**XG boost**	**random forest**
Country characteristics					
Island (dummy)			15 (-)	252	244
Currency union member (dummy)		14 (+)	5 (+)	65	188
Global variables					
US Treasury bill rate	lag		29 (+)	270	364
VIX, end of period	3-year change		23 (+)		539
World average Mineral rents / GDP	5-year change	25 (+)		148	337
World average public external debt / GDP	5-year change		24 (-)	181	384
World nominal GDP (million USD)	5-year change		3 (+)	46	248
World average public debt / GDP	lag	9 (-)			309
World average official development assistance / GDP	lag	16 (+)	27 (+)	38	322

©International Monetary Fund. Not for Redistribution

Annex A: Additional tables

Annex Table 1. Fiscal Crisis: Definitions and Data Sources

	Event	Criterion Minimum two years gap between crises	Thresholds			Literature	Sources	Notes
			AEs	EMEs	LIDCs			
(1)	**Credit Event**	**Default, restructuring, or rescheduling**				Detragiache and Spilimbergo (2001); Chakrabarti and Zeaiter (2014); Reinhart and Rogoff (2011b)	BoC-BoE Sovereign Default Database complemented with information from IMF desks; Cruces and Trebesch (2013); World Bank	The database mainly containsexternal defalts on sovereign debt denominated in foreign currency
		(i) of substantial size (in percent of GDP p.a.); AND		>0.5				
		(ii) defaulted nominal amount grows by a substantial amount (in percent p.a)		≥ 10				
(2)	**Exceptionally large official financing**	(i) **High-access IMF financial arrangement** with fiscal adjustment objective in place (in percent of quota); OR		≥ 100		Baldacci and others (2011)	IMF	
		(ii) **EU program**						
(3)	**Implicit domestic public default**	(i) **High inflation rate** (in pct. of growth of annual average CPI p.a.) OR	≥ 35	≥ 100		Baldacci and others (2011); Sturzenegger and Zettelmeyer (2006); and Fisher, Sahya and Vegh (2002)	IMF (World Economic Outlook)	
		(ii) **Steep increase in domestic arrears** (in first difference of the ratio of 'other account payables (OAP)' to GDP in percentage points)		≥ 1		Checherita-Westphal, Klem, and Viefers (2015); Reinhart and Rogoff (2011a)	Eurostat; OECD (data on other accounts payable)	
(4)	**Loss of market confidence**	(i) **High price of market access** (in basis points of sovereign spreads or CDS spreads) OR						
		(a)Level of spreads (bps)		≥ 1,000 bps		Sy (2004); Baldacci and others (2011)	Reuters Datastream; Bloomberg	
		(b) Annual change in spreads (bps)	≥ 300	≥ 650	na			
		(ii) **Loss of market access**		when market access is lost (after maintaining market access for a 1/4 of the sample time and 2 consecutive years before the loss year)		IMF (2015); Kose and others (2017)	Guscina, Sheheryar, and Papaioannou (2017); Gelos, Sahay, and Sandleris (2004);	

©International Monetary Fund. Not for Redistribution

Annex Table 2. Sample of Countries

Advanced Economies	Emerging Markets		Low Income Developing Countries	
Australia	Albania	Libya	Afghanistan	Myanmar
Austria	Algeria	Malaysia	Bangladesh	Nepal
Belgium	Angola	Maldives	Benin	Nicaragua
Canada	Antigua & Barbuda	Marshall Islands, Rep.	Bhutan	Niger
Cyprus	Argentina	Mauritius	Burkina Faso	Nigeria
Czech Republic	Armenia	Mexico	Burundi	Papua New Guinea
Denmark	Azerbaijan	Micronesia	Cambodia	Rwanda
Estonia	Bahamas, The	Mongolia	Cameroon	São Tomé and Príncipe
Finland	Bahrain	Montenegro	Central African Republic	Senegal
France	Barbados	Morocco	Chad	Sierra Leone
Germany	Belarus	Namibia	Comoros	Solomon Islands
Greece	Belize	Oman	Congo, Dem. Rep. of	Somalia
Iceland	Bolivia	Pakistan	Congo, Republic of	South Sudan
Ireland	Bosnia and Herzegovina	Palau	Côte d'Ivoire	Sudan
Israel	Botswana	Panama	Djibouti	Tajikistan
Italy	Brazil	Paraguay	Eritrea	Tanzania
Japan	Brunei Darussalam	Peru	Ethiopia	Timor Leste
Korea	Bulgaria	Philippines	Gambia, The	Togo
Latvia	Cape Verde	Poland	Ghana	Uganda
Lithuania	Chile	Qatar	Guinea	Uzbekistan
Luxembourg	China	Romania	Guinea-Bissau	Vietnam
Malta	Colombia	Russia	Haiti	Yemen, Republic of
Netherlands	Costa Rica	Samoa	Honduras	Zambia
New Zealand	Croatia	Saudi Arabia	Kenya	Zimbabwe
Norway	Dominica	Serbia	Kiribati	
Portugal	Dominican Republic	Seychelles	Kyrgyz Republic	
San Marino	Ecuador	South Africa	Lao PDR	
Singapore	Egypt	Sri Lanka	Lesotho	
Slovak Republic	El Salvador	St. Kitts and Nevis	Liberia	
Slovenia	Equatorial Guinea	St. Lucia	Madagascar	
Spain	Fiji	St. Vincent and the Grenadines	Malawi	
Sweden	FYR Macedonia	Suriname	Mali	
Switzerland	Gabon	Swaziland	Mauritania	
United Kingdom	Georgia	Syria	Moldova	
United States	Grenada	Thailand	Mozambique	
	Guatemala	Tonga		
	Guyana	Trinidad and Tobago		
	Hungary	Tunisia		
	India	Turkey		
	Indonesia	Turkmenistan		
	Iran	Tuvalu		
	Iraq	U.A.E.		
	Jamaica	Ukraine		
	Jordan	Uruguay		
	Kazakhstan	Vanuatu		
	Kosovo	Venezuela		
	Kuwait	Vietnam		
	Lebanon			

©International Monetary Fund. Not for Redistribution

Annex Table 3. Description and sources of data

Description	Source	Permutations used
Country category variables		
Dummy: Monetary union member	IMF WEO	t
Dummy: Island country	Wikipedia	t
Dummy: Landlocked country	CIA World Factbook	t
Dummy: Small state	Authors' calculations, IMF WEO, WB WDI	t
Dummy: Fragile state	WB (FY2017 list)	t
Dummy: Commodity exporter	IMF WEO	t
Contagion / Crisis history variables		
Contagion: number countries with fiscal crisis start	Medas et al (2018)	
All countries		t; current or past year
Advanced and Emerging Economies		t; current or past year
Emerging and Low Income Economies		t; current or past year
Advanced Economies		t; current or past year
Emerging Economies		t; current or past year
Low Income Economies		t; current or past year
Contagion: number of countries currently in fiscal crisis	Medas et al (2018)	
All countries		t
Advanced and Emerging Economies		t
Emerging and Low Income Economies		t
Advanced Economies		t
Emerging Economies		t
Low Income Economies		t

©International Monetary Fund. Not for Redistribution

Years passed since last fiscal crisis	Medas et al (2018)	t
Historical fiscal crisis frequency	Medas et al (2018)	t
Dummy: Banking crisis start	Laeven and Valencia (2018)	t; t-1
Contagion: number countries with banking crisis start	Laeven and Valencia (2018)	
All countries		t
Advanced and Emerging Economies		t
Emerging and Low Income Economies		t
Advanced Economies		t
Emerging Economies		t
Low Income Economies		t

External sector variables

Net official development assistance (% of GDP)	OECD	t; t-1; fd_t; fd_t-1; 3-year change; W
Current account balance (% of GDP)	IMF WEO	t; t-1; fd_t; fd_t-1; 3-year change
Export of goods and services (% of GDP)	IMF WEO	t; t-1; fd_t; fd_t-1; 3-year change; W
Import of goods and services (% of GDP)	IMF WEO	t; t-1; fd_t; fd_t-1; 3-year change; W
Personal remittances (% of GDP)	World Bank staff estimates based on IMF balance of payments data.	t; t-1; fd_t; fd_t-1; 3-year change; W
Current account without import	IMF WEO	t; t-1; fd_t; fd_t-1; 3-year change
Net foreign direct investment (% of GDP)	IMF WEO	t; t-1; fd_t; fd_t-1; 3-year change; W
Other investment, net (loans, deposits, insurance, pensions, trade credits, SDR, percent of GDP	IMF WEO	t; t-1; fd_t; fd_t-1; 3-year change; W
Portfolio investment, net	IMF WEO	t; t-1; fd_t; fd_t-1; 3-year change; W
Annual percentage change of average USD exchange rate	IMF WEO	t; t-1; fd_t; 3-year average
Annual percentage change of end-of-period USD exchange rate	IMF WEO	t; t-1; fd_t; 3-year average
Openness (10-year average of exports + imports of goods and services / GDP)	IMF WEO	t; W

©International Monetary Fund. Not for Redistribution

Percent change in real exchange rate, period average	IMF WEO	t; t-1; fd_t; 3-year average
Log of PPP-based real exchange rate	IMF WEO, WDI, author's calculation based on Rodrik (2008)	t; t-1; W
RER overvaluation	IMF WEO, WDI, author's calculation based on Rodrik (2008)	t; W
Percent change in total reserve assets, excluding gold (USD)	IFS	t; t-1; fd_t; pc3_t; W
Reserves, in months of imports	WDI	t; t-1; fd_t; fd_t-1; 3-year change; W
Percent change in terms of trade (of goods and services) Index	IMF WEO	t; t-1; fd_t; 3-year average
Trading partner real GDP growth	IMF GEE	t; t-1; fd_t; 3-year average
Trading partner import demand growth	IMF GEE	t; t-1; fd_t; 3-year average
External gross financing needs	author's calculations; IMF WEO; WDI	t; t-1; fd_t; fd_t-1; 3-year change
Value of oil export, percent of GDP	IMF WEO	t; t-1; fd_t; fd_t-1; W
Fiscal variables		
General government expenditures (% of GDP)	IMF WEO; Medas et al. (2018); Abbas et al. (2011)	t; t-1; fd_t; fd_t-1; 3-year change; W
General government primary expenditures (% of GDP)	IMF WEO; Medas et al. (2018); Abbas et al. (2011)	t; t-1; fd_t; fd_t-1; 3-year change; W
Overall balance (% of GDP)	IMF WEO; Medas et al. (2018); Abbas et al. (2011)	t; t-1; fd_t; fd_t-1; 3-year change; W
General government primary balance, percent of GDP	IMF WEO; Medas et al. (2018); Abbas et al. (2011)	t; t-1; fd_t; fd_t-1; 3-year change; W
General government revenues in percent of GDP	IMF WEO; Medas et al. (2018); Abbas et al. (2011)	t; t-1; fd_t; fd_t-1; 3-year change; W
Stock and flow adjustments to public debt	author's calculations	t; t-1; fd_t; fd_t-1; 10-year average; W
Global variables		
Percent change of crude oil price	IMF WEO (GAS Live)	t; t-1; pc3_t; fd_L
Percent change of non-fuel commodity price index	IMF Primary Commodity Prices; Medas et al (2018)	t; t-1; pc3_t; fd_L
Percent change of global food price index	IMF Primary Commodity Prices; Medas et al (2018)	t; t-1; pc3_t; fd_L
US T-Bill rate, Percent	IFS	t; t-1; fd_t; fd_L2
VIX, period average	Bloomberg	t; L2

©International Monetary Fund. Not for Redistribution

VIX, period end	Bloomberg	t; L2
Percent change of VIX, period average	Bloomberg	t; t-1; pc3_t; fd_L
Percent change of VIX, period end	Bloomberg	t; t-1; pc3_t; fd_L
US T-Note 5 year rate Percent, Period Average	Bloomberg	t; t-1; fd_t; fd_L2
US T-Note 10 year rate Percent, Period Average	Bloomberg	t; t-1; fd_t; fd_L2
US T-Note 5 year rate Percent, End of Period	Bloomberg	t; t-1; fd_t; fd_L2
US T-Note 10 year rate Percent, End of Period	Bloomberg	t; t-1; fd_t; fd_L2
World real GDP growth, in percent	IMF WEO	t; t-1; fd_t; 3-year average
Institutions / elections		
Revised Combined Polity Score (single regime score, runs from 1 (full democracy) to -1 (full autocracy))	Center for Systemic Peace	t; t-1; fd_t; fd_t-1; 5-year change; 10-year change; W
Checks and balances index	DPI	t; 5-year change; 10-year change; W
Bureaucracy Quality	PRS Group; WB WGI; author's calculations	t; 5-year change; 10-year change
Corruption	PRS Group; WB WGI; author's calculations	t; 5-year change; 10-year change
Years remaining in current chief executive's term	DPI	t
Legislative election held dummy variable	DPI	t
Executive election held dummy variable	DPI	t
Political Stability and Absence of Violence/Terrorism: Estimate	WB WGI	t; 5-year change; 10-year change
Regulatory Quality: Estimate	WB WGI	t; 5-year change; 10-year change
Demographics		
Population ages 15-64, total	WDI	t; 5-year change; 10-year change; W
Percent change of population ages 15-64, total	WDI	t; W
Urban population (% of total)	WDI	t; 5-year change; 10-year change; W
Age Dependency Ratio, % of working-age population	WDI	t; 5-year change; 10-year change; W
Population density (people per sq. km of land area)	WDI	t; W

©International Monetary Fund. Not for Redistribution

Log of population (relative to US)	IMF WEO	t; t-1; fd_t; 3-year change; 5-year change; 10-year change; W
Private debt variables		
(One-sided) credit gap	GDD	t; t-1; fd_t; fd_t-1; W
Total Debt, loans and securities, (% of GDP)	GDD	t; t-1; fd_t; fd_t-1; 5-year change; 10-year change; W
10-year average credit gap	GDD	t; W
Domestic credit to private sector by banks (% of GDP)	WDI	t; t-1; fd_t; fd_t-1; 5-year change; 10-year change; 3-year change; W
External debt stocks, private nonguaranteed, percent of GDP	WDI	t; t-1; fd_t; fd_t-1; 5-year change; 10-year change; W
WDI Broad Money, % of GDP	WDI	t; t-1; fd_t; fd_t-1; 5-year change; 10-year change; 3-year change; W
Public debt variables		
General government short-term external debt (% of GDP)	IMF WEO; IMF VEE; BIS	t; t-1; fd_t; fd_t-1; W
General government short-term external debt in percent of reserves	IMF WEO; IMF VEE; BIS; IFS	t; t-1; fd_t; fd_t-1; W
Public external debt (% of GDP)	IMF WEO; WDI; US bureau of economic analysis; Haver Analytics;Arslanalp and Tsuda (2012); Quarterly External Debt Statistics	t; t-1; fd_t; fd_t-1; 5-year change; 10-year change; W
Public debt (% of GDP)	GDD	t; t-1; fd_t; fd_t-1; 5-year change; 10-year change; W
Public debt in percent of general government revenue	IMF WEO	t; t-1; fd_t; fd_t-1; W
Public external debt, percent of exports	IMF WEO; WDI; US bureau of economic analysis; Haver Analytics;Arslanalp and Tsuda (2012); Quarterly External Debt Statistics	t; t-1; fd_t; fd_t-1; 3-year change; W
Short-term external debt (% of GDP)	IMF WEO; WB WDI	t; t-1; fd_t; fd_t-1; W
Short-term external debt to reserves	IMF WEO; WB WDI; IFS	t; t-1; fd_t; fd_t-1; W
Debt service		

©International Monetary Fund. Not for Redistribution

General government interest expenses (% of GDP)	IMF WEO; Medas et al. (2018); Abbas et al. (2011)	t; t-1; fd_t; fd_t-1; 5yr_t; 3-year change; W
Amortization of external public debt (% of GDP)	IMF WEO January 2019; WDI	t; t-1; fd_t; fd_t-1; 5-year change; 10-year change; 3-year change; W
Amortization of external public debt in percent of reserves	IMF WEO; WDI; IFS	t; t-1; fd_t; fd_t-1; 5-year change; 10-year change; 3-year change; W
Public debt service to revenue (Approximate), in percent	IMF WEO January 2019; WDI	t; t-1; fd_t; fd_t-1; 3-year change; W
Public debt service to export, in percent	IMF WEO January 2019; WDI	t; t-1; fd_t; fd_t-1; 3-year change; W
Debt service on total external debt, percent of GDP	WDI	t; t-1; fd_t; fd_t-1; 5-year change; 10-year change; W
Debt service on total external debt, percent of export	WDI; IMF WEO	t; t-1; fd_t; fd_t-1; 5-year change; 10-year change; W
Debt service on total external debt, percent of reserves	WDI; IFS	t; t-1; fd_t; fd_t-1; 5-year change; 10-year change; W

Real sector variables

Impact of natural disasters, (% of GDP)	CRED's EM-DAT	t; t-1
log(real GDP per capita) (in PPP dollars), relative to US	IMF WEO	t; t-1; fd_t; 3-year change; 5-year change; 10-year change; W
Percent change of real GDP per capita	IMF WEO	t; t-1; fd_t; 3-year average
Percent change of real GDP	IMF WEO	t; t-1; fd_t; 5-year average; deviation from 5-year average
Percent change of real GDP, deviation from 5-year average	IMF WEO; Medas et al (2018)	t; W
Percent change of nominal GDP	IMF WEO	t; t-1; fd_t; 3-year average
Percent change of period average consumer price index	IMF WEO	t; t-1; fd_t; pc3_t; W
Percent change of end-of-period consumer price index	IMF WEO	t; t-1; fd_t; pc3_t; W
interest-growth differential	author's calculation	t; t-1; fd_t; W
domestic savings, private (current US$)	WDI	t; t-1; fd_t; fd_t-1; 3-year change; mean_t; W

©International Monetary Fund. Not for Redistribution

Log of nominal GDP in USD, relative to US	IMF WEO	t; t-1; fd_t; 3-year change; 5-year change; 10-year change; W
Mineral rents (% of GDP)	WDI	t; t-1; 5-year change; 10-year change; mean_t; W
oil rent (% of GDP)	WDI	t; t-1; 5-year change; 10-year change; mean_t; W
Total natural resources rent (% of GDP)	WDI	t; t-1; 5-year change; 10-year change; mean_t; W
Agriculture, forestry, and fishing, value added (% of GDP)	WDI	t; W
Total debt (public + private)		
Total Debt, in % of GDP	GDD	t; t-1; fd_t; fd_t-1; 5-year change; 10-year change; W
Total external debt, % of GDP	WDI; Lane and Milesi-Ferretti (2007)	t; t-1; fd_t; fd_t-1; 5-year change; 10-year change; W
Volatility (10-year standard deviations)		
real GDP growth	IMF WEO; author's calculations	t
percentage change in terms of trade	IMF WEO; author's calculations	t
percentage change in period-average nominal exchange rate	IMF WEO; author's calculations	t
period-average CPI inflation	IMF WEO; author's calculations	t
percent change in end-of-period exchange rate	IMF WEO; author's calculations	t
stock and flow adjustment to public debt	IMF WEO; author's calculations	t

Note: t = current year value; t-1 = past year value; fd_t = first difference; fd_t = lagged of first difference; W = cross sectional weighted average for all the permutations listed. WEO = World Economic Outlook; DPI = Cruz et al. (2016); GDD = Mbaye et al. (2018); IFS = International Financial Statistics; WDI = World Development Indicators; WGI = Worldwide Governance Indicators.

©International Monetary Fund. Not for Redistribution

Annex B: Cross-validation and hyperparameter tuning

To select (tune) an algorithm's hyperparameter values, I search over a grid of candidate values and select the value that delivers the smallest average log-likelihood loss when cross-validated. My cross-validation algorithm is as follows:

1. From the years included in the training sample, select a year *t** that serves as the evaluation fold.
2. For the given hyperparameter value, estimate the model on the training sample after dropping observations between years *t*-1* and *t*+1.*
3. Using the model estimates from step 2, make predictions for year *t** and compute the log-likelihood loss.

I repeat these steps for each year in the training sample and then take the average log-likelihood loss. The procedure is repeated for each candidate hyperparameter value, so that each hyperparameter value is associated with a different loss function value.

I select the hyperparameter value that minimizes the loss function and use that value to re-estimate the model on the training sample. Note that the hyperparameter tuning and model estimation makes no use of the test sample (steps 1-3 above are done within the training sample). The resulting model is then used to make predictions for observations in the test sample.

Annex Table 4. Tuning parameter grids

Method	**Hyperparameter**	**Tuning grid**		**Selected value (in Section VII)**	**Method reference in *caret* package**
		# values	**range**		
Logit	-	-	-	-	glm, glmStepAIC
Classification tree	-	-	-	-	rpart2
Elastic net	α	11	[0, 1]	1	glmnet
	λ	400	[.003, 10]	.0096	
XG boost	number of trees	6	[10, 85]	70	xgbTree
	maximum depth	5	[1, 14]	14	
	η (learning speed)	4	[.05, .6]	.05	
	γ (minimum gain required for split)	5	[.005, 6]	1.5	
	minimum child weight	5	[1, 40]	11	
Random forest	m_{try}	15	dynamic	38	rf

©International Monetary Fund. Not for Redistribution