New Early-Warning Indicators of Macro-Financial Risk: A Mixture-of-Experts Approach*

Francisco Blasques[1], Alice Joan Brooks[2], Paola Castillo[2], Nepomuk Dunz[2], Erik Feyen[2], Andres Guzman-Cordero[1], Wissam Harake[2], Charl Jooste[2], Siem Jan Koopman[1], Ewa Korczyc[2], Jose Marzluf[2], Gabriele Mingoli[3], Frederico Gil Sander[2], Smriti Seth[2], and Entian Zhang[2]

[1]Vrije Universiteit Amsterdam, Tinbergen Institute, The Netherlands
[2]World Bank, Washington D.C., United States
[3]Aarhus University and Center for Research in Energy (CoRE), Denmark

Abstract

Predicting macro-financial crises in developing economies is a critical but difficult task, complicated by significant data challenges. Machine learning techniques have shown improved predictive performance relative to traditional parametric models, yet this often comes at the cost of lower model explainability, limiting their use for policy makers. To strike a balance between predictive performance and explainability, this paper proposes a Mixture-of-Experts framework that combines ensemble models and traditional logistic regression to generate country-level crisis risk scores which can be decomposed into sub-risks, such as fiscal, financial, and balance of payments, to aid understanding of the drivers of risk within and across countries. Leveraging a global panel dataset spanning from 1990 to 2025 for 161 countries, the Mixture-of-Experts framework is shown to meaningfully outperform standard crisis prediction models in the literature. This paper therefore offers a practical and flexible framework for technical audiences developing early-warning systems in the presence of data gaps.

JEL classification: C45, C82, G18, O11.
Keywords: Mixture of experts, machine learning, country risk, macro-financial crises, early-warning, out-of-sample prediction.

---

* Corresponding authors: efeijen@worldbank.org, wharake@worldbank.org, and s.j.koopman@vu.nl. This paper's findings, interpretations and conclusions are entirely those of the authors and do not necessarily represent the views of the World Bank, their Executive Directors, or the countries they represent. We are grateful to Michael Dupree Carson, Ralph van Doorn, Manuela Francisco, Aart Kraay, James Sampi, Sophie Sirtaine, and Francesco Strobbe for comments and suggestions. All errors and omissions are ours.

1. Introduction

The analysis and prediction of sovereign, financial, and economic crises has long been a key focus of macroeconomic and macro-financial research, as this can help identify and mitigate risks before they escalate. As such, policymakers and other stakeholders such as financial institutions are seeking effective early-warning systems that identify risks and their drivers, both at the global and country levels. However, building such systems presents significant challenges, particularly for emerging market and developing economies (EMDEs), where data limitations often prevail.

While traditional econometric models such as probit, logit, and Vector Autoregressive (VAR) models have long been used to forecast macro-financial risks and economic growth, their limits in dealing with complex datasets and non-linear interactions between variables have sparked interest in machine learning (ML) models and techniques. ML models are well known for their capacity to work with large datasets and deliver strong predictive performance (Cesa-Bianchi et al., 2019). Hence, ML models are increasingly used to predict macro-financial crises as the complexity of this task is eminent. These models can flexibly leverage a diverse range of data inputs to provide more accurate and timely predictions (Bussiere et al., 2015). ML-based models not only excel at integrating multiple data types, but also at dealing with imbalanced data sets, as these are typically encountered for countries with varying degrees of economic development and crisis frequency (Ahelegbey et al., 2016; Giannone et al., 2020). As a result, there is increased focus on ML approaches for macro-financial crisis prediction compared to traditional econometric methods. This is achieving gains in multi-year forecast accuracy (Bholat et al., 2016; Chen and Guestrin, 2016). Furthermore, ML methods may combine several predictors and account for complicated inter-dependencies across variables (De Mol et al., 2008).

Traditional econometric models have historically served as the foundation for anticipating sovereign and financial crises by analyzing economic indicators using well-established statistical methodologies.[2] For example, probit and logit models can predict the probability of a crisis by evaluating factors such as debt-to-GDP ratios, fiscal deficits, current account balances, and interest rate spreads (Kaminsky et al., 1998; Berg and Pattillo, 1999). Early-warning systems, such as those based on the Cox proportional hazards model, predict crisis-timings by taking into account the time-length until an incident happens (Bussiere and Fratzscher, 2006). The dynamic linkages between macroeconomic factors and their combined impact on sovereign risk are also commonly

[2] Traditional econometric models also have generally followed three main econometric approaches. The first revolves around partial or general equilibrium models which are highly theoretical with micro-foundations (Corsetti et al. 1998, Christiano et al. 2002). A second approach involves a cross-sectional analysis of a pool of financial crisis episodes. In these cases, authors estimate explanatory powers for independent variables on dependent binary indicators that capture either crisis episodes or no-crisis episodes. The third approach examines time-series relationships of indicators that capture market pressures and explanatory variables or determinants in one or more countries (see for example Barsutto and Gosh, 2000).

captured using VAR models (Eichengreen et al., 1996). These models are prized for providing interpretable insights into the ways in which certain economic conditions impact the likelihood of a crisis, by means of the impulse response function generated by VAR models.

To better account for non-linear effects between interacting variables and changing economic situations, econometric methodologies have been developed further in ways to incorporate lagged effects, interaction terms, and time-varying coefficients (Manasse and Roubini, 2009; Frankel and Saravelos, 2012). To improve model performance, regime-switching models have also been proposed, which allow for the effect of various economic conditions on crisis probabilities (Davis and Karim, 2008; Demirgüç-Kunt and Detragiache, 1998). Taken together, these traditional models continue to be crucial tools for policy makers, offering trustworthy and understandable frameworks for crisis prediction, despite difficulties in handling high-dimensional, complicated data sets and non-linearities.

More recent research, however, has provided evidence that ML models, such as ensemble models (e.g., random forests and gradient-boosted trees approaches), can outperform parametric models in terms of handling very large data sets, capturing complicated dynamics, and producing accurate forecasts of crises (Hellwig, 2021). For example, tree-based models appear to be better suited for evaluating crises in both developed and developing countries (Wiseman et al., 2021). This finding emphasizes the importance of using a wide range of indicators and appropriate data processing methods to maximize model performance (Batsuuri et al., 2024). However, in data-constrained environments such as low-income countries (LICs), machine learning approaches have often encountered challenges. Research indicates that simpler econometric models, like probit models, tend to be more effective in these settings. A study focusing on LICs found that probit models outperformed both the IMF-World Bank Debt Sustainability Framework (DSF) and random forests models, particularly in small samples (Luckner et al., 2023). These findings suggest that the relative strengths of ML and traditional econometric approaches may depend on data availability and sample characteristics, motivating the exploration of hybrid frameworks that leverage complementary advantages of both approaches to enhance crisis prediction.

Building on this body of work, this paper develops a novel ML approach, based on a Mixture-of-Experts (MoE) architecture, that forecasts macro-financial crises by combining both ML and conventional econometric methodologies. This hybrid framework is designed to strike a practical balance between predictive accuracy and the explainability required for credible policy advice.

The paper is intended for a technical audience—such as economists and data scientists at central banks, finance ministries, and international financial institutions—tasked with developing or enhancing country-level early-warning systems. The primary objective is not to present a definitive, one-size-fits-all model, but rather to propose and validate a flexible methodology for building custom early-warning frameworks.

The proposed MoE framework offers several distinct advantages. First, it directly addresses the common issue of data limitations prevalent in many EMDEs. By using a multi-layer framework modular architecture with multiple "expert" models, the framework can incorporate data of varying frequencies (annual, quarterly, monthly) and coverage, ensuring that predictions can be generated even when some data is missing. Second, the MoE approach enhances interpretability by decomposing a country's overall macro-financial risk into three distinct, policy-relevant sub-risks: fiscal, financial, and balance of payments (BoP). This disaggregation allows to understand not only the level of risk but also its primary drivers, which is essential for targeted interventions. The framework explicitly models the interconnection between these risks in a final, transparent stage, acknowledging that crises rarely occur in isolation.

However, this approach is not without its challenges. The complexity of the multi-layered architecture requires significant computational resources and expertise to implement and maintain. While the framework is designed to be robust to missing data, the quality and availability of input data remain critical constraints on its predictive performance. Furthermore, while the final risk aggregation is transparent, the underlying ML "expert" models can still be partially "black box," making the attribution of risk to specific indicators less direct than in purely econometric models.

This paper demonstrates the application of the MoE framework using a global panel dataset spanning 1990–2025 for 161 countries. The output that the MoE framework provides are country risk scores, which the probability of a macro-financial crisis occurring within the next 12 months. These scores are normalized on a scale from 0 to 10 and can also be broken down into three sub-risk components: fiscal, financial, and BoP. The results show that the presented MoE framework meaningfully outperforms standard crisis prediction models found in the literature. By doing so, the paper offers a practical and extensible blueprint for technical teams to develop more robust and insightful early-warning systems tailored to their specific country contexts and data landscapes.

The outline of this paper is as follows. The data set is introduced in Section 2, which also covers important aspects including the range of variables, data frequencies, and missing observations. With an emphasis on the MoE architecture, Section 3 describes the approach and explains how the model combines econometric and machine learning techniques to handle complicated data sets while preserving interpretable outcomes. Section 4 outlines how the MoE architecture is applied to macro-financial country risk. Section 5 applies the developed MoE framework, showing the empirical performance of the framework in predicting country risk across several sectors. Section 6 offers an analysis of risk score patterns over time and across groups. Section 7 concludes.

2. Data

2.1. Dependent variables: Crises Indicators

The objective of this approach is to generate forward-looking, country-level macro-financial crisis risk scores over a 12-month horizon, broken down into fiscal, financial, and balance of payments risk scores. Three binary target variables are constructed for the model:

1. **Fiscal Crisis Indicator:** In this paper, a fiscal crisis is defined as an episode in which the sovereign fails to meet its contractual debt-service obligations, or restructures those obligations on terms that impose material economic losses on creditors. Operationally, the indicator is coded using the Bank of Canada–Bank of England Sovereign Default Database, which identifies sovereign defaults on debt owed to official and private creditors. A default episode includes missed principal or interest payments beyond any applicable grace period, failure to honor guaranteed debt obligations, or debt exchanges/restructurings that reduce the economic value of creditors' claims. The database reports the stock of government obligations in default by country, year, creditor type, and debt instrument, expressed in nominal U.S. dollars. [3] In constructing the binary crisis variable, a country-year is classified as a fiscal crisis when the BoC–BoE database records a non-negligible stock of sovereign debt in default. To avoid coding very small or technical arrears as systemic fiscal crises, a materiality threshold is applied: defaulted obligations must exceed USD 1 million or 1 percent of GDP. [4] .

2. **Financial Crisis Indicator:** Following Laeven and Valencia (2020), a financial crisis is defined by systemic instability in the banking system, identified either as non-performing loans (NPLs) exceeding 20 percent of total loans, or when bank closures represent more than 20 percent of the banking system's assets. Additionally, the crisis is further confirmed if at least three out of six emergency measures are implemented by the government to stabilize the situation. These measures include imposing deposit freezes or bank holidays, nationalizing banks, making fiscal outlays for bank restructuring, providing extensive liquidity support, offering government guarantees on bank liabilities, and conducting government asset purchases.[5] In addition to the Laeven and Valencia (2020) definition, a financial crisis is also deemed to occur when bank tier 1 capital to risk-weighted assets falls below 8 percent, or when bank tier 1 capital to total assets falls below 3 percent.

[3] Arrear indicators have commonly been key components of fiscal and/or debt crises definitions (Asonuma & Trebesch (2016), Gerling et al (2017)). The intuition being that the accumulation of arrears, on both external and domestic debt, are socially and politically more viable responses to fiscal stresses.
[4] Updated global data on arrear accumulation on sovereign debt is available and accessible via the Bank of Canada. We use arrears data on external debt since external debt is better monitored and more transparent than domestic debt.
[5] The sample is expanded up to 2022 by adopting Laeven and Valencia (2020) criteria as well as relying on World Bank CPIA scores for financial stability being smaller than 2, EIU banking sector risk scores being larger than 88 (corresponding to a D rating in the EIU methodology), or IMF Article IV reports.

3. **The Balance-of-Payments Crisis Indicator** is designed to capture acute external liquidity pressure. We define a Balance-of-Payments (BoP) crisis as a country-year in which the economy's stock of liquid foreign exchange reserves is inadequate relative to near-term external payment obligations. The indicator is coded as one if either of two conditions is met: foreign exchange reserves excluding gold fall below one month of imports, or short-term external debt exceeds four times foreign exchange reserves excluding gold. The import-coverage criterion captures the risk that the country cannot sustain essential external purchases without severe compression of imports, arrears, or exceptional financing. The short-term-debt criterion captures vulnerability to rollover pressures, sudden stops, and refinancing stress, where reserves are insufficient to cover maturing external obligations.[6]. This definition is useful in an early-warning framework because it focuses on the liquidity dimension of external crises. BoP crises often materialize when countries lose access to foreign currency needed to finance imports, service external debt, or defend the exchange rate. Reserve adequacy indicators are therefore natural predictors of external stress: they measure the buffer available to the monetary authority against import needs and short-term foreign-currency liabilities. Using both import coverage and short-term-debt coverage is preferable to relying on a single reserve ratio, because the relevant pressure point differs across countries. In trade-dependent economies, import coverage may be the binding constraint; in financially integrated economies, short-term external debt may be the more relevant source of rollover risk[7].

Composite Crisis Indicator: Given that fiscal, financial, and balance of payments crises often coincide or reinforce one another, a composite indicator is constructed that equals one when a country experiences at least one of these crisis types within a given year and zero otherwise. It captures the joint manifestation of macro-financial distress and serves as the basis for an overall risk score summarizing systemic vulnerability across sectors.

[6] The utility of FX reserves indicators is well established in early warning systems of macro-financial crises. In fact, in a literature survey by Frankel and Saravelos (2012) covering pre-2008 crises, FX reserve holdings were found to be top performers. The authors specifically highlight the role of FX reserves in months of imports and as ratios of short-term debt and of external debt in the 2008 crisis.

[7] Thresholds for FX reserves in months of imports and as a ratio of short-term external debt have varied over time and across countries. Anecdotal evidence of policy objectives suggests FX targeting of a few months of imports (i.e. 3-4 months) and full coverage of short-term external debt. For example, the Guidotti–Greenspan rule, which is an international economic guideline, suggests that a country's FX reserves should equal or exceed its short-term external debt (maturing in one year or less) to ensure liquidity during financial crises. For our purposes, and as we want to capture tail-side events, we apply the above identified stricter thresholds in order to identify a BoP crisis.

Table 1. Summary of fiscal, financial and balance of payments crisis indicators

		Number of countries†	*Coverage Period*	*Time in crisis, % of total period*††
Fiscal		**153**	**1970-2022**	**13.5%**
Financial		**163**	**1970-2022**	**6.0%**
Balance-of-Payments	*Overall*	**112**	***1975-2023***	**12.5%**
	Import Coverage	*100*	*1975-2024*	*11.9%*
	Debt Coverage	*111*	*1975-2023*	*7.6%*

† This is the number of countries with at least one observation over the coverage period
†† This is a ratio of crisis observations over the total of crisis plus no crisis observations (precluding empty observations)

2.2. Independent Variables

A diverse array of data with varying sample frequencies (Appendix 1 provides a detailed overview of indicators and their sources) serves as inputs to assess country risk across the three crisis categories. The data set comprises 60 annual, 44 quarterly, and 29 monthly variables (Table 2), which are categorized into different types of risk features. These categories include "macro-fiscal," "BoP-related," "financial," and "global" variables:

- **Macro-fiscal variables**, available at the country level, encompass measures such as gross domestic product (GDP) and consumer price index (CPI), both in levels and growth rates. This category also includes fiscal and debt indicators expressed in relevant percentages.
- **BoP-related variables**, also at the country level, cover import/export volumes and prices, and various measures related to reserves, assets, and liabilities.
- **Financial risk variables** include bank-related metrics such as the number of branches, accounts, lending and borrowing rates, bank capital, private credit, deposits, and house prices. This category further captures spreads on the emerging markets bond index (EMBI), sovereign credit default swaps (5-year), and Standard & Poor's sovereign credit risk (long-term) measures. Additionally, the data set incorporates selected country-level variables from the International Country Risk Guide (ICRG) provided by the PRS Group.
- **Global variables**, applicable to all countries, include US Federal Reserve rates (across different maturities and spreads), oil prices, and global and US consumer price indices, both in levels and growth rates.

Table 2: Key characteristics of the data set

Number of countries	218	Across 7 regions
Number of country-specific variables	132	
- Annual (A)	60	
- Quarterly (Q)	44	
- Monthly (M)	29	
Number of global economic variables	18	All at frequencies A, Q, M
Percentage missing	~20%	More for lower-middle and low- income countries
Time series range	1950-2025	Most series start in 1990
Crisis indicators	Fiscal	Unable to meet debt obligations to creditors (incl. multilateral institutions) reflected by significant external arrears.
	Financial	Systemic banking distress marked by severe asset quality deterioration, bank failures, policy intervention, and critically weak capital buffers.
	BoP	Severe foreign-exchange reserve inadequacy relative to imports or short-term external debt.
Country income groups	High Upper-middle Lower-middle Low	

Note: All variables are reported in the Appendix.

The availability of most data variables begins in 1990, with coverage and detail increasing in subsequent years. However, data for lower-income countries is often less complete, resulting in missing entries in the input data set. While the MoE architecture demonstrates a degree of robustness to these gaps (as detailed in Section 3), the approach also entails measures to address inconsistencies in the underlying data. This includes enforcing fundamental domain constraints, such as requiring non-negative values for economic flows and ensuring that key ratios remain within empirically plausible ranges.

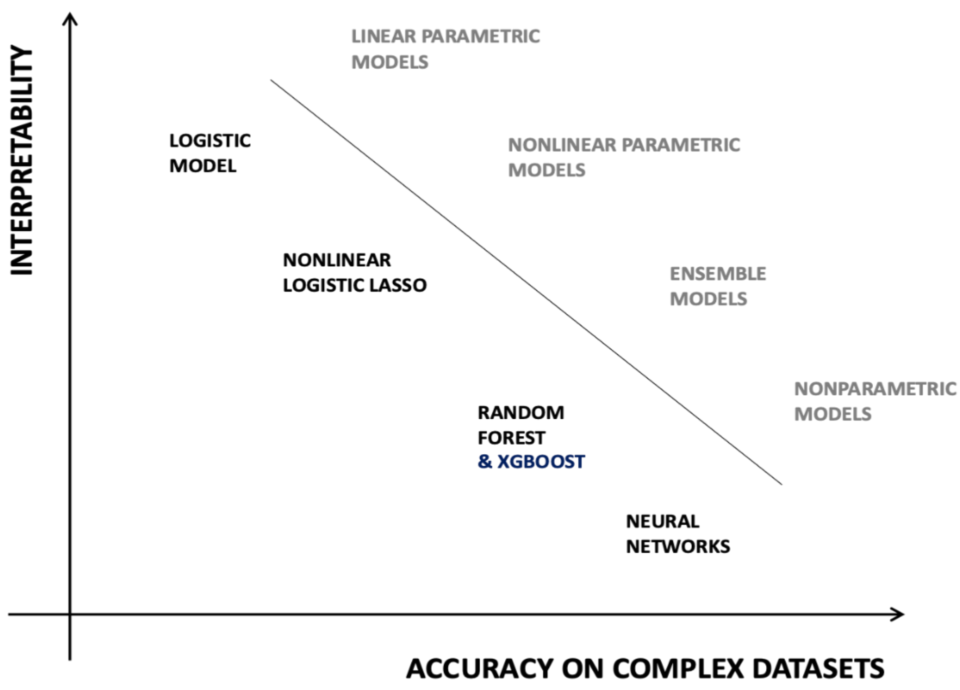
3. The Mixture-of-Experts Approach

3.1. The Trade-off between Predictive Performance and Explainability

It is well known that flexible models, such as ensemble models and non-parametric models can improve data fit and predictive performance for complex data-generating processes and datasets, relative to simpler parametric models. However, model flexibility often comes at the expense of overfitting and explainability or model interpretability. In such flexible models, predictions are more difficult to explain and parameter estimates are more difficult to interpret. In contrast, in settings where interpretability is key, the simplicity of standard parametric models (such as linear or logistic regression) provides researchers with a clear understanding of how inputs drive predictions. The design simplicity may, however, hamper the ability of such models to achieve increased accuracy on complex data sets featuring a variety of inputs.

Figure 1 illustrates this trade-off by plotting the relative positions of alternative model classes along an interpretability versus accuracy on complex data sets chart. It shows that parametric models such as logistic models are easily interpretable. Logistic models with non-linear transformations of features and automatic lasso selection of inputs can accommodate data sets with a larger set of inputs and more complex relationships between inputs and target variables. In contrast, ensemble methods (e.g., random forests (Breiman, 2001)) and nonparametric models (e.g., artificial neural networks (Bishop, 1995)) can achieve higher accuracy on complex data sets but lack a simple and straightforward interpretation of how inputs drive predictions. In recent years, substantial effort has been placed into improving the explainability of flexible ensemble and nonparametric models (Ewald et al. 2024, Benedek et al. 2022). This effort attempts to move ensemble models and nonparametric models such as random forests and neutral networks along the interpretability axis, without sacrificing predictive accuracy.

Figure 1: Trade-off between the accuracy on complex data sets and the interpretability of the currently existing models.



Source: Authors' elaboration.
Notes: In black, the figure shows, along the trade-off line, the position of several models and predictive algorithms that can be used to predict country risk using binary crisis targets. In shaded gray, the figure shows the position of broader model classes, from linear parametric to non-parametric models.

3.2. The Mixture-of-Experts Methodology

The Mixture-of-Experts (MoE) methodology (Jacobs et al. 1991) is designed to achieve both high model interpretability and high accuracy on complex data sets using a layered multi-stage approach. It breaks down the predictive objective into clearly interpretable targets and combines flexible ML methods, such as eXtreme Gradient Boosting (XGBoost), with interpretable logistic models.

Specifically, in the context of country risk, using a single flexible model to input all features and predict various crisis types—fiscal, financial, and balance of payments—can hinder explainability. Generating sector-specific risk scores becomes challenging, and understanding the relationship between individual inputs and specific risk types is impeded. Alternatively, estimating separate models for each crisis type allows for the production of sector-specific risk scores. However, this approach does not yield a comprehensive risk score for the country that consistently considers the interactions between different risk types.

The MoE breaks away from the interpretability-accuracy trade-off by leveraging a layer of experts, which separately predict risk at the sector level, while at the same time employing a layer of gating models. Those gating models can combine the different sector risk scores into a coherent country risk score. Good predictive performance is achieved by employing flexible expert models, which excel in handling complex or incomplete datasets. Transparency regarding the types of risk influencing country risk is maintained through the use of interpretable gating models.

Interpretability: The proposed model leverages the layered structure of the MoE architecture. The expert layer consists primarily on XGBoost models, which efficiently extract non-linear patterns and interactions from high-dimensional macroeconomic and financial data. In addition, the MoE framework is flexible enough to incorporate traditional parametric models as experts, such as logistic models. These expert models are nested into a sector-specific gating layer that ensures that the information is channeled and aggregated into sector-specific financial, fiscal, or balance of payments risk scores. Conceptually, this stage acts as a data-driven dimensionality reduction step that leverages cutting-edge models to ensure the maximum amount of information is transmitted to the final gating model. This final gating layer consists of a logistic model that can link the sector-specific inputs to country risk by ensuring a clear and interpretable trajectory from sector risk to country risk.

Dealing with complex and missing data[8]: EMDE datasets face large gaps in coverage (see Luckner et al., 2023)—but the proposed MoE architecture addresses both. In the expert layer, multiple ensemble models first translate and compress the raw data, each model natively tolerating missing values so that the aggregated output passed to subsequent layers is gap-free. Within this framework, the framework applies tailored imputation rules: (i) vectorization for series that start late; (ii) Multiple Imputation by Chained Equations (MICE; Azur et al., 2011)[9] if data is missing between different data points; (iii) an AR(1)-based forecaster[10] for the most recent missing observations; and (iv) tree-based learners for features that are entirely absent.[11] When country-specific data are especially scarce—as is often the case in low-income economies—experts can fall back on global covariates, ensuring every prediction benefits from at least a common information set. Because the MoE is modular, new sources can be added as separate experts and routed through the intermediate gating layer without disrupting the existing pipeline, preserving robustness while steadily enriching the signal passed to the final decision layer.

4. The Mixture-of-Experts Architecture Applied to Macro-financial Country Risk

4.1. Overview

Our applied MoE model consists of three layers as graphically displayed in Figure 2. The first layer consists of sub-risk specific expert models (the expert layer). The experts in this layer are designed and trained to predict respectively fiscal, financial, and BoP crises. The first gating layer is composed of three sector specific gating models. Each of these gating models receives inputs from their respective experts and provides an aggregated prediction of the sub-risk. The final gating layer then receives the sub-risk scores as input and produces an overall country risk score. Appendix 2 describes methodological details regarding which type of models and model performance metrics are used. Appendix 3 provides a technical discussion of how individual variables can be attributed to country (sub-)risk scores in practice.

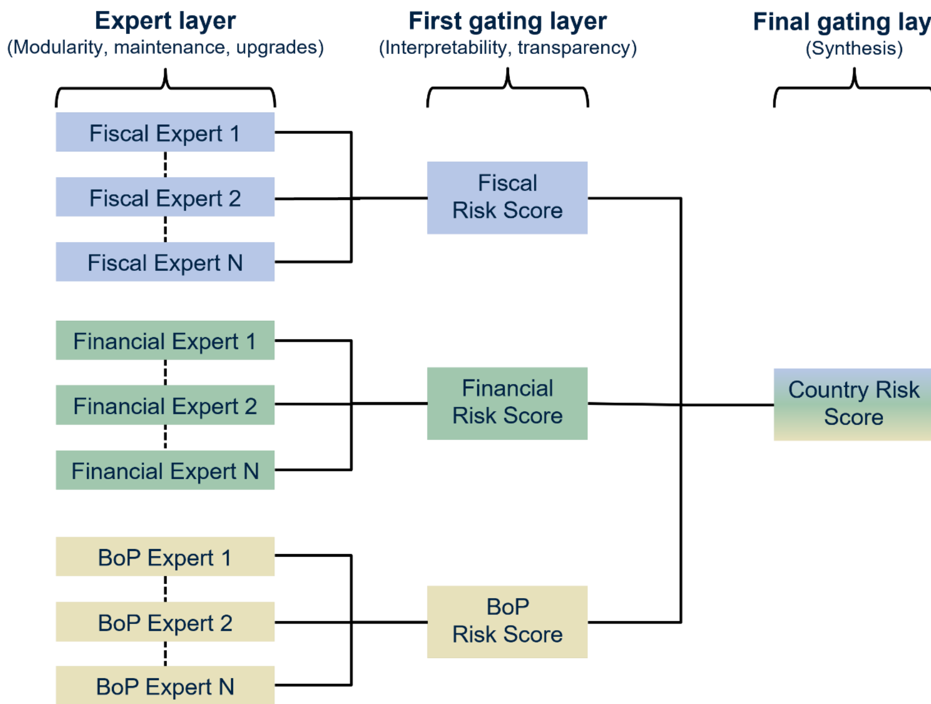
[8] Throughout the paper, "data limitations" primarily refer to gaps in coverage and missing observations. As in much of the macro-financial literature, issues related to data quality (e.g., varying measurement methodologies across jurisdictions), comparability, and reporting timeliness are acknowledged as structural constraints common to macro-financial surveillance and are not specific to the methodology proposed here.

[9] This robust statistical technique fills in missing values by modeling each feature with missing data as a function of other features, creating multiple complete datasets.

[10] This approach assumes that the best predictor of the next observation is the most recent observation, making it particularly useful when there are too few observations to fit a more comprehensive model.

[11] These models are proficient at handling missing data by splitting the data into branches based on the presence or absence of certain features, allowing analysis to continue even when some data is missing.

Figure 2: Mixture of Experts application for country risk



Source: Authors' elaboration.

4.2. The Expert Layer

The expert models in the first layer are assigned the task of predicting the occurrence of sector-specific risks. To the same task, the engine assigns a multiplicity of experts, each one with its specific features, to extract as much information as possible from the data. The task of the first gating layer is then to aggregate this information.

The first distinction among experts is based on data frequency. The target variables for this application are at an annual frequency, but the availability of higher-frequency explanatory variables enables the development of models at higher frequencies. Regular experts are defined as models that rely on both target variables and explanatory variables at an annual frequency. The first layer comprises several different regular experts for each type of risk, each with its own features (Table 3). High-frequency experts are defined as models that utilize quarterly or monthly explanatory variables. The target variable for these models is converted into a higher frequency through an infill procedure. The current iteration of the MoE includes two higher-frequency experts for each sector, specifically quarterly and monthly experts.

The details of the sector-specific experts are as follows: The baseline specification is an XGBoost model with unbalancing weights, given the relatively sparse nature of crisis events, and is AUC regularized. This specification is used for the first regular expert and for the higher-frequency experts. The specifications of the other experts are variations of this baseline. The second regular expert is created by defining "excessive" unbalancing weights, assigning a higher weight to crisis events, and is termed as Focused. The third regular expert is an F1-score regularized XGBoost model. The fourth and final regular expert is a regularized Logistic model with L2 penalty. Because the Logistic model cannot handle missing values, it is trained on a subset of the data, similar to the approach by Luckner et al. (2023), and then extrapolates scores to missing observations through an imputation step that backfills its predictions using information from the tree-based experts, ensuring all experts produce scores for the same set of observations. Fiscal risks are modeled with four yearly experts, while financial and balance of payments risks each employ six experts (four yearly, one quarterly, and one monthly), resulting in a total of 16 experts in the first layer.[12,13]

Table 3: Structure of the experts in the first layer of the MoE architecture

Expert	Frequency	Regularization	Unbalancing Weight	Model
1	Yearly	AUC Regularized	Balanced	XGBoost
2	Yearly	AUC Regularized	Focused	XGBoost
3	Yearly	F1-score Regularized	Balanced	XGBoost
4	Yearly	G2S Regularized	Balanced	Logistic model
5	Quarterly	AUC Regularized	Balanced	XGBoost
6	Monthly	AUC Regularized	Balanced	XGBoost

4.3. The First or Intermediate Gating Layer

The task of the models in the intermediate (or first) gating layer is to take the outputs from the different experts and aggregate their predictions into a final sector-specific risk score. This layer consists of three different gating models, one for each sector. Each of these models is an XGBoost model trained with unbalanced weights and is AUC regularized. Each one of these models takes as input the predictions produced by the different experts, the interaction of these predictions with country-specific features (e.g., country income group), and some of the predictions' summary statistics (e.g., the maximum value or the mean). This structure of inputs allows the model to disentangle the relevance of expert models according to different settings (e.g., country groups). Having the three models of the intermediate gating layer separately responsible for financial, fiscal,

[12] Experts utilizing lagged inputs to capture time dynamics and those incorporating time dummies to account for potential structural breaks did not significantly enhance the model's performance. Consequently, these elements were discarded to achieve a more parsimonious model.

[13] In the course of the MoE model development, random forests, lasso, and elastic net models were also tested as additional experts. However, they did not improve out-of-sample performance and were hence excluded from the MoE model.

and balance of payments risks is crucial for the interpretability of the MoE. These models act as a node linking sub-risks to the final risk score generated in the final gating layer.

Parametric intermediate gating layer models, such as logistic models, demonstrated inferior performance compared to the more flexible XGBoost specification. In an attempt to enhance model performance, the feature space of the intermediate layer was expanded with data from the original dataset, but no evidence of improved accuracy was found. Various methods to incorporate sub-risks into the intermediate gating layer were also explored. Transformations aimed at introducing non-linearity into the relationship, such as squaring sector scores and employing a number of dummies, were considered but ultimately discarded due to weaker performance.

4.4. The Final Gating Layer

The final gating layer, adhering to the MoE approach, consists of a single panel logistic model that is fully interpretable. This regression utilizes fiscal, financial, and BoP risk predictions from the intermediate gating layer of the MoE as independent variables, producing a final composite risk score at the country level for each year. Interaction terms of different sector risks are included to capture cross-effects, along with indicator variables for sectors where estimated crisis probabilities exceed 0.6, capturing non-linear effects. The logistic specification is deliberately retained in this final layer for the interpretability and traceability of the composite risk score. After tree-based stages process high-dimensional macro-financial information into sector-level risk scores, a parsimonious parametric model expresses, in closed form, the marginal impact of each sectoral score on overall crisis probability. This approach facilitates interpretation at the coefficient level and transparent communication about how each risk component contributes to the final risk assessment. The logistic layer therefore strikes the desired balance between predictive accuracy and explainability that underpins the entire MoE philosophy.

5. Empirical results

5.1. Cross-validation and overall model performance

The MoE model uses a strictly chronological, expanding-window cross-validation design, keeping a train, validation, and test set separate. For each evaluation year T, the most recent observations are reserved as the test set, and only earlier years are used for training and hyperparameter selection. Within the training set (years $< T$), hyperparameters are chosen via an expanding-window cross-validation, in which each fold trains on earlier years and validates on strictly later years (fold sizes tailored to sectoral data depth). Importantly, the intermediate sector layer is cross-validated on out-of-sample, one-year-ahead expert scores.[14] At each cutoff, the expert models are

[14] Assume a 2000 model, for example, where the training pool is all years before 2000 (e.g., 1990-1999). Folds are trained on earlier years and validated on strictly later year blocks. Using the financial sector with two-year validation blocks for illustration: Fold 1 trains on 1990-1994 data and validates on 1995-1996 data; Fold 2: trains on 1990-1996

re-estimated using data prior to the cutoff and then used to generate one-year-ahead predictions that are treated as fixed inputs. Intermediate layers are re-estimated on an expanding window and evaluated in three-year blocks. This hierarchical, no-leakage design prevents the intermediate layer from learning from in-sample expert first, reducing overfitting and yielding genuine out-of-sample performance over long time series.

The model's performance to predict macro-financial crises is assessed using AUC. AUC quantifies the overall performance of the classifier by calculating the area under the RoC curve.[15] It provides a single scalar value that summarizes the ability of the model to discriminate between the positive and negative classes, regardless of the chosen classification threshold; for example, AUC = 1 denotes a model that perfectly separates all positive and negative instances; AUC = 0.5 suggests that the model is no better than random guessing, unable to distinguish between positive and negative classes.

The AUCs for the model are 0.94[16], 0.85 and 0.90 for the fiscal, financial and balance of payments risk scores, respectively, and 0.92 for the overall risk score (Figure 3). These results are quite competitive when benchmarked against the literature; for example, Hellwig et al. (2021) reported AUCs ranging from 0.53 to 0.77 for fiscal crises using various models, while an IMF Technical Report (2021) indicated AUCs between 0.55 to 0.78 for fiscal crises, 0.43 to 0.72 for financial crises, and 0.54 to 0.69 for external crises.

Relative to a simple conventional L1-regularized logit[17], the MoE model delivers a superior balance of accuracy and parsimony in crisis detection (Figures 4–6). At very low cutoffs the logit captures a marginally larger share of actual crises, but this gain in recall is offset by a surge in false alarms, sharply lowering precision. By contrast, the MoE sustains a more favorable trade-off: as

---

and validates on 1997-1998 data; Fold 3 trains on 1990-1997 data and validates on 1998-1999 data. The test (holdout) block for this anchor would be 2000-2002. Experts trained only on years before 2000 generate one-year ahead predictions at 2000, 2001, 2002. These out-of-sample scores feed the intermediate layer, which is evaluated on the same three-year blocks. Repeating this procedure at later anchors (e.g., 2003, 2006, etc.) yields additional three-year out-of-sample blocks that are appended to train the final gating model without ever using a block to tune the experts that produced it.

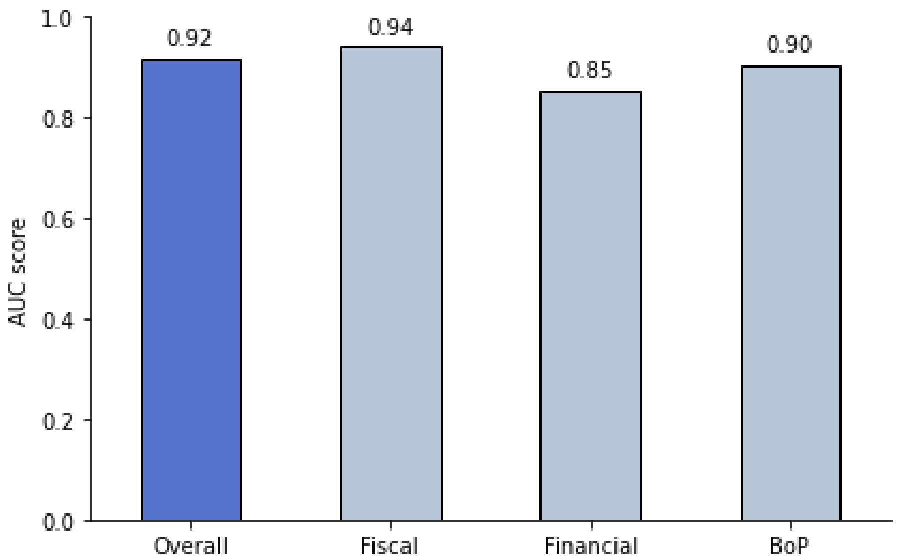
[15] The ROC curve is a commonly used performance metric for classification tasks in machine learning, particularly when dealing with binary classification problems. The curve is a graphical representation of a classifier's performance across different classification thresholds. It plots the True Positive Rate (TPR) against the False Positive Rate (FPR) at various threshold settings. The TPF (also called Recall or Sensitivity) is the ratio of correctly predicted positive observations to all actual positives. It measures the ability of the classifier to detect positive instances. The FPR is the ratio of incorrectly predicted positive observations (False Positives) to all actual negatives. It measures the likelihood of the classifier incorrectly labeling a negative instance as positive.

[16] An AUC of 0.94 means that in 94% of the cases, a randomly chosen crisis country will receive a higher risk score than a randomly chosen stable country.

[17] The explanatory variables are selected balancing empirical relevance from the macro-financial crisis literature (IMF, 2021; Luckner et al., 2024) with data availability with a minimum of 100 countries covered consistently since 2004, which correspond to 26 country-level features and 18 global variables. The inclusion of all available features in the regularized logit was also tested.

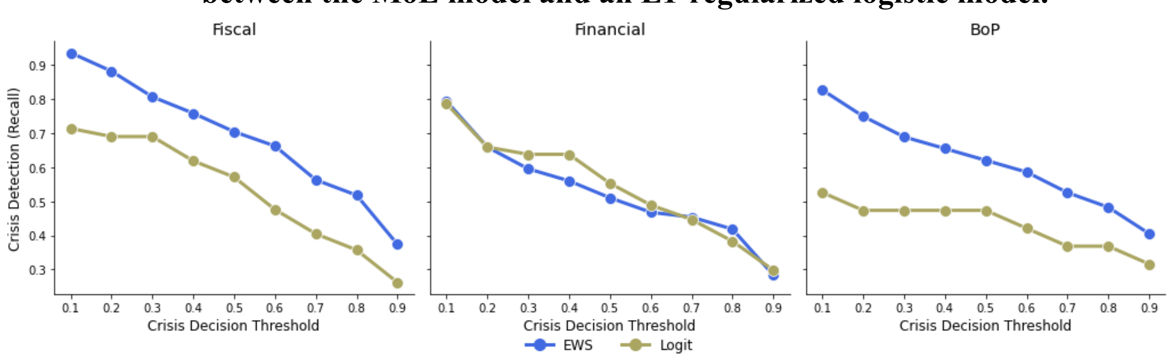
thresholds rise, it sacrifices only a small portion of recall while markedly improving precision, so that its combined F1-score is higher at every cutoff.[18] These patterns underscore the MoE's reliability as an early-warning tool: it flags genuine crises while limiting costly false positives.

Figure 3. Area Under the Curve (AUC) performance



Source: Authors' calculations.

Figure 4. Comparison of crisis detection (recall) across different crisis decision thresholds between the MoE model and an L1-regularized logistic model.



[18] The evaluation framework emphasizes decision metrics at policy-relevant thresholds, such as ROC AUC, precision, recall, F1, and performance at specified TPR/FPR levels, which may directly inform operational choices under class imbalance and align with established crisis prediction literature (e.g., Schularick and Taylor, 2012; IMF, 2021; Bluwstein et al., 2023). While false discovery rate (FDR) and false omission rate (FOR) offer alternative framings, they are algebraically derivable from reported metrics (FDR = 1 – precision; FOR = 1 – negative predictive value) and thus would not change the conclusions drawn from the reported quantities.

Figure 5. Comparison of warning accuracy (precision) across different crisis decision thresholds between the MoE model and an L1-regularized logistic model.

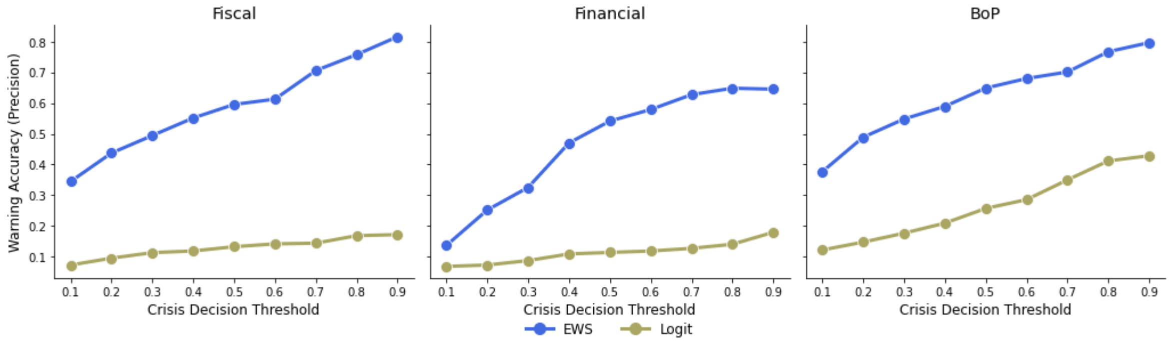
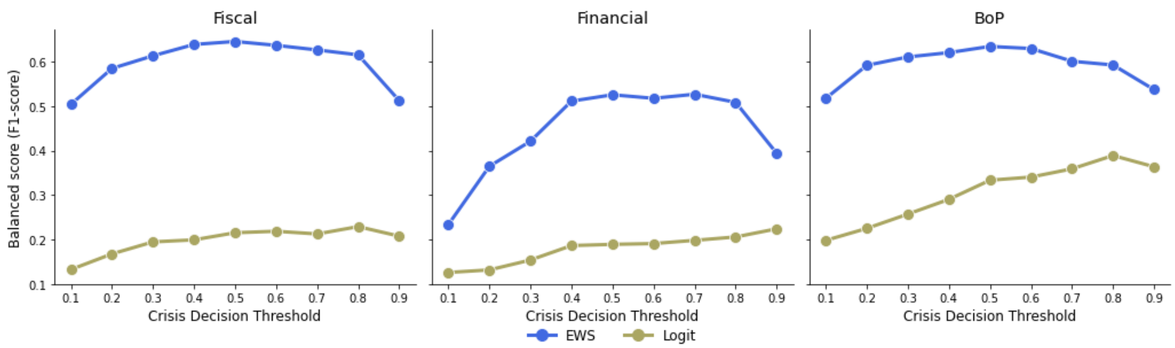


Figure 6. Comparison of balanced scores (F1-score) across different crisis decision thresholds between the MoE model and an L1-regularized logistic model.



Source: Authors' calculations.

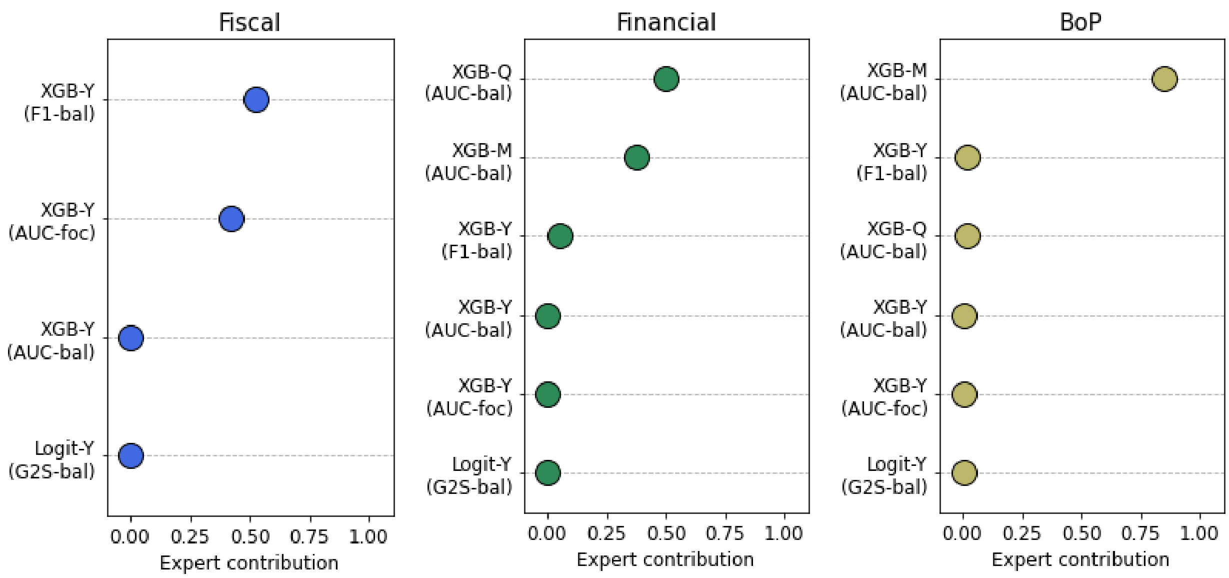
Note: Crisis detection rate (i.e., "recall" or "true positive rate") represents the share of correctly predicted crises among all actual crises. Precision represents the share of predicted crises that actually occurred. The F1-score is a way to measure how well a model does at both finding real crises and not raising too many false alarms. It combines two things: how many real crises the model catches and how often its crisis warnings are actually correct. The crisis decision threshold is the risk score (the probability of a crisis) at which the model classifies an episode as a crisis.

5.2. Model performance at the expert / prediction module level

Figure 7 illustrates the expert signal strengths across the different crises in our mixture of experts framework. For fiscal crisis prediction, yearly XGBoost models optimized for F1-score with balanced weighting (XGB-Y, F1-bal) and AUC with focused weighting (XGB-Y, AUC-foc) provide the strongest signals, accounting for 0.52 and 0.42 of the gating weight, respectively. Financial crisis prediction benefits most from higher-frequency models, with a quarterly AUC-optimized XGBoost contributing 0.50 of the predictive power, complemented by monthly (0.38) and yearly (0.05) variants. Balance of payments crisis prediction likewise concentrates its gating weight on tree-based learners, with a monthly AUC-optimized XGBoost contributing 0.84. Notably, across all risk domains, tree-based XGBoost algorithms substantially outperform traditional logistic regression approaches, which contribute minimally (≤0.01) to the overall predictive power.

This performance advantage likely stems from XGBoost's ability to capture complex non-linear relationships and interactions between macroeconomic variables that linear logistic models alone cannot effectively represent. Given this clear superiority, the second-layer gating network is itself an XGBoost classifier. Using a tree-based learner at the gating stage allows the model to learn sophisticated, time-varying interactions among expert outputs and to handle class imbalance via built-in weighting, yielding noticeably better calibration and AUC than a logistic alternative in validation tests. While not shown in Figure 9, additional specialized experts incorporating interactions with income groups and coalition approaches provided marginal improvements to the performance metrics reported above, further validating the applied hierarchical mixture approach to crisis prediction.

Figure 7. Sub-expert contributions by sector



Source: Authors' calculations.
Notes: Models: XGB = XGBoost; Logit = Logistic. Frequency: Y = Yearly; Q = Quarterly; M = Monthly. Regularization methods: AUC = Area Under the Curve; F1 = F1-score; G2S = General-to-specific. Class imbalance weighting: bal = Balanced; foc = Focused.

6. Risk Scores Over Time and Across Country Groupings

6.1. Global Trends

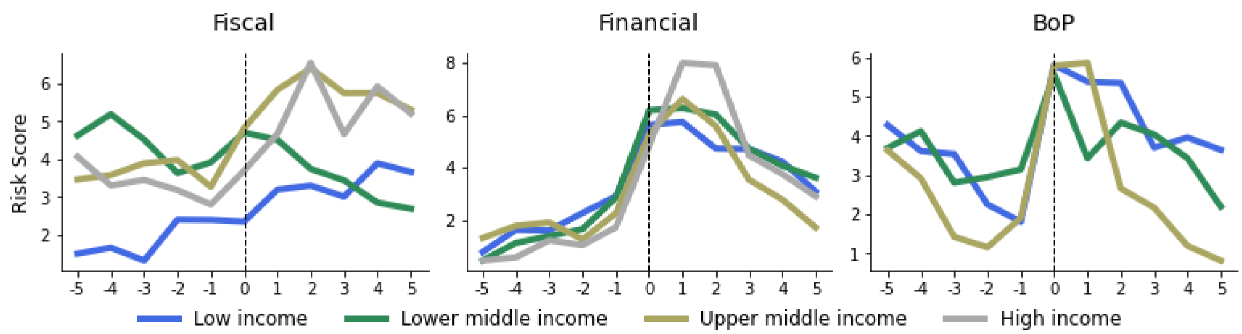
The MoE framework produces country risk scores, normalized on a 0-10 scale, which indicate the probability of a crisis within the next 12 months. These scores can be disaggregated into three sub-components: fiscal, financial, and BoP risk. Examining the dynamics of country risk before and after crisis events provides further insight into the model's ability to detect early signs of macro-financial distress. Figure 10 illustrates the average evolution of these three risk scores around the onset of crises across different country income groups.

Unlike the distinct bell-shaped patterns observed for financial and BoP crises, fiscal risk scores follow more uneven trajectories across income groups. This likely reflects the gradual and heterogenous nature of fiscal stress, which often builds through debt accumulation, weak revenues, or policy drifts rather than sudden shocks. The timing of crisis recognition can also be blurred by policy interventions, debt restructurings, or political factors that delay adjustment and create long periods of elevated risks rather than a shar pre-crisis spike. Institutional differences, such as fiscal frameworks and market access, may further shape how fiscal distress unfolds.

In contrast, Figure 8 shows a pronounced increase in financial risk scores across all income groups leading up to banking crises, underscoring the model's sensitivity to pre-crisis conditions. The persistence of elevated risk levels also varies across income groups. Low-income and lower-middle-income countries tend to exhibit more prolonged periods of heightened financial risk after banking crises. Interestingly, financial risks in high-income countries tend to peak between one and two years after the start of banking crises, reflecting contagion and a slower deleveraging process attributable to their highly interconnected financial systems.

BoP risk scores also show a pronounced pre-crisis build-up, consistent with episodes of foreign exchange reserve depletion or rising short-term external debt burdens that signal liquidity shortages. After crisis begin, balance of payments stress typically declines as FX adjustments, emergency financing, or policy measures help restore reserve adequacy. This surge is more evident in low-income and lower-middle-income countries, where external buffers are thin, while higher income display faster stabilization.

Figure 8: Financial risk dynamics around banking crises



Source: Authors' calculations.
Note: Figure shows crisis episodes that were not preceded by another crisis in the previous five years, where income groups with fewer than three countries are excluded. Fiscal crises: Sample of 61 crises across 5 LICs, 11 LMICs, 22 UMICs, and 7 HICs. Financial crises: Sample of 101 crises across 14 LICs, 25 LMICs, 28 UMICs, and 18 HICs. Balance of payments crises: Sample of 43 crises across 9 LICs, 13 LMICs, and 13 UMICs.

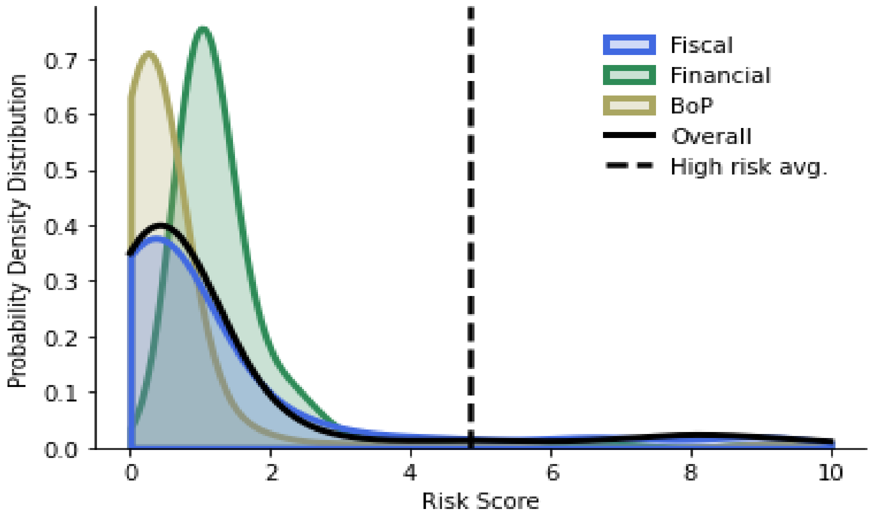
6.2. Recent Risk Score Distributions and Trends

Figure 9 presents the probability density distributions of the overall risk score and its three subcomponents for 2025. This provides a detailed view of the variation and concentration of risk across economies. The overall risk score exhibits a positively skewed distribution, meaning that the majority of countries face a small probability of being in crisis within the next 12 months. However, there is a long right tail, indicating a set of countries that face a higher probability of being in crisis within the next 12 months. Among the top quintile of countries. The average overall risk score stands at 4.85.

All three sub-risk distributions are also right-skewed: the majority of countries sit at relatively safe levels, while a small minority endure acute stress. Fiscal risk has a broad cluster at low values but a long upper tail, highlighting wide disparities in fiscal resilience. Financial risk is also skewed, yet its peak lies slightly higher, suggesting that moderate vulnerability is more common and a residual probability of crisis persists across many countries. Balance-of-payments risk records the lowest average level but the greatest dispersion: most economies have limited external fragility, whereas a few are sharply exposed to foreign-exchange liquidity shocks, stretching the distribution into a long tail.

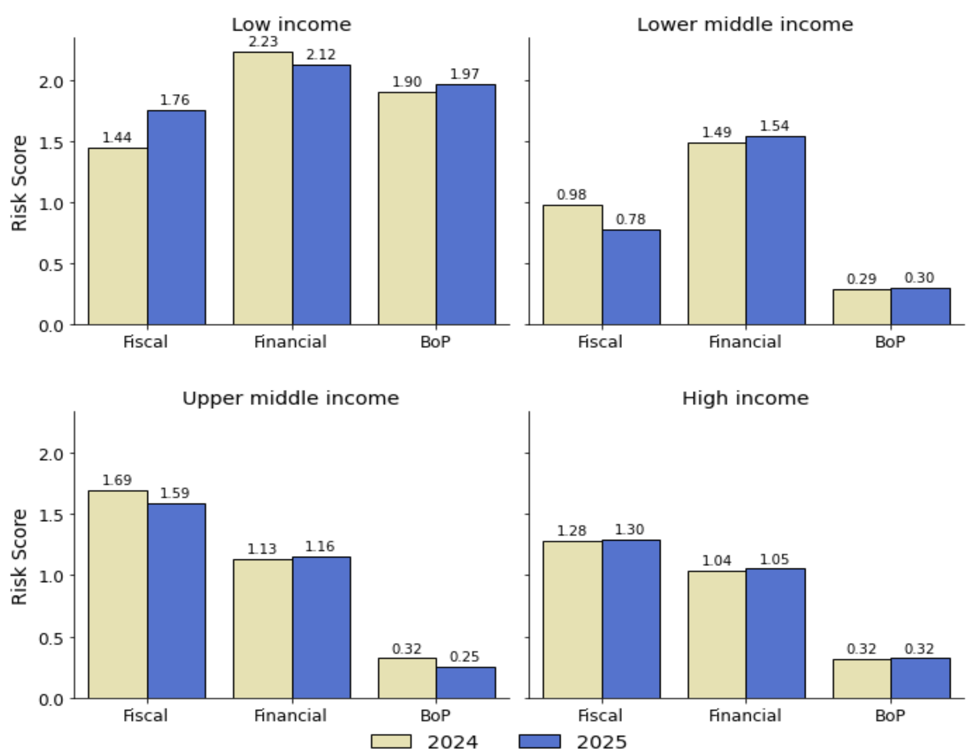
Figure 10 compares macro-financial risks between 2024 to 2025 and reveals an uneven risk landscape across income groups. Fiscal pressure climbed in low-income countries, overtaking upper-middle income countries' average risk levels, which were highest in 2024, whereas fiscal risks in lower-middle income and upper-middle income countries dropped from 2024 to 2025.

Figure 9: Distribution of country risks, 2025



Source: Authors' calculations.
Note: High risk countries are labeled as those in the top 20 percent of overall risk scores, which correspond to 32 countries in 2025.

Figure 10: Recent risk trends across income groups, 2024-2025

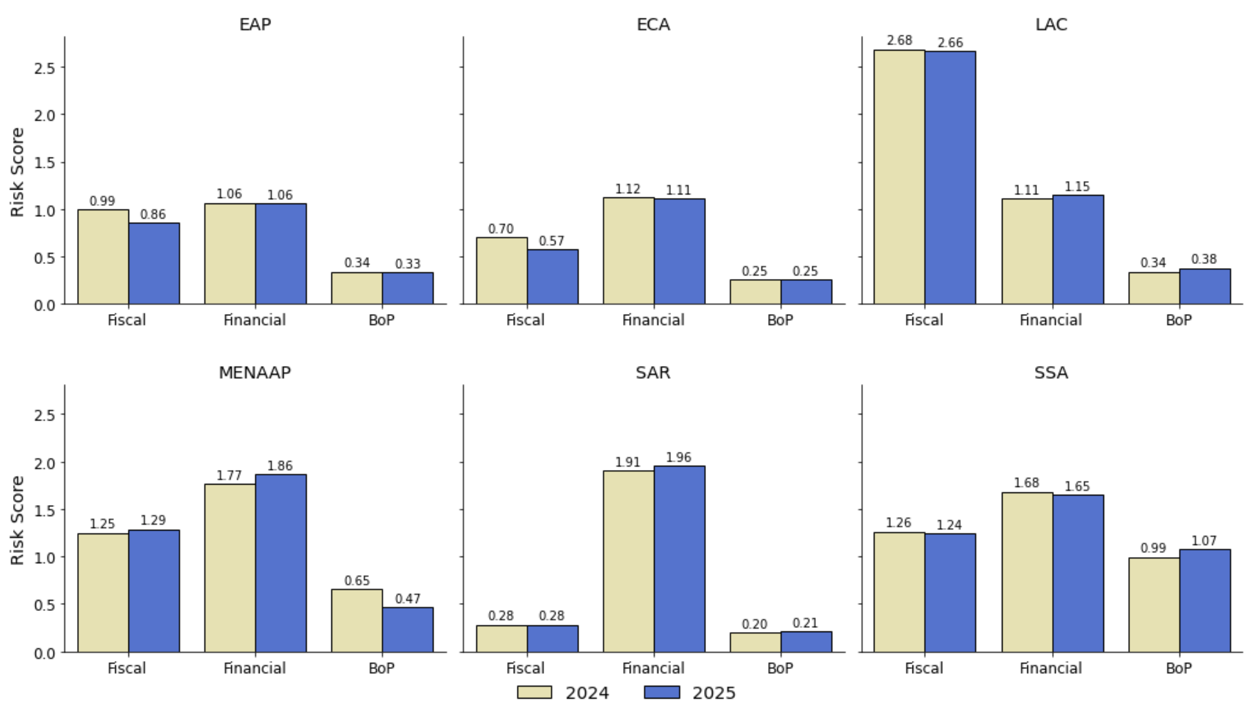


Source: Authors' calculations.
Note: Panel consists of average risk scores across 24 LICs, 50 LMICs, 50 UMICs, and 35 HICs.

Financial risk inched up for all groups except low-income economies, yet those same low-income countries still bear the heaviest financial strain on average. BoP risk also moved unevenly, rising moderately in every income category except upper-middle-income economies; here, too, low-income countries remain most exposed, reflecting intensifying external pressures. Overall, fiscal and financial risks are deepening at the lower end of the income spectrum, whereas fiscal risks have eased in in lower-middle income and upper-middle income countries. BoP risk shows relative stability or only gradual change outside the low-income economies.

Similarly, Figure 11 shows the changes in fiscal, financial, and BoP risk scores across regions from 2024 to 2025, highlighting regional disparities in the evolution of vulnerabilities. Fiscal risks are noticeably higher in Latin America and the Caribbean (LAC), while remaining stable or decreasing in East Asia and Pacific (EAP), Europe and Central Asia (ECA), and South Asia (SAR). Financial risks remain relatively elevated in Middle East and North Africa, Afghanistan & Pakistan (MENAAP), SAR, and Sub-Saharan Africa (SSA). BoP risks exhibit marginal improvements overall, with modest increases in LAC and SSA.

Figure 11: Recent risk trends across regions, 2024-2025

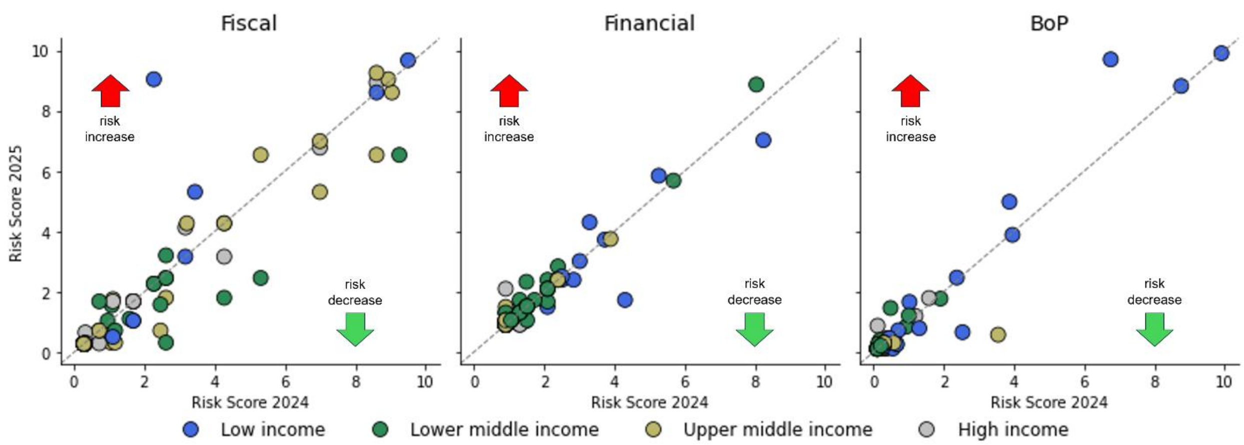


Source: Authors' calculations.

Note: Panel consists of average risk scores across 24 EAP, 28 ECA, 32 LAC, 22 MENAAP, 6 SAR, and 48 SSA countries. EAP = East Asia and Pacific; ECA = Europe and Central Asia; LAC = Latin America and the Caribbean; MENAAP = Middle East, North Africa, Afghanistan and Pakistan; SAR = South Asia; SSA = Sub-Saharan Africa.

Finally, Figure 12 drills down to country-level dynamics by plotting year-over-year changes in fiscal, financial, and BoP for each income group. The 45-degree line in each scatter represents no change; points above it signal worsening conditions, while points below denote improvement. The fiscal-risk panel shows a clear bifurcation: one cluster of countries hovers near the origin, indicating stable or modest shifts, whereas another lies well above the line, underscoring a subset still burdened by substantial legacy fiscal pressures. Financial-risk observations are more tightly packed and generally fall below a score of 5.0, suggesting that most economies keep vulnerabilities in check, though incremental deteriorations are visible. Balance-of-payments risks display a similar concentration at lower scores, but the largest year-to-year swings are concentrated among low-income countries, pointing to their heightened sensitivity to external shocks.

Figure 12: Recent country-level risks shifts, 2024-2025



Source: Authors' calculations.
Note: Panel consists of 24 LICs, 50 LMICs, 50 UMICs, and 35 HICs.

7. Conclusion

This paper introduces a novel Mixture-of-Experts framework for forecasting macro-financial risk, designed to address the specific challenges faced by EMDEs. Intended for a technical audience, the primary objective of this paper was not to present a definitive, one-size-fits-all model, but to propose and validate a flexible methodology for building custom early-warning systems. By integrating the predictive power of machine learning with the structured interpretability of traditional econometrics, this hybrid approach offers a practical solution for macro-financial surveillance, particularly in the presence of data gaps. The MoE framework enhances interpretability by decomposing a country's overall risk into three distinct, policy-relevant sub-risks: fiscal, financial, and BoP. This disaggregation allows policymakers to understand not only the level of risk but also its primary drivers, which is essential for targeted interventions.

The empirical results confirm that this methodology meaningfully outperforms standard crisis prediction models found in the literature, as measured by the AUC. However, there are also persistent challenges inherent in macro-financial surveillance. The complexity of the MoE framework requires significant expertise, and while robust to missing data, its performance is still constrained by the quality and availability of underlying data. Therefore, the model's outputs should be interpreted as probabilistic risk signals that complement, rather than replace, expert judgment. Ultimately, this paper contributes to the crisis prediction literature by offering a practical blueprint for developing more robust and insightful early-warning systems. It demonstrates how hybrid modeling strategies can leverage the complementary strengths of econometrics and machine learning to create tools tailored to specific country contexts. Future research could build on this architecture by integrating additional data sources or developing more granular risk decompositions, further enhancing its predictive capabilities and applicability across different economic environments.

References

Ahelegbey, Daniel F., Billio, Monica, and Casarin, Roberto (2016). “Bayesian Graphical Models for Structural Vector Autoregressive Processes”, Journal of Applied Econometrics, 31:357–386.

Asonuma, T., & Trebesch, C. (2016). Sovereign Debt Restructurings: Preemptive or Post Default, Journal of the European Economic Association, 14, 175-214.

Azur MJ, Stuart EA, Frangakis C, Leaf PJ. (2011). “Multiple imputation by chained equations: what is it and how does it work?” Int J Methods Psychiatr Res. 20(1):40-9. doi: 10.1002/mpr.329. PMID: 21499542; PMCID: PMC3074241.

Basurto, Gabriela and Atish Ghosh (2000, February). “The interest rate-exchange rate nexus in the asian crisis countries.” Technical Report Working Paper No. WP/00/19, International Monetary Fund, Washington, DC.

Batsuuri, Tsendsuren et al., (2024). “Predicting IMF-Supported Programs: A Machine Learning Approach”. IMF Working Paper WP/24/54.

Benedek Rozemberczki and Lauren Watson and Péter Bayer and Hao-Tsung Yang and Olivér Kiss and Sebastian Nilsson and Rik Sarkar. (2022). “The Shapley Value in Machine Learning”, 2202.05594, https://arxiv.org/abs/2202.05594

Berg, Andrew and Pattillo, Catherine (1999). “Are Currency Crises Predictable? A Test”, IMF Staff Papers, Vol. 46 No. 2, pp. 107–138.

Bholat, David, Gharbawi, Tariq, and Weldon, Richard (2016). “Machine Learning for Policymakers: What It Means and What It Doesn’t”, Bank of England Working Paper No. 674.

Bishop, C. (1995). Neural Networks for Pattern Recognition, Clarendon Press, Oxford.

Bluwstein, Kristina, Buckmann, Marcus, Joseph, Andreas, Kapadia, Sujit, Şimşek, Özgür (2023). “Credit growth, the yield curve and financial crisis prediction: Evidence from a machine learning approach”, Journal of International Economics, Vol. 145, 103773.

Breiman, L. (2001). Random Forests, Machine Learning 45: 5-32

Bussiere, Matthieu et al., (2015). “For a Few Dollars More: Reserves and Growth in Times of Crises”, Journal of International Money and Finance, Vol. 52, pp. 127–145.

Bussiere, Matthieu and Fratzscher, Marcel (2006). “Towards a New Early Warning System of Financial Crises”, Journal of International Money and Finance, Vol. 25 No. 6, pp. 953–973.

Cesa-Bianchi, Ambrogio, Martin, Fernando M., and Thwaites, Gregory (2019). “Foreign Booms, Domestic Busts: The Global Dimension of Banking Crises”, Journal of Financial Economics, Vol. 133 No. 3, pp. 629–648.

Chen, Tianqi and Guestrin, Carlos (2016). “XGBoost: A Scalable Tree Boosting System”, Proceedings, 22nd ACM SIGKDD Conference Knowledge Discovery and Data Mining, 785–794.

Christiano, Lawrence J., Christopher Gust, and Jorje Roldos (2002, June). “Monetary Policy in a Financial crisis. ” Working Paper 9005, National Bureau of Economic Research.

Corsetti, Giancarlo, Paolo Pesenti, and Nouriel Roubini (1998). “Paper Tigers? a model of the Asian Crisis.’ Technical Report 6783, National Bureau of Economic Research.

Davis, E. Philip and Karim, Dilruba (2008). “Comparing Early Warning Systems for Banking Crises”, Journal of Financial Stability, Vol. 4 No. 2, pp. 89–120.

De Mol, Christine, Giannone, Domenico, and Reichlin, Lucrezia (2008). “Forecasting Using a Large Number of Predictors: Is Bayesian Shrinkage a Valid Alternative to Principal Components?”, Journal of Econometrics, Vol. 146 No. 2, pp. 318–328.

Demirgüç-Kunt, Asli and Detragiache, Enrica (1998). “The Determinants of Banking Crises in Developing and Developed Countries”, IMF Staff Papers, Vol. 45 No. 1, pp. 81–109.

Eichengreen, Barry, Rose, Andrew K., and Wyplosz, Charles (1996). “Contagious Currency Crises”, Scandinavian Journal of Economics, Vol. 98 No. 4, pp. 463–484.

Ewald, F.K., Bothmann, L., Wright, M.N., Bischl, B., Casalicchio, G., König, G. (2024). A Guide to Feature Importance Methods for Scientific Inference. In: Longo, L., Lapuschkin, S., Seifert, C. (eds) Explainable Artificial Intelligence. xAI 2024. Communications in Computer and Information Science, vol 2154. Springer, Cham. https://doi.org/10.1007/978-3-031-63797-1_22

Frankel, Jeffrey A. and Saravelos, Georgios (2012). “Can Leading Indicators Assess Country Vulnerability? Evidence from the 2008–09 Global Financial Crisis”, Journal of International Economics, Vol. 87 No. 2, pp. 216–231.

Gerling, K., Medas, P., Poghosyan, T., Farah-Yacoub, J., & Xu, Y. (2017). "Fiscal Crises", IMF Working Papers 2017/086. International Monetary Fund.

Giannone, Domenico, Lenza, Michele, and Primiceri, Giorgio E. (2020). “Economic Predictions with Big Data: The Illusion of Sparsity”, Econometrica, Vol. 88 No. 5, pp. 1973–1989.

Hellwig, Klaus-Peter (2021). Predicting Fiscal Crises: A Machine Learning Approach. IMF Working Paper WP/21/150.

Jacobs R.A., M. I. Jordan, S. J. Nowlan and G. E. Hinton. (1991), "Adaptive Mixtures of Local Experts," in Neural Computation, 3(1): 79-87, March 1991, doi: 10.1162/neco.1991.3.1.79.

Kaminsky, Graciela, Lizondo, Saul, and Reinhart, Carmen M. (1998). “Leading Indicators of Currency Crises”, IMF Staff Papers, Vol. 45 No. 1, pp. 1–48.

Luckner, Clemens, Horn, Sebastian, Kraay, Aart, and Ramalho, Rita (2023). “Predicting Debt Distress in Low-Income Countries”, World Bank Working Paper.

Manasse, Paolo and Roubini, Nouriel (2009). “‘Rules of Thumb’ for Sovereign Debt Crises”, Journal of International Economics, Vol. 78 No. 2, pp. 192–205.

Schularick, Moritz, and Taylor, Alan (2012). "Credit Booms Gone Bust: Monetary Policy, Leverage Cycles, and Financial Crises, 1870-2008." American Economic Review 102:1029–61.

Van Rijsbergen, C.J., 1979. Information Retrieval, 2nd edn. Newton, MA.

Wiseman, Kevin et al., (2021). How to Assess Country Risk: The Vulnerability Exercise Approach Using Machine Learning, Technical Notes and Manuals, /21/03. International Monetary Fund.

8. Appendix

1. Macro-Financial Indicators and Data Sources

Type	Indicators	Frequency	Source
Macro-fiscal	CPI (index)	Annual, Quarterly, Monthly	IMF IFS
	CPI Previous period, Percent	Quarterly, Monthly	IMF IFS
	CPI Corresponding period previous year, Percent	Annual, Quarterly, Monthly	IMF IFS
	Export price deflator / Import price deflator	Annual	World Bank MPO
	External public debt repay to exports (%)	Annual	World Bank MPO
	GDP growth rate %	Annual	World Bank MPO
	Gross Domestic Product at Market Price, Volume, Millions 2015 real USD	Annual	World Bank MPO
	Real per capita GDP	Annual	World Bank MPO
	Bond issuance total volume (% of GDP)	Annual	FinStats
	Sovereign interest to revenue (%) (proxy for default)	Annual	World Bank MPO
	Primary budget balance (% of GDP)	Annual	World Bank MPO
	Domestic sovereign debt (% of GDP)	Annual	World Bank MPO
	External sovereign debt (% of GDP)	Annual	World Bank MPO
	Total sovereign debt (% of GDP)	Annual	World Bank MPO
	Domestic public debt repayment (% of GDP)	Annual	World Bank MPO
	External public debt repayment (% of GDP)	Annual	World Bank MPO
	Total public debt repayment (% of GDP)	Annual	World Bank MPO
	Public Expenditure (% of GDP)	Annual	World Bank MPO
	Public Revenues (% of GDP)	Annual	World Bank MPO
	Domestic interest on pub. Debt	Annual	World Bank MPO
	Export price deflator	Annual	World Bank MPO
	Imports, Implicit Deflator	Annual	World Bank MPO
	Gross Domestic Product at Market Price, Value, Millions USD	Annual	World Bank MPO
	Primary balance required to stabilize debt (% of GDP)	Annual	World Bank MPO
	Private debt, loans and debt securities (% of GDP)	Annual	IMF Global Debt Database
Exter	Official Reserve Assets and Other Foreign Currency Assets, US Dollars	Annual	IMF IRFCL
	Reserves as percentage of GDP (%)	Annual	IMF IRFCL

	Reserves growth rate %	Annual, Quarterly	IMF IRFCL
	Net FDI flows to GDP (%)	Annual	World Bank MPO
	Imports of G&S to GDP (%)	Annual	World Bank MPO
	Exports of G&S to GDP (%)	Annual	World Bank MPO
	FDI Liabilities GDP % (using annual GDP)	Quarterly	IMF IFS
	FDI Liabilities growth rate %	Quarterly	IMF IFS
	FDI Assets GDP % (using annual GDP)	Quarterly	IMF IFS
	FDI Assets growth rate %	Quarterly	IMF IFS
	Foreign Exchange rate	Annual, Quarterly, Monthly	Bloomberg
	Budget balance (% of GDP)	Annual	World Bank MPO
	IIF monthly portfolio flows, US million	Monthly	IIF
	IIF monthly portfolio flows growth rate %	Monthly	IIF
	IIF Net capital flows, USD billion	Monthly	IIF
	IIF net capital flows growth rate %	Monthly	IIF
	IIF Reserve operations, USD billion	Monthly	IIF
	IIF reserve operations growth rate %	Monthly	IIF
	Official Reserve Assets and Other Foreign Currency Assets, US Dollars	Monthly	IMF IRFCL
	FDI Liabilities Other Investments GDP % (using annual GDP)	Quarterly	IMF IFS
	FDI Liabilities Other Investments growth rate %	Quarterly	IMF IFS
	FDI Assets Other Investments Assets GDP % (using annual GDP)	Quarterly	IMF IFS
	FDI Assets Other Investments Assets growth rate %	Quarterly	IMF IFS
	Exchange rate LCU / US$, Value, LCU	Annual	World Bank MPO
	FDI Liabilities Portfolio Investments GDP % (using annual GDP)	Quarterly	IMF IFS
	FDI Liabilities Portfolio Investments growth rate %	Quarterly	IMF IFS
	FDI Assets Portfolio Investments GDP % (using annual GDP)	Quarterly	IMF IFS
	FDI Assets Portfolio Investments growth rate %	Quarterly	IMF IFS
	Official Reserve Assets and Other Foreign Currency Assets, US Dollars	Quarterly	IMF IRFCL
	Short-term total external debt as percentage of reserves	Quarterly	IMF IRFCL, WB QEDS
	FX growth rate %	Annual, Quarterly	Bloomberg
Financial	Deposit rate	Annual, Quarterly, Monthly	IMF IFS
	Deposits to total loans	Annual, Monthly	IMF FSI
	Customer Deposits to Total (Non-interbank) Loans, Percent	Quarterly	IMF FSI
	Real deposit rate growth %	Annual, Quarterly, Monthly	IMF IFS
	Lending rate	Annual, Quarterly, Monthly	IMF IFS

	Account per thousand adults, commercial banks	Annual	FinStats
	Number of branches per 100,000 adults, commercial banks	Annual	FinStats
	ODC Liabilities to Non-residents (% of GDP)	Annual	IMF IFS
	ODC Liabilities to Non-residents /GDP % rolling sum of 4 last GDPs)	Quarterly	IMF IFS
	Net foreign liabilities / Total bank assets (%)	Monthly	IMF IFS
	Private credit / GDP (%)	Annual	IMF IFS
	Private credit / GDP (%) GDP rolling sum of 4 last GDP	Quarterly	IMF IFS
	Private credit as percentage of sector assets	Annual, Quarterly, Monthly	IMF IFS
	House price to income ratio	Quarterly	OECD
	House price to rent ratio	Quarterly	OECD
	Net interest margin	Annual	FinStats
	Bank capital to assets	Annual	FinStats
	Stock Market Capitalization (% of GDP)	Annual	FinStats
	Return on assets	Annual	FinStats
	Return on equity	Annual	FinStats
	Private credit to deposits	Annual	FinStats
	Credit to Government and SOEs / Total bank assets (%)	Annual, Quarterly, Monthly	IMF IFS
	numeric S&P long-term foreign currency issuer credit rating	Quarterly	Bloomberg
	Numeric S&P long-term foreign currency issuer credit rating	Annual	Standard and Poor's
	numeric S&P Issuer long-term Local currency credit rating	Quarterly	Bloomberg
	Numeric S&P Issuer long-term Local currency credit rating	Annual	Standard and Poor's
	Interest rate spread	Annual, Quarterly, Monthly	IMF IFS
	Sovereign CDS 5-year spread	Annual, Quarterly, Monthly	Bloomberg
	Composite risk 50 Max risk; 0 Min risk	Monthly	ICRG
	EIU Banking sector risk score, 1-100 higher more risk	Monthly	EIU
	EMBI spread	Annual, Quarterly, Monthly	Bloomberg
	Economic risk (50 Max risk; 0 Min risk)	Monthly	ICRG
	Liabilities, Direct Investment, US Dollars	Quarterly	IMF IFS
	Assets, Direct Investment, US Dollars	Quarterly	IMF IFS
	Financial risk 50 Max risk; 0 Min risk	Monthly	ICRG
	Net open position in foreign exchange to capital, Percent	Annual, Quarterly, Monthly	IMF FSI
	Non-performing loans to total gross loans, Percent	Annual, Quarterly, Monthly	IMF FSI
	Liabilities, Other Investment, Other Liabilities (with Fund Record), US Dollars	Quarterly	IMF IFS
	Assets, Other Investment, US Dollars	Quarterly	IMF IFS

	Liabilities, Portfolio Investment, US Dollars	Quarterly	IMF IFS
	Assets, Portfolio Investment, US Dollars	Quarterly	IMF IFS
	Political risk 50 Max risk; 0 Min risk	Monthly	ICRG
	Tier 1 capital to risk-weighted assets, Percent	Annual, Quarterly, Monthly	IMF FSI
Global	US Federal Funds Effective Rate	Annual, Quarterly, Monthly	Bloomberg
	VIX index	Annual, Quarterly, Monthly	Bloomberg
	US 10-year government yield	Annual, Quarterly, Monthly	Bloomberg
	US 5-year government yield	Annual, Quarterly, Monthly	Bloomberg
	US 2-year government yield	Annual, Quarterly, Monthly	Bloomberg
	US government 10Y-2Y spread	Annual, Quarterly, Monthly	Bloomberg
	TED spread 3- month	Annual, Quarterly, Monthly	Bloomberg
	Move index	Annual, Quarterly, Monthly	Bloomberg
	1st Crude Oil, WTI	Annual, Quarterly, Monthly	Bloomberg
	US CPI (index)	Annual, Quarterly, Monthly	IMF IFS
	US inflation rate (YoY), Percent	Annual, Quarterly, Monthly	IMF IFS
	US inflation rate (QoQ/MoM), Percent	Annual, Quarterly, Monthly	IMF IFS
	Non-Fuel Price Index, 2016 = 100	Annual, Quarterly, Monthly	Primary commodity prices IMF
	Food and Beverage Price Index, 2016 = 100	Annual, Quarterly, Monthly	Primary commodity prices IMF
	Food Price Index, 2016 = 100	Annual, Quarterly, Monthly	Primary commodity prices IMF
	Non-Fuel Price growth rate %	Annual, Quarterly, Monthly	Primary commodity prices IMF
	Food and Beverage Price growth rate %	Annual, Quarterly, Monthly	Primary commodity prices IMF
	Food Price growth rate %	Annual, Quarterly, Monthly	Primary commodity prices IMF
	Oil Price growth rate %	Annual, Quarterly, Monthly	Primary commodity prices IMF
Other	CPIA	Annual	World Bank Internal
	Population, total	Annual	WB WDI

2. Methodological details

The methodologies used in this paper is well developed and described. Below we provide a basic outline of each of the methodologies with technical aspects related to each described in the references in the Methodology section of the paper.

Logistic regression

We use logistic regression and express the crisis of type $i = (fiscal, financial, external)$ as:

$$P\left(\text{default}_{i,t}\right) = \frac{1}{1 + \exp\left(-\left(\beta_0 + \beta_1 x_{i1,t} + \beta_2 x_{i2,t} + \cdots + \beta_k x_{ik,t}\right)\right)}$$

Where x is an explanatory variable from k number of variables. Note that in our framework we also allow for cross-terms and regional and income dummies in the gating part of the model.

To mitigate overfitting, we use a L2 regularization (ridge regression), which adds a penalty term to the loss function:

$$L(\mathbb{B}) = -\sum_{t=1}^{N}\left[\widehat{\text{default}}_{i,t} \log\left(\widehat{\text{default}}_{i,t}\right) + \left(1 - \text{default}_{i,t}\right)\log\left(1 - \widehat{\text{default}}_{i,t}\right)\right] + \frac{\lambda}{2}\sum_{j=1}^{k}\beta_j^2$$

Where $L(\mathbb{B})$ is the loss function, $\widehat{\text{default}}_{i,t}$ is the predicted crisis probability, $\mathbb{w}$ is the model coefficients, λ is the regularization parameter, and p is the number of features. This approach helps to manage multicollinearity. λ controls the strength of the penalty, with larger values resulting in stronger regularization (i.e., a simpler model with higher bias but lower variance). The values of λ is determined via cross-validation.

Boosted trees

eXtreme Gradient Boosting (XGBoost) is an ensemble method that combines many decision trees in sequence to produce a predictor. The trees are split on each feature space (the explanatory variables, $\mathbb{x}$) to predict the probability of default. The model is trained iteratively, with each new tree learning to correct the errors for the previously grown ensemble of trees. After M trees, the final prediction is a weighted sum of the predictions from each individual tree.

For each training iteration m, we learn a new tree h_m that minimizes a loss function. The final model is:

$$\hat{y}_{i,t} = \sum_{m=1}^{M} \alpha_m h_m(\mathbb{x})$$

Where α_m are the learning rates (weights) that scale each tree's contribution. Note that this equation is the average of the decision trees, and applies to random forests as well as boosted trees. The latter builds trees sequentially with each tree correcting the errors of the previous tree, while random forests build trees independently (called bagging). The benefit of using XGBoost is that it typically offers higher predictive accuracy on large and complex datasets. The XGBoost employs regularization (penalties on tree complexity) to avoid overfitting.

Evaluation metrics

The literature identifies several evaluation metrics to assess models and to help with tuning the parameters. Below we list the main ones in the literature that are applied to our framework.

Our performance metrics to assess and select each model is based on an F1 score that balances precision vs. recall or the Area Under the ROC (AUC).

The F1 score (see Van Rijsbergen, 1979) is the harmonic mean of precision and recall:

$$F1 = 2.\frac{\text{precision. recall}}{\text{precision} + \text{recall}}$$

Where **recall** is the model's ability to correctly identify defaults (or countries that actually defaulted).

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}$$

A low recall means that the model misses many actual defaults (false negatives).

Precision measures the accuracy of the model when it predicts default – or, all of the countries predicted to default, how many actually defaulted. We can state this as follows:

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}}$$

Low precision means that we have many false alarms (or false positives) where we predict a crisis but a crisis does not materialize. The F1 score is thus an attempt to balance precision and recall.

The AUC measures the model's ability to rank observations from most likely to default to least likely. A perfect AUC is 1, while a random guess is an AUC of 0.5.

The F1 score emphasizes classification quality (useful for imbalanced datasets), while the AUC assesses discrimination power.

From our methods we then have the following estimates for each default type:

1. XGBoost (F1 and AUC) for yearly data
2. Logistic regression (regularization)

3. Feature importance and risk attribution

Feature Importance

Feature importance quantifies each feature's contribution to the model's predictions. There are several ways to measure feature importance. For an ensemble algorithm like the XGBoost used in both the expert and intermediate gating layers, metrics such as Gain, Cover, and Frequency can be calculated for each feature and averaged across all trees, allowing for a ranking of features by their importance in the model. These feature importance scores can help understand which features are most influential in the model's predictions. Gain, cover, and frequency can be described as follows:

- Gain: This measures the improvement in accuracy brought by a feature to the branches in which it is included across all trees in the model. It calculates the change in accuracy when a feature is used to split a node. Higher gain values indicate features that contribute more to reducing the error in the model.
- Cover: This metric calculates the number of samples affected by a feature when it is used for splitting. Features with higher cover values split more samples across all trees, indicating features that may frequently help with segmenting the data.
- Frequency: Frequency (or count) measures how often a feature is used to split data across all trees. Features that appear frequently are likely valuable for the model, although this metric doesn't consider the improvement in accuracy.

Risk Attribution

While feature importance aims to rank features according to how influential they are in the model's predictions, risk attribution aims to understand the exact impact that each feature has on the risk score prediction of a country. At any given moment, this impact may be:

- Positive: raising the country's risk score above the average risk score.
- Negative: reducing the country's risk score below the average risk score.
- Insignificant: not affecting the country's risk score.

At any given moment, the risk score of a country corresponds to the average risk score plus the sum of all significant risk contributions, calculated for each feature. Techniques used in the literature include SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-agnostic Explanations), among others. SHAP is based on Shapley values from cooperative game theory and calculates the contribution of each feature by considering all possible combinations of feature subsets. LIME follows the surrogate model approach and employs local linear models around each prediction to approximate the behavior of the original complex model in that small area of the feature space.

In this paper, we employ a numerical approximation for feature importance that deals well with the diverse nature of the features, from categorical to numerical, discrete to continuous, and mutable to immutable. Specifically, we note that while certain features cannot change, i.e., they are immutable (e.g., a country's location, continent, neighboring countries), other features change in accordance with economic performance and context, i.e., they are mutable. The risk attribution is calculated jointly for all immutable features that characterize a country (e.g., regional location) $Z^1, \dots, Z^p$, and separately for mutable features that can potentially change over time, contained in $X_t^1, \dots, X_t^q$.

Immutable country features

For every country i at time t, the joint effect of all immutable features is obtained as follows:

1. Calculate panel risk score mean

$$ERS = \mathbb{E}\left[RS_t\left(Z^1, \dots, Z^p, X_t^1, \dots, X_t^q\right)\right]$$

2. Calculate country i risk at average mutable inputs

$$RSE = RS_{it}\left(Z_i^1, \dots, Z_i^p, \mathbb{E}[X_{it}^1], \dots, \mathbb{E}\left[X_{it}^q\right]\right)$$

3. Obtain effect

$$ERS - RSE$$

Mutable country features

For every country i, the individual effect of each mutable feature j at time t, is obtained as follows:

1. Calculate panel risk score mean

$$RS_{it} = RS_{it}\left(Z_i^1, \dots, Z_i^p, X_{it}^1, \dots, X_{it}^q\right)$$

2. Calculate country i risk at average mutable inputs

$$RSE_{it}^j = RS_{it}\left(Z^1, \dots, Z^p, X_{it}^1, \mathbb{E}\left[X_{it}^j\right], \dots, X_{it}^q\right)$$

3. Obtain effect

$$RS_{it} - RSE_{it}^{j}$$

We note that the feature attributions calculated in this way correspond intuitively to ones obtained in a linear model. Specifically, in the linear regression

$$RS_{it} = \alpha_1 Z_i^1 + \cdots + \alpha_p Z_i^p + \beta_1 X_{it}^1 + \cdots + \beta_q X_{it}^q$$

The calculated effects correspond to $ERS - RSE = \alpha_1(Z^1 - \mathbb{E}Z^1) + \cdots + \alpha_p(Z^p - \mathbb{E}Z^p)$ for the total effect of the immutable features, and $RS_{it} - RSE_{it}^{j} = \beta_j\left(X_{it}^{j} - \mathbb{E}X_{it}^{j}\right)$ for the effect of feature j at time t in country i.