TECHNICAL NOTES AND MANUALS

How to Assess Country Risk: The Vulnerability Exercise Approach Using Machine Learning

Approved by Sanjaya Panth

Strategy, Policy, and Review Department

INTERNATIONAL MONETARY FUND

INTERNATIONAL MONETARY FUND

How to Assess Country Risk: The Vulnerability Exercise Approach Using Machine Learning

Approved by Sanjaya Panth

Prepared by an interdepartmental team led by Kevin Wiseman, and comprising Pranav Gupta and Andrew Tiffin (AFR), Andrew Swiston (APD), Juliana Gamboa Arbelaez, Klaus Hellwig, Andrew Hodge, Paulo Medas, Marialuz Moreno Badia, Roberto Perrelli, and Yuan Xiang (all FAD), Maksym Ivanyna (ICD), Sean Simpson (iLab), André Leitão Botelho and Wan Li (ITD), Silvia Iorgova (MCM), Suman Basu (RES), Chuqiao Bi, Jorge Chan-Lau, Sandile Hlatshwayo, Chengyu Huang, Lamya Kejji, Agustin Roitman, Weining Xin, Harry Zhao, and Yunhui Zhao (all SPR), Le Xu (SPR Summer Intern), and Daria Ulybina (STA); under the supervision of Daria Zakharova and Wojciech Maliszewski (SPR). Sharon Eccles (SPR) provided excellent administrative assistance.

ACKNOWLEDGMENTS

We are grateful for the support of this project by the iLab through the AI/ML Innovation Challenge and Catalyst series. We would also like to gratefully acknowledge the support from and the discussions with Michal Andrle, Alberto Behar, Angana Banerji, Fabian Bornhorst, Eugenio Cerutti, Marcos Chamon, Kirpal Chauhan, Qianying Chen, Mali Chivakul, Federico Diaz Kalan, Florence Dotsey, Aquiles Farias, Vikram Haksar, Yuko Hashimoto, Niko Alfred Hobdari, Plamen Iossifov, Tetsuya Konuki, Miguel Lanza, Emilia Magdalena Jurzyk, Maxym Kryshko, Nan Li, Sandra Lizarazo, Albert Touna Mama, Jimmy McHugh, Alexis Meyer Cirkel, Chifundo Moya, Nkunde Mwase, Rajesh Nilawar, Liam O'Sullivan, Marijn Otte, Mamoon Saeed, Jasmin Sin, Shannon Staley, Fabian Valencia, Tristan Walker, Hans Weisfeld, Jason Weiss, and Weijia Yao.

©International Monetary Fund. Not for Redistribution

©2021 International Monetary Fund

Cataloging-in-Publication Data
Joint Bank-Fund Library
Names: International Monetary Fund.
Title: How to Assess Country Risk: The Vulnerability Exercise Approach Using Machine Learning
Other titles: Technical Notes and Manuals (International Monetary Fund)
Series/volume #: TNM/21/03
Description: Washington, DC : International Monetary Fund | Periodic | Some issues also have thematic titles.
Classification: LCC HC10.W79
HC10.80

ISBN: 978-1-51357-421-9

DISCLAIMER: The views expressed in this paper are those of the author(s) and do not necessarily represent the views of the IMF, its Executive Board, or IMF management. The opinions contained in this document are the sole responsibility of the authors.

JEL Classification Numbers:	C14, C36, C38, C52, C53, E32, E37, F32, F47, G01, H63
Keywords:	Risk Assessment, Supervised Machine Learning, Prediction, Sudden Stop, Exchange Market Pressure, Fiscal Crisis, Debt, Financial Crisis, Economic Crisis, Economic Growth
Corresponding Authors' E-Mail Addresses:	Daria Zakharova (DZakharova@imf.org) Wojciech Maliszewski (WMaliszewski@imf.org) Kevin Wiseman (KWiseman@imf.org) Andrew Tiffin (ATiffin@imf.org) Roberto Perrelli (External Sector) (RPerrelli@imf.org) Klaus Hellwig (Fiscal Sector) (KHellwig@imf.org) Jorge Chan-Lau (Real Sector) (JChanLau@imf.org)

Publication orders may be placed online, by fax, or through the mail:
International Monetary Fund, Publication Services
P.O. Box 92780, Washington, DC 20090, U.S.A.
Tel. (202) 623-7430 Fax: (202) 623-7201
E-mail: publications@imf.org
www.imfbookstore.org
www.elibrary.imf.org

©International Monetary Fund. Not for Redistribution

CONTENTS

©International Monetary Fund. Not for Redistribution

Figures

Tables

©International Monetary Fund. Not for Redistribution

EXECUTIVE SUMMARY

The IMF's Vulnerability Exercise (VE) is a cross-country exercise that identifies country-specific near-term macroeconomic risks. As a key element of the Fund's broader risk architecture, the VE is a bottom-up, multi-sectoral approach to risk assessments for all IMF member countries. Assessments reflect the judgement of country teams informed by consistent, cross-country quantitative models as well as country-specific context.

The VE modeling toolkit is regularly updated in response to global economic developments and the latest modeling innovations. Earlier models evolved organically, assessing advanced economies, emerging markets, and low-income countries separately and looking at different types of risks. The new generation of models presented here closes gaps in risk and country coverages from previous models, while improving consistency and comparability of risk assessments across countries.

The new generation of VE models presented here leverages machine-learning (ML) algorithms. Macroeconomic risk assessment is a challenging task: crises are infrequent and almost always involve some elements of surprise. They tend to feature interactions between different parts of the economy and non-linear relationships that are not well measured in "normal times." ML tools can often better capture these relationships. They can also be more robust to outliers, noise, and the diversity of experiences across countries.

The performance of machine-learning-based models is evaluated against more conventional models in a horse-race format. The models assess the near-term risk of a crisis in the external, financial, fiscal, and real sectors. In each sector, rigorous performance metrics are used to compare new tools against traditional approaches. It turns out that random forest-based models, which are popular modern ML methods that average over many decision trees, outperform other options in most cases. In other cases, the signal extraction approach, a robust non-parametric method designed for macro-crisis detection, performs best. These winning models represent a new generation of models at the core of the VE.

The paper also presents direct, transparent methods for communicating model results. ML techniques can sometimes appear to be a black box due to their complexity and infrequent (though rapidly growing) use in economics. Communication tools, developed to inform country teams about the model assessments, help take the last step from predicting to informing.

©International Monetary Fund. Not for Redistribution

THIS PAGE INTENTIONALLY LEFT BLANK

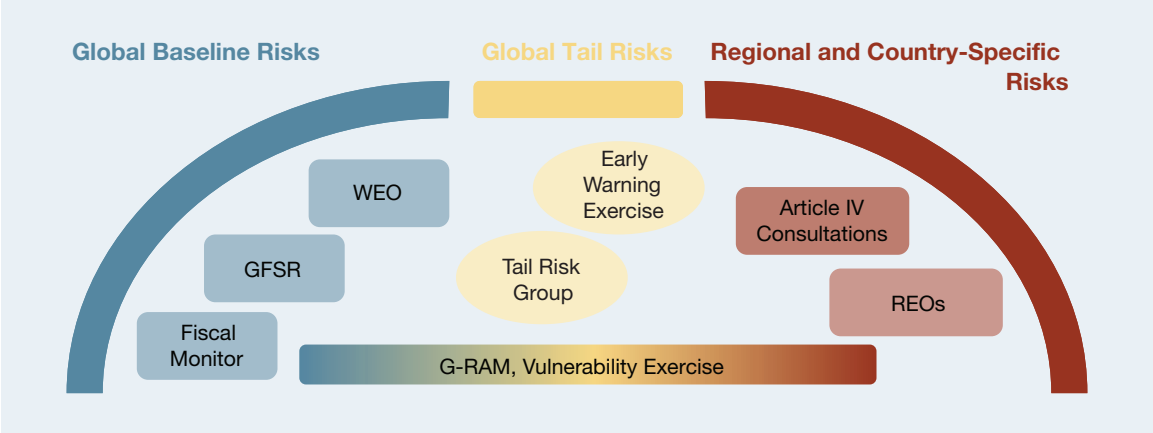
©International Monetary Fund. Not for Redistribution

INTRODUCTION

The Vulnerability Exercise (VE) is a multi-sectoral assessment of the factors that make a country vulnerable to a crisis. While crises are not easy to predict, the likelihood of any individual crisis is shaped by a country's underlying domestic vulnerabilities and global shocks. As part of the exercise, every country is assessed for vulnerabilities in its fiscal, external, financial, and non-financial corporate and household sectors, based on domestic vulnerabilities and global conditions.[1] Assessments reflect the judgement of country teams informed by consistent, cross-country quantitative models as well as country-specific context. The assessments are used to identify emerging risk areas and to inform policies that reduce vulnerabilities and mitigate risks. They also help guide resource allocation, for example in capacity development.

The VE is part of a comprehensive risk assessment framework at the Fund. This framework is multimodal, encompassing a variety of global, regional, and country-specific risks. It spans near-baseline and tail risks as well as methodologies ranging from simple indicators to large structural models, quantitative empirical work to narrative and even experiential approaches.[2] Within this architecture the VE helps to integrate these approaches. It is a bottom-up, country-based approach which yields regional and global insights. It combines cutting-edge quantitative empirical assessments with country-specific details and expert judgment.

FIGURE 1. Risk Architecture at the IMF



Source: Ahuja, Syed, and Wiseman (2017).[3]

The VE framework has evolved continually to better serve institutional needs. The exercise began with emerging markets (VEE) in 2001, subsequently updated in 2007, 2013, and 2017.[4] An exercise for advanced economies (VEA) was first developed in 2009 with a large suite of models, which have evolved organically—modules are regularly added, improved, and sometimes replaced. An assessment for low-income economies (VE-LIC) was added in 2011.[5] From an initial coverage of 43 countries under the VEE, the exercise expanded to cover 158 countries.

---

1 In addition, countries are also assessed for fragility to potential trigger events, susceptibility to spillovers and contagion, as well as policy implementation weaknesses related to political instability or political gridlock that could impede adequate response to an emerging crisis. See Ahuja and others (2017).

2 See Robinson (2014) and Ahuja and others (2017) for a more detailed description.

3 Unless otherwise specified, all sources are "IMF staff calculations".

4 See IMF (2001, 2007).

5 See Dabla-Norris and Gunduz (2014) and IMF (2011).

©International Monetary Fund. Not for Redistribution

The VE's organic growth revealed areas still in need of improvement. External risks were well covered for Emerging Markets (EMs) but less so for new frontier and low-income countries. Risk coverage in advanced economies was broad but pointillistic, making meaningful aggregation a challenge. Differences in risk assessment methods across income groups made results difficult to compare. Country coverage was still incomplete.

The new generation of models presented here addresses these shortcomings. The models assess risks in the external, fiscal, financial, and real sectors for all IMF member countries. They are also applied consistently across country groups—based on identical, quantitative crisis definitions and assessed within a single model when effective to improve evenhandedness. They also represent a step forward in risk assessment modeling, leveraging the recent advances in modeling macroeconomic risks with ML tools.

Machine-learning tools are well-suited to the challenges of macroeconomic risk assessments. Crises are rare and almost always involve novel features (otherwise they would have been anticipated). When they materialize, they often reveal economic relationships which are not regularly observed in normal times, making it challenging to reach firm conclusions based on any single country's history. Looking across countries can help shed additional light on crises but the degree of heterogeneity across even similar countries is daunting. Without sophisticated designs, classical estimation methods, like logistic regressions, may be poorly suited to capture the interactions, nonlinearities, and high degree of cross-country heterogeneity in crisis assessment. ML tools have become increasingly popular to address these issues in economics and other fields and have shown promise specifically in the case of crisis prediction. But as with other empirical methods that rely on past observations, ML models cannot predict crisis events in which the shocks or transmission channels have not been seen before.

This paper evaluates machine-learning-based and conventional risk assessment models in a horse race format. In each sector, models are rigorously evaluated based on their out-of-sample performance using modern evaluation methods. The best performing models are presented below in a separate section for each type of crisis, which summarizes their forecasting accuracy and key features of the estimation including key explanatory variables and interactions. The models presented here constitute a new generation of risk assessment models at the core of the VE.[6] They are complemented by a diverse suite of models that provide additional perspectives using alternative methods and risk definitions.[7]

Sectoral risk assessments complement related public IMF risk assessment tools. There are a number of related risk assessment tools at the Fund that are published in staff reports or multilateral surveillance products like the Global Financial Stability Report (GFSR) or External Sector Report (ESR). These tools include debt sustainability assessments, external balance assessments, and the Growth-at-Risk models, some of which consider longer time horizons than the VE and offer a deeper dive into country-specific features, though sometimes at the expense of narrower country coverage. The models in the VE offer a snapshot across all countries and sectors at a single point in time with an emphasis on crisis risk, cross-country consistency, and model performance.

[6] See Weisfeld and others (2020) for a sectoral crisis-based approach for low-income countries using machine learning tools, a direct predecessor to the assessments presented here.

[7] See Basu and others (2017), and the online technical appendices of Ahuja and others (2017) especially for advanced economies.

©International Monetary Fund. Not for Redistribution

MACHINE LEARNING AND CRISIS FORECASTING[8]

Early Warning Systems (EWSs) have long been a common feature in country surveillance. Both private- and public-sector institutions have repeatedly emphasized the development of models to anticipate crises, especially in the wake of the emerging-market turbulence of the 1990s. The traditional early-warning literature has typically relied on two approaches—discrete-choice (logit or probit) regressions (see, for example, Eichengreen and Rose, 1998) or the signal extraction approach pioneered for the Fund by Kaminsky and Reinhart (1999)—the latter has served as a core element of the VE for more than a decade. These approaches have a number of advantages, including most notably their ease of interpretation and widespread acceptance. But they have often suffered from two key problems: i) a frequently large gap between in-sample fit and out-of-sample predictive performance (in discrete-choice regressions); and ii) difficulty in coping with a large number of predictors (in discrete-choice regressions) and their potential interactions (in signal extraction approach). Modern ML can help with both.

Definitions of machine learning vary, but a distinguishing feature of machine-learning applications is their focus on prediction. ML has amassed a significant and rapidly-expanding amount of practical experience in the design of experiments that focus on finding generalizable predictive patterns applicable to new data observations rather than, as in the case of statistics, drawing population inferences from a sample. As a result, models in ML applications are often evaluated based on how well they generalize to out-of-sample observations. Of course, all models can be used and evaluated this way (including standard OLS) and these techniques often feature implicitly in the applied predictive modeling familiar to most econometricians—although not always consistently applied in practice. Non-parametric models in particular often lend themselves to an increased focus on generalization, as the model structure is not specified *a priori* but is instead determined from data. In this regard, the (non-parametric) signal extraction approach is a close precursor to some newer decision tree-based ML models. Indeed, Berg and others (2005) find that the signal extraction approach generalizes well, i.e., offers superior out-of-sample performance, compared to standard logit regressions. Furthermore, more complex non-parametric models often have parameters ("hyperparameters") that govern the model structure, which is helpful in the context of focusing on generalization, as these parameters could be chosen in a way to maximize the model's ability to predict out-of-sample data (Annex IV).

Machine-learning techniques tend to be better at handling complex interactions among many predictors. Accurate crisis prediction is a very challenging problem—the rare onset of a crisis is likely shaped by the (nonlinear) interaction of a range of economic drivers, and there is no theoretical consensus on how these drivers come together to trigger a crisis in any specific instance. This makes it difficult to specify a suitable model *a priori*. Instead, a useful model should be able to efficiently sift through a broad range of potential independent variables, identifying the relationships, thresholds, and interactions that are most informative when making a prediction. Complex ML algorithms, e.g., random forests (Annex II), are designed to explore the dataset more completely, finding key predictive relationships and interactions. For some problems, best predictive models are simple, boiling down to the signaling approach if individual variables considered in isolation can sufficiently capture complex interactions. For other problems, simple models are insufficient.

Machine learning is becoming better established in the field of crisis prediction. As discussed above, ML is hardly new. Many popular ML techniques were clearly anticipated in the non-parametric statistical literature of the 1960s. But the accelerating availability of brute-force computing power has made these

[8] Section prepared by Andrew Tiffin and Kevin Wiseman.

©International Monetary Fund. Not for Redistribution

techniques more and more accessible, including for crisis prediction. The 2000s, for example, saw a marked pick-up in the exploration of simple non-linear decision trees for crisis forecasting. Led by Ghosh and Ghosh (2003), this wave of models examined a variety of crises and types of explanatory variables, but many of the models were small and also subject to the out-of-sample performance issues noted above.[9]

Machine-learning tools are gaining popularity in the wake of the Global Financial Crisis, especially for advanced-economy banking crises. Key contributions include Savona and Vezzoli (2015), Alessi and Detken (2018), Joy and others (2017). Manasse and others (2016) is another important example, possibly the first random forest-like model in crisis forecasting, which examines EM banking crises. Blumstein and others (2020) offers the first incorporation of Shapley values—a method of determining additive contributions to a risk assessment from a model's input series—into their analysis (see Annex VI). Shapley values address one of the major shortcomings of complex non-linear models (i.e., they can be difficult to explain), a major task of modern ML which will be important for wider use of these models in risk assessment and elsewhere. As the field has evolved, papers have begun to perform more extensive horse races, such as a looser comparison in Alessi and others (2015) and a stricter one in Holopainen and Sarlin (2017). The latter also provides a good overview of a variety of modeling methods and collects references.

No single algorithm dominates in the crisis prediction problem, as implied by the "no free lunch theorem." The theorem's implications are reflected in the findings of other authors and those reported here. For instance, Beutel and others (2018) find that the logit regression performs the best in predicting systemic banking crises, while the signal extraction approach does very well in predicting sudden stops. And the analyses presented in this note show that in the case of financial, fiscal, and real sector crises, random forest models are the preferred modeling approach. Early applications of neural network models, one of the preferred approaches for high dimensional data, have yet to prove advantageous (Oh and others, 2006; Nag and Mitra, 2002). More generally, data availability restricts the use of Big Data models in macroeconomic applications.

[9] See summary in Chamon and others (2007). For a summary of models leading up to the global financial crisis and their performance through those events, see Frankel and Saravelos (2012).

©International Monetary Fund. Not for Redistribution

MODELS AND ESTIMATION STRATEGY[10]

This section discusses the broad choice of crisis definitions, explanatory variables, modelling strategy, and overall results. Detailed descriptions of individual models are provided in subsequent sections.

Crisis definitions. Definitions are based on quantitative criteria drawn from the literature rather than determined by IMF country teams.[11] While country teams' judgment is central for the final assessments (especially when idiosyncratic factors play a significant role in shaping crisis developments), the quantitative approach in defining the crises ensures consistency across countries—critical for increased coverage and evenhandedness—and comparability with the literature.

Explanatory variable selection. Variable selection is based on an extensive literature review to identify potential crisis contributors. The improved robustness of some of the new models allows for a wider coverage of variables than in past studies, including global and political factors. Data coverage, of course, is a significant challenge for models with a large number of countries. To maximize coverage, all models are estimated on annual data.[12]

Model selection. Models are selected based on "horse race" exercises, comparing a broad set of models over the same sample period, based on the same crisis definitions and explanatory variables:

- **Set of models.** A range of models are considered, including the signal extraction method (see Annex I), logit models (classic, Bayesian, and penalized versions such as Elastic Net, Ridge, and Lasso), and more innovative ML algorithms, including a number of tree-based models such as random forest (RF) and its implementation variations (balanced random forest, RUSBoost, ADABoost, and gradient boosting),[13] as well as support vector machines (SVM).[14] Tree-based models can be seen as an extension of the signal extraction method and are discussed in more detail in Annexes II and III. They have shown promise in recent work in the crisis forecasting literature.
- **Performance assessment.** Models are evaluated based on out-of-sample predictive performance—an improvement upon many current VE models that tend to be evaluated only on their in-sample goodness of fit. The authors of the models presented here primarily evaluate them through classic backtesting, simulating a "real time" forecasting exercise beginning in the early 2000's and rolling forward both the estimation and the testing periods. This method is intuitive but only tests the model against recent crises—practically just the Global Financial Crisis (GFC) in some sectors. For this reason, backtesting is complemented with "cross-fold validation" in these cases, sequentially

[10] Section prepared by Andrew Tiffin and Kevin Wiseman.

[11] See Laeven and Valencia (2018) for banking sector crises, Basu and others (2017) for sudden stops, and Medas and others (2018) for fiscal crises.

[12] Though models are estimated on annual data, higher-frequency variables are included and can be updated more often.

[13] Boosting is a general technique that improves the accuracy of an ensemble of learning models (learners) by training subsequent learners on the errors made by previous learners. Commonly used boosting algorithms are the adaptive boosting (ADABoost) and the random under-sampling boosting (RUSBoost) (see Annex III).

[14] Support vector machines (SVMs) are a popular ML classification technique for small and medium-sized datasets. After performing a non-linear transformation of the features included in the analysis, they estimate a linear discriminant function. Neural Networks, perhaps the most famous machine learning technique, were not successful in limited experiments, as their effectiveness tends to be most pronounced for significantly larger datasets.

©International Monetary Fund. Not for Redistribution

estimating the model on all but one subset ("fold") of the sample and evaluating the model on the hold-out fold (Annex IV).[15] All models are evaluated by Area Under the Curve (AUC, a standard measure of predictive performance in machine learning models discussed in Annex V) but traditional sum-of-errors and probability-oriented approaches are also used when appropriate.

- **Estimation sample.** For each sector, the possibility of bringing all countries under a single model is assessed, provided this choice yields similar or better model performance.

Model results. Results indicate that ML models are particularly well-suited to crisis forecasting, with the more complex tree-based methods adding additional value in some cases and the robustness of signal extraction winning out in others. Pooling country income groups often improves predictive performance, and new models have yielded fresh insights regarding underlying vulnerabilities and their interaction with other factors:[16]

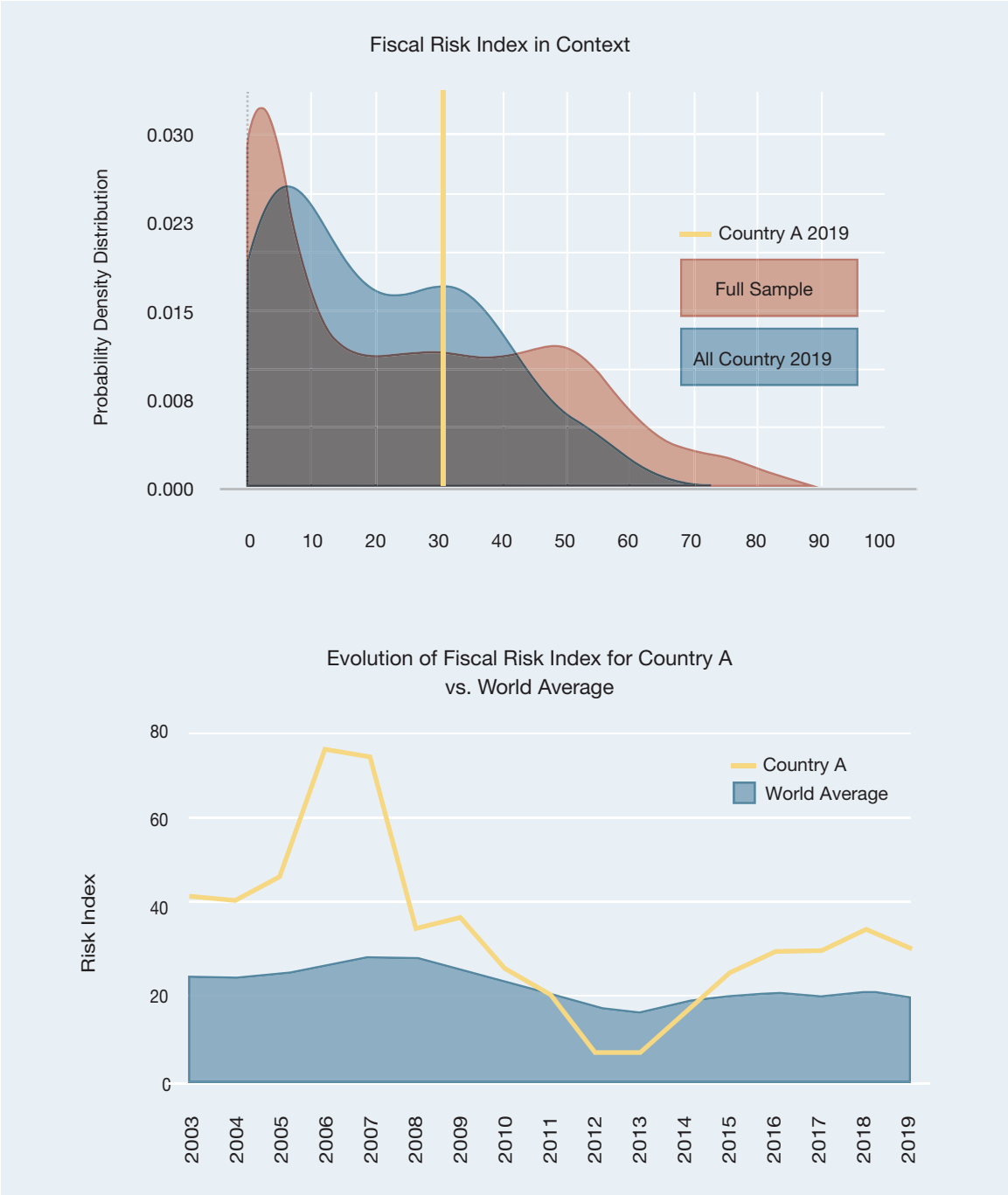
- **Machine-learning models regularly outperform classical econometric methods.** Tree-based models are the most successful in out-of-sample prediction for the financial and fiscal sectors. For the external sector, the signal extraction approach is most effective for sudden stops and Exchange Market Pressure (EMP) events in advanced economies (AEs), while a RF model is better for EMP events in emerging markets (EMs) and low-income countries (LICs).
- **Pooling all countries improves the performance of fiscal and financial models.** Pooling all country groups typically improves forecasting for crises in LICs and AEs while providing similar out-of-sample performance for EMs. Intuitively, the experience of lower-income EMs in the 1980's and 1990's proves informative for assessing vulnerabilities for frontier LIC markets. For AEs, crises are quite rare, but the experience of EMs nonetheless helps inform the models about the types of risks that are most important. For assessing external sector risks, modeling by conventional income groups is preferred (identical crisis definitions and sets of explanatory variables still significantly improve comparability for this group).
- **The results highlight the importance of variable interactions in assessing crisis risks.** Interactions between global and domestic risk factors are the most notable examples. For financial crises, vulnerabilities from external debt, debt growth, and the external exposure of the banking sector all depend critically on global interest rates. In the external sector, vulnerabilities from foreign liability growth are amplified by generalized rapid domestic expansion in credit.

[15] To ensure that the time series properties of the model are respected, folds are chosen to be "time-blocks" of all observations in groups of consecutive years. The use of block bootstrapping over time periods is well established in the econometric literature, See Politis and Romano (1994) and Lahiri (2003).

[16] These results are consistent with recent work on crisis forecasting in LICs, as in Weisfeld and others (2020).

©International Monetary Fund. Not for Redistribution

FIGURE 2. Fiscal Crisis Risk – Country A, 2019



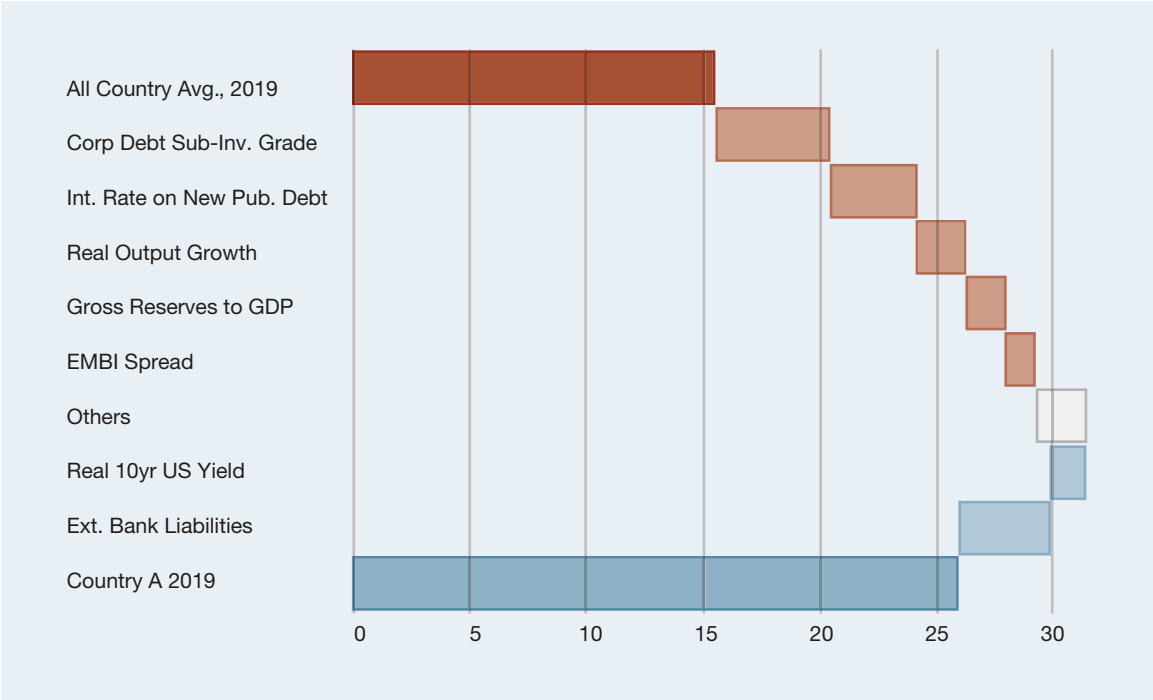
©International Monetary Fund. Not for Redistribution

COMMUNICATING RESULTS

Communication tools are an essential part of any machine-learning model. Since assessments of macroeconomic risks are designed to inform policymakers and guide policy decisions, it is necessary that the results are interpretable. The rich patterns in the data ML models capture are not necessarily easy to interpret, especially since economists and policymakers are not yet familiar with the models. Fortunately, the rapidly growing field of interpretable machine learning provides a number of interpretation and visualization tools to facilitate translating model insights into policy guidance.[17]

Machine-learning results need to be set in context. An explanation for a model result is necessarily contrastive—it explains why risks are assessed as lower or higher than, for example, some average risk of crisis or the risk of crisis last year. Even when a model result has a direct interpretation as a probability of a crisis, it may still not be clear whether, for example,
5 percent is high. It depends on the background frequency of crises and the crisis definition used. For this reason, topline charts presented here set risk assessments in the context of the distribution of assessments for other countries, as well as the evolution of the risk assessment over time.

FIGURE 3. Contribution to Risk Index



The contribution of each explanatory variable can be captured by Shapley Values. ML models capture greater richness and complexity in their assessments. As a result, an explanation for these assessments can be complex, but it is still possible to provide an easy-to-understand overview. A first-order explanation decomposes a risk assessment into additive contributions of variables. Shapley values, an ML concept borrowed from game theory, are an increasingly popular, consistent, and robust method for assigning contributions (see Annex VI). These contributions can be collected into waterfall charts

[17] See, for example, Hall and Gill (2019), or Molnar (2019) for a continuously improving online book.

©International Monetary Fund. Not for Redistribution

(Figure 3) to digestibly summarize these effects. In the Figure, the difference between a particular country's risk assessment and the average across countries for the same time period is presented in terms of variables that increase (in red) or decrease (in blue) the risk: the country's final risk index (on the bottom) equals the sum of the individual contributions of these variables.

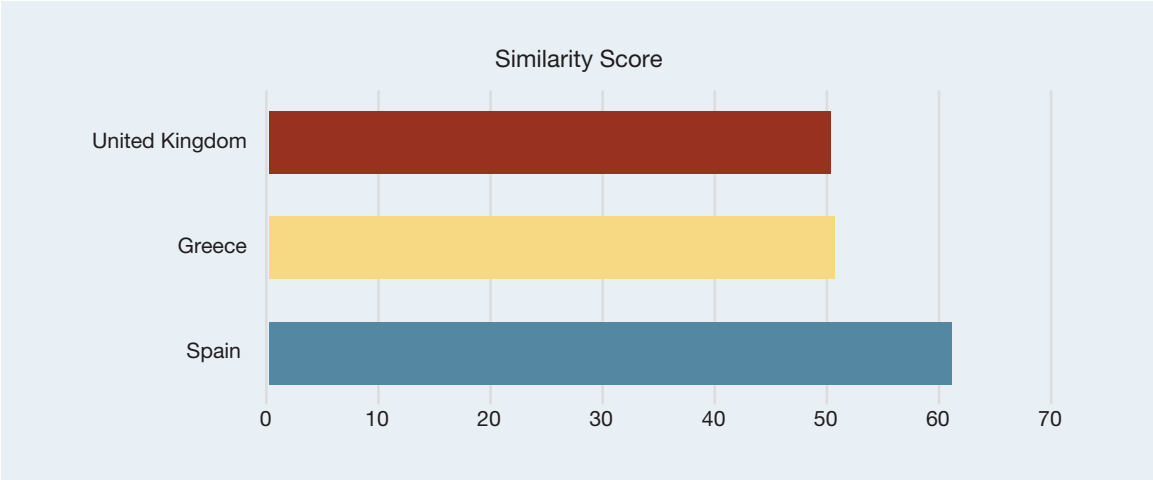
FIGURE 4. Distribution of Variable with Largest Risk Contributions

Model-based attribution of risk to data can be complemented with a look directly at the data. Figure 4 presents box plots for the distributions of key risk indicators across crisis and non-crisis cases. For each variable, the left plot presents the empirical distribution of the crisis cases (the short horizontal lines are the 5th, 25th, 75th, and 95th percentiles, if possible[18]), and the right plot presents that of the non-crisis cases. The values for the selected country are indicated by the yellow long horizontal lines. These charts allow country teams to quickly verify the data and absorb the historical association of that data with crisis incidence. They provide assurance that the attributions are well grounded. Box plots directly reflect the signal extraction methodology and most directly capture the intuition for this model but may not line up with the results of more complex models when interaction effects are important.

[18] More specifically, the bottom and top short horizontal lines are constrained for better visualization: if the 5th percentile is too small, the bottom line plots the 25th percentile minus 150 percent of the interquartile range (which is defined as the 75th percentile minus the 25th percentile); if the 95th percentile is too large, the top line plots the 75th percentile plus 150 percent of the interquartile range. In all cases, the median is the line between the dark blue and light blue boxes.

©International Monetary Fund. Not for Redistribution

FIGURE 5. Countries with a Similar Financial Sector Risk Profile with Ireland in 2005



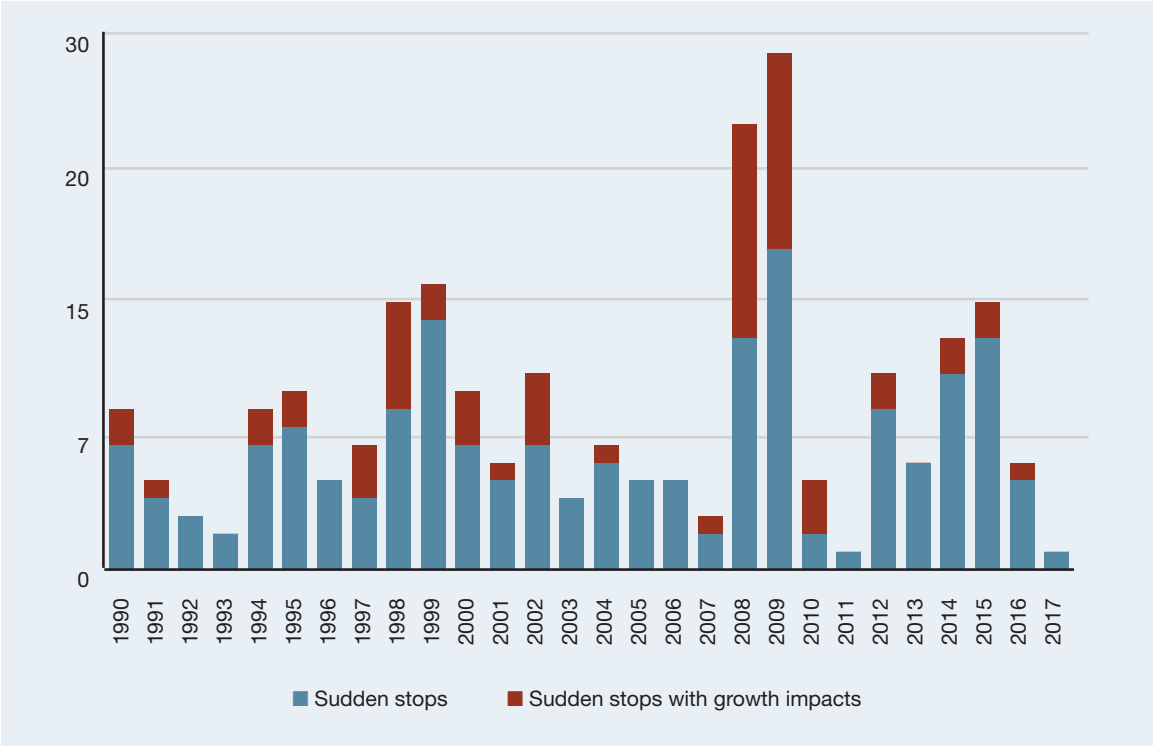
Random forest models can be used to identify countries that face a similar constellation of risks. The models in this note are estimated on thousands of examples, too many to sort through individually to find useful guides. Fortunately, "similarity scores" identify current and past country cases that are the most similar to the country in question in view of the model, based on the number of instances when the RF model places two countries in the same final "node" of the tree (see Annex II). Similarity scores are a useful tool for analysts and policymakers, as they help identify country cases with similar risk profiles that might offer insights on emerging risks and policies to mitigate them. For instance, as indicated in Figure 5, Ireland's financial sector risk profile in 2005 looked similar to that in the United Kingdom, Greece, and Spain—countries that would also experience financial sector stress during the Global Financial Crisis.

©International Monetary Fund. Not for Redistribution

EXTERNAL SECTOR MODEL[19]

Crisis definitions. The external sector crisis model estimates the risk of external sector crisis events when there is a sudden switch in investor preferences from domestic to foreign assets. How such a switch in preferences translates into domestic macroeconomic outcomes depends on the structure of the economy. In this exercise, two different definitions of external crises are explored, sudden stops and exchange market pressure events.

FIGURE 6. Frequency of Sudden Stops



Sudden stops in capital flows. These are the most impactful external crisis events, typically occurring in EMs with capital accounts open enough for inflows to accumulate, but with domestic financial markets that are not sufficiently developed for sudden outflows to be easily insured against. Sudden stops are defined as occurring when net private capital inflows as a percentage of GDP are at least 2 percentage points lower than in the previous year and two years before, as well as when the country gets approved to tap large IMF financial support.[20] Sudden stops are often followed by severe real economic consequences. Of particular interest are those with sizeable growth impacts, i.e., large growth declines resulting from binding financial constraints throughout the economy caused by sudden stops in private capital flows.[21]

[19] Model developed by Suman S. Basu, Roberto A. Perrelli, and Weining Xin. For a full treatment, see Basu and others (Forthcoming).

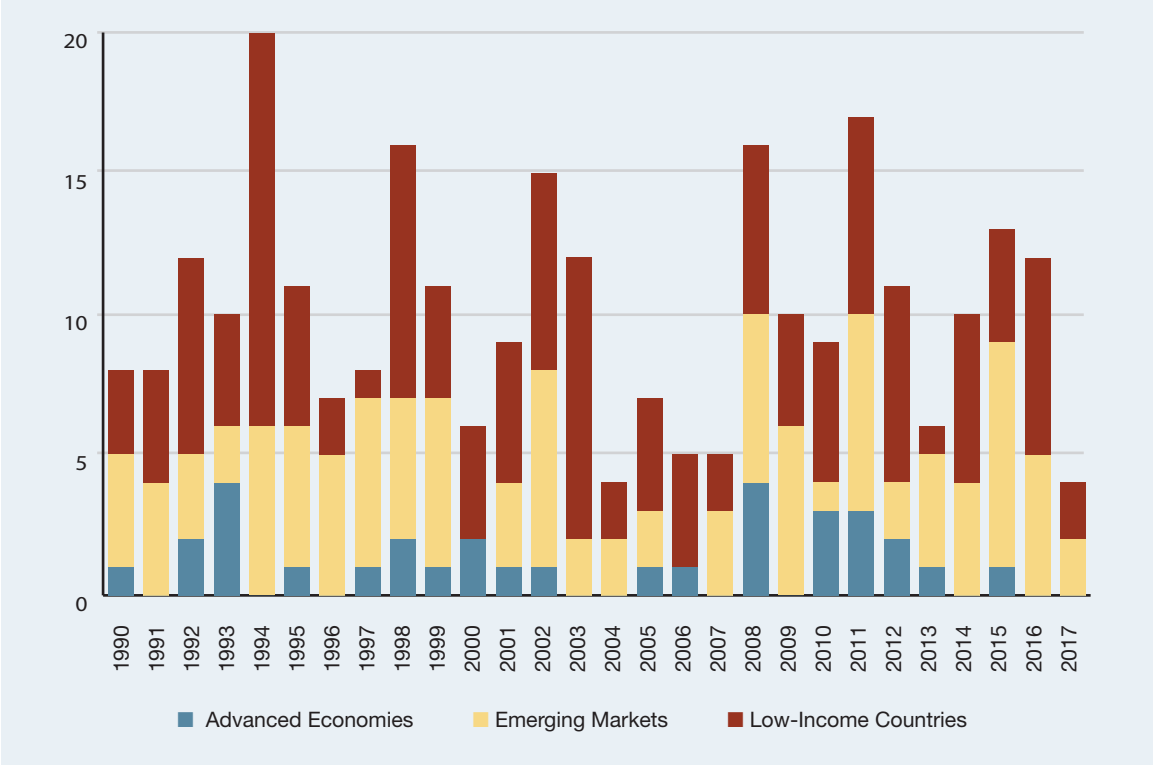
[20] Hereafter defined as IMF arrangements with agreed amount at least five times as large as the respective country's quota at the IMF. This criterion attempts to capture counterfactual situations in which sudden stops were prevented by large IMF financial support. One caveat with this approach is that some programs aim to address structural imbalances over an extended period rather than reflect a temporary shift in investor preferences from domestic to foreign assets. In addition, lags between program requests and the actual programs may weaken the identification of the crisis.

[21] In line with the theoretical literature on external crises (e.g., Mendoza, 2002).

©International Monetary Fund. Not for Redistribution

Large growth declines are defined as occurring when the changes in growth relative to the previous five-year average growth rate lie in the lower 10th-percentile of the whole panel.[22] Hereafter these episodes are called sudden stops with growth impact (SSGI). The data sample covers 53 EMs and spans the period between 1990 and 2017. There are 183 sudden stops accounting for 12.3 percent of the sample, and 61 SSGIs, accounting for 4.1 percent of the sample. The definition of sudden stops with growth impacts gives rise to several clusters of crises in the mid-1990s, late 1990s, early 2000s, and late 2000s.

FIGURE 7. Frequency of EMP Events



Exchange market pressure (EMP) events capture episodes of sudden exchange rate depreciation or reserves depletion for all economies. They may occur owing to a sudden switch in investor preferences from domestic to foreign assets, even if the realized capital outflow is not large.[23] In the spirit of early papers in the empirical literature on currency crises, an EMP index is constructed combining degrees of exchange rate depreciation and international reserves loss.[24] The index is defined as a weighted average of the annual percentage depreciation in the nominal exchange rate and the annual decline in reserves as a percentage of the previous year's GDP.[25] EMP events are defined as occurring when the index lies in the lower 15th-percentile of the whole panel, as well as when the country gets approved for large

[22] Adapting the definition of Basu and others (2017).

[23] Such events are especially relevant for economies which are financially closed (so that exchange rate misalignment may occur, but the potential outflows are limited by the small capital inflows in prior periods), which have active crisis management (so that reserves are used to absorb a sudden stop in capital outflows), and which have few binding financial constraints (so that exchange rate misalignment may be quickly corrected without generating severe negative consequences for domestic output and credit).

[24] For example, Eichengreen and others (1995), Kaminsky and Reinhart (1999), and Berg and others (2005).

[25] The weights are chosen so that the variance of the two components in the pooled sample is the same.

©International Monetary Fund. Not for Redistribution

IMF support.[26] The EMP database covers 192 countries during the 1990-2017 period with an average incidence of 6.3 percent, with significantly lower frequency among AEs. EMP events in EMs and LICs are more heterogeneous events than sudden stops, spanning events in financially-closed economies (e.g., India in 1991) and the absorption of large capital outflows using reserves (e.g., China in 2016). In AEs, there are two clusters of EMP events: in the early 1990s; and during the GFC, followed by the European debt crisis.

Explanatory variable selection. As in all sectoral models, the external sector crisis model evaluates the contribution of vulnerabilities from any sector to an external crisis. Variable selection is based on whether the variable is associated with clear economic channels and interpretable mechanisms according to different generations of the academic literature on external crises.[27] The final model is based on 79 variables listed below, including variables selected according to three generations of crisis models, as well as those capturing current account shocks, political shocks and contagion effects. There is substantial overlap between this set and the variables used in the old VEE model, providing some continuity in the assessment of vulnerabilities contributing to external risks. For tree-based models, missing data are imputed using the sample median to preserve the information value of the available data while minimizing the role of imputed series on the risk assessment (see Annex VII for details).

Model selection. A number of econometric and ML models are evaluated for their ability to anticipate a crisis one to two years in advance:

- **Testing procedure.** Each model is evaluated using classic backtesting starting in 2007 through the end of the sample. As this back-testing exercise risks evaluating the model on extremely rare crisis events out-of-sample, it is complemented by a single cutoff backtesting, in which model is estimated based on data through 2007 and applied to the rest of the sample. The sum of errors—defined as the sum of fractions of missed crises and false alarms—is the main evaluation metric, while the AUC is also calculated for reference.

[26] As previously described, this is to capture counterfactual situations in which sharp exchange rate depreciations or large declines in reserves are prevented by large IMF support.

[27] First-generation models of Krugman (1979) and Flood and Garber (1984), second-generation models pioneered by Obstfeld (1996), and third-generation models (Dornbusch and others, 1995; Mendoza, 2002).

©International Monetary Fund. Not for Redistribution

TABLE 1. External Crisis: Explanatory Variables

FIRST GENERATION
- Fiscal balance (% of GDP)
- 5-year change in M2/GDP
- Reserves/M2
- Reserves/GDP
- Dummies for hard peg and float
- Dummy for parallel market

SECOND GENERATION
- Change in unemployment rate
- Real GDP growth

THIRD GENERATION: FLOWS AND MISMATCH
- Share of non-investment grade debt
- Current account balance/GDP
- Amortization
- FX share of public debt
- Debt service/exports
- FX share of household and non- financial corporate credit

POLITICAL SHOCKS
- Political violence
- Successful coup

THIRD GENERATION: DEBT SHOCKS
- External debt/GDP and exports
- Private external debt/GDP
- Bank external debt/GDP
- Private credit/GDP
- Non-bank private external debt/GDP
- Total and external Public debt/GDP
- Cross-border bank-to-bank liabilities/GDP
- Household liabilities/GDP
- Foreign liabilities/ Domestic credit

THIRD GENERATION: BUFFERS
- EMBI spread (level and growth)
- Corporate sector returns
- Default probability
- Interest coverage ratio
- Price-Earnings ratio
- Bank returns
- Share of non-performing loans
- Banks' capital-asset ratio
- Loan-to-deposit ratio
- Primary gap/GDP
- Inflation

THIRD GENERATION: BURSTING BUBBLES
- REER acceleration
- Real house price acceleration
- Real stock price acceleration
- Changes in all debt/ GDP in debt shocks

THIRD GENERATION: GLOBAL SHOCKS
- FFR (level and growth)
- VIX
- US NEER change
- US yield spread
- TED spread

LAW OF ONE PRICE
- 5-year cumulative inflation

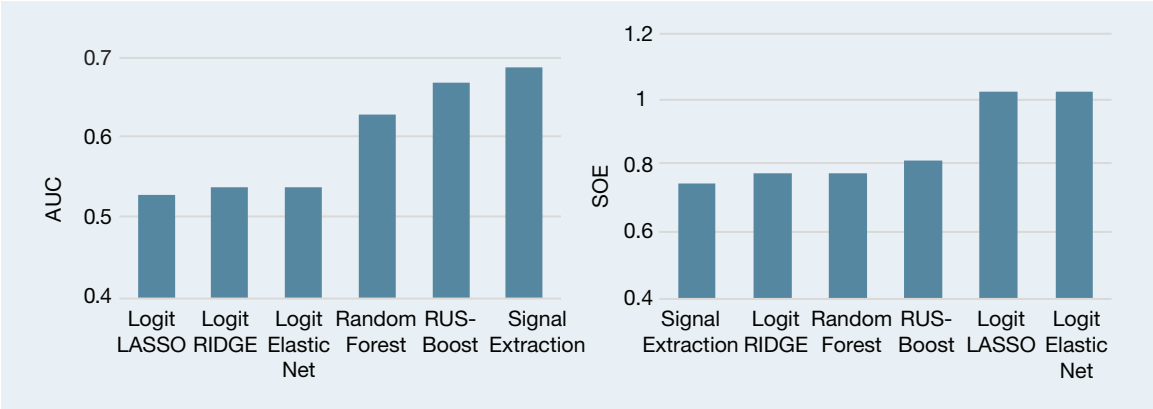
THIRD GENERATION: MEDIUM-TERM (5-YR) BUILDING BUBBLES
- Private sector credit growth
- Housing price growth
- Stock price growth
- REER growth
- External debt/ GDP growth
- Cross-border bank-to-bank liabilities to GDP growth
- Contribution of finance to GDP
- Contribution of construction to GDP

CURRENT ACCOUNT SHOCKS
- Real growth in exports
- % change in ToT
- % change in non-fuel commodity TOT
- Absolute oil balance/GDP
- % change in oil price

CONTAGION
- Change in export partner growth relative to 5-year trend
- Bank-to-bank Liabilities to AEs with financial crisis/GDP
- Frequency of banking crises in AEs
- Similarity to last year's crises

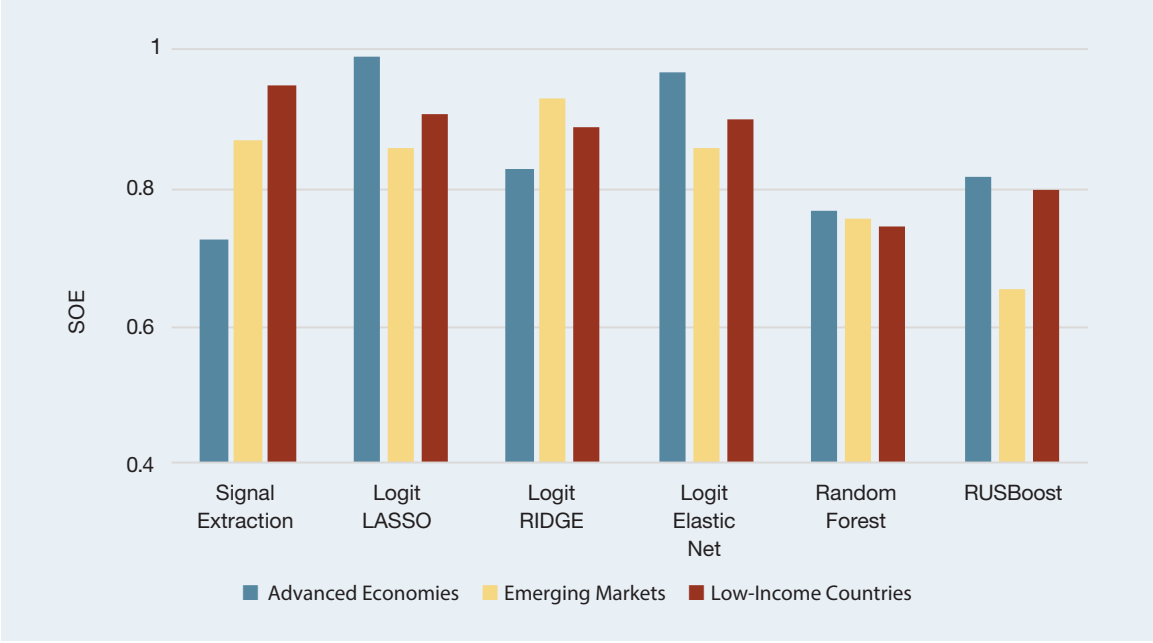
©International Monetary Fund. Not for Redistribution

FIGURE 8. External Sector Model Performance: SSGI



- **The signal extraction model performs the best in predicting SSGI.** The above graph presents the averaged performance across the backtesting for the signal extraction model and several other popular classification techniques, including three types of penalized logit (RIDGE, LASSO and Elastic Net) and two types of tree-based ensemble the models (RF and RUSBoost). The signal extraction method is superior in terms of both out-of-sample AUC and sum of errors.

FIGURE 9. External Sector Model Performance: EMP Events



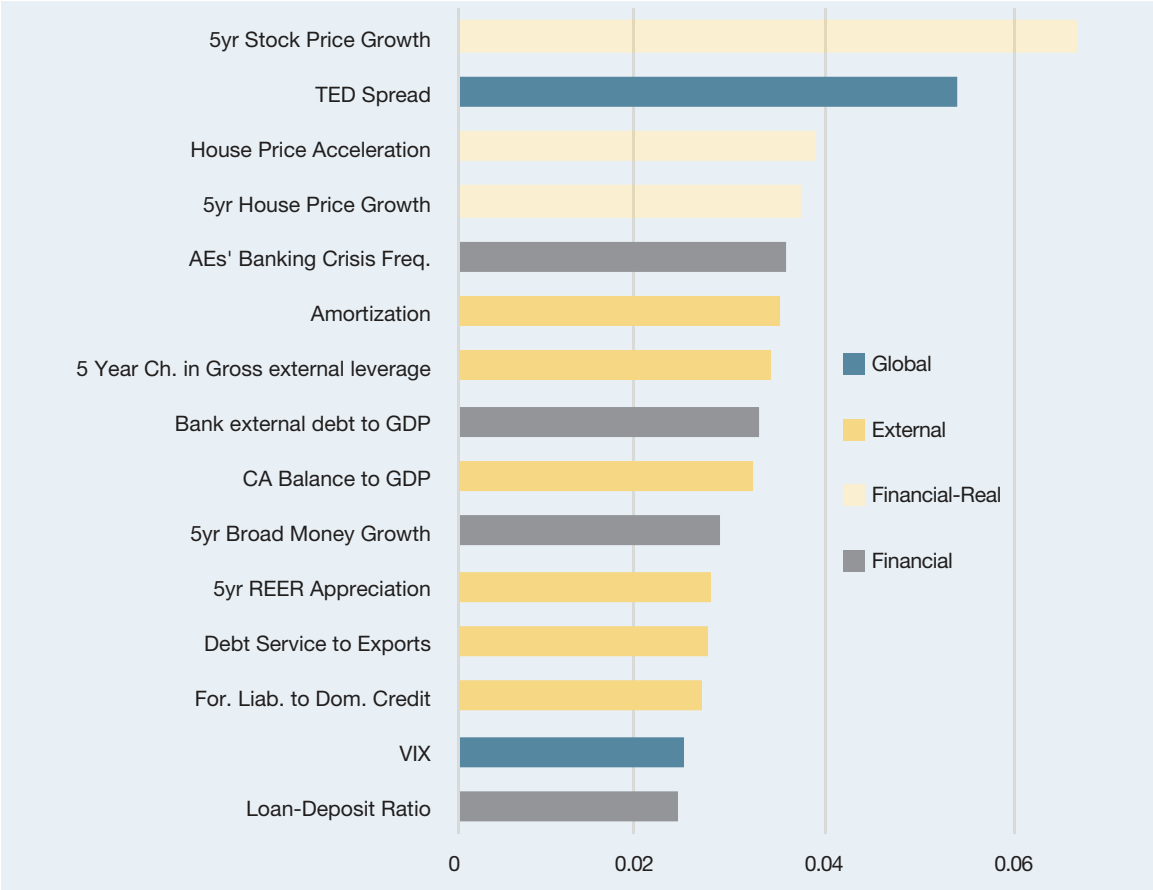
- **No single model consistently performs the best in predicting EMP events.** In backtesting, the signal extraction model performs the best (in terms of sum of errors) for AEs, RUSBoost performs the best for EMs, and RF model performs the best for LICs. Implied by these backtesting results, a signal extraction model is applied to predict EMP events in AEs and two RF models estimated within the income groups are applied to predict EMP events in EMs and LICs, respectively.[28]

[28] Because the algorithm used to calculate Shapley values does not work for RUSBoost, the second-best RF model in backtesting for EMPEs in EMs is chosen to use in practice for predicting EMPEs in EMs.

©International Monetary Fund. Not for Redistribution

- **An RF model estimated on a pooled sample of EMs and LICs is applied to oil exporters.** Due to the small sample size for oil exporters, a separate RF model is estimated on a pooled sample including EMs, LICs, including oil exporters, and applied to oil exporters only to assess their vulnerability to an EMP event. The model performance is comparable to the separate EM and LIC models discussed above.

FIGURE 10. SSGI Model Variable Importance



Model results. Model results indicate that both SSGIs and EMP events are driven by a variety of factors:

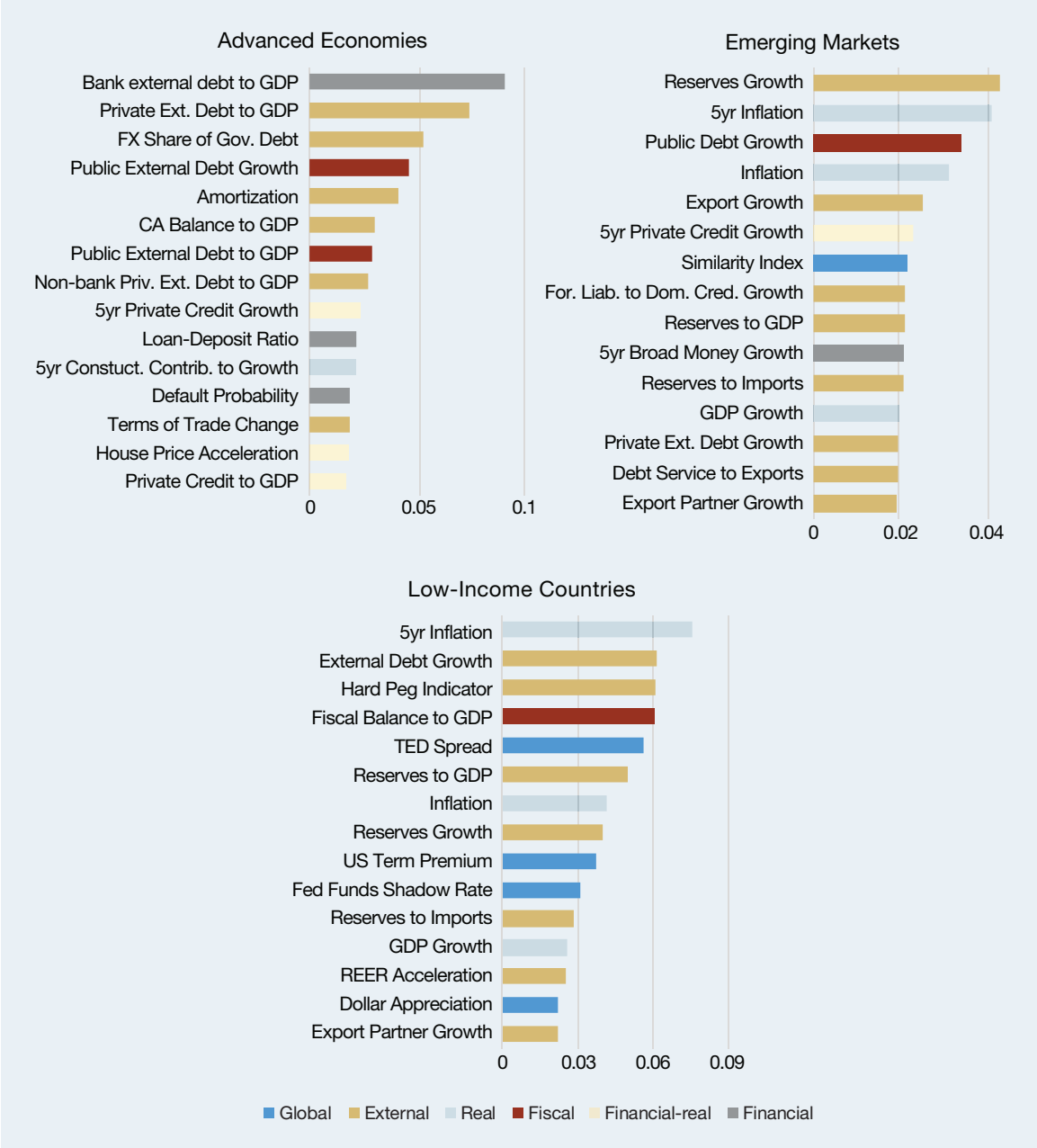
- **SSGIs are driven by variables from all sectors.** Estimated on the full SSGI sample (1990-2017), the "winning" signal extraction model puts emphasis on asset bubble building and busting, reflected in stock price and housing price variables.[29] In addition, global shocks and contagion effects (proxied by TED spread and the frequency of banking crises in AEs) play an important role, particularly in explaining GFC events. Country-specific external variables are also highly important: half of the top 15 are external variables, capturing external debt shocks and current account shocks.

[29] The variable importance is measured by signal-to-noise ratio, which is equivalent to Shapley values—our standardized way of measuring variable importance—in the case of a signal extraction model.

©International Monetary Fund. Not for Redistribution

- **There is cross-country variation in drivers of EMP events.** In AEs, the external sector plays an important role, among which banking sector external exposure is the leading factor. In EMs, reserves growth, export growth, and domestic credit growth—which perform well in predicting currency crises in the literature—help predict EMP events, while none of the global shocks or contagion effects appear among key drivers. In LICs, medium-term inflation plays an important role, and global shocks take up four among the top fifteen predictors, including the TED spread, US term premium, Fed rate as well as dollar appreciation.

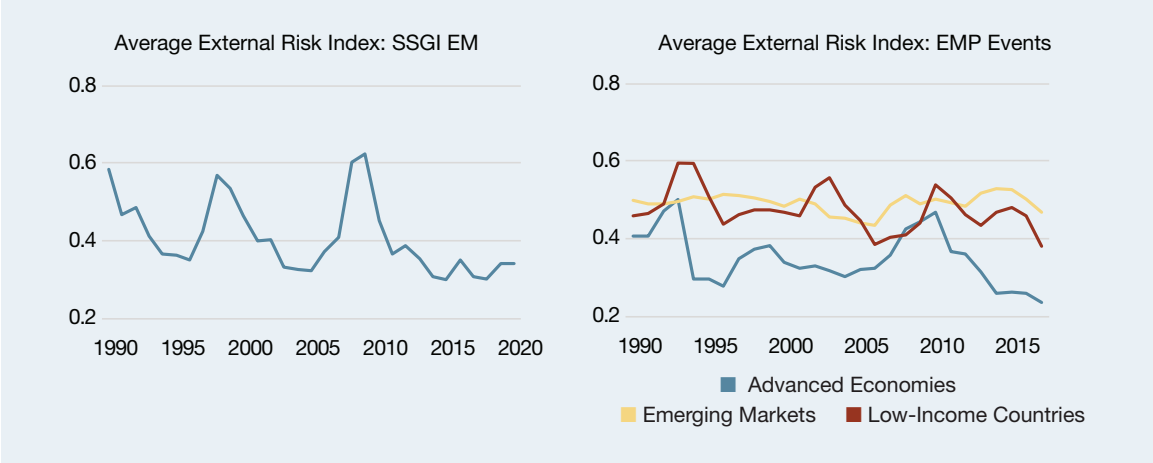
FIGURE 11. EMP Model Variable Importance



©International Monetary Fund. Not for Redistribution

Aggregate risk moves broadly in line with the frequency of past crises, both for SSGI and EMP. The aggregate risk is tracked by the average external risk index constructed based on the "winning" models. For SSGI, there is a spike at the end of 1990s, a quiet period in the mid-2000s, a spike for the GFC, and another lull ever since. For EMP, in AEs, there is one spike in the early-1990s, a lull for a decade, a spike for the GFC, and a quiet period ever since; in EMs, the risk is sustained through the three decades, with one spike for the GFC and one in mid-2010s; in LICs, the risk follows a similar trend with that in AEs but at a higher level.

FIGURE 12. Historical Risk Indices Over Time

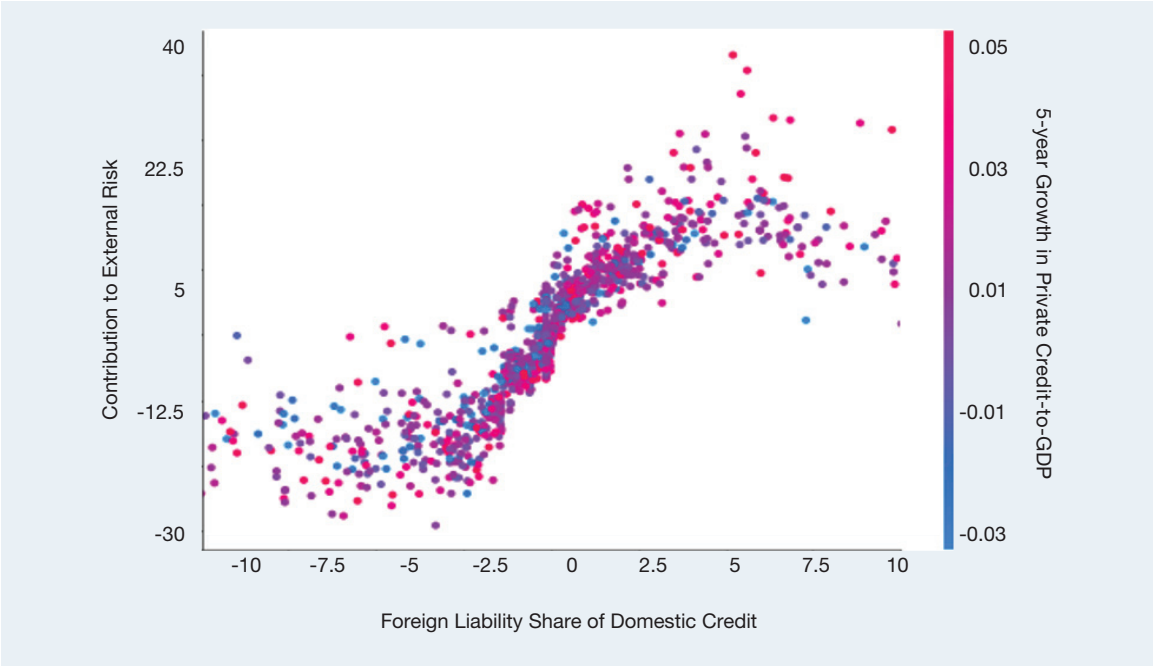


Variable interactions. One key feature of tree-based models is their ability to identify interactions between any pair (or indeed higher-order tuples) of variables. Such interactions cannot be captured by signal extraction models, and including all possible interaction terms would cause the loss in stability and robustness in traditional econometric models. The RF model for EMs and LICs, also used for oil exporters, uncovers interesting and valuable interactions.[30] Figure 13 plots the Shapley values associated with changes in foreign liability to domestic credit ratio for every observation in the dataset. Since Shapley values assign importance to a single variable by compressing information about its potential interactions with other variables, there is vertical dispersion in the points. It indicates that the same value for ratio growth will contribute differently to external risk depending on values of other variables. As plotted in Figure 13, these contributions tend to increase with the 5-year cumulative change in private credit-to-GDP ratio (blue dots in Figure 13 represent observations with low credit growth and purple and pink dots reflect higher growth). Currency mismatches—captured by the foreign liability share—thus matter more in the presence of a large ongoing private sector credit boom.

[30] The results are from the pooled EMP RF model for EMs and LICs.

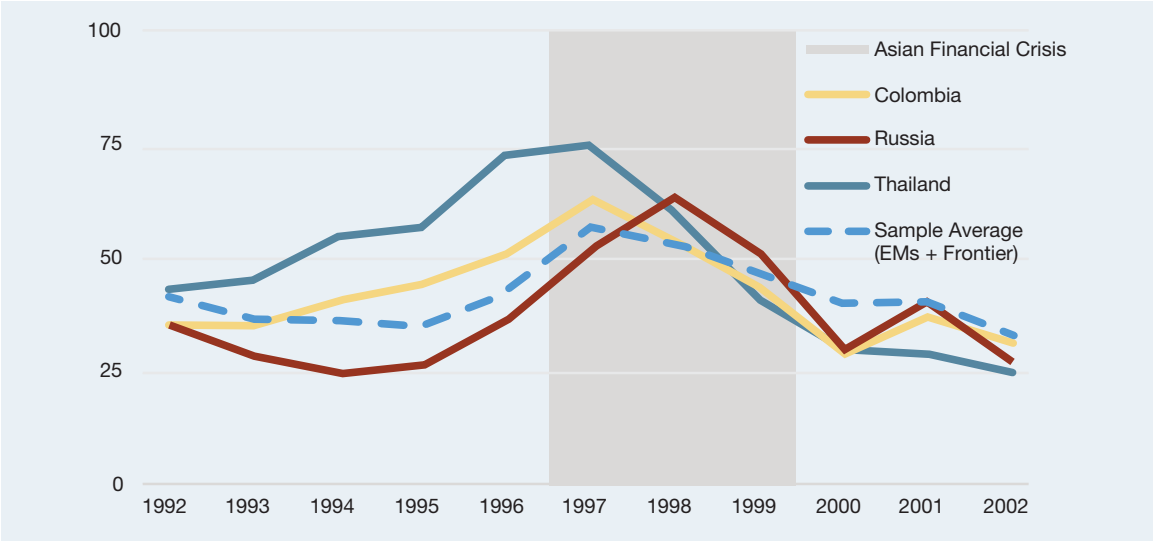
©International Monetary Fund. Not for Redistribution

FIGURE 13. External Risk Interactions



Illustrative country cases. The model captures well the factors underlying the Asian Financial Crisis (Figure 14). Thailand's risks were already above the sample average in the early 1990s and increased dramatically between 1993 and 1996. Notably, the external sudden stop model also picked up the threat of contagion to other countries like Colombia and Russia. While lower in magnitude than Thailand's spike, both saw rising risks of a sudden stop prior to the start of the Asian Financial Crisis and prior to the period when the crisis eventually spilled over to these countries. Of course, while the Asian Financial Crisis played a significant role, Colombia and Russia also had domestic imbalances and fragilities that increased their vulnerability to crisis.

FIGURE 14. External Sudden Stop Index and the Asian Financial Crisis, Selected Countries



©International Monetary Fund. Not for Redistribution

FISCAL SECTOR MODEL[31]

Crisis definitions. The fiscal sector crisis model estimates the risk of a fiscal crisis event in line with Medas and others (2018). A country is classified as being in a fiscal crisis in any given year if any of the four criteria is met: (1) occurrence of sovereign default or debt restructuring; (2) exceptionally large official financing; (3) high inflation or accumulation of domestic arrears (implicit default); and (4) loss of market access or spikes in sovereign yields (Table 1). The crisis dataset covers 188 economies from 1980 to 2017, ensuring that all country types are reflected in model estimation and offering a large set of observations for model training and testing.

TABLE 2. Fiscal Crisis: Definitions

EVENT	CRITERION (Minimum two years gap between crises)	THRESHOLDS AMs	EMs	LICs
1. Credit Event	**Default, restructuring, or rescheduling**			
	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	
	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
	ii. **Steep increase in domestic arrears** (in first difference of the ratio of 'other account payables (OAP)' to GDP in percentage points)		≥ 1	
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	
	b. Annual change in spreads (bps)	≥ 300	≥ 650	n.a.
	ii. **Loss of market access**			

On average, countries have undergone two fiscal crises since 1980 but there is large heterogeneity. At one end, LICs have experienced more than three crises on average while AEs have seen less than one. The duration of crises also varies significantly, with EMs and LICs showing the longest episodes—on average five years. Historically, fiscal crises tend to come in waves with the largest concentration taking place in the 1990s (Figure 15).

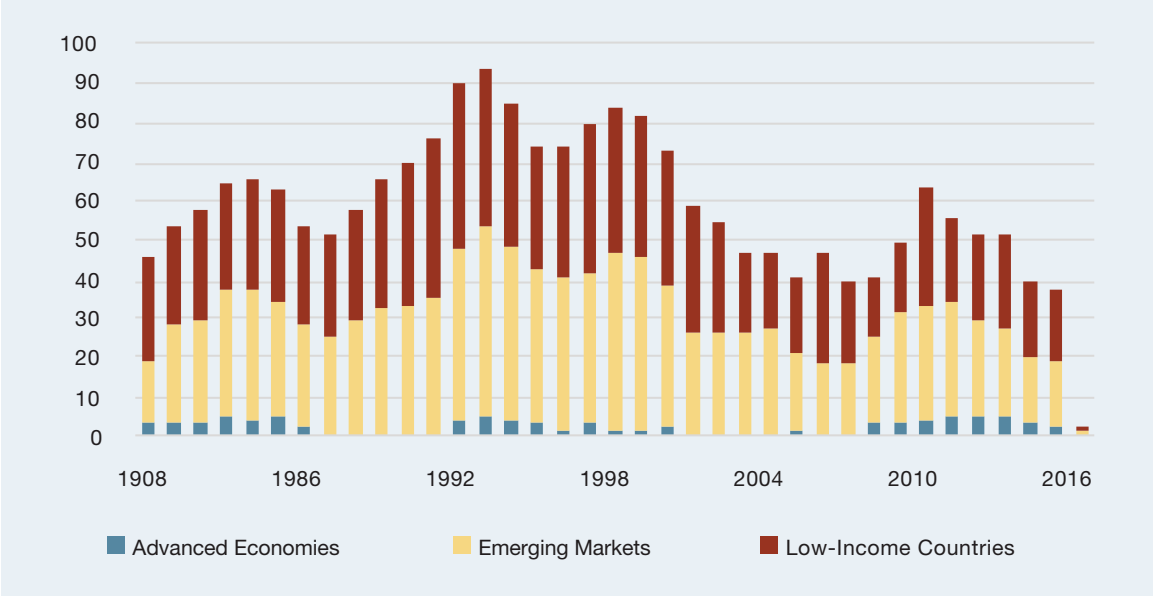
[31] Model developed by Klaus Hellwig, Marialuz Moreno Badia, Pranav Gupta, and Paulo Medas. For a full treatment, see Moreno Badia and others (2020), and Hellwig (2020).

©International Monetary Fund. Not for Redistribution

Variable selection. The fiscal sector crisis model evaluates the contribution of vulnerabilities from any sector to fiscal crises, albeit with clear emphasis on vulnerabilities typically seen as contributing to fiscal risks. The final model is based on 106 variables (see Annex VIII).[32] The variables are grouped into 11 categories:

- *Fiscal* variables, including government expenditures, revenues, and overall budget balance.
- Three groupings on debt: *public debt*, *private debt,* and *total debt*.
- Variables linked to economic activity, like economic growth and inflation, are grouped under *Real* sector.

FIGURE 15. Countries with Fiscal Crises, 1980-2017



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

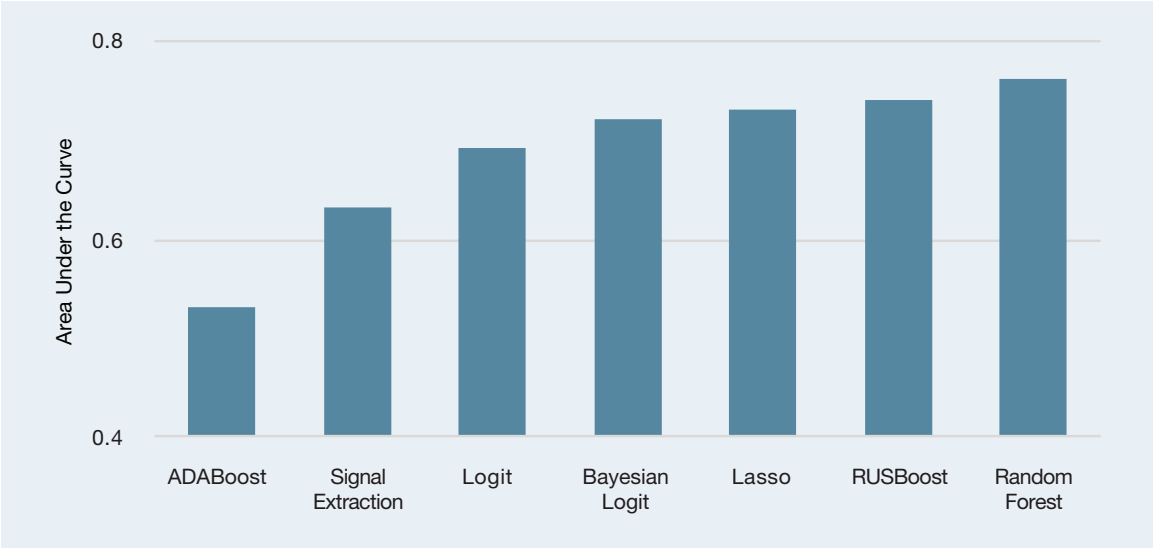
- Two categories on external variables: country-specific *external* sector variables, which include current account, imports, exchange rate and FDI, among others; and *global* variables, such as commodity prices, US interest rates, and global GDP growth.
- *Volatility*: variables that capture volatility (measure as a backward-looking standard deviation) in key economic aggregates, like economic growth, terms of trade, and inflation.
- *Institutions*, including indicators on governance and electoral systems.
- Other groupings include *crisis history* and *country grouping* (dummies for country categories).

To ensure full country coverage and to avoid losing the information value of the available data, missing data is imputed using the sample median, consistent with other sectors.

[32] The analysis also uses several permutations of the variables—for example, levels, lags and first differences for the same variable. In total, 290 variables are used.

©International Monetary Fund. Not for Redistribution

FIGURE 16. Fiscal Sector Model Performance



Model selection. A number of econometric and ML algorithms are evaluated for their ability to anticipate a crisis one to two years in advance:[33, 34]

- **Testing procedure.** Each model is evaluated using classic back-testing starting in 2000 and then repeatedly re-estimated adding a year of data to the sample in each iteration. Accuracy is assessed using the likelihood score, but the AUC is also calculated for reference.[35]
- **RF model consistently performs the best or among the best for a pooled sample of all income groups.** This conclusion is supported by both the log likelihood ratio and AUC, comparing the RF model to Logit, Signal Extraction, Elastic Net, and gradient boosted trees (Figure 16). The model estimated on a pooled sample performs better than estimated on isolated income group subsamples: to the extent that the heterogeneity across countries is relevant for crisis prediction, it is adequately captured by the tree-based model.

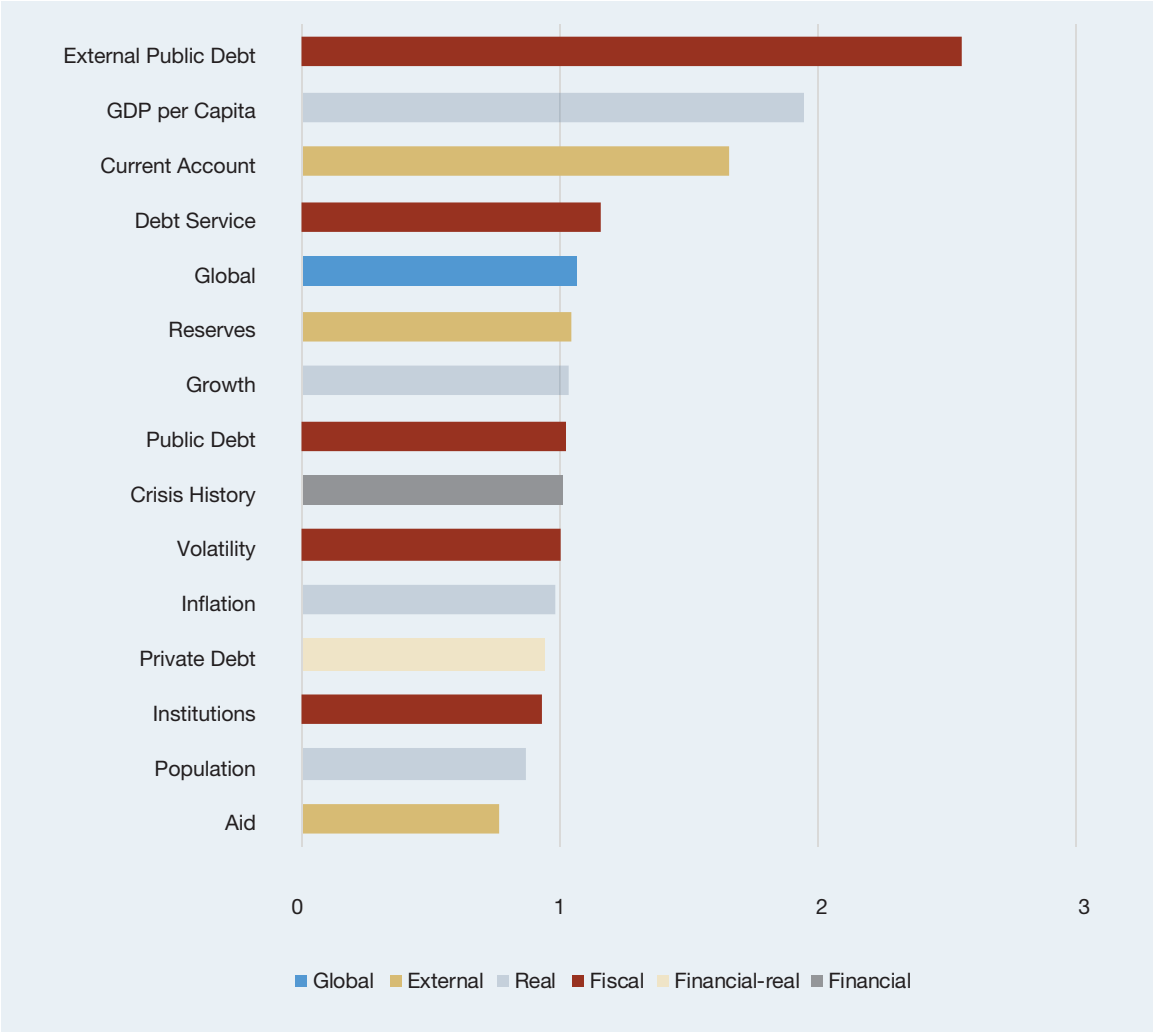
[33] Other boosting methods, including adaptive boosting and RUS boosting, were also explored initially.

[34] For the logit model, a stepwise forward variable selection algorithm is used.

[35] For ease of interpretability, we report the likelihood ratio (LR), which puts this measure in relation to the score obtained from the training-sample average frequency—a "naïve" benchmark.

©International Monetary Fund. Not for Redistribution

FIGURE 17. Fiscal Model Variable Importance



Model results. Estimated on the full sample (1980–2017), the RF model puts emphasis on variables from a variety of sectors. The most important predictor, after accounting for the time passed since the last crisis and the level of development, is a country's level of external public debt, expressed as both a ratio to GDP and to exports. Also among the top are GDP per capita, reserve coverage of imports, inflation volatility, and the quality of bureaucracy (a measure of governance). To summarize the relative importance of all 290 variables (106 series plus lags, differences, and other permutations), Figure 17 aggregates the importance measures by category.[36] Country-specific external sector variables and public-debt-related variables are the most important predictor groups, followed by real variables.

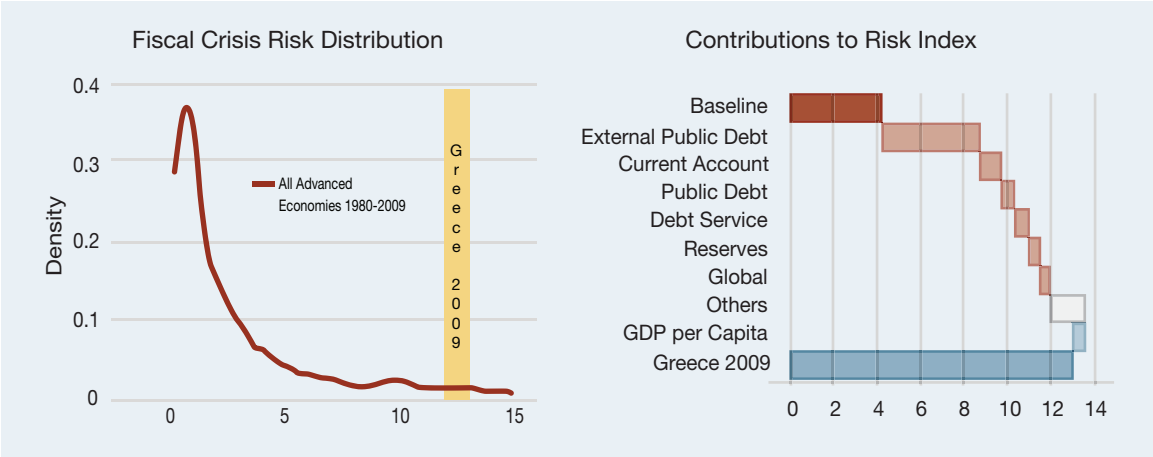
Illustrative country cases. Greece experienced a twin crisis during the GFC with the fiscal crisis occurring in 2010. Figure 18 presents an out-of-sample forecast for Greece in 2009, using a model estimated on data through 2007. The model points to the stock of public debt, especially external debt, as the key driver of risk, supplemented by current account pressures and in particular debt service. The model also sees Greece's high level of development as a mitigating factor. The key role of debt stocks is an

[36] This aggregation can give a more accurate account of economic significance than the importance of individual predictors, because the importance of variables such as external debt is distributed to its various permutations.

©International Monetary Fund. Not for Redistribution

important step forward in this model. More traditional models often have difficulty differentiating high debt carrying capacity countries from excessive debt burden countries, but the ML tools used here help to disentangle these effects. The model demonstrates good predictive performance, although with the benefit of some key revised fiscal series.[37] A 13 percent probability of default constitutes a far-right tail event for an advanced economy based on the experience through 2007 and would have strongly flagged this risk.

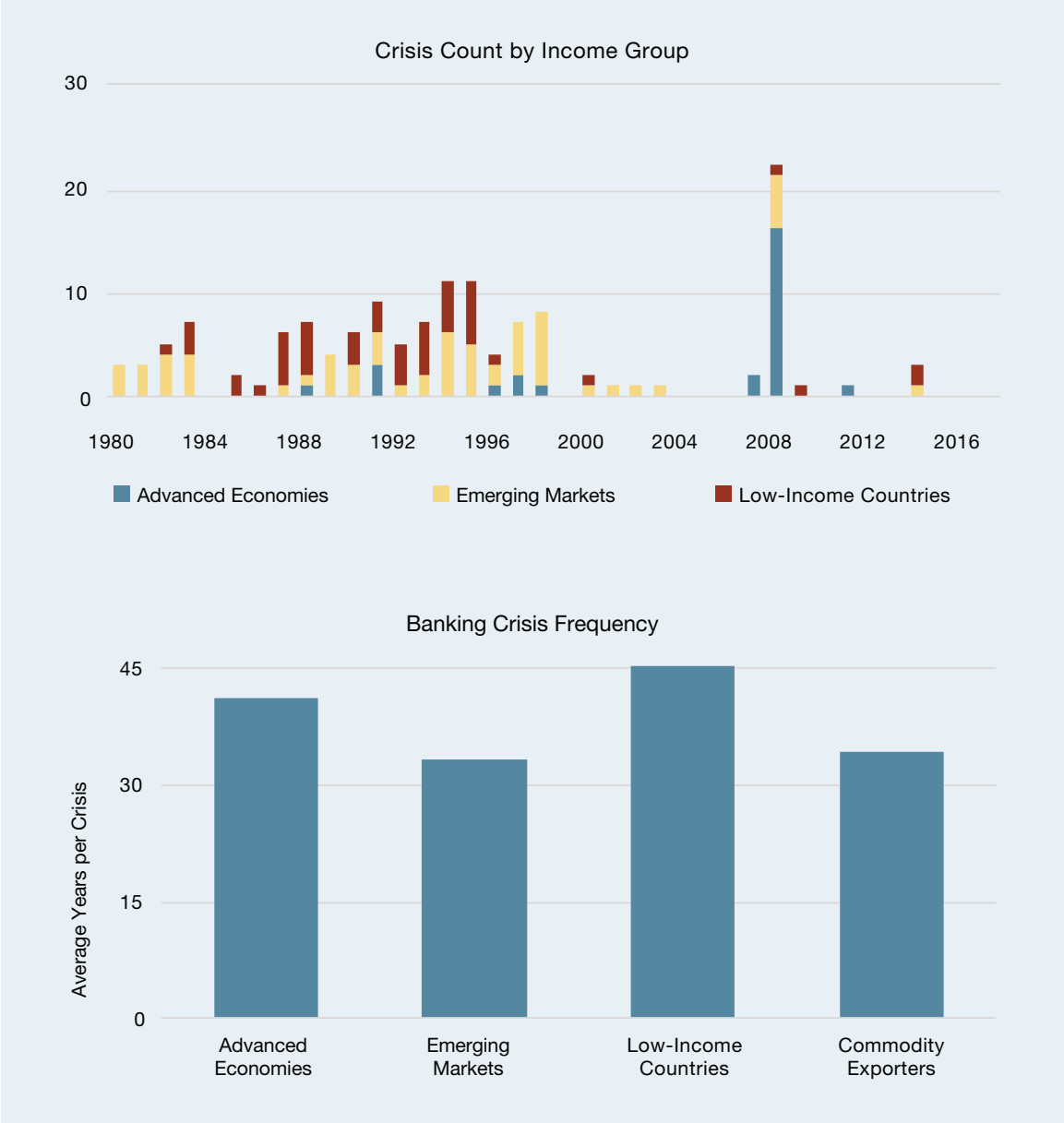
FIGURE 18. Fiscal Risk in Greece, 2009



[37] Note that the model is estimated with some revised fiscal series, which were not available at the time.

©International Monetary Fund. Not for Redistribution

FIGURE 19. Bank Crisis History and Frequency



©International Monetary Fund. Not for Redistribution

FINANCIAL SECTOR MODEL[38]

Crisis definitions. The financial sector crisis model estimates the risk of banking sector crisis events documented in Laeven and Valencia (2018).[39,40] Crises meet two sets of criteria: evidence of significant financial distress (e.g., sizable bank runs, bank losses, or bank liquidations) and significant policy interventions (e.g., emergency liquidity support, public recapitalization, or nationalization). The crisis dataset covers 117 countries from 1980 to 2017, covering all country types and offering a large set of observations for model training and testing. These crisis dates are widely used in the literature, placing the model well in the broader context of research on financial crises.[41] As shown in Figure 19, crisis frequencies are broadly consistent across income groups, averaging one crisis every 30 to 45 years. Crisis patterns have varied considerably over time, with crises in the 80's and 90's occurring frequently in EMs and LICs and rarely in AEs, while crises since 2000 being concentrated on the GFC and in AEs.

Explanatory variable selection. As in all sectoral models, the financial sector crisis model evaluates the contribution of vulnerabilities from any sector to banking crises, albeit with clear emphasis on vulnerabilities typically seen as contributing to financial risks. The final model is based on 31 variables, including 7 financial sector variables and 9 variables reflecting the intersection of the financial and private non-financial sectors (listed below). Data availability posed a significant constraint on the variable selection process, given the long time-span and wide country coverage of the model sample. There are many informative variables with short histories, as with the now commonly reported Financial Soundness Indicators (FSIs), or limited country coverage as with housing price series. Missing data is imputed with the median value.

[38] Model developed by Jorge A. Chan-Lau, Silvia Iorgova, Kevin Wiseman, and Le Xu.

[39] See Laeven and Valencia (2018), which includes crises identified through 2017.

[40] Like the financial sector model here, the Fund's Financial Sector Assessments (FSAs) model assesses risks from the financial sector. The FSAs, however, address narrative risks customized to the country in question, are quantified by shock scenario simulations, and examine the entire financial sector. The work presented here seeks to capture all-cause risk in a purely empirical framework and focusses on banking crises due to data availability.

[41] See, for example, Manasse, Savona, and Vezzoli (2016) for a Random Forest-based example.

©International Monetary Fund. Not for Redistribution

TABLE 3. Financial Crisis: Explanatory Variables

GLOBAL	FINANCIAL	FINANCIAL-REAL	REAL
• Real 10yr US Yield, level and gap • 3mo T-bill Rate HP gap	• External Bank Liabilities to GDP • Loan to Deposit Ratio • Capital Adequacy Ratio • Avg. Bank Prob. Default • Financial Inclusion — Access • Financial Inclusion — Efficiency • Real Deposit Rate Growth • CB Liquidity Support, level and growth • Credit to GDP	• Corp Debt Sub-Inv. Grade • Total Debt Growth • House Price-to-Income • BIS Credit Gap[42] • House Price-to-Rent • Equity Price Gap • Avg. Non-Bank Prob. Default • Avg. Corp. Prob. Default	• Output per Capita (PPP) • Real Output Growth • HP Output Gap

EXTERNAL	FISCAL
• External Debt to Exports • Gross Reserves to GDP, Imports • Debt Service to CA Credits	• Int. Rate on New Pub. Debt • LT Bond Yield Growth

Model Selection. A number of econometric and ML models are evaluated for their ability to anticipate a crisis one to two years in advance:

- **Testing procedure.** Each model is evaluated using classic backtesting. Models are estimated based on data through 2000, and then repeatedly re-estimated adding a year of data to the sample in each iteration. For each estimation the model is applied to three out-of-sample years of data. This backtesting exercise risks evaluating the model exclusively on its performance on a single event—the GFC—so it is complemented by a cross-validation based on groups of consecutive years.
- **The balanced RF model consistently performs the best or among the best.** Figure 20 presents the averaged performance (in terms of the out-of-sample AUC) across the backtesting and cross validation exercises for the balanced RF model and several other popular classification techniques, including a standard Logit, two types of penalized logit (Ridge and Lasso), the signal extraction methodology, two "boosted" variants of RF models (ADABoost and RUSBoost), and Support Vector Machines (SVM). The SVM performs very well in the backtesting exercise but is nearly uninformative in cross-validation and is generally more unstable in robustness tests. The classic logit, by contrast, does roughly as well as the RF in cross validation but poorly in backtesting. The RF model performs best in cross-validation and second-best in backtesting, with similar performance in both contexts. Robustness checks for performance of one- and three-year horizons as well as with data sets with 84, 160, and 949 variables, were conducted with similar results.

[42] Due to some limitations of the BIS credit gap data, future work could consider the External Sector Report (ESR) credit gap, which is from the desk economists through the External Balance Assessment (EBA) process. More details can be found in Baba and others (2020).

©International Monetary Fund. Not for Redistribution

FIGURE 20. Financial Sector Model Performance

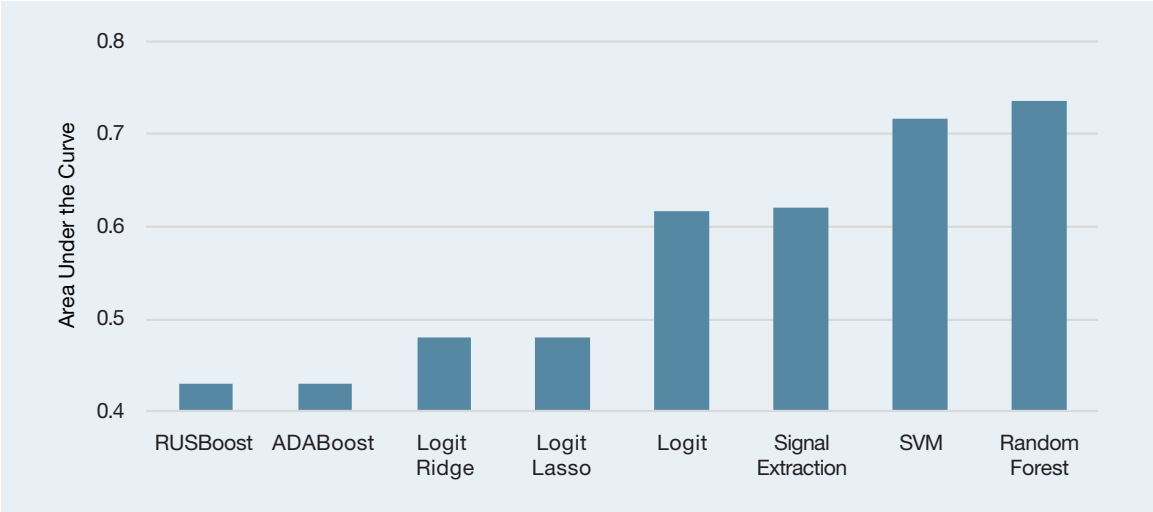
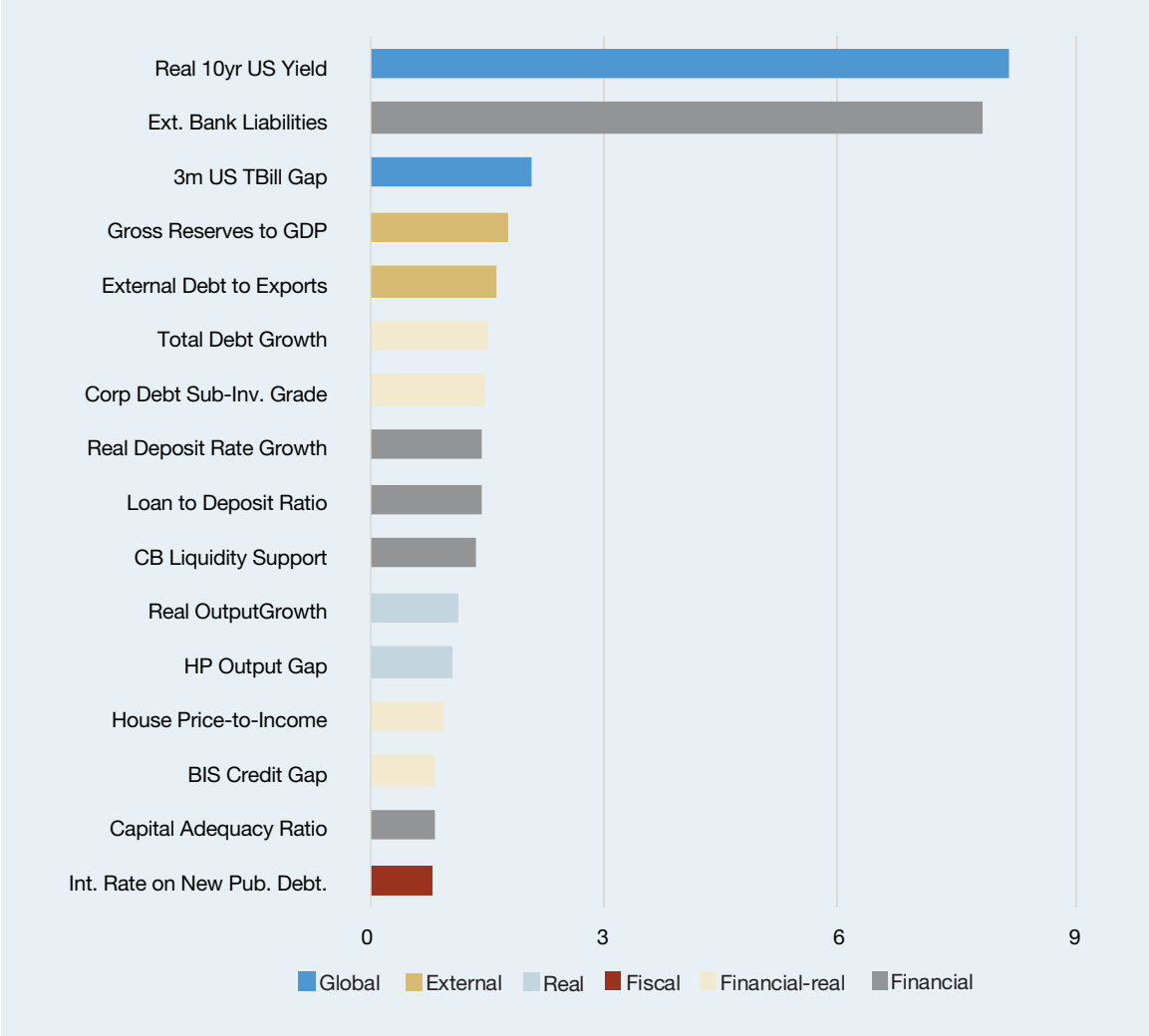


FIGURE 21. Financial Model Variable Importance

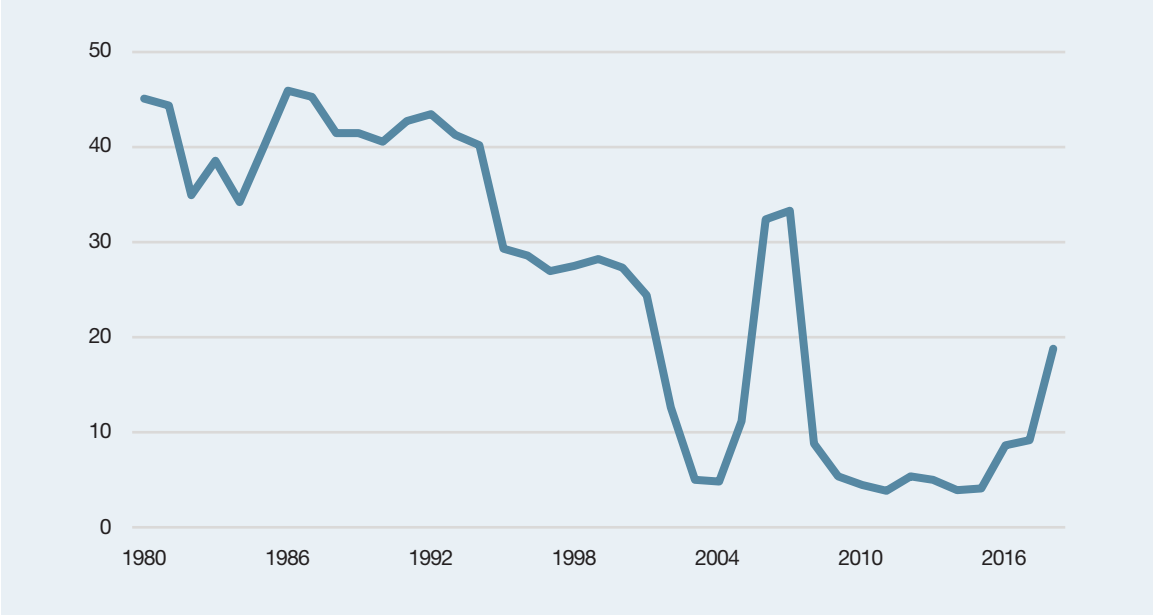


©International Monetary Fund. Not for Redistribution

Model Results

- **Risk of financial crises is driven by variables from all sectors.** Estimated on the full sample (1980-2015), the balanced RF model puts emphasis (as measured by the average absolute Shapley value) on the real US long-term interest rates (a proxy for global financial conditions), followed by external liabilities of the banking sector as a percent of GDP (a measure of financial depth). Financial depth serves to explain the low rate of crises seen in financially shallow countries, while the global interest rates are associated with the high rates of crises in the 1980s and 1990's. Many indicators that play a role in explaining GFC events fill out the top 15 predictors, including housing prices and credit gaps. External indicators also play an important role in the model.

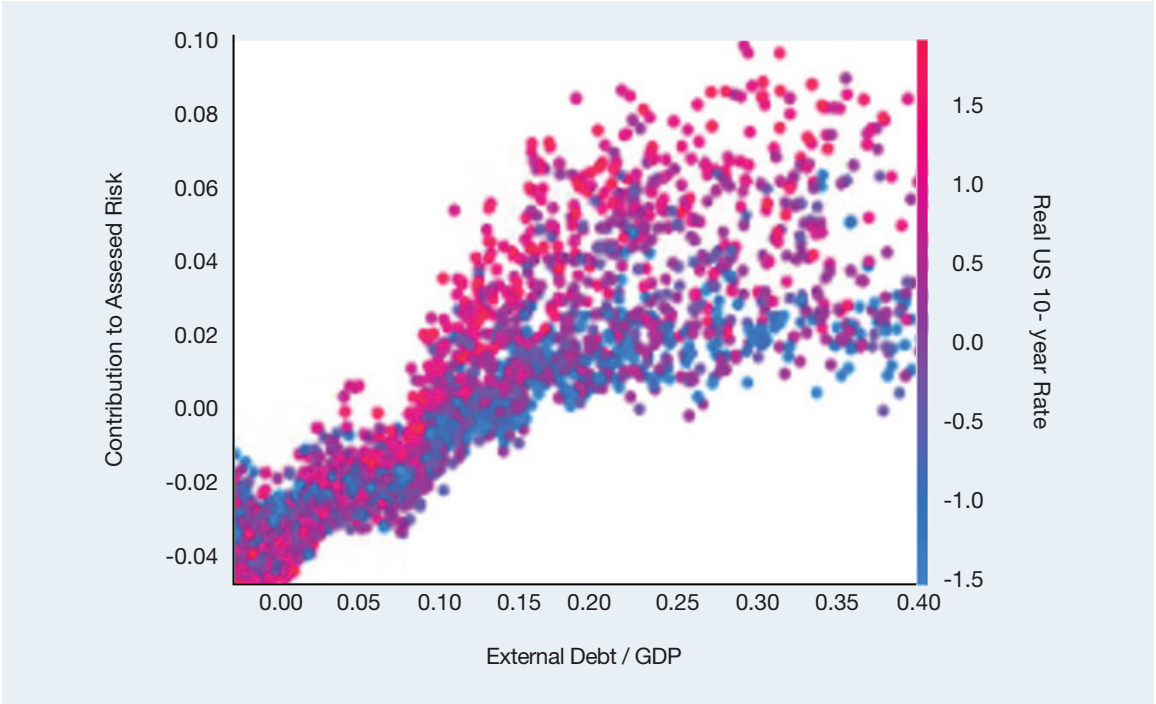
FIGURE 22. Historical Evolution of Average Risk Index



- **Aggregate risk moves broadly in line with the frequency of past crises.** The aggregate financial risk—tracked by the average financial risk index constructed based on the model—shows sustained levels through the mid-1990s, a lull in the mid-2000s, a spike for the GFC, and another lull ever since. The recent uptick of the risk index is driven by global variables, especially the 3-month US T-Bill gap measured as a deviation from a one-sided HP trend. By end-2018 this gap had edged up above 1%, the fourth highest observation and highest since 2006.

©International Monetary Fund. Not for Redistribution

FIGURE 23. Global-Local Variable Interaction in Financial Sector Model



Variable interactions. The importance of interactions comes through clearly in the financial sector when investigating the interaction between domestic vulnerabilities and global financing conditions. The chart above plots the Shapley values associated with external debt of every observation in the dataset. Blue dots in the picture represent observations with low Real US 10-year Bond rates—a measure of external financing conditions—and purple and pink dots reflect higher rates. Risks rise as external debt exceeds 15 percent, but the increase in risks is magnified under higher global interest rate conditions.

Illustrative country cases. Figure 24 shows the evolution of financial risk ratings for Iceland and Ireland in the run-up to the GFC.[43] Both Ireland and Iceland saw sizable increases in financial risk during this period and were near the top of the distribution by 2007, with Ireland already having shown signs of vulnerability in 2005.

[43] Ratings reported here are pseudo out-of-bag ratings. For each observation the model is re-estimated on the full sample in 1980-2015 but with the observation and the two adjacent observations (previous and subsequent years) dropped, and then this model estimation is applied to the observation.

©International Monetary Fund. Not for Redistribution

FIGURE 24. Financial Crisis Risk Index, Iceland and Ireland, 2005 and 2007

©International Monetary Fund. Not for Redistribution

REAL SECTOR MODEL[44]

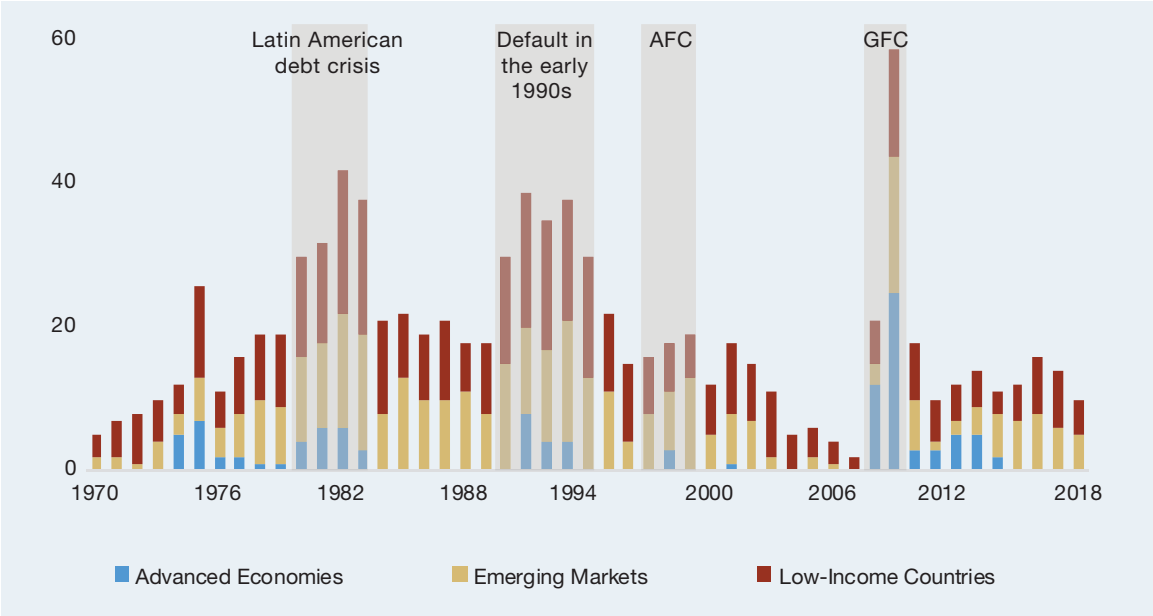
Crisis definitions. Real sector crisis episodes capture a mix of sharp slowdowns in economic activity, as well as more moderate but prolonged episodes, typically described as V- and U-shaped recoveries. The real sector crisis definition aims to take a pluralistic view of what constitutes a crisis in terms of depth and duration, but also in terms of comparator countries. Several crisis criteria are consistently applied across countries and a real sector crisis is defined when a majority of these criteria are satisfied.

- **Crises are defined based on four different GDP series and four different thresholds.** The four series are i) a country's annual growth rate, ii) its cumulative growth rate over the past three years, iii) its growth performance relative to the most recent five-year average, and iv) its average GDP level relative to the previous three-year average. For each of these, the focus is on GDP per working-age person. Values of these series are flagged as being in a crisis if they fall below the 10th percentile of observations in one of the following groups: i) all countries in the sample, ii) all countries in the same income group – advanced economies (AEs), emerging markets (EMs), and low income and developing economies (LICs) according to the WEO classification in 2020, iii) by income group according to the WEO classification in 1980 with an additional category of countries with a population below one million, and iv) countries in the same tercile of the total sample for year-on-year growth volatility.
- **These four series and four thresholds lead to sixteen crisis criteria, and the ultimate crisis definition appears consistent with historical real sector crises episodes.** Each indicator assesses whether the point-in-time value of one of the series is below one of the thresholds. A country in one particular year is recorded as experiencing a real sector crisis whenever nine or more indicators signal a crisis. The crisis count by income group is provided in Figure 25. Under our definitions, there are four clusters for real sector crises: i) during early 1980s in LICs and EMs, corresponding to the Latin American debt crisis; ii) during early 1990s in LICs and EMs; iii) during late 1990s in EMs, corresponding to Asian Financial Crisis; and iv) during late 2000s in all income groups, corresponding to the GFC. Countries on average experienced two real sector crises over the sample period, with each crisis lasting two years on average.

[44] Model developed by Jorge A. Chan-Lau, Maksym Ivanyna, Andrew Swiston, Andrew Tiffin, and Yunhui Zhao.

©International Monetary Fund. Not for Redistribution

FIGURE 25. Real Sector Crisis History



Explanatory variable selection. The selection of explanatory variables is guided by a review of the academic and practitioners' literature on economic crises. These include global variables such as short- and long-term US interest rates (as a proxy for global financing conditions) and oil price. Domestic real sector variables are included if they are perceived to presage a downturn, including measures of overheating such as inflation and output gap. External variables indicate domestic vulnerability to or the incidence of external shocks, including real effective exchange rate, terms of trade, external debt and reserves. Fiscal variables are also included to capture possible risks from fiscal distress, including debt to revenue and the fiscal deficit. See Table 4 for a complete list. The explanatory variables are constructed using either the variables' raw values, or applying transformations such as differences, growth rates, and one-sided Hodrick-Prescott trends. Missing data is imputed with median imputation to retain the information in the available data while making the imputations as neutral as possible.

©International Monetary Fund. Not for Redistribution

Model selection. The model selection proceeds in two stages. The first stage assesses logit, signal extraction (SE), SVM, and RF models in a backtesting exercise.

- **Testing procedure.** The analysis is conducted for the 1980-2016 sample period with the first training window covering 1980-2000 and the testing period comprising the next two years, 2001-2002. The second training period extends the window by one year, to 1980-2001, and tests the model using the crisis observations two years ahead; and so on. Hyperparameters are tuned at each time horizon using AUCs in a time-block cross validation exercise on the training set. Because crisis episodes in the samples are imbalanced, i.e., there are too few crisis observations, the data set is balanced using the Synthetic Minority Over-sampling Technique (SMOTE) (Chawla and others, 2002). The results are summarized in the next two charts.

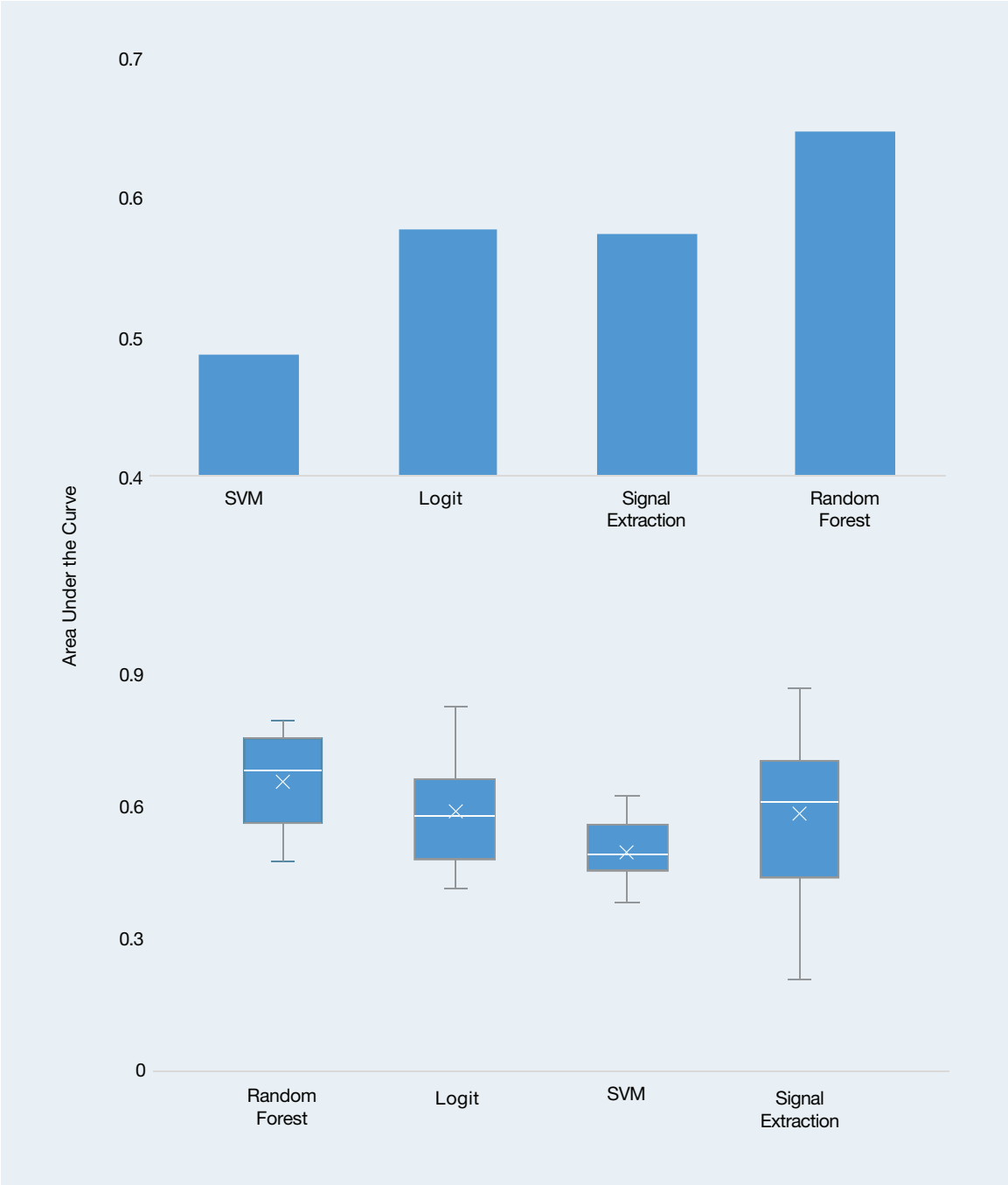
TABLE 4. Real Crisis: Explanatory Variables

GLOBAL	FINANCIAL	REAL
• FFER Shadow Push Factors • Real 10yr US Yield • Trade-Weighted Dollar Index	• Capital Adequacy Ratio • Loan to Deposit Ratio • Private Debt and Loans to GDP • Real Deposit Rate • Short-Term Nominal Deposit Rate	• HP Output Gap • Inflation • Natural Disasters, Material Impact • PPP Income Relative to US • Real GDP Growth, Export Weighted • Real GDP Growth
EXTERNAL	**FINANCIAL - REAL**	**FISCAL**
• Reserves as Percent of ARA Metric • Current account balance, % of GDP • Exports of goods and services • Net non-FDI capital inflows, % of GDP • Exchange Rate (National Currency per US Dollar) • Real Effective Exchange Rate • External Debt to Exports • Oil Price, Growth Rate • Terms of Trade	• Equity Price Growth • 5yr House Price Growth • Corp Debt Sub-Inv. Grade • BIS Credit Gap • Total Debt Growth	• Debt Revenue • Fiscal Balance to Fiscal Revenue • Interest Rate on New Pub. Debt. • Interest Expense to Revenue

- **Random forest models usually outperformed other models.** Goodness of fit measures are more volatile for the signal extraction and logit regression models. SVM performs poorly. Building on the expanding window analysis, the RF class is selected as the core model for predicting real sector crises. The final model is trained using 1980-2016 data. Selected results are presented below.

©International Monetary Fund. Not for Redistribution

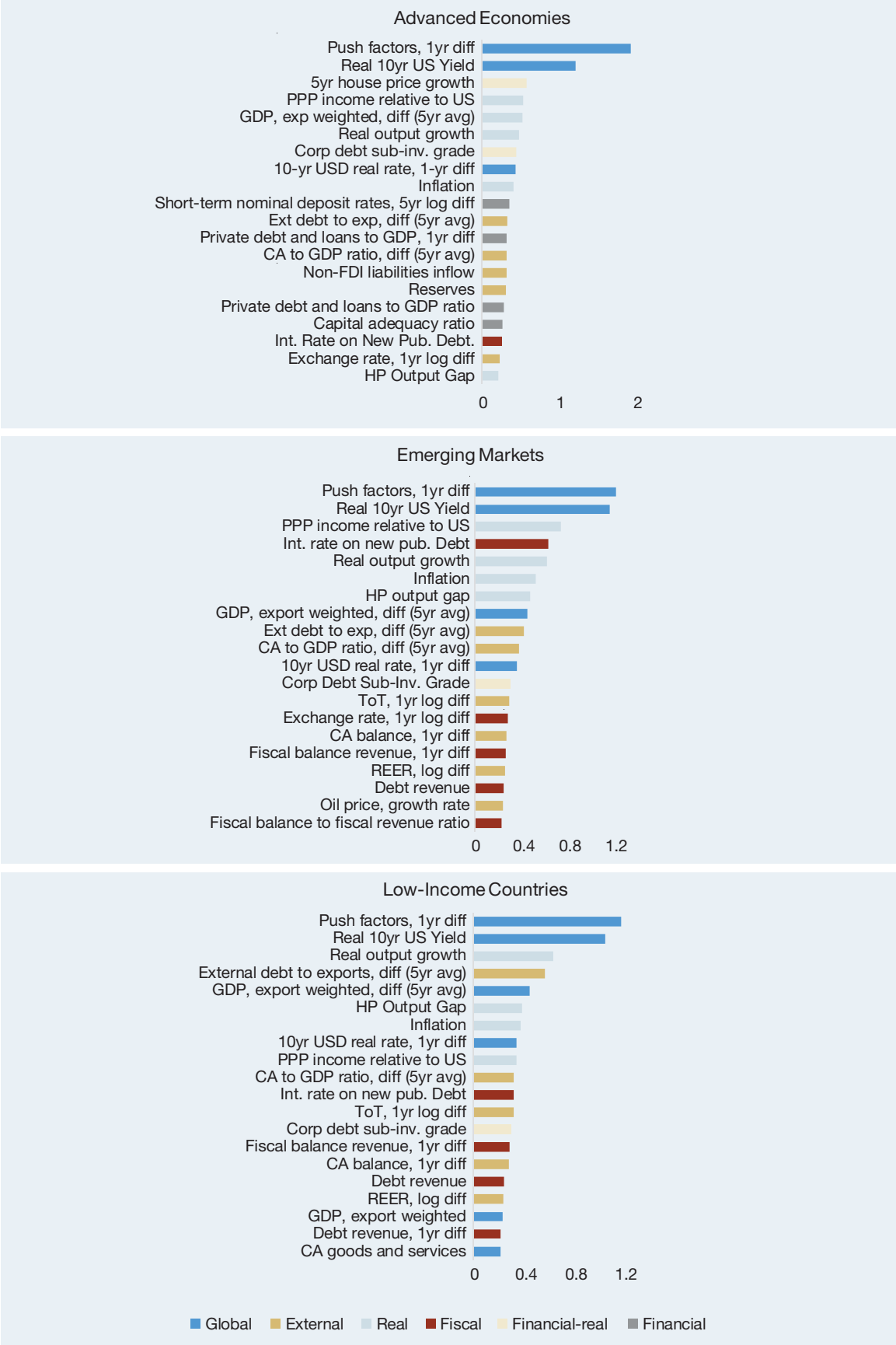
FIGURE 26. Real Sector Model Performance



Model results. Variable importance is calculated using Shapley values. While global variables are important for predicting crisis episodes, there are differences among AEs, EMs, and low income and developing economies (LICs). In AEs, the most important variables are global variables associated with the policy stance in the US, followed by variables in the real and financial sectors. For LICs, in contrast to AEs and EMs, fiscal sector variables are also important besides global and real sector variables.

©International Monetary Fund. Not for Redistribution

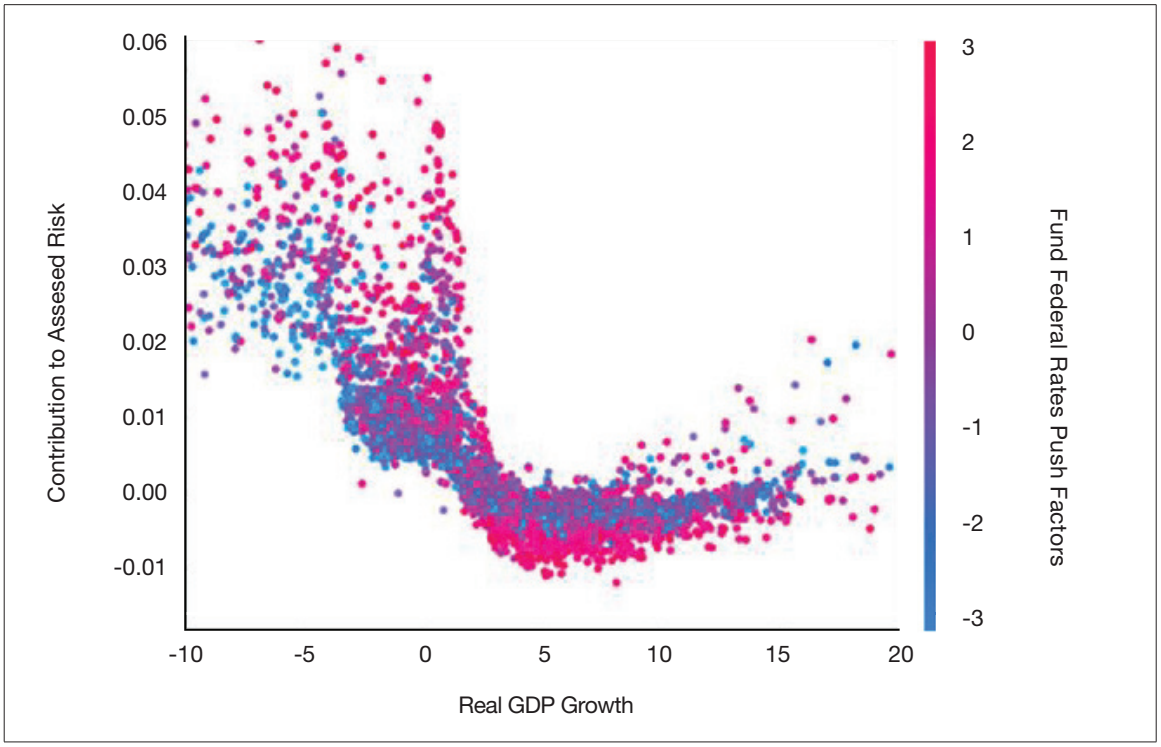
FIGURE 27. Real Model Variable Importance



©International Monetary Fund. Not for Redistribution

Variable interactions. As with other sectoral models, the real sector ML model also features important nonlinear interactions not captured by traditional models. Such interactions come through clearly in the real sector when investigating the interaction between domestic and global risk factors. Figure 28 plots the Shapley values associated with domestic real GDP growth of every observation in the dataset. Blue dots represent observations with low federal funds shadow rates (a measure of global financial conditions), where purple and pink dots reflect higher rates. As the figure shows, for a given level of negative domestic GDP growth, domestic growth becomes a more important predictor of real sector crises when the federal funds shadow rate is higher, suggesting that tighter global financial conditions amplify the impact of negative domestic growth on the likelihood of a real sector crisis.

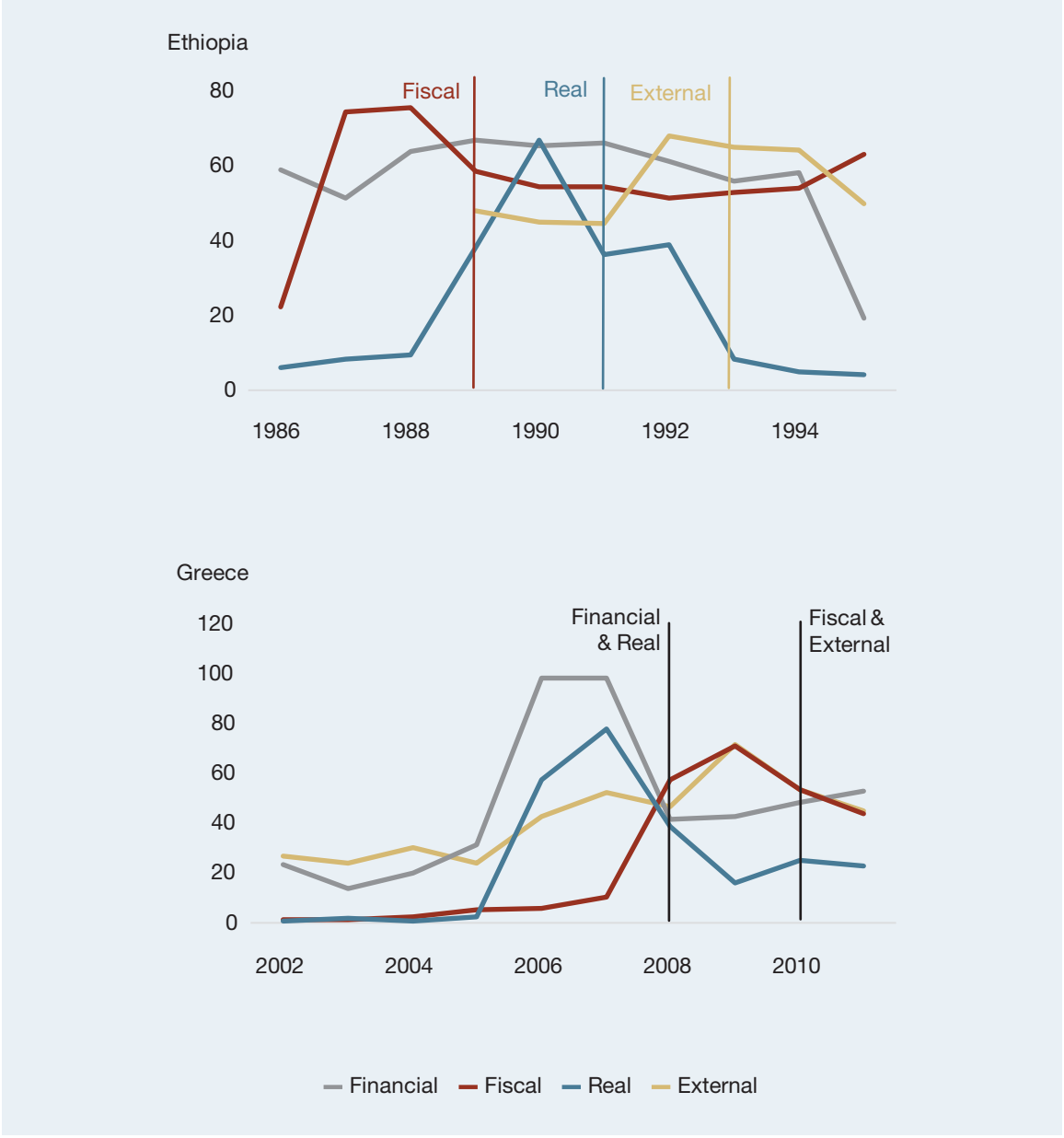
FIGURE 28. Nonlinear Interactions in Real Sector Model



Illustrative country cases. Figure 29 shows the evolution of real crisis risk index for Ethiopia (from late 1980s to early 1990s) and Greece (in 2010s), as well as risk indices from other sectoral models discussed earlier. In both cases, all four models perform well in anticipating the sectoral crisis realizations: the risk index spikes 1-2 years ahead of each corresponding crisis. The figure also reveals a pattern (displayed in some other countries as well) in terms of the sequencing of different types of crises: in LICs where financial sectors are less developed and fiscal sectors play a relatively large role, real sector crises are typically preceded by fiscal crises; by contrast, in AEs, real sector crises could come hand-in-hand with financial crises, triggering subsequent crises in real and external sectors.

©International Monetary Fund. Not for Redistribution

FIGURE 29. Crisis Risk Indices in Four Sectors, Ethiopia and Greece



©International Monetary Fund. Not for Redistribution

CONCLUSION AND NEXT STEPS

The machine-learning models presented here represent the next generation of risk assessment models in the VE. Those models with the best horse race performance will underpin country risk assessments in the four sectors explored here. The results from these models are complemented by a number of the models that were already in use (Ahuja and others, 2017). The additional models, representing alternative methods and areas of focus, help provide a more comprehensive and granular representation of the risk conjuncture.

Machine-learning techniques offer new opportunities in risk assessment. Beyond the crisis risk assessment models presented here is a wide horizon of techniques that economists are only beginning to apply to risk assessment. Multi-classification models, which can simultaneously assess the possibility of different combinations of crises, offer a clearer understanding of how the crises studied here overlap or contribute to each other. There is also room to move into bigger data, higher frequency series, and less structured information for a short-horizon crisis risk assessment, leveraging more advanced ML techniques. Beyond forecasting techniques, recent progress in causal identification with ML techniques (Athey and others, 2019; Chernozhukov and others, 2017) promises better measures of the consequences of crises (Tiffin, 2019), and can perhaps identify successful prevention methods.

©International Monetary Fund. Not for Redistribution

REFERENCES

Ahuja, Ashvin, Murtaza Syed, and Kevin Wiseman, 2017, "Assessing Country Risk—Selected Approaches—Reference Note," IMF Technical Notes and Manuals, 17/08 (Washington: International Monetary Fund).

Alessi, Lucia, and Carsten Detken, 2018, "Identifying Excessive Credit Growth and Leverage," *Journal of Financial Stability*, Vol. 35, pp. 215–225.

Alessi, Lucia, Antonio Antunes, Jan Babecky, Simon Baltussen, Markus Behn, Diana Bonfim, Oliver Bush, Carsten Detken, Jon Frost, Rodrigo Guimaraes, Tomas Havranek, Mark Joy, Karlo Kauko, Jakub Mateju, Nuno Monteiro, Benjamin Neudorfer, Tuomas Peltonen, Paulo Rodrigues, Marek Rusnák, Willem Schudel, Michael Sigmund, Hanno Stremmel, Katerina Smidkova, R. Van Tilburg, Borek Vasicek, and Diana Zigraiova, 2015, "Comparing Different Early Warning Systems: Results from a Horse Race Competition among members of the Macro-Prudential Research Network," Unpublished manuscript.

Athey, Susan, Mohsen Bayati, Guido Imbens, and Zhaonan Qu, 2019, "Ensemble methods for causal effects in panel data settings," *AEA Papers and Proceedings*, Vol. 109, pp. 65-70.

Azur, Melissa, J., Elizabeth A. Stuart, Constantine Frangakis, and Philip, J. Leaf, 2011, "Multiple Imputation by Chained Equations: What is it and how does it work?" *International Journal of Methods of Psychiatric Research*, Vol. 20(1), pp 40–49.

Baba, Chikako, Salvatore Dell'Erba, Enrica Detragiache, Olamide Harrison, Aiko Mineshima, Anvar Musayev, and Asghar Shahmoradi, 2020, "How Should Credit Gaps Be Measured? An Application to European Countries," IMF Working Paper 20/6 (Washington: International Monetary Fund).

Basu, Suman S., Marcos Chamon, and Christopher Crowe, 2017, "A Model to Assess the Probabilities of Growth, Fiscal, and Financial Crises," IMF Working Paper 17/282 (Washington: International Monetary Fund).

Basu, Suman S., Roberto A. Perrelli, and Weining Xin, Forthcoming, "External Crisis Prediction Using Machine Learning: Evidence from Three Decades of Crises Around the World," IMF Working Paper (Washington: International Monetary Fund).

Berg, Andrew, Eduardo Borensztein, and Cathy Pattillo, 2005, "Assessing Early Warning System: How Have They Worked in Practice?" *Staff Papers,* International Monetary Fund, Vol. 52(3), pp. 462–502.

Beutel, Johannes, Sophia List, and Gregor Von Schweinitz, 2018, "An Evaluation of Early Warning Models for Systemic Banking Crises: Does Machine Learning Improve Predictions?" Discussion Paper No. 48, Deutsche Bundesbank.

Breiman, Leo, 2001, "Random Forests," *Machine Learning*, Vol. 45(1), pp. 5-32.

Chamon, Marcos, Paolo Manasse, and Alessandro Prati, 2007, "Can We Predict the Next Capital Account Crisis?" *Staff Papers,* International Monetary Fund, Vol. 54(2), pp. 270–305.

Chawla, Nitesh, Kevin Bowyer, and Lawrence Hall, 2002, "SMOTE: Synthetic Minority Over-Sampling Technique," *Journal of Artificial Intelligence,* Vol. 16, pp. 321–357.

©International Monetary Fund. Not for Redistribution

Chernozhukov, Victor, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, and Whitney Newey, 2017, "Double/Debiased/Neyman Machine Learning of Treatment Effects," *American Economic Review*, Vol. 107(5), pp. 261–65.

Dabla-Norris, Era, and Yasemin Bal Gunduz, 2014, "Exogenous Shocks and Growth Crises in Low-Income Countries: A Vulnerability Index," *World Development,* Vol. 59, pp. 360–378.

Dornbusch, Rudiger, Ilan Goldfajn, and Rodrigo O. Valdes, 1995, "Currency Crises and Collapses," *Brookings Papers on Economic Activity*, Vol. 2, pp. 219–293.

Eichengreen, Barry, and Andrew K. Rose, 1998, "Staying afloat when the wind shifts: External factors and emerging-market banking crises," *National Bureau of Economic Research*, No. 6370.

Eichengreen, Barry, Andrew K. Rose, and Charles Wyplosz, 1995, "Exchange Market Mayhem: The Antecedents and Aftermath of Speculative Attacks," *Economic Policy*, Vol. 10(21), pp. 249–312.

Flood, Robert P. and Peter M. Garber, 1984, "Collapsing Exchange-rate Regimes: Some Linear Examples," *Journal of International Economics*, Vol. 17(1-2), pp. 1–13.

Frankel, Jeffery, and George Saravelos, 2012, "Are Leading Indicators of Financial Crises Useful for accessing Country Vulnerability? Evidence from the 2009-09 Global Financial Crisis," *Journal of International Economics*, Vol. 87(2), pp. 216–231.

Gelos, R. Gaston, Ratna Sahay, and Guido Sandleris, 2004, "Sovereign Borrowing by Developing Countries: What Determines Market Access?" IMF Working Paper 04/211 (Washington: International Monetary Fund).

Ghosh, Atish R., and Swati R. Ghosh, 2003, "Structural Vulnerabilities and Currency Crises," *Staff Paper*, International Monetary Fund, Vol. 50(3), pp. 481–507.

Guscina, Anastasia, Malik Sheheryar, and Michael Papaioannou, 2017, "Assessing Loss of Market Access: Conceptual and Operational Issues," IMF Working Paper 17/246 (Washington: International Monetary Fund).

Hall, Patrick, and Navdeep Gill, 2019, *An Introduction to Machine Learning Interpretability, 2nd Edition* (Sebastopol: O'Reilly Media, Inc.)

Hellwig, Klaus-Peter, 2020, "Predicting Fiscal Crises: A Machine Learning Approach," Unpublished manuscript.

Holopainen, Markus, and Peter Sarlin, 2017, "Toward Robust Early-warning Models: A Horse race, Ensembles and Model Uncertainty," *Quantitative Finance*, Vol. 17(12), pp. 1933–1963.

International Monetary Fund, 2001, "Early Warning Systems in Fund Work," SM/01/306, October (Washington).

———, 2007, "Assessing Underlying Vulnerabilities and Crisis Risks in Emerging Market Countries–A New Approach," SM/07/328, September 17 (Washington).

———, 2010, "The IMF-FSB Early Warning Exercise: Design and Methodological Toolkit," IMF Policy Paper (Washington).

———, 2011, Managing Volatility: A Vulnerability Exercise for Low-Income Countries," Available from www.imf.org/external/np/pp/eng/2011/030911.pdf.

———, 2013, "Early Warning System Models: The Next Steps Forward," May, Chapter 4, (Washington).

©International Monetary Fund. Not for Redistribution

Joy, Mark, Marek Rusnak, Katerina Smidkova, and Borek Vasicek, 2017, "Banking and Currency Crises: Differential Diagnostics For Developed Countries," *International Journal of Finance and Economics*, Vol. 22, pp. 44–67.

Kaminsky, Graciela L., and Carmen M. Reinhart, 1999, "The Twin Crises: The Causes of Banking and Balance-of-Payments Problems," *American Economic Review*, Vol. 89(3), pp. 473–500.

Krugman, Paul, 1979, "A Model of Balance-Of-Payments Crises," *Journal of Money, Credit and Banking,* Vol. 11(3), pp. 311-325.

Laeven, Luc, and Fabián Valencia, 2018, "Systemic Banking Crises Revisited," IMF Working Paper 18/206 (Washington: International Monetary Fund).

Lahiri, Soumendra N., 2003, *Resampling Methods for Dependent Data* (New York: Springer).

Manasse, Paolo, Roberto Savona, and Marika Vezzoli, 2016, "Danger Zones for Banking Crises in Emerging Markets," *International Journal of Finance and Economics*, Vol. 4, pp. 360–381.

Medas, Paulo, Tigran Poghosyan, Yizhi Xu, Juan Farah-Yacoub, and Kerstin Gerling, 2018, "Fiscal Crises," *Journal of International Money and Finance*, Vol. 88, pp. 191–207.

Mendoza, Enrique G., 2002, "Credit, Prices, and Crashes: Business Cycles with a Sudden Stop. In Preventing Currency Crises in Emerging Markets," in *Preventing Currency Crises in Emerging Markets*, ed. by Sebastian Edwards and Jeffrey A. Frankel (Chicago: University of Chicago Press), pp. 335–392.

Molnar, Christoph, 2019, "Interpretable Machine Learning: A Guide for Making Black Box Models Explainable," https://christophm.github.io/interpretable-ml-book/.

Moreno Badia, Marialuz, Paulo Medas, Pranav Gupta, and Yuan Xiang, 2020, "Debt Is Not Free." IMF Working Paper 20/1 (Washington: International Monetary Fund).

Nag, Ashok, and Amit Mitra, 2002, "Forecasting Daily Foreign Exchange Rates Using Genetically Optimized Neural Networks," *Journal of Forecasting,* Vol. 21(7), pp. 501–511.

Obstfeld, Maurice, 1996, "Models of Currency Crises with Self-Fulfilling Features," *European Economic Review*, Vol. 40(3-5), pp. 1037–1047.

Oh, Kyong Joo, Tae Kim, and Chiho Kim, 2006, "An Early Warning System for Detection of Financial Crisis using Financial Market Volatility," *Expert System*, Vol. 23(2), pp. 83–98.

Politis, Dimitris N., and Joseph P. Romano, 1994, "The Stationary Bootstrap," *Journal of the American Statistical Association*, Vol. 89(428), pp. 1303–1313.

Robinson, David, 2014, "The IMF Response to the Global Crisis: Assessing Risks and Vulnerabilities in IMF Surveillance," IMF Working Paper 14/09 (Washington: International Monetary Fund).

Savona, Roberto, and Marika Vezzoli, 2015, "Fitting and Forecasting Sovereign Defaults using Multiple Risk Signals," *Oxford Bulletin of Economics and Statistics*, Vol. 77(1), pp. 0305–9049.

Shapley, Lloyd S., 1953, "A Value for N-Person Games," in *Contributions to the Theory of Games II*, ed. by Harold Kuhn, and Albert Tucker (Princeton: Princeton University Press), pp. 307–317.

Tiffin, Andrew, 2019, "Machine Learning and Causality: The Impact of Financial Crises on Growth," IMF Working Paper 19/228 (Washington: International Monetary Fund).

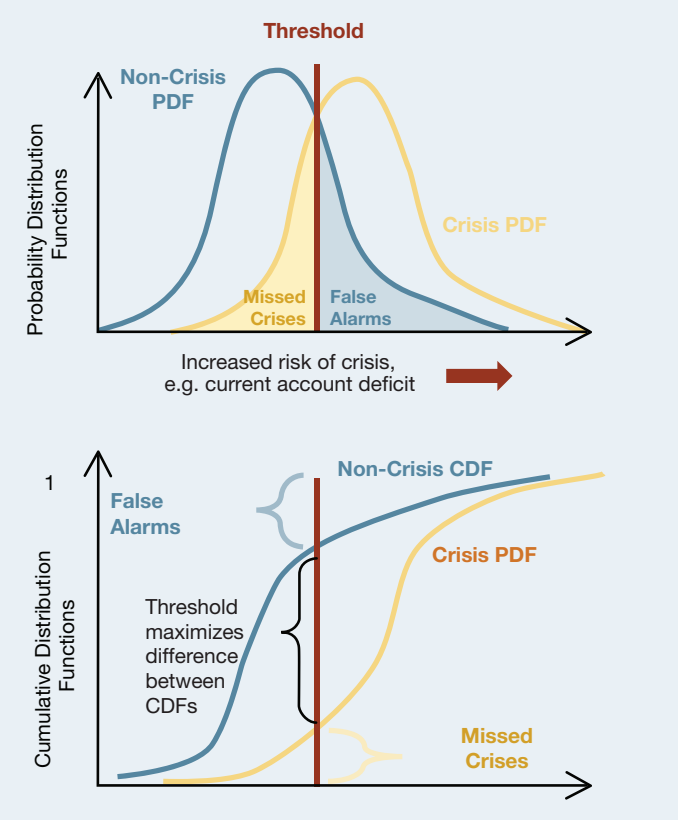
©International Monetary Fund. Not for Redistribution

Varian, Hal R., 2014, "Big Data: New Tricks for Econometrics," *Journal of Economic Perspectives*, Vol. 28(2), pp. 3–28.

Weisfeld, Hans, Irineu de Carvalho Filho, Fabio Comelli, Rahul Giri, Klaus Hellwig, Chengyu Huang, Fei Liu, Sandra Lizarazo Ruiz, Alexis Meyer Cirkel, and Andrea Presbitero, 2020, "Predicting Macroeconomic and Macro Financial Stress in Low-Income Countries." IMF Working Paper 20/289 (Washington: International Monetary Fund).

©International Monetary Fund. Not for Redistribution

ANNEX I. SIGNAL EXTRACTION METHOD[45]



The signal extraction model establishes thresholds on an indicator-by-indicator basis. Observations are assessed based on the number of indicators for which they exceed the threshold, and the signaling importance of these indicators.

For each indicator, a threshold is defined to flag elevated vulnerability. The threshold is chosen to minimize the sum of the percentage of crises missed and the percentage of non-crises falsely flagged as a crisis (false alarms). This is equivalent to maximizing the difference between the cumulative distribution functions of the crisis and non-crisis samples (see figure). Crises are relatively rare in the data, so this definition captures the notion that missing a crisis observation is much more costly than issuing a false-alarm (e.g., if crises are 5 percent of the sample, missing one crisis is as costly as issuing 19 false-alarms), though in general alternative weights can be chosen at the modeler's discretion. Countries receive a 1 if their value of the indicator falls on the risky side of the threshold and a zero otherwise.

Indicator results are aggregated by their signaling power. Ones and zeros for each indicator are typically averaged with weights given by their signal to noise ratio – defined as (1-z)/z, where z is the sum of the fractions of false alarms and missed crises. When there is an extensive literature on the relative importance of different crisis indicators, judgment can also be used to determine the weights for aggregation.

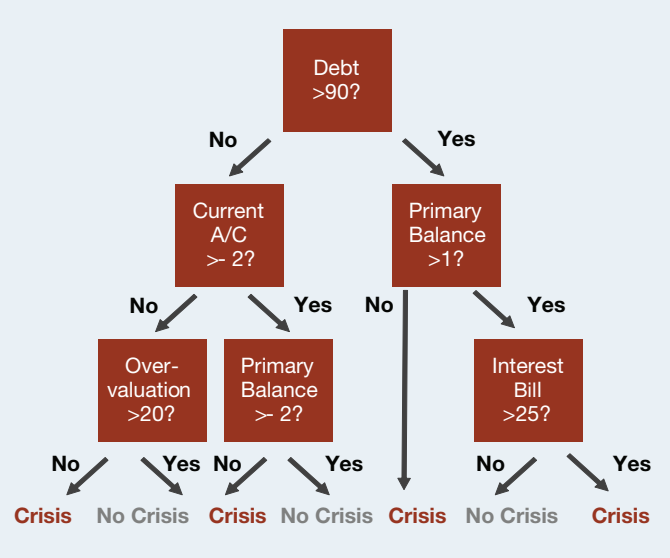
The model is well-tailored to heterogeneous pools of macroeconomic data. The use of thresholds keeps results robust to outliers as the center of the distribution determines the risk assessment. The aggregation procedure also easily accommodates missing data, allowing the inclusion of additional indicators where available without limiting country coverage.

The Signal Extraction Model is a machine learning approach. With no data generating process underlying it, the signal extraction approach (a central element of the VE for more than a decade), is a machine learning method hiding in plain sight. Indeed, it bears a strong resemblance to boosted trees methods (Annex IV) with very short trees. Its popularity is due in part to its success in real time forecasting in the late 90's and early 2000's (Berg et. al, 2005).

[45] This annex is an edited version of the box that appeared in the previous methodology note by Ahuja and others (2017).

©International Monetary Fund. Not for Redistribution

ANNEX II. FROM DECISION TREES TO A RANDOM FOREST[46]



Tree-based methods provide an intuitive, easy-to-implement way of modeling non-linear relationships. At core, these methods are based on the notion of a *decision tree*; which aims to deliver a structured set of yes/no questions to predict a particular outcome (e.g., the likelihood of a crisis in the next two years). One of the key attractions of decision trees is that they can take an extremely complex, non-linear problem, with a wide range of potential predictor variables, and reduce it to a procedure that is easily understood by a non-technical user. Imagine a flowchart, where each level is a question with a yes or no answer. Following the chart, and answering the questions one by one, eventually the chart will give you a solution to your initial problem. That is a decision tree. The challenge is to come up with the right questions.

A traditional econometric method would usually center around a logit or probit model. But decision trees take a very different approach. Rather than fitting a (transformed) linear regression, they are focused instead around the repeated partitioning of the predictor space into two sets, starting with an initial split that decreases the prediction error the most: i.e., the algorithm considers every possible split on every possible predictor variable, and chooses the one split on the one variable that best separates the sample into the two most dissimilar subsamples (based on the predicted outcome). These binary partitions then continue until the termination of the tree, and are recursive—i.e., each subsequent split only considers the subsample under which it falls, rather than the whole dataset. The result is an efficient set of yes/no questions that can quickly narrow down the likelihood of an outcome falling into a one category ("crisis in two years") or another ("no crisis in two years").

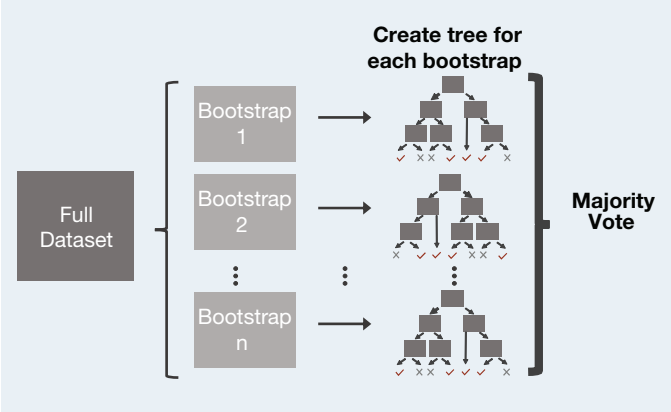
Decision trees are computationally efficient and work well for problems where there are important nonlinearities and interactions. They also are well suited to cope with missing data. Trees tend not to work as well if the underlying relationship is linear, but even in these cases they can reveal aspects of the data that are not apparent from a traditional linear approach (Varian, 2014). A number of papers in the 2000's applied this method to crisis forecasting, beginning with Ghosh and Ghosh (2003). See the related work section for more examples.

The Random Forest (RF) algorithm modifies the decision-tree approach to minimize the problem of overfitting. One problem with decision trees is that they often provide models that fit the training sample well, but perform poorly when making out-of-sample predictions. A common solution is to shorten or "prune" the tree by imposing a penalty for an overly long/complex structure. The ideal degree

[46] Annex prepared by Andrew Tiffin.

©International Monetary Fund. Not for Redistribution

of complexity is then chosen using cross-validation techniques. Instead of pruning, however, the RF algorithm (Breiman, 2001a) modifies the decision-tree approach—seeking instead to improve the model's predictive ability by growing numerous (unpruned) trees and combining the results.



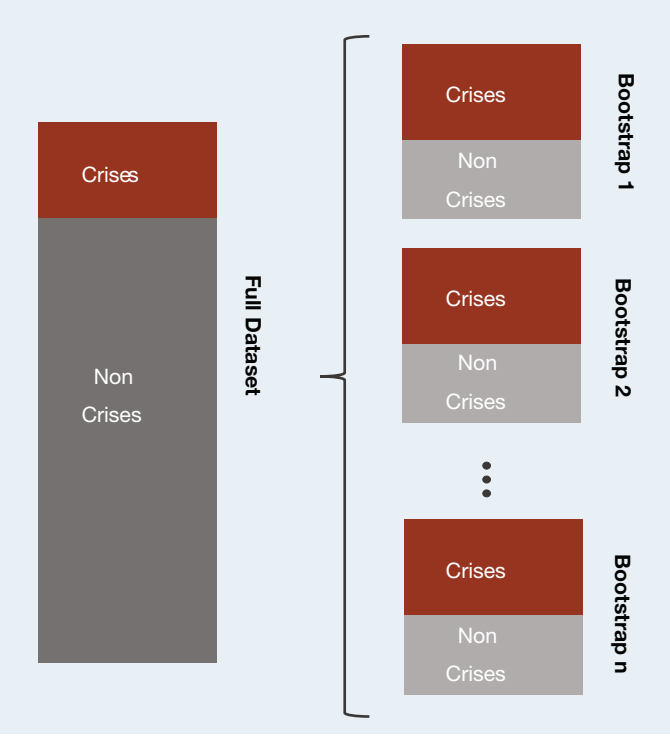
The first Random Forest modification is the use of bootstrap aggregation (or "bagging"). In bagging, an individual tree is built on a random sample of the dataset, roughly two thirds of the total observations—the remaining one-third are referred to as out-of bag (OOB) observations and can be used to gauge the accuracy of the tree. This is repeated hundreds or thousands of times. When asked to predict the most likely outcome of a new instance, then, the RF algorithm will feed that instance through each of these individual trees, and will aggregate their predictions; say by taking the majority vote. The fact that none of the trees is pruned means that each individual tree is a weak model that will have a hard time distinguishing the dataset's underlying *signal* from simple statistical *noise*. However, by building a large ensemble of (weak) individual trees, the algorithm is essentially exploiting the law of large numbers to average out the noise, leaving only the signal.

The second modification is to take a random sample of the set of predictors at each split. In the case of highly correlated predictors, and particularly in the event of a single driving predictor, bagging by itself can be insufficient, as it may simply produce multiple versions of the same tree. To get around this problem, RF introduces an added element of randomization—at each split, the algorithm only considers a random subset of the available set of predictors (usually the total number of predictors divided by three). By randomizing the predictor space, the RF algorithm effectively guarantees that the trees that go into the final collection will be relatively diverse. Again, each tree on its own will be a weak model, as it is grown on a deliberately limited dataset. But the essence of the RF approach is that, by combining a large number of (uncorrelated) weak models, we can end up with an aggregate prediction that is surprisingly strong.

Random Forest is one of the most popular and successful general-purpose algorithms currently available. Random Forests require almost no input preparation, as they can handle a range of different predictor types (binary, categorical, numerical) without any need for scaling. The procedure implicitly incorporates an element of feature selection, and provides a decent indicator of the relative importance of each variable. Random Forests are also quick to train and can be applied to a wide range of modeling tasks, ranging from classification, to regression, to cluster analysis. Moreover, the predictive performance of Random Forests is usually impressive. Although there is no single algorithm that will dominate in all applications, Random Forests will usually do well and will often take significantly less time and effort to train than most alternative candidates. This is why Random Forests are often used as a benchmark model.

©International Monetary Fund. Not for Redistribution

ANNEX III. BEYOND RANDOM FORESTS: BALANCED FOREST AND BOOSTING[47]



Many classification algorithms have difficulty coping with imbalanced datasets. These are datasets where one class constitutes a very small minority of the data. This is a potential issue for crisis prediction, as crises are rare and those assessing risk are often more interested in accurately predicting the onset of a crisis, rather than predicting a non-crisis (if crises only make up 2 percent of the sample, accurately predicting a non-crisis is trivial, if not particularly helpful). Although the Random Forest algorithm is relatively robust to this problem, it nonetheless still tries to minimize the overall error rate, so that the larger class will tend to get a low error rate while the smaller class will have a larger error rate. (More concretely, with an extremely imbalanced dataset, there is a significant chance that an individual bootstrap sample will contain few or even none of the minority class, resulting in a tree with poor performance when predicting the minority class).

Balanced Random Forests. One popular solution is to modify the Random Forest algorithm so that each bootstrap replication has a less skewed composition. In the "Balanced Random Forest" procedure, this is done by downsampling; where instances of the majority class are deliberately underrepresented to bring their frequency closer to that of the minority class. While a general disadvantage of downsampling is a potential loss of information, this is less of a concern for Random Forests, as repeated bootstrapping ensures that the entire sample is covered. The result is a model that has all the advantages of a Random Forest, but which also pays sufficient attention to the prediction of rare events.

Boosting. An alternate approach to predicting rare events entails the process of "boosting." Similar to Random Forests, this entails the aggregation of a large number of weak models (trees). But rather than averaging over all models at once, boosting is a sequential ensemble algorithm, in which the models are constructed one at a time, and in which each model aims to learn from the mistakes of the previous one.

[47] Annex prepared by Andrew Tiffin.

©International Monetary Fund. Not for Redistribution

For example, **gradient boosting (XGBoost)** starts out by training an initial decision tree on the full dataset. Taking the predictions of the first tree, a second tree is then trained to predict the errors from the first. A third tree is then trained to predict the residual errors from the second, and so on. The final prediction is then the sum of the individual predictions from all of the trees.

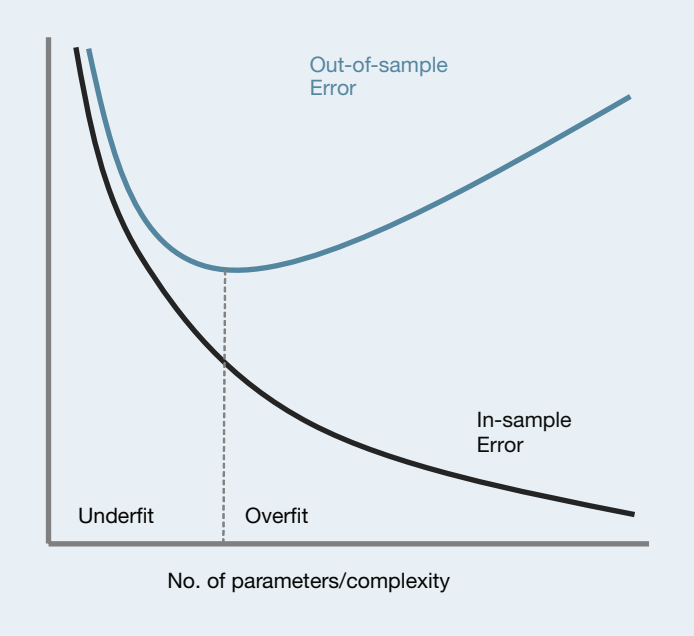
Adaptive boosting (ADABoost) takes a slightly different approach. Instead of predicting the errors of the previous model, each iteration tries instead to generate predictions based on a reweighted dataset, where the weights are determined by the previous model. More weight is given to instances that the previous model handled poorly, and less is given to those it handled well. The final prediction is a weighted sum of all the models, with weights determined by each model's accuracy.

Both boosting approaches are well suited to imbalanced datasets, as increased attention is given to the minority class at each successive iteration, given that instances from this class are often misclassified. As a further modification, however, the same downsampling techniques used to improve Random Forests can also be used to further improve boosting.

Random Under Sampling Boosting (RUSBoost) takes the adaptive boosting approach, but trains each model on a randomly downsampled version of the (reweighted) dataset; where again, the majority class is underrepresented in order to balance the frequency of the two classes and increase the attention that the overall model gives to accurately predicting rare events.

©International Monetary Fund. Not for Redistribution

ANNEX IV. PREDICTING OUT-OF-SAMPLE PERFORMANCE: CROSS VALIDATION[48]



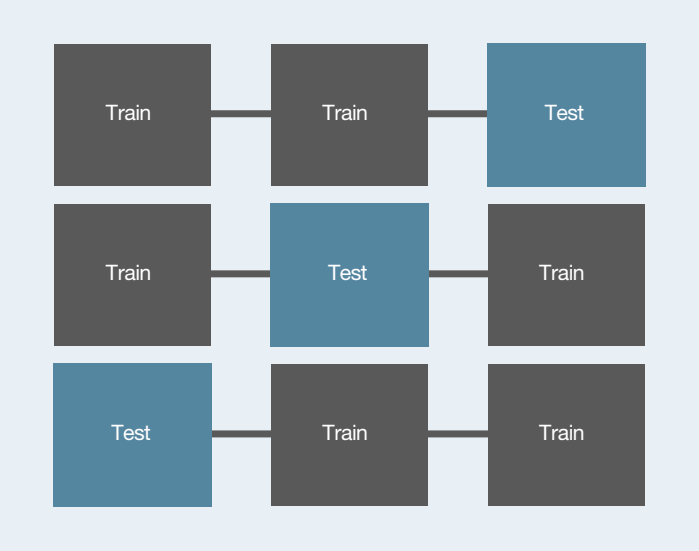
Fitting is easy. Prediction is hard. And crisis prediction is *particularly* hard, given that the likelihood of a crisis is shaped by the influence (and interaction) of a broad range of potential economic factors. In such circumstances, focusing on a model's in-sample fit is insufficient. Indeed, within the machine-learning literature it is stressed that in-sample fit generally tells us little of value, other than the number of parameters in a model (it is always possible to improve in-sample fit by adding more parameters). The key danger is that a model with a supposedly good in-sample fit may in fact be modeling the idiosyncratic noise within a particular dataset. When taken out of sample, then, the model will perform poorly. Such a model is said to be "overfit". This is the core issue that machine learning seeks to address—in fact, the entire field of machine learning centers around the design of experiments that evaluate how well a model trained on one dataset will predict new data.

Estimating future performance: Holdout validation. The process of predicting how a model will perform on new data is called model *validation*. In *holdout validation*, the data is split into a training and testing set. The model is generated using only the training set, and is then asked to make predictions using inputs from the test set. The *validation error* is the difference between these predictions and the actual test-set outcomes, and serves as a gauge of likely out-of-sample performance going forward. This error can be used to help choose between different models, or indeed, between different versions of the same type of model (e.g., a penalized regression with a heavy penalty weight vs. one with a lighter penalty weight.)

[48] Annex prepared by Andrew Tiffin.

©International Monetary Fund. Not for Redistribution

Estimating future performance: Cross validation. As an alternative approach, *cross validation* takes advantage of the entire data set. The basic idea is simple: (i) First divide the entire dataset into K folds (say, K=3), take one of those folds and set it aside as a test set. (ii) Using the remaining (2) folds as a training set, estimate the model, and then use the test set to determine the model's prediction error. (iii) Repeat this procedure using all combinations of the test and training sets, producing an array of (3) validation errors associated with that particular model, which then provides a gauge of its average out-of-sample performance. Once again, this metric can be used to help choose between different types of models. For models with pre-determined settings (e.g. the penalty weight on a penalized regression), cross validation can help determine the setting that optimizes likely out-of-sample performance. In machine-learning parlance, these settings are called "hyperparameters," and are "tuned" to minimize the cross-validation error.

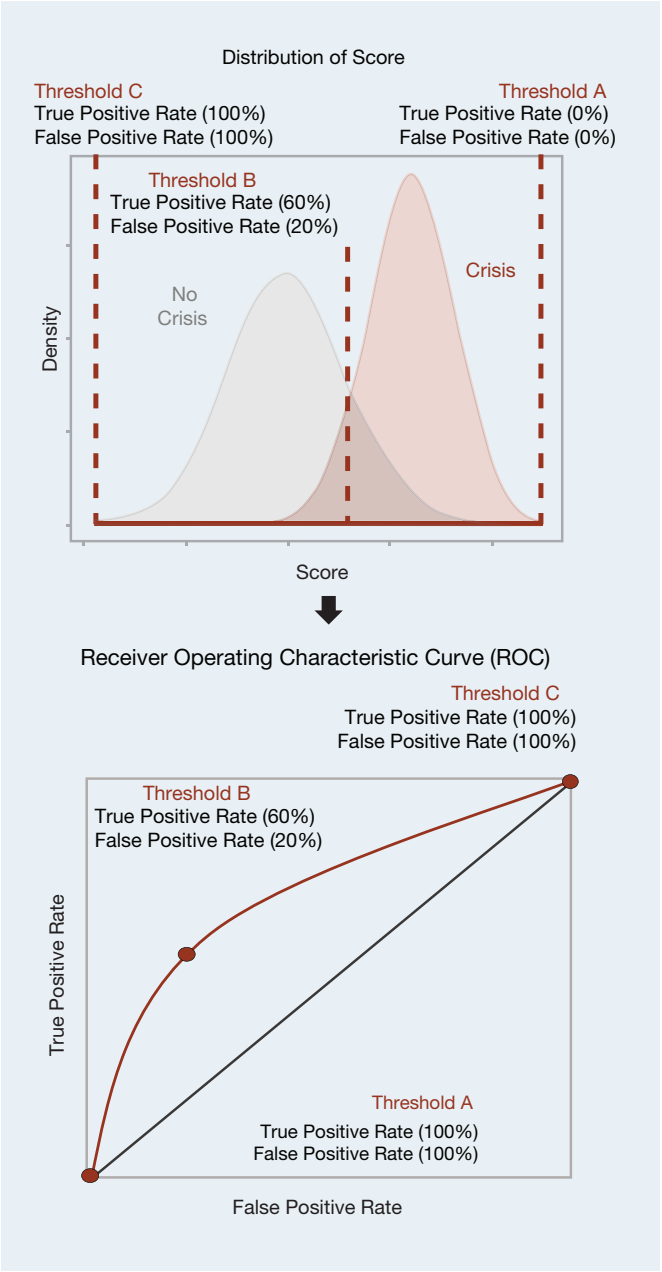


©International Monetary Fund. Not for Redistribution

ANNEX V. COMPARING CLASSIFIERS: AREA UNDER THE CURVE (AUC)[49]

In machine-learning classification models, a standard measure of model accuracy is AUC or Area Under the Curve. In this case, the "curve" is the Receiver Operating Characteristic (ROC) curve, which has its origins in Radar engineering in World War II, but is generally applicable to any model that seeks to predict a discrete (binary) outcome. Early warning modelers consider a range of different crisis-prediction models, and the AUC provides a guide as to each model's ability to distinguish between countries that have gone on to a crisis versus those that have not.

The ROC curve illustrates a model's accuracy across a range of different thresholds. Most classification algorithms will compute a classification score, and will then arrive at a prediction (crisis/no crisis) based on whether that score is above a certain threshold. Different thresholds will give different results. An extremely high threshold, for example, will fail to predict any crises, regardless of a country's circumstances. In that case, although the false positive rate will be zero, the true positive rate will also be zero. The ROC curve plots the tradeoff between true positives and false positives as the threshold changes.



[49] Annex prepared by Andrew Tiffin.

©International Monetary Fund. Not for Redistribution

The area under the ROC curve (AUC) then provides a broad summary of a model's performance, which does not depend on the actual threshold chosen. Ranging from 0 to 1, the higher the AUC, the better the model's overall performance. Moreover, the metric is sufficiently general to allow a comparison between very different types of algorithm, as it only requires the model to rank observations according to their likelihood of falling into one category or another. Intuitively, the AUC answers the following question: *"If two observations were chosen at random from the sample, one from each category, what is the probability that the model will rank them correctly*." An AUC of 0.5 is poor, as it suggests that the model is little better than flipping a coin. An AUC of 1 suggests that the model is perfect, as it can correctly distinguish between crisis and non-crisis cases 100 percent of the time. Within the machine-learning literature, a "good" model will typically aim to achieve an AUC of 0.8 or higher, but clearly this depends on the problem at hand. Predicting crises out of sample is difficult, and so the AUC for many crisis-prediction models may be lower. Nonetheless, comparing AUC scores will allow us to identify the most accurate model possible.

©International Monetary Fund. Not for Redistribution

ANNEX VI. ASSIGNING THE BLAME: SHAPLEY VALUES[50]

Prediction is hard. Explaining a prediction is even harder. This is particularly problematic in the case of complex models, where certain variables (e.g., oil prices) may only matter for certain types of countries (e.g., oil exporters), or where the impact of a particular variable (external debt) may depend on the value of another variable (global interest rates), or indeed where the final prediction is the result of an *ensemble* of thousands of sub models. In such circumstances there is often some tension between a model's *accuracy* and its *interpretability*. In response, various methods have been proposed to help users interpret the predictions of complex models.

One simple method of determining the importance of a variable is to turn it into noise and see what happens. Take a particular variable, randomly reassign all of the observations to different country-years and see how the fit of the model deteriorates. A large loss in explanatory power would indicate that the variable is important in identifying crisis risk. This is a global method which evaluates variables at the level of the model, but it can also be used to understand individual assessments. For each country-year pair, one can compare the model's risk assessment with the expected risk assessments when that variable is randomly replaced by the value from a different observation. This calculation tells us the marginal importance of the variable for an individual assessment and is easy to understand.

This type of assessment breaks down when interaction terms are important. Imagine that two variables are strong predictors of a crisis only when both are high, as in the external debt and global interest rates example above. An observation in which both are high would see a big fall in its risk assessment under either an average external debt level or an average level of global interest rates. Under this method the implied effect of the two variables adds up to more than the total risk assessment, because the model (and the world) is not additive. One promising solution to this problem has its origins in cooperative game theory.

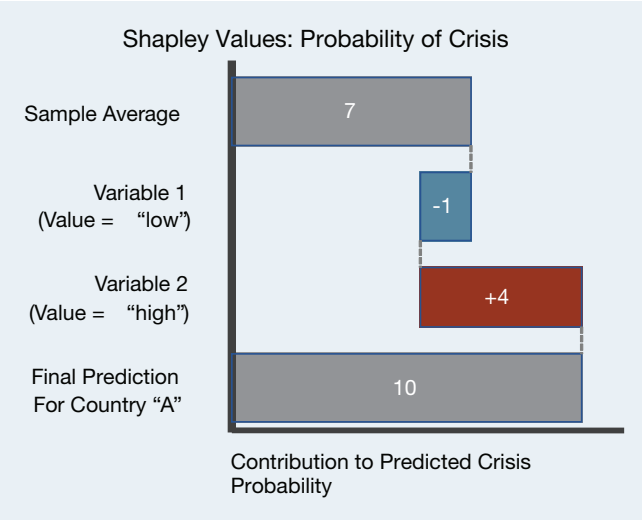
Shapley values initially came out of a core question in coalitional game theory: in a team of multiple players with differing skill sets, what is the fairest way to allocate a collective payoff? One solution is to imagine the players joining in sequence, and then keeping track of their marginal contribution. But what if some players, say Alice and Bob, have similar skill sets? Then, it might be the case that Alice would have a higher marginal contribution if she joined the group before Bob, as she would be the first one to provide their overlapping skill set. When Bob joined, his marginal contribution would be lower. The Shapley Value concept was developed in response to this problem and can be understood as *finding each player's marginal contribution, averaged over every possible sequence in which the players could have been added to the group.* To take the simplest example, suppose Alice and Bob are sharing a taxi, and Alice lives on the route to Bob's house. Their marginal

	Alice Pays	Bob Pays
Alice goes first	$6	$36
Bob goes first	$0	$42
Shapley	$\frac{\$0 + \$6}{2} = \$3$	$\frac{\$36 + \$42}{2} = \$39$

[50] Annex prepared by Andrew Tiffin.

©International Monetary Fund. Not for Redistribution

contribution to the cost of the taxi ride will depend on the order in which their claims are considered. The Shapley Value, however, will average these contributions over all conceivable sequences (in this case there are only two) to arrive at the fairest possible allocation.



This concept can be used to explain the contributions of different variables to an individual prediction. The "payoff" is the actual prediction for a particular instance *less* the average prediction for the entire dataset. The "players" are the values of each variable that fed into that prediction, which "collaborate" to produce the payoff. Shapley Values divide this prediction (payoff) among the variables (players) in a way that fairly represents their contributions across all possible subsets of variables. In the case of a crisis prediction model, for example, where a country's crisis probability is higher than for the sample as a whole, Shapley Values will indicate which variables prompted the model to assign a higher probability for that country (compared to the sample average) and will provide a quantitative guide as to each variable's relative contribution to that country's prediction. Since Shapley values divide credit for joint work among contributing variables, the value for a particular variable (e.g., external debt) can be different for two countries with the same value of that variable but different interacting variables.

©International Monetary Fund. Not for Redistribution

ANNEX VII. FILLING IN THE GAPS: DEALING WITH MISSING DATA[51]

Missing data is a regular feature of most empirical analysis. And this is a particular problem for the VE, where the goal is to provide coverage over the broadest possible range of countries, over the longest available time period, and over a wide variety of potential predictors. Faced with missing data, the researcher has four general options.

1. ***Drop all observations with missing variables.*** This frequently entails an unacceptable loss of information, severely curtailing the coverage of the model and rendering the resulting predictions somewhat less reliable.

2. ***Drop all variables with missing observations.*** Again, this generally entails an undesirable loss of information, particularly if the variable is strongly related to the outcome. The decision on whether to drop a variable will then depend on an assessment as to how many observations are missing, and whether the remaining observations can nonetheless provide useful (non-misleading) information.

3. ***Simple Imputation.*** This approach typically replaces a variable's missing observations with a non-informative proxy—usually the variable's mean or median for numeric variables, or the mode for qualitative variables.

4. ***Model-Based Imputation.*** In this approach, the replacement for the missing observation is instead constructed using information from the entire dataset. In essence, in order to complete the dataset for the main VE model, the researcher is using a second background model—using all available information to provide plausible values for the dataset's missing observations. These models can be simple (e.g. linear regression) or complex (e.g. k-nearest neighbor, random forest, etc.). And in cases where more than one variable contains missing observations, the models are often employed iteratively in a chain—where the incomplete variables are filled in one-by-one, and where the imputations for each newly completed variable are added to the data for modelling the next incomplete variable.52

The first approach ("complete case analysis ") would dramatically reduce sample sizes and model performance. Dropping all incomplete observations would eliminate 80% of the sample for the EMP AE model, 89% for EMP EM, almost 98% for EMP LIC, and more than 98% for the financial sector model. This results in poor model performance, for example in the fiscal sector complete case analysis would reduce the fiscal model's AUC by 10 percent for low income countries where data availability is most challenging.

Most of the VE models use a combination of the second and third approach. In the fiscal model, for example, variables are dropped if more than one-third of their observations are missing, and any remaining gaps are imputed using the variable's median value. The median is preferred, as it is robust to outliers. More importantly, the simple-imputation approach is less likely to complicate communication of the model's results. For example, suppose that a country were to have a missing value for international reserves in a particular year. Using the sample median as a proxy would effectively prevent reserves from contributing to that country's crisis score in that year (as the value would not differ from the sample and so would be less likely to generate a significant Shapley value).

[51] Annex prepared by Andrew Tiffin.

[52] For an introduction to this literature, see Azur and others (2011).

©International Monetary Fund. Not for Redistribution

The breakdown would focus instead on the *remaining* (complete) variables that push that country's predicted score away from the sample average. If, on the other hand, a model-based approach were used, the breakdown would be shaped by the *predicted* value of reserves, and this might then feature as a key determinant of the final score. The desk would then have the task of understanding not only how the primary crisis model interprets that country's (predicted) level of reserves, but also how the secondary background model arrived at the prediction in the first place.

There is no ideal way of dealing with missing values, and different approaches have different strengths. But, given the nature of the available data, and for the purposes of the VE—where the ultimate object is to aid desks in their assessment of crisis risks—the VE teams have generally opted for a simpler, more robust approach. That said, each model faces a specific missing-data challenge, and the details of their chosen solutions are provided in their respective working papers.[53]

[53] For example, the signaling approach chosen for the external model is proof against missing values, but imputation is nonetheless employed to ensure consistent Shapley values.

©International Monetary Fund. Not for Redistribution

ANNEX VIII. FISCAL SECTOR EXPLANATORY VARIABLES

FISCAL

- General government expenditures in percent of GDP
- General government primary expenditures in percent of GDP
- Overall balance in percent of GDP
- General government primary balance, percent of GDP
- General government revenues in percent of GDP

GLOBAL

- Percent change of crude oil price
- Percent change of Non-fuel price
- Percent change of food price
- US T-Bill rate Percent
- VIX Index Period Average
- VIX Index Period End
- Percent change of VIX Index Period Average
- Percent change of VIX Index Period End
- US T-Note 5-year rate Percent, Period Average
- US T-Note 10-year rate Percent, Period Average
- US T-Note 5-year rate Percent, End of Period
- US T-Note 10-year rate Percent, End of Period
- World real GDP growth, in percent
- Geometric average of the last 3-year world GDP growth

EXTERNAL

- Net official development assistance in percent of GDP
- Current account balance in percent of GDP
- Export of goods and services in percent of GDP
- Import of goods and services in percent of GDP
- Personal remittances in percent of GDP
- Current account without import
- Net foreign direct investment in percent of GDP
- Other investment, net (loans, deposits, insurance, pensions, trade credits, SDR, percent of GDP
- Portfolio investment, net
- Percent change of exchange rate (NC units per U.S. dollar, period average Units)
- Exchange Rate, end of period
- Average of the last 10 year of the sum of export and import of goods and services
- Percent change in real exchange rate, period average
- Log of PPP exchange rate
- PPP overvalue
- Percent change in total reserves excluding gold in national currency
- Percent change in total reserves (number of months of imports)
- Percent change in terms of trade (of goods and services) Index
- Trading Partner Growth (Real GDP, 2005=100, local currency, Weighted by trade exports to all economies)
- Percent change of trading Partner Import Demand (Imports volume of goods and services, 2005=100, Weighted by trade exports to all)
- Net official development assistance in percent of GDP
- Current account balance in percent of GDP

DEBT-PRIVATE

- One sided credit gap based on the GDD loans and securities
- Total Debt, loans and securities, in percent of GDP

DEBT-PUBLIC

- General government short-term external debt in percent of GDP
- Public external debt in percent of GDP
- Public debt in percent of GDP
- Public debt in percent of general government revenue
- Public external debt to export
- General government interest expenses in percent of GDP
- Amortization of external public debt in percent of GDP
- Public debt service to revenue in percent
- Public debt service to export in percent

TOTAL DEBT

- Total debt in percent of GDP

©International Monetary Fund. Not for Redistribution

CRISIS HISTORY	EXTERNAL	VOLATILITY
• Number of crisis	• Export of goods and services in percent of GDP	• Standard deviation of real GDP growth
• Number of Crisis, Advanced and Emerging Economies	• Import of goods and services in percent of GDP	• Standard deviation of the change in terms of trade
• Number of Crisis, Emerging and Low-Income Economies	• Personal remittances in percent of GDP	• Standard deviation of the percent change in exchange rate (end of period)
• Number of Crisis, Advanced Economies	• Current account without import	• Standard deviation of the inflation
• Number of Crisis, Emerging Economies	• Net foreign direct investment in percent of GDP	• Standard deviation of the percent change in exchange rate (end of period)
• Number of Crisis, Low Income Economies	• Other investment, net (loans, deposits, insurance, pensions, trade credits, SDR, percent of GDP)	
	• Portfolio investment, net	

COUNTRY CATEGORY	REAL	INSTITUTIONS, ELECTIONS
• Dummy: Monetary Union	• Percent change in Material impact - all natural disaster hazards	• Revised Combined Polity Score (single regime score, runs from 1 (full democracy) to -1 (full autocracy))
• Dummy: Advanced economies	• Percent change of real GDP per capita	• Checks and balances index
• Dummy: Emerging market economies	• Percent change in real GDP	• Bureaucracy Quality
• Dummy: Low income economies	• Growth deviation from past 5-year average	• Corruption
• Dummy: Small developing state	• GDP growth rate	• Years remaining in current chief executive's term
• Dummy: Fuel exporter	• GDP growth rate relative to the past 5-year average growth rate	• Legislative election held dummy variable
• Dummy: Fragile state	• Geometric average of the last 3-year GDP growth	• Executive election held dummy variable
• Dummy: Fuel exporter or VELIC commodity exporter	• Percent change of Consumer Price Index, period average	
	• Percent change of CPI, end of period Units	
	• R minus G	
	• Log GDP per capita (PPP), relative to US	
	• Log GDP per capita (PPP)	
	• Log GDP in USD	
	• Log of population	

©International Monetary Fund. Not for Redistribution

TNM/21/03

International Monetary Fund
Strategy, Policy, and Review Department
700 19th Street NW
Washington, DC 20431
USA
T. +(1) 202.623.8554
F. +(1) 202.623.6073

ISBN-13: 978-1-51357-421-9