INTERNATIONAL MONETARY FUND

Predicting IMF-Supported Programs: A Machine Learning Approach

Tsendsuren Batsuuri, Shan He, Ruofei Hu, Jonathan Leslie and Flora Lutz

WP/24/54

2024
MAR

WORKING PAPER

 WP/24/54

IMF Working Paper
Finance Department

Predicting IMF-Supported Programs: A Machine Learning Approach

Prepared by Tsendsuren Batsuuri, Shan He, Ruofei Hu, Jonathan Leslie and Flora Lutz*

Authorized for distribution by Carlo Sdralevich
March 2024

ABSTRACT: This study applies state-of-the-art machine learning (ML) techniques to forecast IMF-supported programs, analyzes the ML prediction results relative to traditional econometric approaches, explores non-linear relationships among predictors indicative of IMF-supported programs, and evaluates model robustness with regard to different feature sets and time periods. ML models consistently outperform traditional methods in out-of-sample prediction of new IMF-supported arrangements with key predictors that align well with the literature and show consensus across different algorithms. The analysis underscores the importance of incorporating a variety of external, fiscal, real, and financial features as well as institutional factors like membership in regional financing arrangements. The findings also highlight the varying influence of data processing choices such as feature selection, sampling techniques, and missing data imputation on the performance of different ML models and therefore indicate the usefulness of a flexible, algorithm-tailored approach. Additionally, the results reveal that models that are most effective in near and medium-term predictions may tend to underperform over the long term, thus illustrating the need for regular updates or more stable – albeit potentially near-term suboptimal – models when frequent updates are impractical.

RECOMMENDED CITATION: T. Batsuuri, S. He, R. Hu, J. Leslie and F. Lutz. 2024. “Predicting IMF-Supported Programs. A Machine Learning Approach.” IMF Working Paper WP/24/54, International Monetary Fund, Washington D.C.

JEL Classification Numbers:	E65, E66, F47, G01, C45, C53
Keywords:	Early warning systems; IMF Lending; Machine Learning

* The authors would like to thank Carlo Sdralevich, Greetje Everaert, Heikki Hatanpaa, Andreas Bauer, Jeannie Khaw, Andrew Swiston, Qianying Chen, Nicolas Ernesto Magud and the participants at the FIN Economists’ Group Seminar for their valuable comments. We are also grateful to the Strategy, Policy & Review – Risk Unit team for their consultations regarding data and technical aspects. We would like to especially thank Maksym Ivanyna, Clément Marsilli, Kevin Wiseman, Ritong Qu, and Xin Weining for sharing their valuable advice and insights on the technical aspects of machine learning modeling. The study also greatly benefitted from comments received during the interdepartmental review process.

WORKING PAPERS

Predicting IMF-Supported Programs: A Machine Learning Approach

Prepared by Tsendsuren Batsuuri, Shan He, Ruofei Hu, Jonathan Leslie and Flora Lutz[1]

[1] The authors would like to thank Carlo Sdralevich, Greetje Everaert, Heikki Hatanpaa, Andreas Bauer, Jeannie Khaw, Andrew Swiston, Qianying Chen, Nicolas Ernesto Magud and the participants at the FIN Economists' Group Seminar for their valuable comments. We are also grateful to the Strategy, Policy & Review – Risk Unit team for their consultations regarding data and technical aspects. We would like to especially thank Maksym Ivanyna, Clément Marsilli, Kevin Wiseman, Ritong Qu, and Xin Weining for sharing their valuable advice and insights on the technical aspects of machine learning modeling. The study also greatly benefitted from comments received during the interdepartmental review process.

Contents

1. Introduction

The IMF was created in the aftermath of World War II to establish a framework for economic cooperation with the aim to build a more prosperous global economy. Since its establishment, the IMF has functioned as a lender of last resort by providing financial assistance for eligible member countries experiencing balance of payment (BoP) needs. The IMF's near-universal membership and its reliable financing through an array of instruments underline its important and well-recognized role at the center of the global financial safety net (GFSN). Amidst a turbulent and shock-prone global environment, it is crucial to anticipate member countries' financing needs and to understand its key drivers to ensure that the IMF remains adequately funded.

Based on this motivation, we analyze the capabilities of a wide set of machine learning (ML) techniques to predict member countries' future use of IMF-supported funding arrangements. Recent studies have highlighted the usefulness of machine learning techniques in the context of developing macroeconomic early warning systems stemming from their ability to detect non-linearities and select relevant features from large predictor sets.[2] Our focus lies in uncovering the potential of ML-based algorithms to capture the complex, nonlinear dynamics that likely drive the commencement of IMF-supported financing arrangements: a facet inadequately explored in existing studies that have analyzed the use of IMF-supported financing arrangements using traditional econometric methods such as logistic regression.[3] This project aims to address this gap in the literature and focuses on three main questions: (1) Can ML techniques improve forecasts of IMF resource use compared to traditional econometric methods and what methods perform best? (2) Which factors are most indicative of a country's future use of IMF resources and how sensitive are they to specific models? (3) How can international institutions like the IMF navigate the complexities of training and utilizing ML-based models for predictive purposes to achieve optimal performance and maintain relevance over time?

Our analysis provides three key insights. Firstly, ML-based techniques outperform traditional econometric methods in out-of-sample predictions of new IMF-supported arrangements. In line with previous studies, we find that decision tree-based algorithms like the random forest and extra tree are among the best performing methods. Secondly, the models consistently identify key predictors that align well with existing studies and demonstrate considerable agreement across different algorithms. Our results particularly stress the relevance of including variables related to a broad set of sectors, including the external, fiscal, real, and financial sectors as well as institutional factors like membership in regional financial arrangements. We further reveal non-linear relationships between key predictors and demonstrate the robustness of prediction performances to alternative feature sets. Thirdly, our empirical findings underline the importance of data processing decisions and regular model updates. We find that the best-performing models select different feature sets, sampling methods, and missing data imputation techniques: a finding that highlights the usefulness of a flexible, algorithm-tailored approach. Moreover, models with the most accurate near and medium-term (5-7 years) forecasts can experience reduced performance in the long term (beyond eight years), underscoring the need for regular updates or the use of more stable – though possibly less near-term optimal – models when updates are impractical.

Our empirical strategy is organized into four stages: (i) data processing, feature selection, and imputation, (ii) model selection, data splitting, and sampling, (iii) model training and evaluation, and (iv) model analysis. The

[2] See Samitas et al. (2020), Badia et al. (2022), Weisfeld et al. (2020), Hellwig (2021), Jarmulska (2022), Fouliard et al. (2021), and Malladi (2022).

[3] See, for example, Poulain & Reynaud (2017), Maeder et al. (2019), Hills, Nguyen, & Sab (2021), and Agbloyer et al. (2023).

remaining sections of this paper describe each stage of this modeling pipeline in turn. In Section 2, we summarize our data processing steps including the combination of source datasets, feature selection, variable transformations, and missing value imputation. Our analysis includes data from 1982-2021 for 189 countries and covers essentially all IMF-supported arrangements (GRA and PRGT) excluding emergency financing instruments and undrawn arrangements. We include a broad set of predictors that encompass the external, fiscal, financial, and real sectors as well as structural variables. To maintain a large number of observations, we impute missing data using different imputation methods including K-nearest neighbors' imputation as well as imputation based on mean and median values by income-level country categorization.

Section 3 defines the models that are considered in our analysis. These comprise a broad set of ML classification models including the: logistic regression, regularized logistic regression, kernelized support vector machine (SVM), K-nearest neighbors (KNN), ensemble models such as random forest (RF), extra tree, and boosting techniques (XGBoost, RUSBoost, ADABoost), and a deep learning model (recurrent neural network (RNN)). We then follow standard procedure and split the data into a training set covering the years 1982-2017 which is used to estimate the models and an out-of-sample, hold-out test set covering the years 2018-2021 used to evaluate model performance. Because IMF-supported arrangements are rare as a proportion of the total sample, we address this observed class imbalance in the dependent variable through the use of various over-sampling and under-sampling techniques including Synthetic Minority Oversampling Technique (SMOTE), Adaptive Synthetic (ADASYN) technique, and random under-sampling. Each method produces an even split between observations with and those without IMF-supported arrangements when training the models.[4] In this stage, we also split the training set into four subsets – i.e., folds – each composed of an in-sample training and out-of-sample validation set based on an expanding time window.

In Section 4, we implement a model horse race that compares the predictive performance of all considered models on the hold-out test set constituting stage three of our modeling process. Model hyperparameters are tuned by selecting the best hyperparameter setting for each model type judged based on the highest average area under the receiver operating characteristic curve (ROC-AUC)[5] calculated across the four validation sets in the cross-validation procedure as defined in the second stage. Models with the best hyperparameter setting are refitted on the entire training sample and evaluated using the hold-out test set. The results of the out-of-sample performance horse race indicate that the random forest, extra tree, and recurrent neural network achieve the highest ROC-AUC scores on the hold-out test set while all ML-based algorithms outperform traditional logistic regression.

We further assess model performance by examining the trade-off between false alarms (type 1 error) and missed arrangements (type 2 error). The machine learning models we consider produce output values that are interpretable as predicted probabilities. To categorize these predicted probabilities as a binary prediction – specifically, whether or not a country is expected to start a new IMF-supported arrangement – we compare the predicted probabilities to a designated classification threshold. A country is predicted to engage in a new IMF-supported arrangement whenever its associated predicted probability exceeds this threshold. As a result, lower thresholds will reduce the number of missed arrangements while increasing the likelihood of false alarms. We

[4] Sampling techniques are applied only to the training sample after splitting the data into training and validation/test sets so the under- or over-sampled observations won't affect the sample that should be evaluated as a proxy to represent the model performance on future predictions. It can also help with to avoid data leakage issues, which occur when information from outside the training set is used to train the model.

[5] ROC-AUC refers to the receiver operating curve – area under the curve, a widespread metric used to assess prediction performances of classification algorithms. We provide a detailed description of the metric in Section 3.

evaluate this trade-off across models by comparing the relationship between each model's precision and recall at various classification thresholds through plotting precision-recall curves.[6] The extra tree model has the best combined precision and recall at most classification thresholds followed by the random forest. The RNN and XGBoost outperform the random forest at some combinations of higher recall and lower precision. We plot histograms of the predicted probabilities for three representative models[7] and show that the distribution of the predicted probabilities varies greatly across models, underscoring the need for an algorithm-tailored approach when setting decision thresholds. Country-level predictions highlight the potential usefulness of setting adaptive thresholds based on income groups or changes in a country's predicted probability over time in order to reflect fundamental differences across countries. The reported predictions illustrate that the representative models predict IMF-supported arrangements observed during the hold-out test period well and there is considerable agreement in predictions across models. We also present an ensemble approach that combines the prediction results across the representative models which could mitigate outliers of individual algorithms and provides further insights about the dispersions across models.

In Section 5, we turn to a more in-depth analysis of the representative models. Firstly, we examine the role of various explanatory variables in predicting the future use of IMF resources. Specifically, we use Shapley values to analyze feature importance and explore non-linear interactions between predictors.[8] Our findings confirm that the influential factors align with established economic theories and span various types of indicators including: fiscal variables like public debt to revenue and general government revenue, external variables such as the current account balance and external debt, real variables such as per capita income relative to the US, financial variables like private credit, and structural variables – most notably access to regional financing arrangements (RFAs). These key predictors remain consistent across algorithms with minimal variation in the selected features across models.

We then evaluate model robustness via two exercises. Our first exercise involves adding variables that have been deemed important by earlier contributions but were not selected by our models to be included in the optimal feature sets. We find that ensemble, non-linear models like the random forest maintain their prediction performance, showcasing resilience even when key variables might not have been included in the baseline model. On the other hand, linear models such as regularized logistic regressions display significant performance variations thereby underlining their vulnerability to the potential omission of important variables. Our second robustness exercise evaluates the temporal stability of model performance. We train the random forest, RNN, and regularized logistic regression models on sub-samples consisting of different time periods and then compare the models' out-of-sample performance in subsequent years. The models' out-of-sample prediction shows a general decline in performance over time. This decline emphasizes the changing nature of IMF-supported program use that is influenced by either supply changes like IMF policy adjustments or country demand shifts such as the emergence of alternative funding sources e.g., the rise of China as a lender of last resort (Alfaro & Kanczuk, 2019). Additionally, we note a performance dip during global crises (Asian Financial Crisis, Global Financial Crisis, and COVID-19 pandemic), which underscores the inherent difficulty in predicting such events (Aikman et al., 2021). We conclude our study by using the binary prediction results as input to a linear regression

[6] In our context, precision refers to the proportion of a model's predicted new IMF arrangements that were true new IMF arrangements in actuality, while recall refers to the proportion of all true new IMF arrangements that were correctly predicted by the model.

[7] We select three models as representatives, i.e., regularized logistic regression, random forest, and recurrent neural network, and focus analyses on them,

[8] Shapley value is an approach from coalitional game theory and has been widely used to evaluate feature importance in prediction problems. More details are discussed in Section 5.

to forecast the size of IMF-supported arrangements. While our results indicate that this procedure can provide some suggestive evidence regarding the size of expected IMF-supported arrangements, future work could explore the capabilities of ML-based techniques in this context.

This study relates to two main strands of the literature. First, it relates to a literature assessing the determinants of IMF-supported financial arrangements. Using binary response models, these studies highlight different drivers of IMF lending including country-specific variables such as gross international reserves, GDP growth, and external debt (Knight & Santaella, 1997; Trudel, 2005; and Cerutti, 2007), global variables such as oil prices, world interest rates, and the global financial cycle (Elekdağ, 2006; McGettigan & Reynaud, 2017; and Poulain & Reynaud, 2017), and institutional factors (Bird & Rowlands, 2001; Vreeland, 2004; and Maeder et al., 2019). While these contributions study the determinants of IMF lending among a large panel of advanced and emerging economies, others focus on concessional IMF-supported arrangements (Hills et al., 2021) or the repeated use of IMF-supported programs (Iseringhausen et al., 2019). Linear binary response models are generally found to have some but limited predictive power.[9] Most closely related is the study by Agbloyor et al. (2023) which uses machine learning models to predict IMF-supported arrangements. Apart from traditional macroeconomic factors, they highlight the importance of structural, energy and health-related, and social factors for prediction purposes. Compared to this study, we take further steps by introducing several additional dimensions of analysis including feature sets, sampling techniques, and imputation methods, studying larger feature sets and longer samples, investigating predictor interactions and the size of IMF-supported arrangements.

Second, this study relates to a recent literature that identifies economic early warning indicators and assesses the ability of machine learning techniques to predict crises. These studies show that machine learning techniques can deliver significant improvements in the accuracy of out-of-sample predictions compared to standard econometric approaches. Different studies have focused on the prediction of different types of crises including fiscal crises (Hellwig, 2021; Badia et al., 2022; Cerovic et al., 2018; Jarmulska, 2022), financial crises (Fouliard et al., 2021), balance-of-payment crises (Weisfeld et al., 2020), the recent Covid-19 recession (Malladi, 2022), and financial contagion risk (Samitas et al., 2020). The IMF's Vulnerability Exercise (VE) leverages machine learning algorithms to identify near-term country-specific macroeconomic risks in the fiscal, external, financial, and non-financial corporate and household sectors (IMF, 2022; Hacibedel & Qu, 2022). As in this study, machine learning-based models are evaluated against more conventional models in a horse race format.[10] In contrast to these studies, this analysis focuses on the use of IMF resources rather than crises.

2. Data Processing

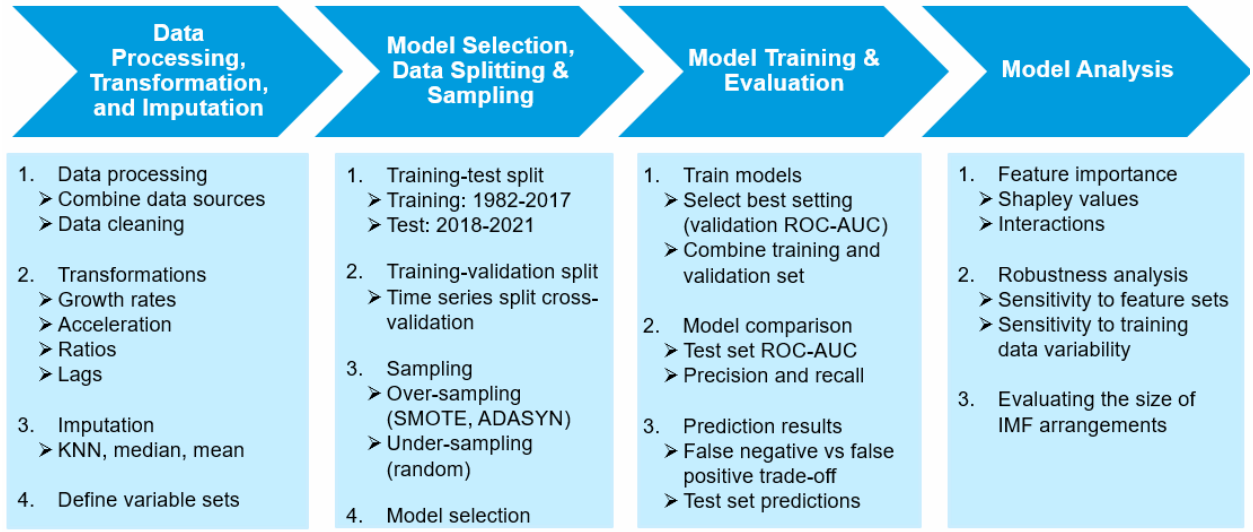
Our empirical strategy is summarized in Figure 1 and includes four main steps: (i) data processing, feature selection, and imputation, (ii) model selection, data splitting, and sampling, (iii) model training and evaluation, and (iv) model analysis. In the first stage described in this section, we combine source datasets, select features, and convert variables that are in units of national currency to ratios (i.e., as a proportion of GDP, exports, imports, fiscal revenues, or IMF quotas) or percentage changes to avoid inconsistencies in currency units and the impact of changing price levels. We generate transformations for each variable including growth rates, accelerations,

[9] The predictive power of binary response models has been evaluated by Cerutti (2007), Iseringhausen et al. (2019) and the 2022 Review of the Adequacy of the Funds Precautionary Balances, among others.

[10] Hellwig (2021) and Weisfeld et al. (2020) also provide a comparison of different ML techniques and conventional models, including random forests, elastic nets, and decision trees.

and lags. Missing values are imputed using three different imputation methods, K-nearest neighbors' imputation, mean imputation by country income-level group, and median imputation by country income-level group.

Figure 1: Process Flow



Sample Selection and Dependent Variable Definition

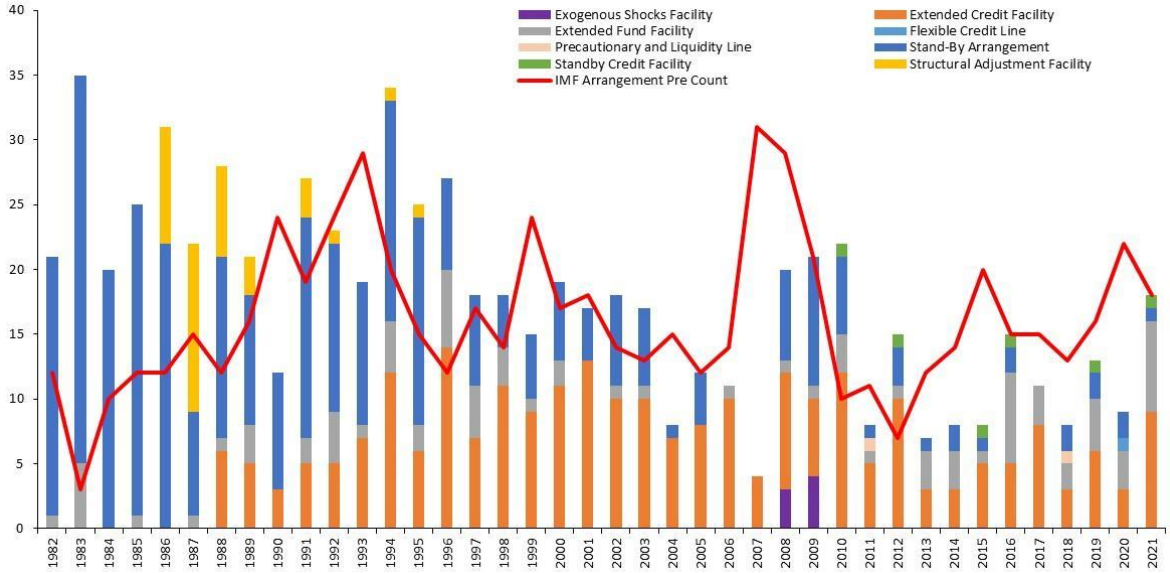
Our sample of IMF-supported arrangements consists of almost all arrangements that were approved between 1982 and 2021.[11] Specifically, the analysis covers 189 countries and includes 710 arrangements, of which 409 are GRA arrangements and 301 are PRGT arrangements. The programs were requested by 129 different member countries, and the number of programs observed per country ranges from 0 to 13. We exclude emergency financing instruments – which includes 90 Rapid Credit Facilities (RCF) and 46 Rapid Financing Instruments (RFI) – for three main reasons. First, the design and purpose of emergency financing programs fundamentally differs from a full-fledged economic program. Specifically, emergency financing programs only include a one-time disbursement, are provided without ex-post program-based conditionality and aim to respond to situations in which a fully-fledged economic program is not necessary (transitory and limited need) or feasible (due to e.g., policy design, capacity and other implementation constraints). Second, the definition of the dependent variable, excluding two periods after program commencement (see Figure 3), does not align with the average duration of emergency programs. Third, emergency financing programs are highly concentrated following the outbreak of the Covid-19 pandemic which could distort prediction results for the testing set.[12] We

[11] While a longer sample period increases the overall number of observations, it also tends to increase the share of missing observations. Given the fact that several variables have been reported since the early 1980, we selected 1982 as the cut-off year.

[12] Emergency financing instruments provide prompt financial assistance to member countries facing urgent balance of payments needs caused by sources including domestic instability, exogenous shocks and fragility. Compared to Upper Credit Tranche (UCT) arrangements, they are provided without ex-post program-based conditionality or reviews and only include one single disbursement. The total amount of the approved RCFs and RFIs was about SDR 24 billion, compared to SDR 615 billion for the approved arrangements included in our sample. Annex I provides a figure including all IMF arrangements approved since 1982, including RCFs, RFIs and precautionary instruments. It is important to note, however, that countries receiving emergency financing will face larger repayments to the fund going forward, thereby increasing countries' demand for UCT fund arrangements. We capture this effect by including fund credit outstanding as a predictor variable in our robustness analysis and find that this is indeed an important factor.

also exclude any arrangements that ended without an actual balance of payments need, i.e., any undisbursed arrangements.[13] All arrangements since 1982, excluding emergency financing loans and undrawn arrangements, are summarized in Figure 2.

Figure 2: Dependent Variable and Composition of Arrangements



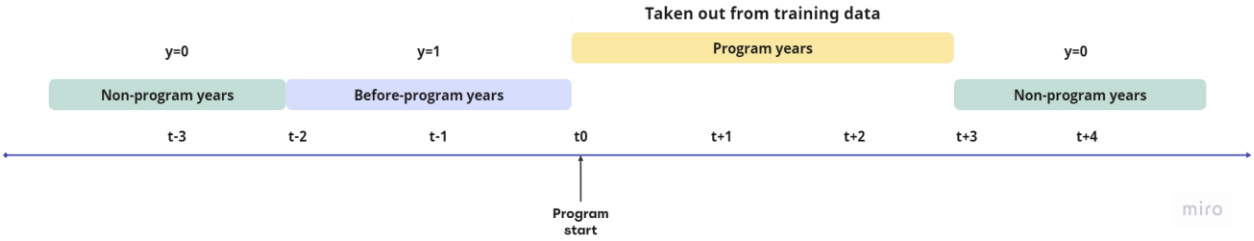
Notes: Number of IMF arrangements per year and lending facilities excluding emergency financing instruments (RCFs and RFIs) as well as undrawn arrangements.

The dependent variable in the baseline specification is a binary variable equal to one for a country in each of the two years prior to the approval of a new IMF-supported arrangement, i.e., t-1 and t-2. To increase the distinction between pre-program and non-program years, we exclude the year of approval as well as the following two years from the sample, i.e., periods t, t+1, and t+2, as illustrated in Figure 3. The yearly sum of the resulting pre-arrangement observations representing our dependent variable to be predicted is shown by the red line in Figure 2.[14] It is important to note that this project focuses on the prediction of materialized IMF arrangements which not only requires a positive demand for an arrangement (i.e., an actual balance-of-payment need) but also a mutual agreement on the terms and conditions of the arrangement. Program requests which do not result in the approval of a new arrangement, e.g., due to disagreement regarding program conditions, are not observed and hence treated as a non-program observation in the analysis.

[13] 112 of the 130 undrawn arrangements were officially classified as precautionary arrangements which are designed to provide financial support to meet actual or potential balance of payments needs of countries with sound policies that may have some remaining vulnerabilities.

[14] This definition is equivalent to the definition used in the vulnerability exercise at the IMF. Our dependent variable definition is also based on the fact that the average duration of programs in our sample is 2.4. An alternative definition used in the literature is a binary variable equal to one in the approval year of a new arrangement and zero for the non-program years (program years except the approval year are excluded).

Figure 3: Dependent Variable Definition



Feature Selection and Missing Value Imputation

We include a broad set of predictors spanning the external, fiscal, financial, and real sectors. In addition, our sample includes several global variables such as the VIX index, U.S. treasury yields, and the federal funds rate. We also include structural variables such as corruption and polity scores from the International Country Risk Guide (ICRG), costs of natural disaster hazards, and population growth. Because countries that have access to other layers of the global financial safety net might be less likely to request IMF arrangements due to their alternative options, we account for these additional factors by including dummy variables that indicate access to central bank currency swap lines and regional financing arrangements (RFA) as well as a measure of political closeness to the US based on UN assembly votes data.[15] In the robustness analysis, we further show that including credit outstanding with the IMF as an explanatory variable can further improve prediction performance significantly.

The complete set of predictors is summarized in Annex I, Table I.1. Given that around one third of the predictors have a share of missing values larger than 50 percent, we impute missing values to maximize coverage across countries and over time. Specifically, we compare the performance of three different imputation techniques: K-nearest neighbors' imputation as well as mean and median imputation by three country groups defined as advanced economies (AEs), low-income developing economies (LIDCs), and EMDEs that are non-LIDCs.[16]

We further generate transformations for each predictor variable including growth rates, acceleration, and lags. The full feature set, including all transformations, comprises 1014 features. To provide an equal starting point for model comparison, we define six feature sets on which we test all models as summarized in Table 1. While the first three sets remove transformations and use increasing missing value thresholds to reduce the number of features, sets 4-6 are defined using feature selection techniques based on the parameter estimates in the Lasso regularized logistic regression and the recursive feature elimination of the random forest model.[17]

[15] We use the dataset provided by Bailey et al. (2017) and proxy the political closeness to the US as follows: (ideal point distance U.S. – ideal point distance if country x) * (-1).

[16] Mean and median imputation were done at a coarse level to avoid dropping countries that have some variables missing across all years while still excluding the use of data points from inherently different economies (e.g., AEs sovereign bond spreads for emerging markets spreads). We noticed a change in posterior distributions, i.e., the mean and median imputed series show a sharp increase in the density at the country group specific means and medians while the general shape of the KNN imputed series remains closer to the base series. We thus conducted the Kolmogorov-Smirnov test, which rejected the null of statistically similar distributions for most variables. See Annex I, Figure I.2 for more details.

[17] It is important to note that the selected feature elimination steps do not address multicollinearity which is almost surely present in the datasets. Future research could explore optimal feature selection and multicollinearity more comprehensively. Importantly, however, multicollinearity should not cause any issue as long as the model is used for prediction purposes rather than causal inference.

Table 1: Variable Sets

Variable Set	Number of Features	Transformations Included	Missing Value Threshold	Other Criteria	Time Period
Full Set	1014	1-year growth rate, 2-year growth rate, 3-year growth rate, 5-year growth rate of levels and ratios, acceleration, first and second lags	None	Drop variables denoted in local currency units and levels	1982-2021
Set 1	480	1-year and 5-year growth rates of levels and ratios, acceleration, first lags	85% (eliminates 5 base variables*)	Drop highly correlated variables	1982-2021
Set 2	277	1-year and 5-year growth rates of levels, first lags	70% (eliminates 9 base variables*)	None	1982-2021
Set 3	109	1-year growth rate of levels	50% (eliminates 9 base variables*)	None	1982-2021
Set 4	76	Not applicable	Not applicable	Lasso	1982-2021
Set 5	70	Not applicable	Not applicable	Random Forest – Recursive Feature Elimination	1982-2021
Set 6	40	Not applicable	Not applicable	Random Forest – Recursive Feature Elimination	1982-2021

*Notes: Base variables are the variables which have been used to calculate the transformations included in the full dataset. Set 5 and 6 both rely on the random forest recursive feature elimination procedure but select a different number of features (70 vs. 40).

3. Models and Evaluation Procedures

In the second stage of our modeling process, we first specify the various ML classification models that we consider in our analysis. We then establish the cross-validation process that is used to tune each model's hyperparameters as well as define the split between the data that will be used to train the models and the out-of-sample test data that will be used to evaluate the models. Finally, we define the application of four different over-sampling and under-sampling techniques that we use to address class imbalance in our dependent variable.

Model Selection

In our analysis, we compare a broad set of traditional econometric and ML classification models including the standard logistic regression, regularized logistic regression with different penalty terms, K-nearest neighbors, kernelized support vector machine, ensemble models including random forest, extra tree, boosting techniques (XGBoost, ADABoost, RUSBoost), and a deep learning model (a recurrent neural network (RNN)). The logistic regression model serves as the baseline representing a traditional econometric model which is widely used in the existing literature to predict IMF arrangements and study their determinants. Penalty terms are added to the cost function of the linear model, i.e., L1 (Lasso), L2 (Ridge), or both (Elastic Net with weights for L1 and L2) to reduce the number of coefficients and control model complexity.[18]

We compare our baseline logistic model with the alternative ML classifier models listed above, which are structured to better recognize nonlinearities and/or specialize in capturing sequential and temporal patterns in the data. The K-nearest neighbors (KNN) algorithm calculates the value of a test data point based on the closest training sample data points in the feature space. Kernelized support vector machines (SVM) involve polynomials or interactions of features and find the decision boundary between two classes in a higher dimension of the feature space. Predictions are made based on the distance of a test data point to the training data points that define the decision boundary, i.e., support vectors. A recurrent neural network (RNN) is a variant of standard neural network models, which are constructed as a series of nested non-linear functions that map inputs into a chosen output via optimizing function parameters to achieve an objective. RNNs are specialized versions of this architecture that learn patterns in sequences of ordered data by calculating a 'memory' state based on the other observed feature values in the sequence.

We also consider various decision tree-based methods. The random forest is an ensemble machine learning model introduced by Breiman (2001) that constructs numerous decision trees during the training process and aggregates their outcomes for improved accuracy and overfitting prevention. Each tree in the forest is trained on a random subset of data and considers a random set of features at each split, which introduces variability and enhances the model's generalization capabilities.[19] To classify observations, a random forest model takes the majority vote from all trees as its prediction. In contrast, boosting models like XGBoost build trees sequentially with each tree attempting to correct the errors of the previous one thereby resulting in a focus on difficult-to-classify instances. The fundamental difference between random forest and boosting models therefore lies in the random forest's parallel and diversified approach compared to boosting's sequential and error-corrective nature.

[18] The lasso penalty term is the sum of the absolute values of the coefficients while the ridge penalty term is the sum of the squared values, with a hyperparameter alpha to control for the strength of pushing coefficients towards 0. Using lasso regularization can make some coefficients equal to zero and thus can help reduce the number of features, while the ridge regularization can only make coefficients close to zero but not exactly equal to zero.

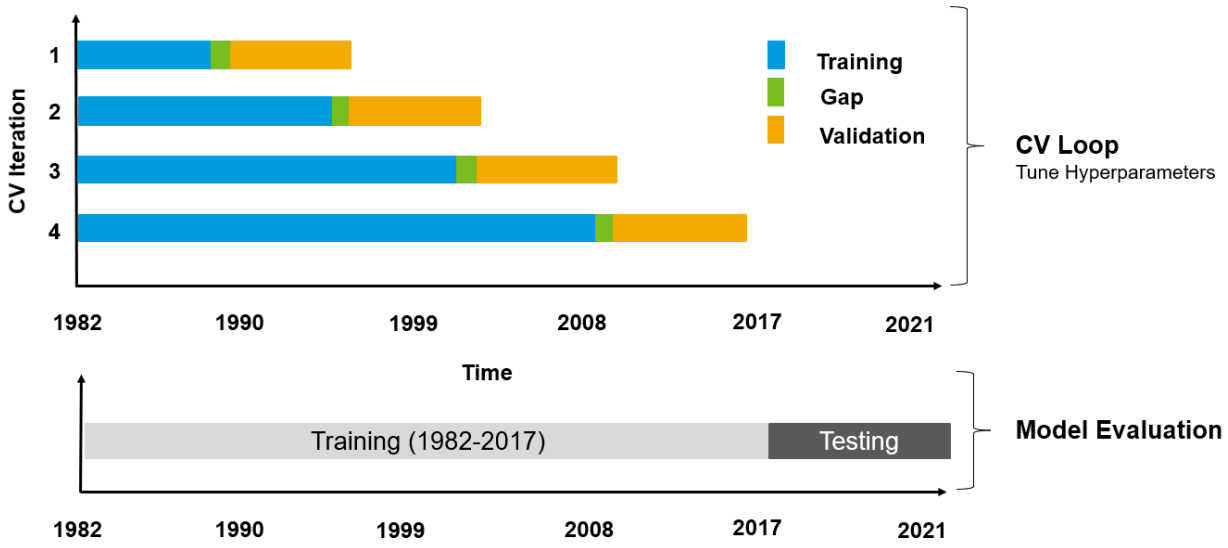
[19] In the context of machine learning, 'trees' refer to decision trees, which map features of data to outcomes, and 'forests' refer to ensembles of decision trees, which work together to improve predictive performance.

Data Splits and Cross-Validation Procedure

To assess predictive performance across models, we split our data sample into a training set covering the years 1982-2017 and an out-of-sample, hold-out test set covering the years 2018-2021. In section 4, the models will be estimated by using the training data set through a cross-validation procedure and then evaluated based on their out-of-sample predictions on the test data set.

Each of the ML models are governed by various hyperparameters such as the strictness of the penalty terms in the regularized logistic regressions or the number of layers in the recurrent neural network. We therefore implement a cross-validation procedure to select the best hyperparameter setting for each of the model types. In the analysis that follows, we compare the different models by using their best hyperparameter setting as determined in the cross-validation process. Specifically, we use the expanding-window gap k-fold time-split approach to account for the time dependence in the data and to avoid data leakage (Burman et al., 1994).[20] We split the training data into four subsets, referred to as folds, that are each further split into an in-sample training subset and an out-of-sample validation subset based on an expanding time window. The training set for the first fold starts with the first seven years in our sample (1982-1988), and the validation set for the first fold consists of the following seven years after including a one-year gap (1990-1996). In each of the following three folds, the training period is iteratively expanded by adding the next seven years, while the size of the validation sets is kept constant as the subsequent seven years following the training set after a one-year gap.[21] Importantly, this approach also reflects how the model will be used in real-time, i.e., using past data to predict the future.

Figure 4: Data Partitions with Time-Series Split



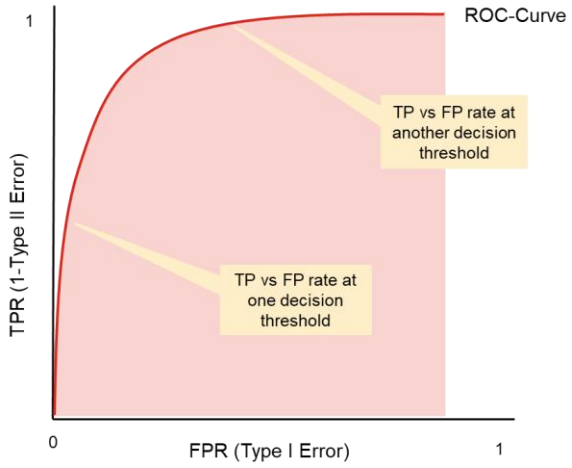
[20] Alternative approaches, such as stratified k-fold cross validation, increase the size of training sets and place equal weights on all observations, but, by shuffling the dataset randomly, ignore the temporal dependence of data and thereby fail to ensure the out-of-sample condition.

[21] The training and validation periods for each iteration are the following: (1) 1982-1988 as the first training period and 1990-1996 as the validation period, with 1989 as the gap year, (2) 1982-1995 for training and 1997-2003 for validation, with 1996 in between, (3) 1982-2002 and 2004-2010 with 2003 in between, and (4) 1982-2009 and 2011-2017 with 2010 in between.

In each fold, candidate versions of each model type with different hyperparameter settings are estimated using that fold's training data and then performance metrics are calculated based on the estimated model's predictions on that fold's out-of-sample validation data. For each model type, the candidate hyperparameter setting that achieves the best average out-of-sample performance on the validation data across the four folds is selected as the best hyperparameter setting that will represent the model in the rest of the analysis. The cross-validation procedure to set model hyperparameters as well as the split between the complete training and test set are illustrated in Figure 4.

We utilize the area under the Receiver Operating Characteristic (ROC-AUC) curve averaged across the four validation sets in the cross-validation procedure as the performance metric to select the best hyperparameter setting for each model type. As shown in Figure 5, the Receiver Operating Characteristic curve plots a classifier model's false positive rate (FPR) relative to its true positive rate (TPR) across varying classification thresholds. For a given model, a high ROC-AUC value therefore indicates a superior classifier with a high true positive rate and a low false positive rate. We selected the ROC-AUC score as our principal evaluation metric because it summarizes false positive rates and true positive rates across various classification thresholds and therefore provides a holistic evaluation of model performance in a single, comparable evaluation metric.[22] While we do not consider alternative evaluation metrics during the hyperparameter tuning stage, we provide a detailed comparison of precision-recall curves and analyze trade-offs between false positive and false negative rates when assessing model performance in Section 4. The resulting best hyperparameter settings for each model type are summarized in Annex II, Table II.1.[23]

Figure 5: ROC Curve and Decision Threshold



[22] An ROC curve closer to the upper left is more favorable. ROC-AUC scores are a widely used evaluation metric for classification algorithms. While the ROC curve is among the most widely used performance evaluation metric for classification algorithms, root-mean squared errors, i.e., the square root of the averaged squared difference between the target value and predicted value, are the most popular metric for continuous prediction tasks.

[23] If two hyperparameter settings give similar average ROC-AUC scores during the cross-validation process, we select the setting with a lower standard deviation of ROC-AUC score across folds.

Class Imbalance and Sampling Methods

Given that IMF-supported arrangements are typically only requested by a small subsample of member countries each year, we face an imbalanced classification problem. Specifically, our sample includes 16 percent minority observations (i.e., pre-arrangement periods) versus 84 percent majority observations (i.e., non-pre-arrangement periods). The limited number of minority class observations makes it difficult for a classifier model to distinguish between observations of either class and generates a bias toward correctly identifying the majority class at the expense of missing minority cases. For example, in our case, a model that simply predicts all observations as a non-pre-arrangement period would achieve 84 percent accuracy without attempting to predict the occurrence of a single pre-arrangement period.

To address this class imbalance challenge, we either under-sample majority class observations (non-pre-arrangement periods) or over-sample minority class observations (pre-arrangement periods) during model training so that there is an equal amount of pre-arrangement and non-pre-arrangement observations.[24] Specifically, we compare the performance of four different sampling techniques in our analysis: (i) no sampling, (ii) random under-sampling, (iii) SMOTE, and (iv) ADASYN.[25] As we discuss in more detail in Section 4, our results indicate that model performance crucially depends on the sampling method, and best performing sampling techniques differ across algorithms.

4. Prediction Results

In this section, we compare the performance of machine learning-based models with the traditional logistic regression model in a horse-race format. As described earlier, we evaluate models based on their out-of-sample performance measured by ROC-AUC scores. We also discuss the trade-off between false negative and false positive alarms as well as the dispersion and agreement across model types at a country level.

Horse Race and Performance Comparison

Our analysis indicates that machine learning models consistently outperform traditional econometric methods on the hold-out test set. The main results are summarized in Figure 6 where the first four columns report the ROC-AUC scores for each model in each of the four validation sets, and the last column reports the results for the

[24] Under-sampling randomly or selectively reduces the observations in the majority class, i.e., the non-pre-arrangement periods, while the over-sampling replicates the observations in the minority class, i.e., the pre-arrangement periods, or creates artificial observations, so the proportion of non-pre-arrangement periods and pre-arrangement periods reach equality (50-50) in the training sample after the sampling techniques are performed. Thus, under-sampling will reduce the size of the training set, whereas over-sampling will increase it.

[25] SMOTE and ADASYN are two over-sampling techniques that create artificial observations based on the existing minority class. SMOTE considers K nearest neighbors from the minority class of a given minority observation in the feature space and generates an artificial sample on the line connecting a randomly selected neighbor and the original minority observation. The number of generated data samples is the same for each original minority observation. ADASYN takes into account the neighbors from the majority class and determines the number of generated samples for each original minority observation. See technical details and more sampling techniques in He & Garcia (2009). It is worth noting that SMOTE and ADASYN fail to reflect the sequential characteristics of time series data. Future research could explore structure preserving oversampling techniques.

hold-out testing set, i.e., the period we are focusing on for performance evaluation.[26] While logistic-based models achieve an ROC-AUC score of roughly 0.81 with their out-of-sample predictions of pre-arrangement periods on the test set, the ROC-AUC values achieved by the other models range from 0.82 to 0.86. Interestingly, the predictive performance of the logistic regression model and its regularized variants with different penalty terms (i.e., Lasso, Ridge, and Elastic Net) are very similar and each exhibited their best performance with the Lasso feature set. This is not only true for the ROC-AUC scores on the test set, which are identical at a three-digit level, but also for alternative metrics such as precision and recall.[27]

We find that tree-based models are most successful in out-of-sample prediction. The random forest and extra tree achieve the highest test set ROC-AUC score of 0.86 followed by the RNN and XGBoost models, which achieve scores of approximately 0.84. The finding that the random forest and extra tree models outperform other models aligns well with previous studies that highlight the superior prediction performance of tree-based models (see for example, Agbloyor et al., 2023; Hellwig, 2021; IMF, 2021 among others). More generally, the ROC-AUC scores illustrate the overall success of the algorithms in predicting IMF-supported arrangements. Within the machine learning literature, models that achieve an ROC-AUC of 0.8 or higher are typically considered to be 'good' classifiers, depending on the context of the prediction problem. Existing studies that compare machine learning-based algorithms for crisis prediction purposes achieved ROC-AUC scores in the range of 0.69 to 0.77.[28] Given the complexity of our prediction problem, we therefore see high potential in the ML models we considered and our modeling process more generally to provide an early warning system for IMF-supported programs.

Figure 6 also highlights significant variations in relative model performance over time. Model performance is best across all model types in 1997-2003: a period characterized by relatively stable global economic growth. By contrast, ROC-AUC scores decline significantly in the validation set ranging from 2004-2010, i.e., a period that includes the onset of the global financial crisis. Interestingly, we do not observe a decline in ROC-AUC scores during the recent COVID-19 pandemic, which may be a consequence of the fact that emergency financing loans – which were the IMF's key instrument to provide prompt financial assistance to member countries during the pandemic – were excluded from our sample. Regarding variations across model types, we observe that the KNN is the worst performing model class across all validation sets but scores above several competitor algorithms in the hold-out test set. This substantial improvement when estimated using the largest training set could be due to the nature of the KNN algorithm in which not-previously-seen observations are directly matched to their closest counterparts in the training data in order to make predictions, thereby increasing the value of a large set of training data representing a broad range of experiences. The random forest model (together with boosting methods) displays the most robust performance and ranks in the top-three across all validation and the test set. The random forest model's superior performance could be attributed to its ensemble approach, which mitigates overfitting by combining multiple decision trees. On the other hand, boosting methods (e.g., XGBoost) improve model performance by sequentially training weak models and giving more weight to misclassified samples in each iteration. This iterative process focuses on correcting errors and enhances model accuracy. The variability in performance across different folds of different tree-based models could be attributed to the heterogeneity in the

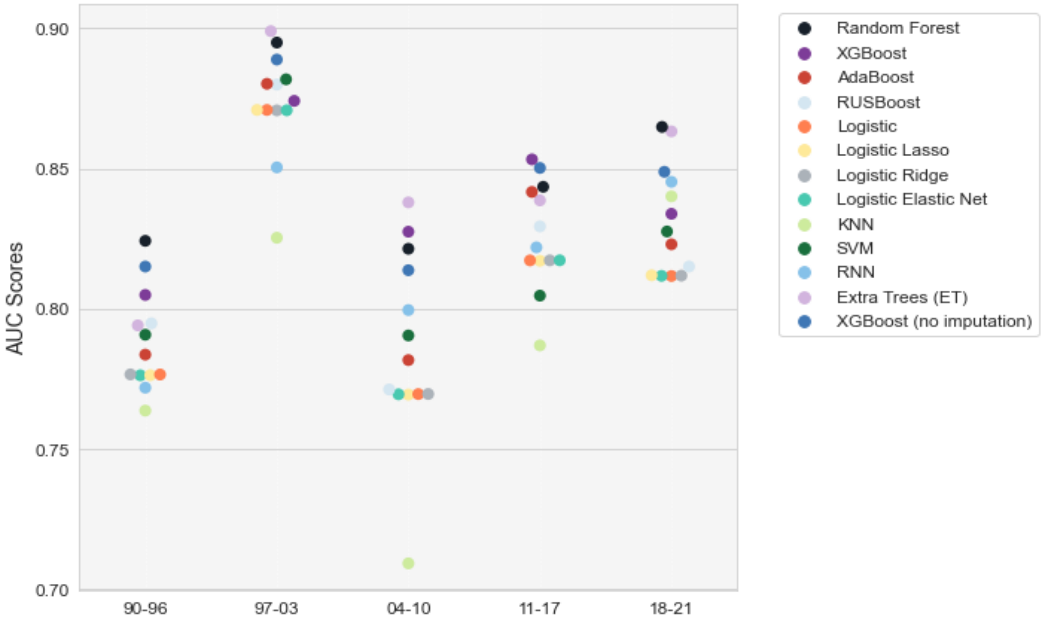
[26] In our analysis, we consider six possible feature sets, three different missing value imputation methods, as well as four different class imbalance sampling methods (see Sections 2 and 3). In Figure 6 and throughout the remaining stages of model evaluation and analysis, each model type is therefore represented by the feature set, imputation method, sampling method, and hyperparameter setting combination that achieves the best performance during the cross-validation procedure.

[27] The almost identical performance of the Logistic models is due to the fact that they rely on equivalent variable sets, i.e., they do not impose any additional regularization. As discussed in more detail in section 2, this approach was chosen to provide an equal starting point for model comparison.

[28] See Cerovic et al. (2018), Weissfeld et al. (2020), and Hellwig (2021), among others.

approved programs across each fold. It is also worth noting that, while the RNN scores below the traditional logistic-based models in the first and second validation sets, its ranking improves consistently as the size of the training set is expanded. This outcome corresponds with conventional experience regarding the tendency of deep learning models to improve their performance as the number of available data observations increases.

Figure 6: Horse-Race Results

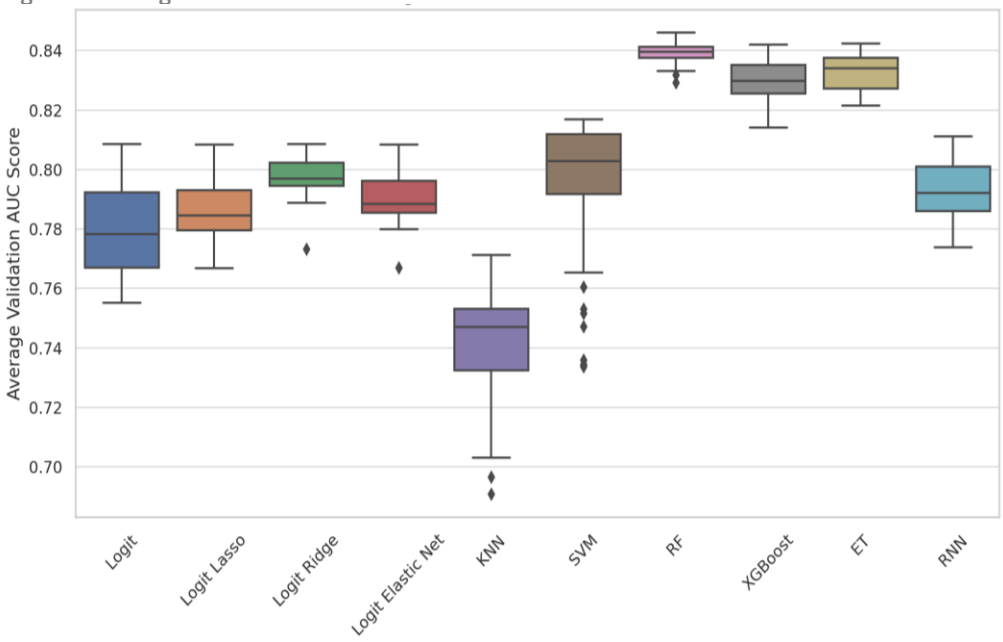


Our findings also highlight that data processing decisions and feature selection matter. Table 3 summarizes the missing value imputation methods, sampling techniques, and selected variable sets of the best performing specification of each model type. We find that the linear models consistently prefer the mean imputation while tree-based models perform best when using KNN imputation. Other machine learning models like the RNN and SVM perform best with median imputation. These results indicate the importance of assessing different imputation techniques when evaluating model performance. Similarly, we find significant variations in the best performing sampling methods across model types. While linear models as well as the RNN perform best using random under-sampling, tree-based models seem to prefer over-sampling methods. With regard to feature sets, all models tend to perform better with smaller feature sets and model-selected feature sets consistently demonstrate superior performance compared to pre-defined feature sets based on missing value thresholds and transformations. Figure 7 further quantifies how much data processing decisions can matter for forecasting accuracy. Specifically, it shows the range of AUC scores for each model across all combinations of tested variable sets, imputation and sampling methods. While data processing decisions matter across all tested algorithms, we find that they are of particular importance in linear settings such as the logistic-based models. The prediction performance of the random forest, on the other hand, seems to be most robust with respect to different data processing decisions. Figures III.1-4 in Annex III provide additional details regarding the importance of individual data processing decisions across models.

Table 3: Summary of Best Performing Models

Prediction Model	Imputation Method	Variable Set	Sampling	Test ROC-AUC Score
Random Forest	KNN	Set 5	SMOTE	0.86
Extra Tree	KNN	Set 6	SMOTE	0.86
XG Boost	None	Set 5	No Sampling	0.85
XG Boost	KNN	Set 5	SMOTE	0.84
Ada Boost	Mean	Set 6	SMOTE	0.82
RUS Boost	Mean	Set 5	ADASYN	0.81
RNN	Median	Set 4	Under-Sampling	0.84
KNN	KNN	Set 5	Under-sampling	0.84
SVM	Median	Set 6	ADASYN	0.83
Logistic	Mean	Set 4	Under-sampling	0.81
Logistic Lasso	Mean	Set 4	Under-sampling	0.81
Logistic Ridge	Mean	Set 4	Under-sampling	0.81
Logistic Elastic Net	Mean	Set 4	Under-sampling	0.81

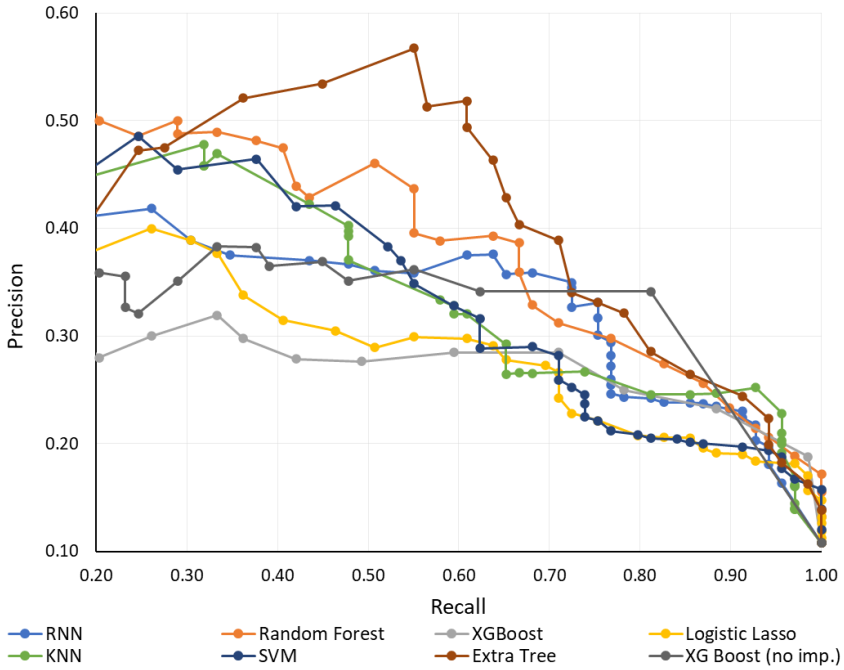
Figure 7: Range of Validation AUC Scores Across Models



Trade-off Between False Negatives and False Positives

To complement the ROC-AUC scores, which provide an aggregated view of overall model performance during the cross-validation and the hold-out test periods, we also evaluate the models by analyzing the trade-off between false negative and false positive predictions across different classification thresholds. First, we compare precision and recall in the test period (2018-2021) across various classification thresholds for every model. As discussed previously, in our context, precision refers to the proportion of a model's predicted pre-arrangement periods that were true pre-arrangement periods, while recall refers to the proportion of all true pre-arrangement periods that were correctly predicted by the model. Higher precision and higher recall imply better performance, with precision of 1 and recall of 1 constituting perfect classifiers. The resulting precision-recall curves, presented in Figure 8, confirm the strong performance of tree-based random forest and extra trees models. Specifically, the extra tree outperforms other models by a significant margin in terms of precision when recall is between 0.3 and 0.7, indicating a lower false positive rate of the extra tree at a given level of recall (false negative rate) in this interval. On the other hand, the recurrent neural network and XGBoost models outperform the random forest and extra tree at some higher levels of recall (above 0.7). It is further interesting to note that the KNN achieves the highest precision score when recall is at 0.93.

Figure 8: Precision-Recall Curves for 2017-2021



The trade-off between precision and recall is also evident in Figure 9, which shows histograms of the models' predicted probabilities for the test set. For illustrative purposes here and in analyses hereafter, we select three representative model types that encompass a linear econometric method, a tree-based model, and a deep learning model: the regularized logistic regression with the lasso penalty term, the random forest model, and the recurrent neural network. The dashed line in figure 8 indicates the classification threshold that maximizes F1

scores.[29] These thresholds vary significantly across models ranging from 0.35 in the random forest to 0.72 in the RNN. It is also interesting to note the differences in the distributions of predicted probabilities across model types. Specifically, the RNN tends to output very low predicted probabilities for most countries without pre-arrangement periods while the probabilities are more evenly distributed for the random forest and the logistic lasso regression. The predicted probabilities for true negatives (true positives) range from 0 to 0.96 (0.16 to 0.97 for true positives) for logit model with lasso penalty term, 0 to 0.84 (0.09 to 0.92) for the random forest, and 0 to 0.99 (0.01 to 0.98) for the RNN. Accordingly, a correct classification of all true negative events by the random forest requires a classification threshold of at least 85 percent implying a significant share of false negative alarms.

As we discuss in more detail in the next section, some of the true negative observations with high predicted probabilities are associated with emergency financing instruments which were treated as no program cases in the definition of our dependent variable. True positive observations with low probabilities, on the other hand, appear to include countries with limited data availability and countries that are regarded to have comparably strong economic fundamentals but made drawings under precautionary facilities.

Agreement and Dispersion of Prediction Results across Models

We now turn to the country-level predictions obtained by our three representative models and compare the prediction results with the approved IMF-supported arrangements in the relevant years. The results are shown in Figure 10. For illustrative purposes, the maps show the 2021 predictions (upper three graphs) and compare the prediction results with the actual active arrangements as of end-October 2023 (lower graph).[30]

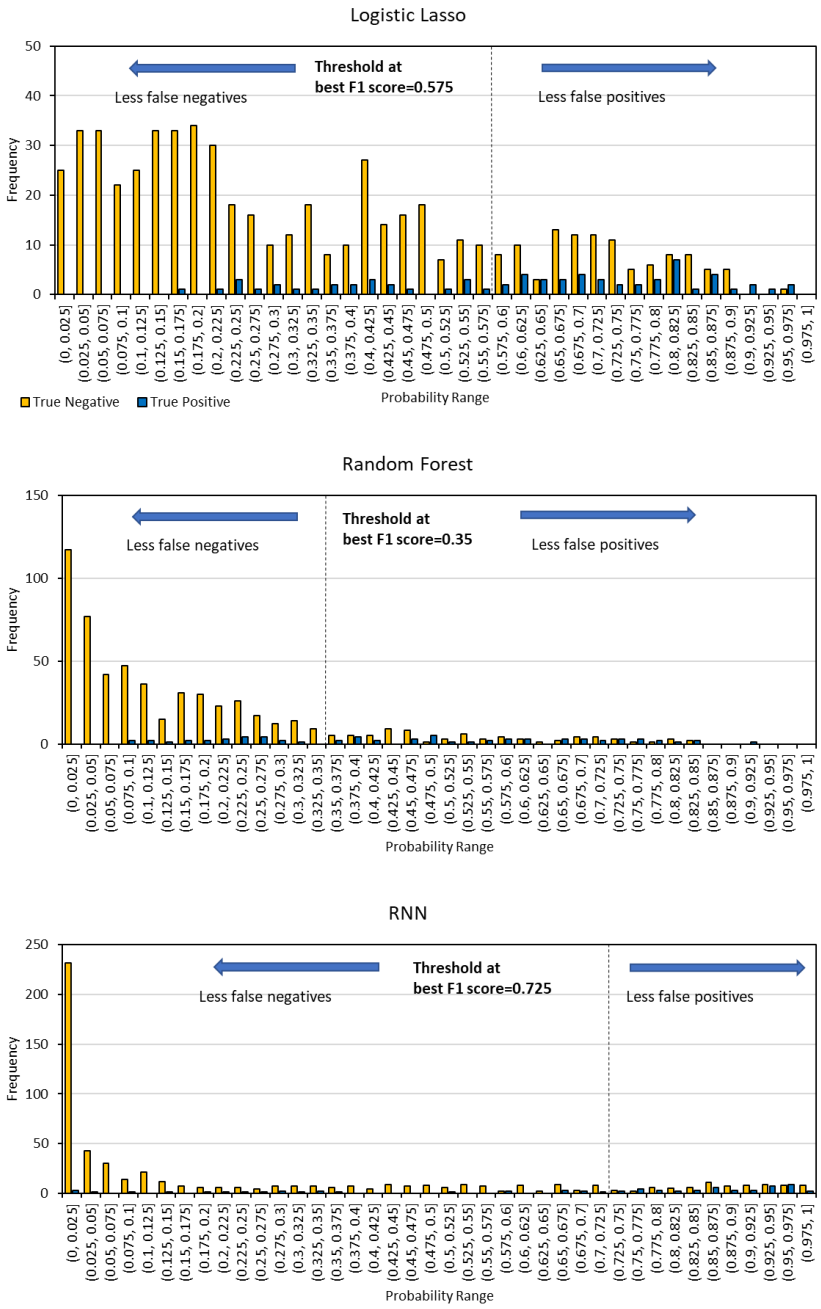
The maps exhibit considerable agreement across models. All models consistently identify most African economies as likely future IMF resource users as well as several economies in the Middle East and South America. All models ascribe a medium to high likelihood to the 18 countries which approved arrangements in 2023 confirming the success of the algorithms. Each model correctly assigns a high probability to five countries that had a program approved in 2023, with some consensus for Pakistan, Côte d'Ivoire, Niger, Senegal, and Sri Lanka between different pairs of models. Remarkably, the RNN also indicates a medium likelihood for the United States and Finland seeking IMF assistance. At the same time, the logistic lasso model suggests a smaller number of countries surpassing the medium likelihood threshold (50^{th} percentile) and is the only model that assigns a probability below the 50^{th} percentile threshold to Argentina and Ukraine.

In part, the differences across models could be driven by the way thresholds are selected. Specifically, by comparing country-level predictions across models using the percentiles of the predicted probabilities does not account for the different shapes of distributions of predicted probabilities. For example, the RNN places a great number of its country predictions close to zero, some close to 1, but only relatively few in between. Therefore, the 50-90th percentiles of the RNN predicted probabilities would have a very large range and countries that have quite low predicted probabilities in absolute terms - like 0.06 for the US - nevertheless end up between the 50th and 90th percentiles. While current thresholds are based on the percentile distributions of model-generated probabilities, it may be crucial to adjust these percentiles considering the varying shapes of distributions of predicted probabilities and the recall-precision trade-offs exhibited by the models to optimize model accuracy.

[29] The F1 score is an alternative evaluation metric that combines precision and recall scores using their harmonic mean.

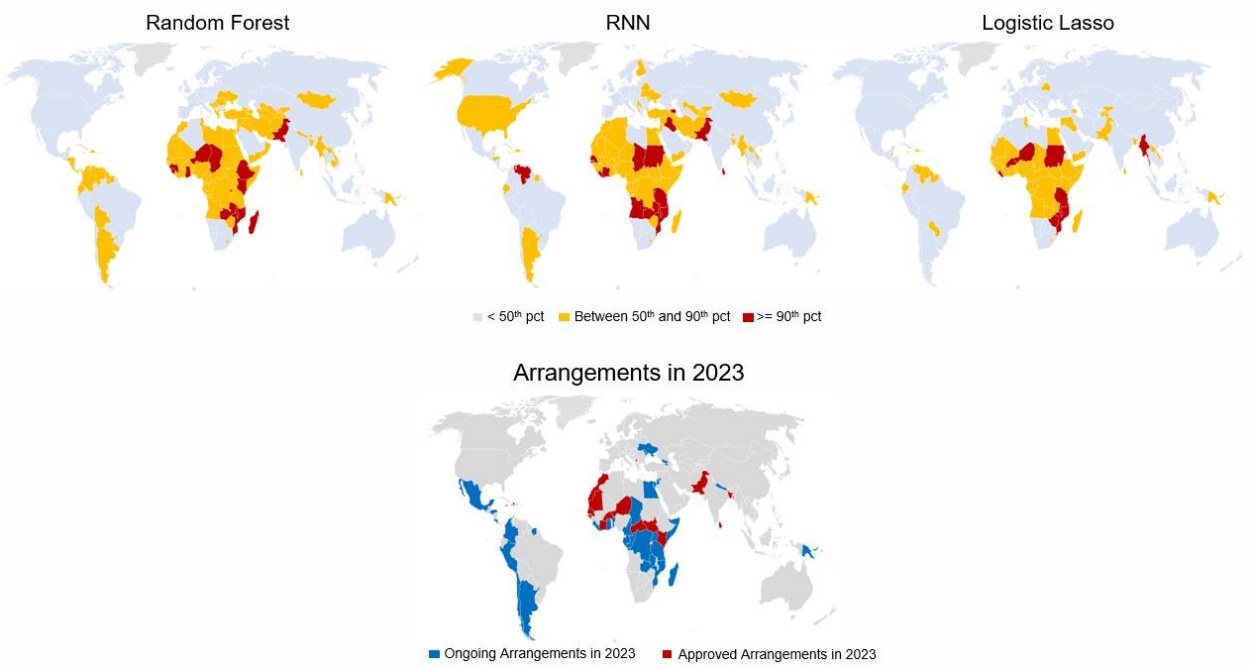
[30] Note that due to our definition of the dependent variable a high predicted probability for 2021 implies that countries are expected to request a new IMF arrangement in the following two years, i.e., 2022 and 2023.

Figure 9: Histograms of Predicted Probabilities for 2018–2021



Notes: The dashed lines on the graphs refer to the thresholds that give the best F1 score.

Figure 10: 2021 Predictions of the RF, RNN and Logistic Lasso model and Realized Arrangements

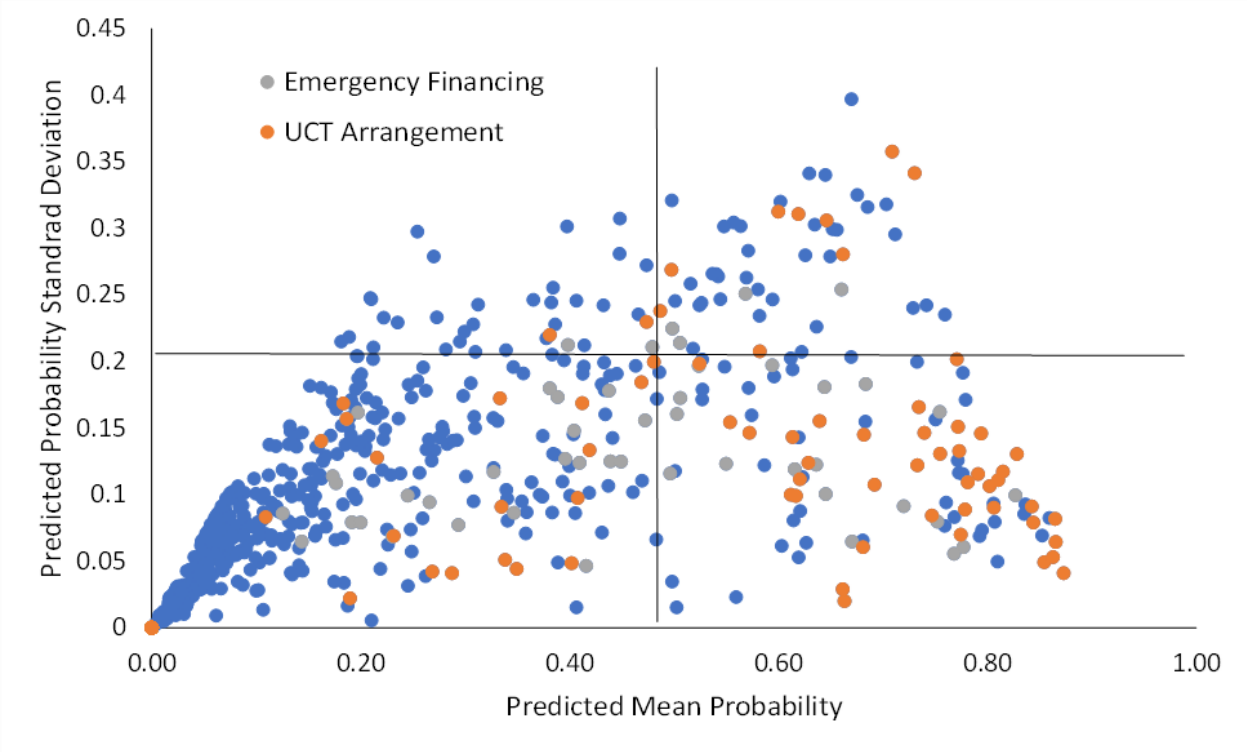


Notes: Percentiles are computed using data spanning from 2018 to 2021 for all countries. On the map, countries with percentiles at or above the 90th percentile are depicted in red, those with percentiles below the 90th percentile but at or above the 50th percentile are represented in yellow, and those with percentiles below the 50th percentile are displayed in light blue.

After examining the country-level predictions for different models individually, we also consider an ensemble approach by combining the prediction results of the three representative model types. To create the ensemble method, we calculate the average and the standard deviation of the predicted probabilities across the three models. Compared to a single algorithm, this ensemble approach can provide further insights regarding the prediction dispersion – i.e., agreement or disagreement across models – and can also serve to mitigate outliers of individual algorithms.

The results are exhibited in the scatter plot in Figure 11 in which each dot represents a country-year observation in the test set (2018-2021) plotted by its mean predicted probability across models on one axis and the standard deviation of the predicted probabilities on the other axis. Blue dots represent non-pre-arrangement periods, orange dots indicate observed arrangements, and gray dots represent emergency financing instruments (excluded from our pre-arrangement dependent variable). A dot with a mean probability close to 1 and standard deviation close to 0 indicates a high level of agreement among models that the country will request an IMF-supported arrangement. The ensemble approach performs well at predicting arrangements in the test set although there is large dispersion across models for some countries. Specifically, the ensemble approach results in an average predicted probability greater than 0.5 for 46 out of the 69 approved UCT arrangements and 33 out of the 65 approved emergency financing instruments between 2018-2021. The mean predicted probability for countries with an actual UCT arrangement or actual emergency financing approved during the test period are 0.59 and 0.51, respectively, compared to an overall mean predicted probability of 0.28 across all observations.

Figure 11: Predictions 2018-2021, Mean and Standard Deviation



5. Model Analysis

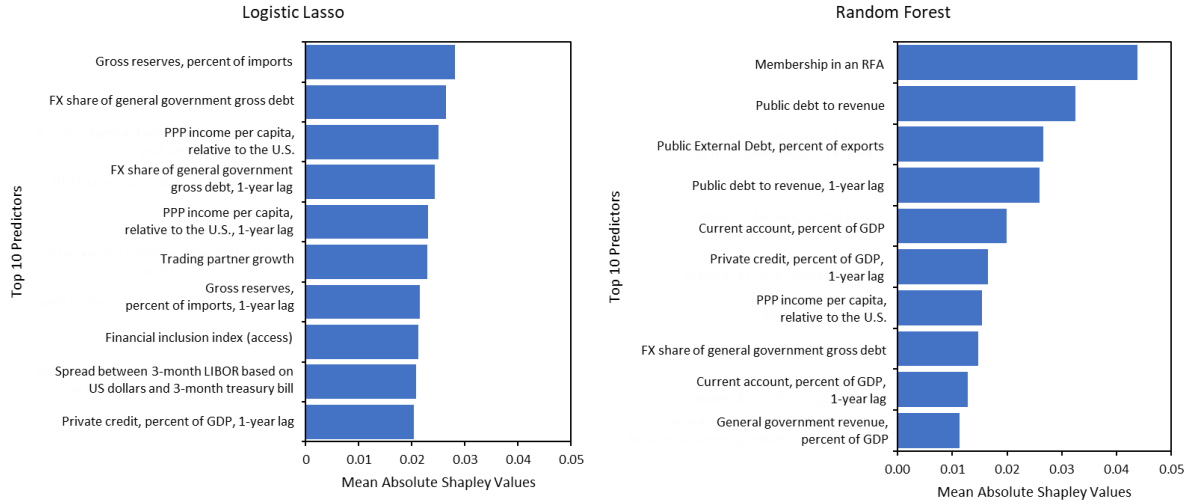
In the final stage of our modeling process, we first examine feature importance and the contribution of different features to model predictions using Shapley values. We then perform two robustness exercises to evaluate the sensitivity of prediction performances with regard to different feature sets and time periods.

Feature Importance and Predictors of IMF Arrangements

Understanding which features are most influential for determining model predictions can provide insights that can help improve existing mechanisms of tracking country economic conditions as well as the development of early warning systems. In order to determine which features drive the observed model predictions and thus are most indicative of the future use of IMF resources, we calculate Shapley values. Shapley values, a concept from coalitional game theory (Shapley, 1953), are used to allocate the contribution of each player (or feature in the context of machine learning) to the overall outcome of a game (or prediction). They are calculated by averaging the marginal contribution of a feature across all possible combinations of features. This approach ensures that the order in which features are added to a model does not bias the attribution of importance. In a predictive model, Shapley values quantify how much each feature contributes to the difference between the actual prediction and the average prediction of the model. When applied to a dataset, the sum of the Shapley values for all features of a single observation explains the deviation of that specific prediction from the average prediction. The Shapley value for a feature is determined by considering all possible subsets of features, thereby capturing the feature's interaction with others and its unique contribution. It's important to note, however, that

Shapley values do not imply causation. While they indicate the importance and contribution of each feature to model predictions, they do not determine whether a feature causes a certain outcome.

Figure 12: Mean Absolute Shapley Values for Top 10 Predictors

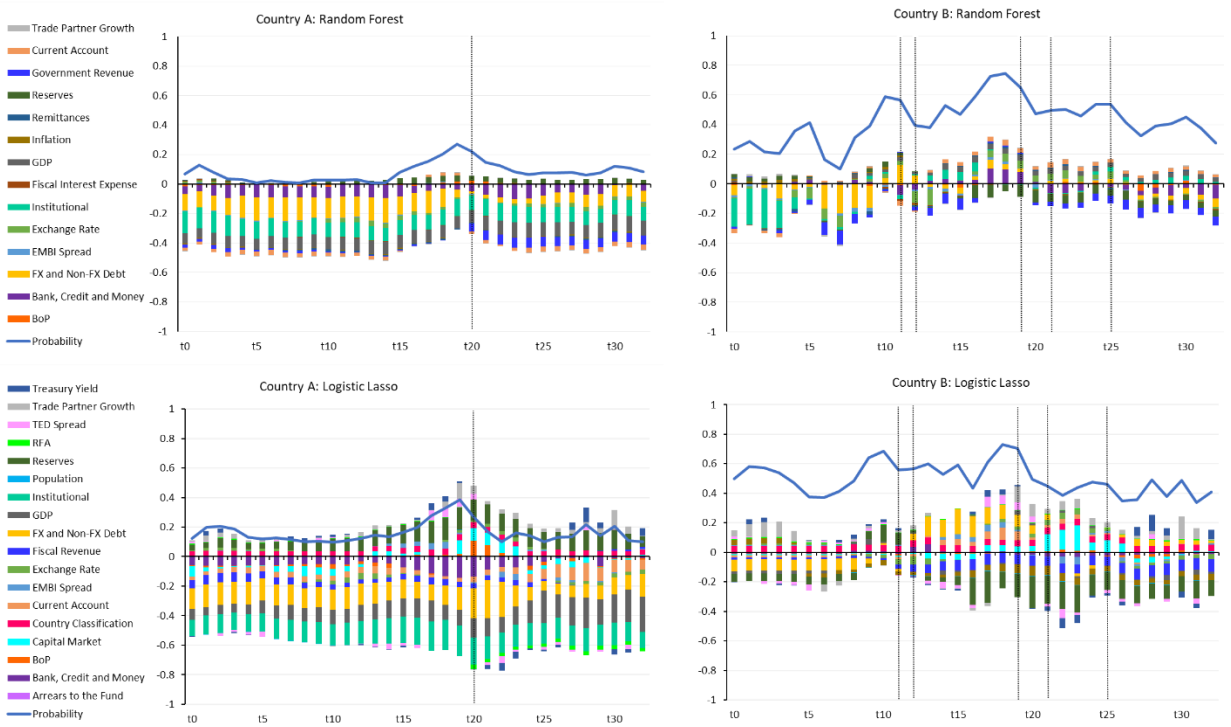


We compare feature importance across two representative models, the traditional linear logistic lasso model and the best performing ML-based algorithm i.e., the random forest.[31] The top 10 features in the logistic lasso model and the random forest model are summarized in Figure 12. To evaluate global feature importance, we calculate mean absolute Shapley values for each feature across all observations. Overall, the most influential features correspond well with economic theory and represent a variety of sectors (including real, fiscal, financial, external sectors, and structural variables). Several features are influential in both the logistic lasso and random forest models including the foreign share of general government gross debt, PPP income per capita, and private credit in percent of GDP. Other variables are influential in only one of the models. For instance, membership in a RFA has the greatest absolute Shapley value in the random forest model followed by features related to debt and the current account. The logistic lasso model assigns the highest Shapley value to gross international reserve in percent of imports. Trading partner growth, the financial inclusion index (access), and TED spreads are additionally among the ten most influential features.

Finally, we present two case studies to evaluate key drivers of observed IMF-supported programs as well as out-of-sample prediction performances at the country level. We select one country with a single program and one with repeated use of IMF-supported financing during the sample period. To ensure truly out-of-sample results, the selected countries were excluded from the training set for this analysis. The results are displayed in Figure 13, highlighting key drivers, predicted probabilities, and actual program start dates.

Figure 13: Case Studies – Out-of-sample Performance and Feature Importance

[31] We currently do not display feature importance results for the recurrent neural network representative model due to external software incompatibilities that prevent producing the Shapley value results comparable with the other models.



Notes: The dashed lines on the graphs refer to actual program start dates.

Four observations are worth noting. Firstly, both models assign a consistently higher probability to Country B, a frequent user of IMF resources, reflecting its ongoing balance of payments needs and the resulting higher likelihood of seeking IMF support. Secondly, the models successfully predict program starts, showing an increase in predicted probabilities before program start dates. This finding demonstrates the model's capability to capture the use of IMF resources from a diverse set of member countries. Thirdly, the drivers for an IMF-supported program differ between Country A and B. For Country A, reserves, trade partner growth, and capital market variables positively affect the likelihood of observing an IMF-supported program while institutional and debt related indicators reduce County A's likelihood. Conversely, for Country B, variables related to bank, credit, money, debt, and the current account increase the probability of observing IMF-supported programs, while fiscal revenue and reserves-related factors are the most important stabilizing factors. Lastly, it is interesting to note that probabilities assigned by the logit model fluctuate less and tend to be higher during non-program years compared to probabilities assigned by the random forest.

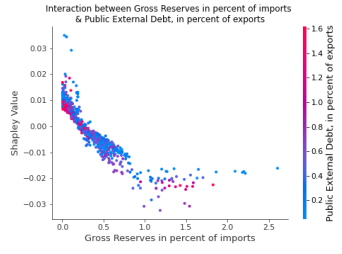
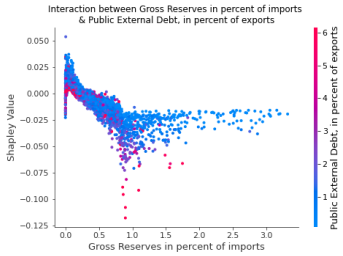
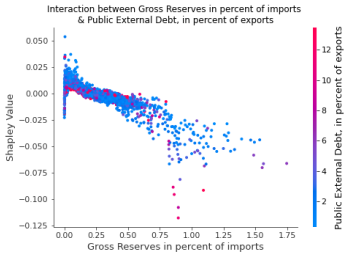
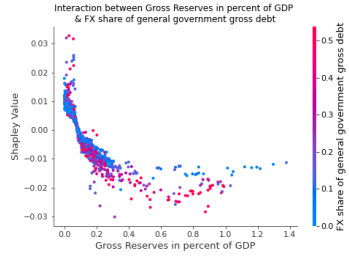
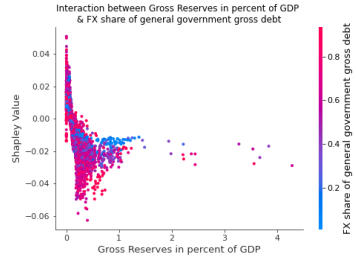
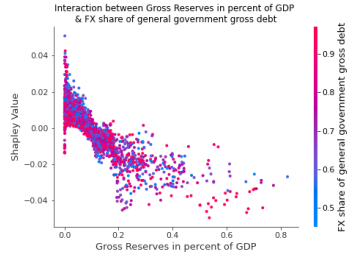
Nonlinearity and Predictor Interactions

A key advantage of machine learning techniques is their ability to handle nonlinear relationships between dependent and independent variables as well as hidden interactions among predictors. In contrast to linear models like the logit model, machine learning algorithms like the random forest are designed to show non-linearities and interactions between variables without being explicitly instructed to do so. Due to the nonlinear nature and the complex interactions of a range of economic factors, the use of IMF resources is inherently challenging to forecast. Moreover, there is no theoretical consensus on how and which factors come together to trigger (or delay) the commencement of an IMF-supported arrangement. It is therefore critical to further improve

our understanding of the interactions and nonlinear relationships between explanatory and response variables to strengthen the current system of program monitoring and supervision.

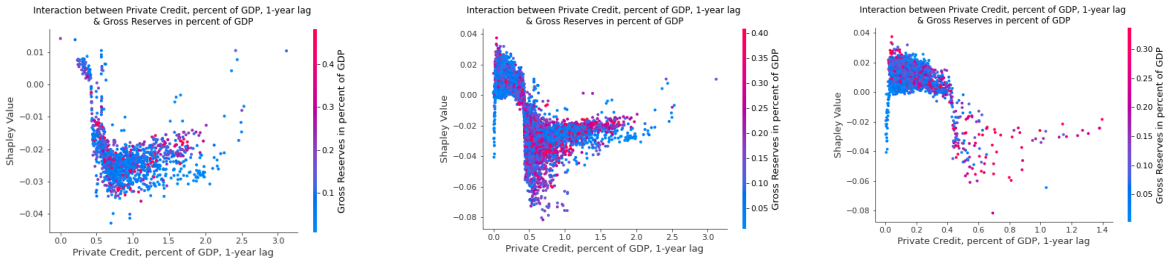
Figure 14 illustrates the interactions between selected predictors in the random forest model and their contributions to the likelihood of observing an IMF-supported program. We observe differences in the distribution of Shapley values and overall magnitudes across three different country groups: Advanced Economies (AEs), Emerging Market Economies (EMs), and Low-Income Developing Countries (LIDCs). This underscores that identical variable values can have disparate impacts on the likelihood of observing a Fund-supported program across different income groups. For instance, possessing the same ratio of reserves to imports (~25%) results in a negative Shapley value for AEs, but a positive Shapley value for EMs and LIDCs. It also indicates that values should not be interpreted in a siloed way but other factors, such as the country's income group, should be considered in tandem. The map's color shows how individual variables interact with others. Interaction plots for AEs, EMs, and LIDCs are provided for the most important predictors selected by the random forest.

Figure 14: Feature Interactions

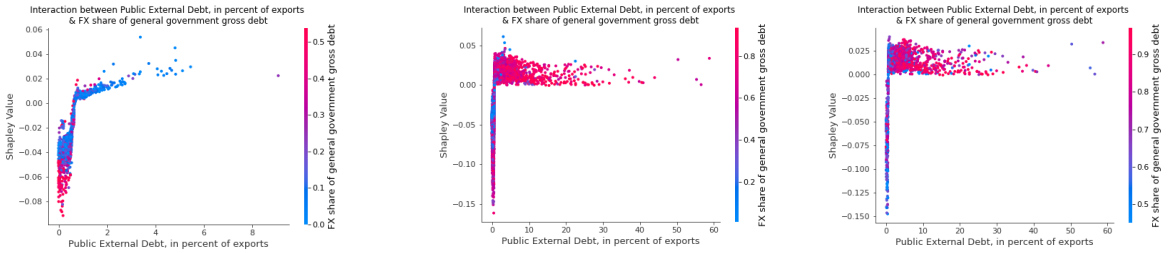
Advanced Economies (AEs)	Emerging Market Economies (EMs)	Low Income Developing Countries (LIDCs)
		
Reserves to import: As expected, higher reserves, expressed as a percent of annual imports, tend to decrease the probability of a country seeking financial support from the IMF. However, reserve levels at which Shapley values turn negative differ among income groups. For EMs and LIDCs, reserves below 50 percent of annual imports amplify the likelihood of an IMF-supported program, whereas for AEs, the threshold is at 25 percent. For LIDCs, higher reserve to import ratios reduce the likelihood of a new program less compared to AEs (steepest) and EMs as indicated by the flatter Shapley curve. It is interesting to note that, countries with low reserves to imports and high external public debt ratios are less likely to start a program than countries with low reserves to imports but low public external debt, potentially pointing to binding financial constraints when debt levels are already high.[32]		
		

[32] Throughout this table, values of the variables presented as a percentage of different metrics (such as GDP, exports, imports, etc.) are expressed in decimals. Thus, a value of 0.5 indicates the variable is 50% of the referenced metric. For example, if reserves are 0.5 of exports, this means reserves are equal to 50% of the export value.

Reserves to GDP: Here we show the relationship between gross reserves in percent of GDP and the FX share of general government gross debt. AEs with a high level of reserves and a high FX share of government debt tend to have lower probability for entering IMF-supported programs than countries with a similar level of reserves but a lower FX share of public debt. This relationship is reversed for low levels of reserves, indicating that a higher FX share of public debt increases the likelihood of entering an IMF-supported program if reserves are low. For EMs and LIDCs, low reserves significantly increase the probability of engaging in a fund-supported program, but there is no clear interaction with the FX share of government debt.

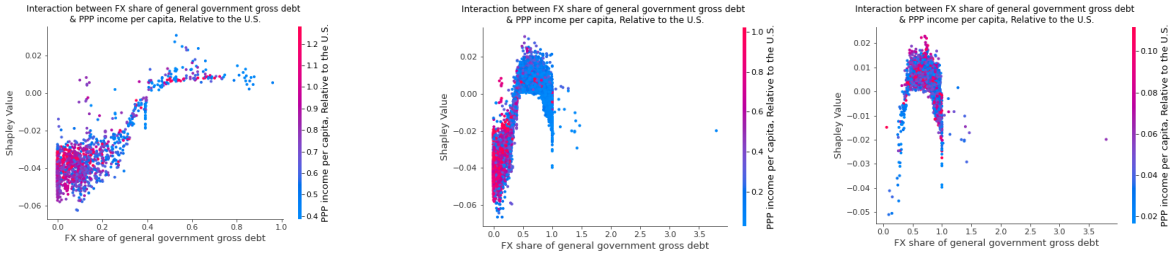


Private credit to GDP: The impact of private credit to GDP on the likelihood of entering an IMF-supported program displays distinct patterns across income groups. Low levels of private credit are associated with a higher likelihood of program commencement across all groups, but this relationship is particularly pronounced for EMs and LIDCs. While this result may seem counterintuitive at first[33], low levels of private credit are often associated with an underdeveloped private financial system, and, more broadly, lower income levels. Additionally, higher private credit to GDP could provide an indication for more prosperous economic conditions, suggesting a diminished necessity for financial support. In EMs, the data also shows that countries with higher private credit tend to have a higher stock of reserves, a trend that partially also extends to LIDCs, albeit less clearly. This observation is consistent with recent studies showing that international reserves can provide insurance against sudden capital outflows in the presence of high outstanding private external debts (see e.g., Lutz and Zessner 2023).

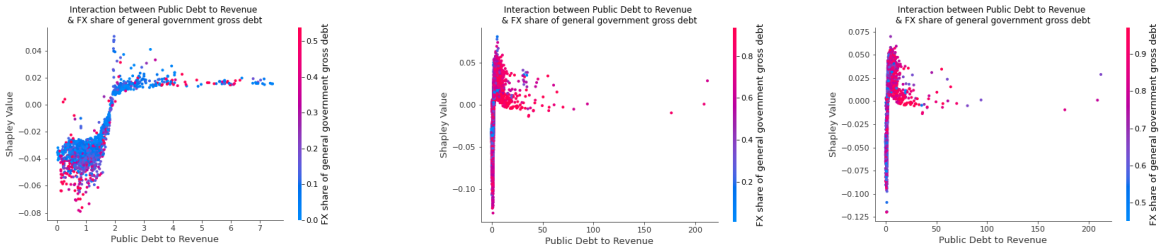


Public external debt to exports: A high level of public external debt generally raises the probability of countries seeking IMF financing across all income groups. However, overall magnitudes and thresholds where Shapley values turn positive vary between groups. Public external debt appears to be most relevant for EMs with low debt levels being associated with the most negative Shapley values and higher debt levels being associated with the most positive Shapley values. Notably, once public external debt to exports surpasses a certain small threshold, the likelihood of program commencement abruptly increases and remains elevated but roughly constant thereafter. The graph also shows that while advanced economies with substantial public external debt tend to have a low FX share of government debt, emerging markets and low-income developing countries with high public external debt often have a high FX share of public external debt. This indicates that EMs and LIDCs predominantly borrow in foreign currency rather than domestic currency, potentially making them more susceptible to exchange rate risks.

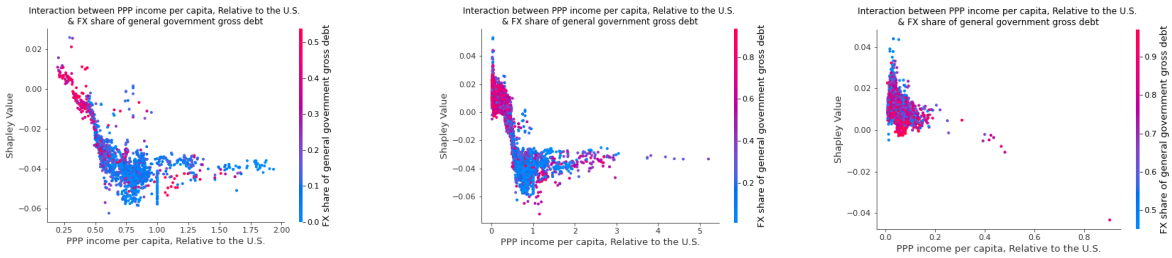
[33] Conventional wisdom could suggest that a higher private credit to GDP ratio could indicate a credit bubble that eventually invites macro-financial conditionality in an IMF-supported program.



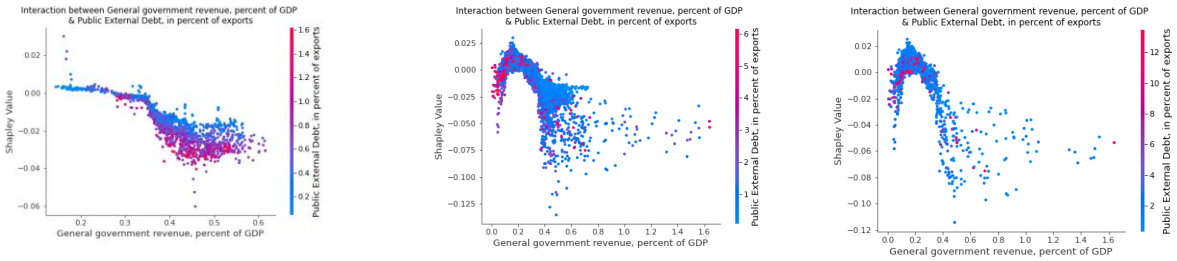
FX share of general government debt: As expected, there is a positive relationship between countries' use of IMF financing and the FX share of public debt. For Advanced Economies (AEs), which are represented in the first column, there's a distinct trend of low FX shares of government debt suggesting less reliance on foreign-denominated debt. Notably, there's an inverted U-shaped relationship for both EMs and LIDCs. Within EMs, countries with higher per-capita income tend to have both a low FX share of debt and a lower likelihood of seeking IMF-supported programs. In contrast, most LIDCs, irrespective of their income, consistently exhibit comparably high FX shares of public debt, predisposing them to seek IMF financing.



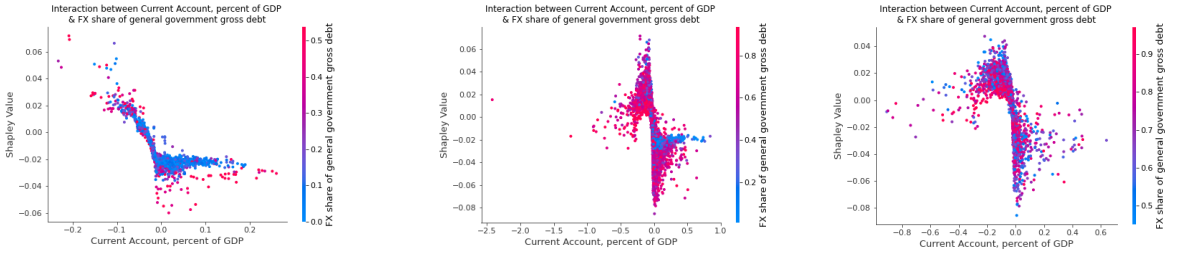
Public debt to revenue: Higher public debt-to-revenue increases the likelihood of using IMF financing across all income groups, with noticeable thresholds where Shapley values escalate sharply —around 150% for AEs and 20% for both EMs and LIDCs. Beyond these thresholds, EMs and LIDCs exhibit a more pronounced response regarding the likelihood of using IMF financing than AEs as indicated by the higher absolute magnitude of Shapley values. Notably, AEs with low public debt-to-revenue ratios tend to have a high FX share of general government debt while FX exposure tends to be elevated across all EMs.



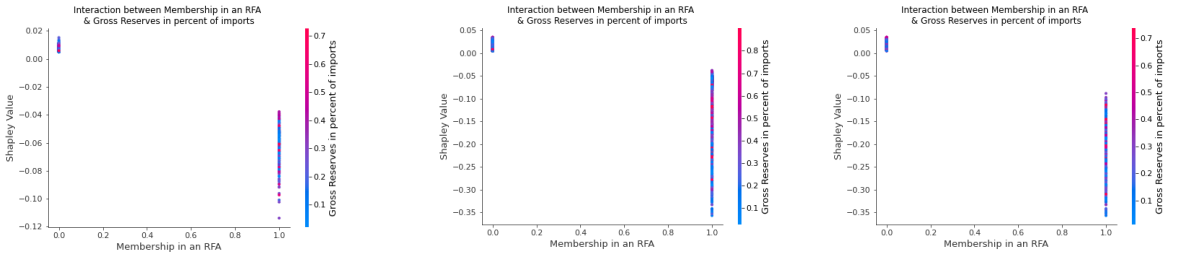
PPP income per capita, relative to the US: This set of graphs shows the variability of income per capita within income groups and how income levels contribute to the likelihood of program commencements. Within each group, low PPP income per capita, relative to the US, is associated with a higher likelihood to observe a fund-supported program. We further find that the FX share of government debt is highly correlated with income levels: Lower income levels are associated with a higher FX share of general government debt, even within each income group. For LIDCs, there is less variability in income per capita as well as the FX share of government debt across country, year observations.



General government revenue to GDP: Higher government revenue to GDP is associated with a lower likelihood of IMF financing requests, a trend observed across all income groups. This inverse relationship is particularly pronounced in EMs, with Shapley values ranging from approximately -0.13 to 0.25, highlighting the significant impact of government revenues on the use of IMF financing in these economies. It is noteworthy that AEs with high government revenues and high levels of external public debt tend to have a lower probability of requesting IMF financing than countries with lower public external debt levels. This pattern is likely attributed to their enhanced access to global financial markets, providing them with alternative financing options. Similarly, for both EMs and LIDCs, the data also indicates that for equivalent levels of general government revenue to GDP, a higher ratio of external public debt to GDP is associated with a decreased likelihood of observing an IMF-supported program. This relationship could be due to both access to global markets and the inability to seek IMF assistance when already grappling with substantial debt burdens.



Current account to GDP: In AEs, the likelihood of observing a fund-supported program does not significantly increase until current account deficits fall below 5% of GDP, as indicated by the Shapley values remaining negative above this point. This trend differs in EMs and LIDCs, where even small current account deficits are linked to a higher probability of program commencement. Additionally, it's noteworthy that the maximum share of government debt held in foreign currency is substantially higher in LIDCs, nearing 100%, compared to around 90% in EMs and just over 50% in AEs, implying a higher exposure to currency risk in LIDCs. AEs with current account deficits and a higher FX share of government debt show a stronger likelihood of observing a fund-supported program, highlighting the intricate interplay between fiscal vulnerabilities and external imbalances. Interestingly, we find evidence for a reversed relationship for countries with current account surpluses.



Membership in an RFA: As expected, a membership in a RFA reduces the likelihood of observing IMF-supported programs across all income groups but overall magnitudes differ significantly, both across and within income groups. Specifically, we find that Shapley values range from -0.11 to -0.03 in AEs, -0.35 to -0.04 in EMs and -0.35 to -0.1 in LIDCs. For EMs and LIDCs, a RFA membership seems to mitigate countries' use of IMF financing somewhat less if the level of international reserves is high.

Robustness Analysis

Evaluating the Impact of Adding Additional Variables

The prediction results presented thus far focused on the best performing variable set for each model, i.e., the set maximizing the ROC-AUC scores during the cross-validation process. As discussed earlier, our results indicate that the selected variable sets generally differ across models, but model-selected sets (set 4-6) consistently outperform pre-defined sets based on missing value thresholds and transformations (sets 1-3 in Table 1). Moreover, all algorithms tend to prefer smaller feature sets. Importantly, however, model-selected variable sets exclude several variables that have been identified as key drivers for countries' demand for IMF resources and economic distress in earlier work.[34]

In this section, we evaluate the robustness of the prediction performance of the RF, RNN and logistic lasso model by adding additional variables that were not included in the selected variable set but have been identified as important by existing studies in the early warning literature. We add the variables individually and ex post (after model fitting) to the best performing sets and evaluate the robustness of the test ROC-AUC scores. Specifically, we add the following 21 variables: reserves in percent of the ARA metric, bank capital adequacy ratios, net non-FDI liability inflows, access to a central bank swap line (dummy), federal funds rate, 10-year treasury yields, terms of trade, debt service to revenue, foreign liabilities to domestic credit, short term deposit rates, bank external liabilities as a percent of GDP, food price inflation, oil price inflation, international risk aversion index, fiscal interest expenses to revenue, consumer prices and a set of country classification dummies (PRGT, AE, non-LIDC EMDE, LIDC). We also evaluate if IMF credit outstanding can further improve prediction performances as existing liabilities with the Fund itself could impact countries' use of new IMF resources.[35]

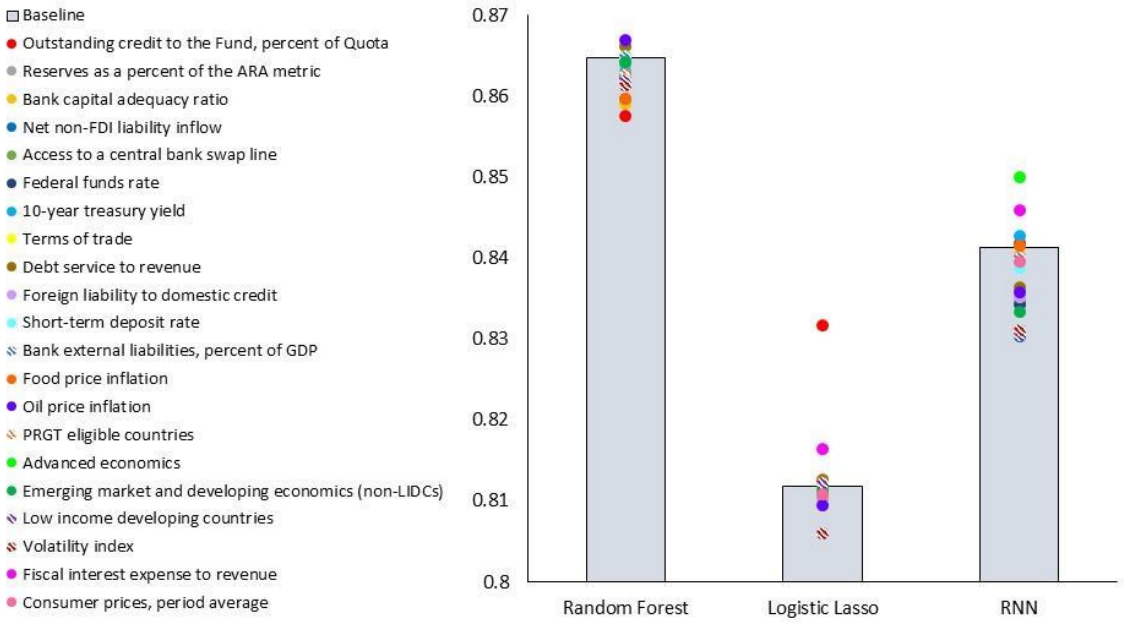
The results are summarized in Figure 15. Overall, it turns out that while non-linear models like the random forest maintain their prediction performance, linear models such as logistic lasso regression exhibited notable performance variability. We observe the most significant increase in the test ROC-AUC score for the logistic lasso model when IMF credit outstanding is added to the model (orange dot), while the RF reports a decline in model performance. Fiscal interest expenses to revenue (fluorescent magenta dot) increase model performance for both the logistic lasso and the RNN (already included in the set selected originally by the random forest). The biggest improvement in prediction performance of the RNN is observed when a dummy for advanced economies is added to the model (RF and logistic lasso ROC-AUC scores decline marginally). These findings underscore that non-linear ensemble models like the random forest might be more robust to possible omissions of important variables compared to simpler, linear models.[36]

[34] See, for example, IMF (2021), Hills et al. (2021), IMF (2022), Agbloyor et al. (2023) among others.

[35] Importantly, we find that the inclusion of credit outstanding at the baseline stage does not change the optimal hyperparameters selected for the random forest model.

[36] It is important to note that we have not done extensive search on optimal feature sets but only experimented with a couple of sets. Future research could explore optimal feature selection and the implications of omitted variable biases or multicollinearity more comprehensively. Given our sole focus on prediction (rather than causal inference), however, omitted variables or multicollinearity should not cause any issue. For operational purposes, we suggest including credit outstanding in the prediction exercise to account for debt rollover and repeated use of fund resources.

Figure 15: Robustness of Model Performances



Evaluating Model Sensitivity to Training Data Variability and Performance Over Time

As explained in the *Data Processing and Empirical Strategy* section, we use a time series split method for cross-validation, progressively increasing the training set by seven years in each iteration (fold) and assigning the next seven years as the out-of-sample validation set for each fold. Subsequently, the best hyperparameter for each model was selected based on the average ROC-AUC score from their out-of-sample performance across various training and test sets.

In this section, we extend our analysis to examine how different training datasets influence model predictions as a robustness check. Specifically, we apply the best-performing random forest (RF), RNN, and logistic lasso model to distinct training datasets from the different folds to evaluate their out-of-sample performance over the entire subsequent period (and not just the immediate seven years post-training).[37] This exercise provides insights into the temporal dynamics of prediction accuracies, particularly during global crises and long-term performances, far beyond the initial validation period. It also allows us to assess the sustained predictive capability of the models and understand any potential degradation in performance over time, ensuring a comprehensive evaluation of their robustness and reliability.

Figure 16 plots average ROC-AUC scores across the models as well as the differences in the scores between the RF and logistic lasso model as well as between the RF and RNN in bars. As depicted in the left chart of Figure 15, we observe higher ROC-AUC scores for models trained on more recent data compared to those trained on earlier periods. The difference between prediction performances across folds is particularly visible in years immediately following an update of the training set, underscoring that the models are best at making near-

[37] There are five training and out-of-sample test periods, including (1) 1982-1988 as the first training period and 1989-2021 as the out-of-sample test period, (2) 1982-1995 for training and 1996-2021 for test, (3) 1982-2002 for training and 2003-2021 for test, (4) 1982-2009 for training and 2010-2021 for test, and (5) 1982-2017 for training and 2018-2021 for test.

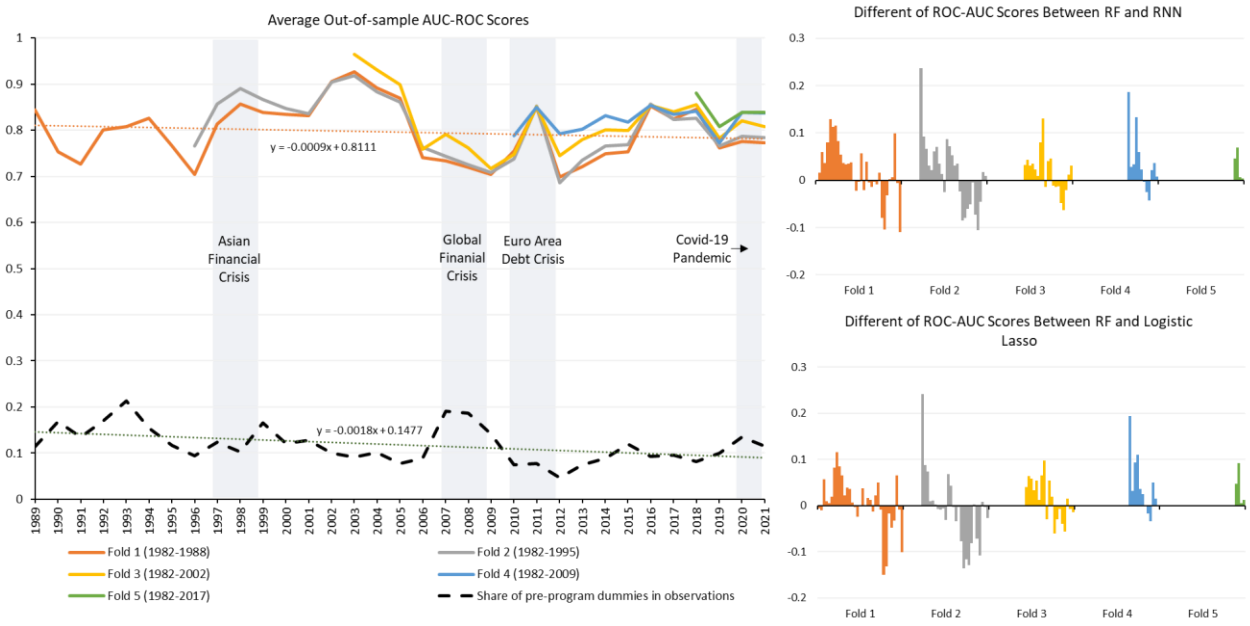
term predictions. Regarding long-term predictions, we notice that average out-of-sample ROC-AUC scores tend to decline over time. This decline in performance could be partly driven by structural breaks and non-stationarities, influenced by e.g., changes in program supply (e.g., changes in Fund policies regarding program requirements, as evidenced by shifting program instruments) or changes in demand factors (e.g., shift in program demand due to factors like the rise of alternative funding sources over time), which the models are unable to account for.[38] Regular model updates are hence required to ensure maximal prediction performance.

We also observe noticeable dips in model performance during periods of global financial distress including the Asian financial crisis, the global financial crisis, the European debt crisis, and the Covid-19 pandemic. These episodes are highlighted by the shaded gray area in Figure 15. Across all highlighted periods, we observe a decline in prediction performance approximately two years prior to the onset of the crisis periods, in line with the definition of our pre-arrangement dependent variable (equal to one in the periods t-2 and t-1). For example, following the onset of the Global Financial Crisis, models failed to predict the surge in IMF-supported programs, particularly among advanced economies, underscoring the limitations of models in predicting sudden shifts (Aikman et al., 2021) that may not have been reflected in countries' macroeconomic fundamentals prior to the shock. The decline in prediction performance is evident across all folds and model classes. Interestingly, however, the decline in prediction performance was relatively modest during the recent Covid-19 pandemic, likely driven by the predominant use of emergency financing instruments (treated as no-program cases in our dependent variable definition) but a relatively modest update in other IMF arrangements. Overall, these findings underscore the fact that periods of global financial distress remain inherently difficult to predict.

Finally, the right two charts in Figure 15 show the differences in prediction performance between the RF and the logistic lasso model or the RNN. Although the RF outperforms competitor models in near-term predictions in each fold, as indicated by the positive differences in ROC-AUC scores, its performance tends to deteriorate more rapidly as training data becomes more outdated. Over time, differences in ROC-AUC scores shift from positive to negative values in later years within each fold, suggesting that the RF model may be more prone to overfitting the training data. Nonetheless, this decline in performance generally becomes more apparent only when the training data is outdated by at least 8-10 years, still positioning the RF model as a strong performer until that point.

Figure 16: Average Test AUC-ROC Scores of RF, RNN, and Logit with Lasso for Different Folds

[38] Please note that we focus here on trends observed across time within each fold rather than individual years. Naturally, ROC AUC are expected to be higher for later folds because the model is trained on a longer sample and hence more 'robust'.



Evaluating the Size of IMF Arrangements

So far, our analysis focused on binary classification models, i.e., if a country is expected to engage in an IMF-supported arrangement in the following two years. However, the size of the approved IMF-supported arrangements varied greatly during the last three decades ranging from 14.5 to 3211.8 percent of quota in our sample.[39] In this section, to assess the magnitude of expected IMF-supported arrangements, we propose a simple two-step procedure that uses the binary prediction results from the most successful classifying algorithm (random forest) to assess potential implications for IMF resources.

Specifically, we run a regression relating the log approved size of IMF arrangements of country *i* in year *t* to a set of explanatory variables lagged by one period. As before our sample ranges from 1982-2021 and is split into a training set (1982-2017) and a hold-out test set (2018-2021). For illustrative purposes, this section focuses on arrangements approved under the GRA account but could be easily extended to the full set of IMF-supported arrangements (including the PRGT and RST account).[40] We drop all observations where no IMF arrangement was approved for a country in a year which leaves us with 585 data points in the training set. The set of explanatory variables includes the 40 variables selected by the random forest recursive feature elimination procedure (set 6). As a robustness check, we further evaluate prediction performance if the additional variables, identified as important by the literature, are added to the model. We evaluate model performance based on the out-of-sample root mean-squared error.

The findings are summarized in Figure 17. The blue area shows the sum of approved amounts per year and model predictions for the testing sample (2018-2021) are shown by the black lines. We compare four different specifications including two different sets of explanatory variables (set 6 vs. set 6 plus the additional variables)

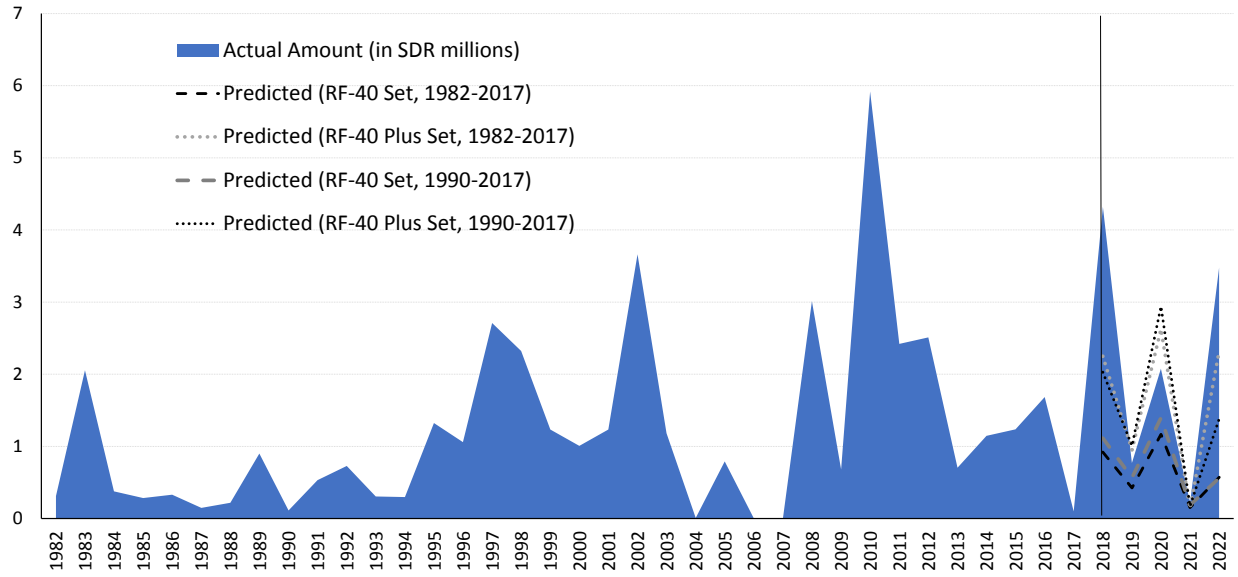
[39] As before, the sample ranges from 1982-2021 but only includes drawn GRA arrangements.

[40] The size of approved arrangements under the GRA and PRGT account differs significantly in our sample with GRA arrangements being on average 2.5 times larger than arrangements approved under the PRGT account.

and two different time frames in the training set (1982-2017 vs. 1990-2017). While the model based on set 6 underestimates the size of approved arrangements in every year, the model using set 6 plus tends to predict a larger use of IMF resources and overestimates the sum of approved arrangements in some years (2019 and 2020). A major difference among the specifications is that set 6 plus also includes credit outstanding to the fund, improving the model's ability to predict arrangements with large access levels significantly. The best performing model, in terms of out-of-sample root MSE as well as adjusted r-square uses the RF-40 Plus variable set and includes observations from 1990-2017. The full results are provided in Annex IV, Table IV.1.

While coefficients of the regression analysis should be interpreted cautiously due to the possibility of model misspecifications (including in particular omitted variables biases), the following observations are worth noting. Firstly, only a small set of predictors included in the set 6 are significant at a 10 percent level, including the dollar appreciation, the FX share of public debt, private credit and PPP income. Interestingly, most of these variables are also among the top-10 predictors identified by several classification algorithms including the random forest and the logit model. Secondly, several of the additional variables added in the set 6 plus turn out to be significant, including total fund credit outstanding, reserves in percent of the ARA metric, swap line access, the federal funds rate, the federal funds rate and the VIX index.

Figure 17: Actual and Predicted Amounts (in SDR million)



Several limitations of the presented two-set approach are worth noting. As discussed in earlier sections of this paper, one major advantage of ML-based and deep learning models is that they can detect non-linearities without requiring explicit programming to do so. Given the wide range of approved access levels, with some outliers at the upper end of the distribution, non-linearities are clearly important in this context. Other drawbacks include issues related to model misspecifications, such as multicollinearity or omitted variables biases. We see the use of ML-based algorithms to predict the size of IMF arrangements as another promising area for future research.

6. Conclusion

In this study, we provide a comprehensive analysis of the predictability of countries' demand for IMF arrangements. Using a large data set, including thirty-five years and nearly all IMF member countries, and a broad set of ML-based algorithms, we highlight the potential of machine learning techniques to provide early warnings before program commencement. We find that ML models consistently outperform econometric methods in out-of-sample prediction of new IMF arrangements. In line with earlier work, ensemble models and recurrent neural networks are found to be among the most successful classifiers. We further contribute to the discussion regarding influential factors for predicting future IMF resource use, highlighting the importance of variables related to various sectors (external, fiscal, real, financial) as well as institutional factors such as membership in a regional financing arrangement and existing fund credit outstanding. There is considerable agreement across algorithms in the set of selected predictors and prediction performances of ML-based models are largely robust to alternative feature sets. On a more technical level, we show that data processing decisions (imputation and sampling methods) can have significant implications for model performance and different models could perform best under different settings, calling for a flexible, algorithm-tailored approach.

In summary, our research underscores the appeal of ML-based techniques for predicting the use of IMF resources due to their ability to recognize nonlinearities and to capture sequential and temporal patterns in the data. Nevertheless, it is essential to acknowledge several limitations. Firstly, we observe noticeable dips in model performances during periods of global financial distress, underscoring the fact that these events remain inherently difficult to predict. Secondly, prediction performances depend on reliable and near real-time data. This is clearly a challenge in some environments where national accounts data are published with significant lags and noise, including vulnerable economies which are often most reliant on IMF support. Furthermore, the models may miss other factors that could deter countries from demanding a fund-supported program, such as changes in risk perception or stigma related to requesting an IMF-supported program. This is because the models primarily rely on economic fundamentals and institutional factors to make predictions, potentially overlooking shifts in these non-economic factors. Natural language processing techniques could be one way to help address data challenges and publication lags. Finally, it is worth noting that this paper concentrates on classification models and provides only a brief discussion relating to the overall size of approved arrangements. We see the use of ML-based algorithms to predict the size or repeated use of IMF-supported arrangements as another promising area for future research.

Bibliography

Agbloyor, E. K., Pan, L., Dwumfour, R. A., & Gyeke-Dako, A. (2023). We are back again! What can artificial intelligence and machine learning models tell us about why countries knock at the door of the IMF?. *Finance Research Letters*, 57, 104244.

Aikman, D., Galesic, M., Gigerenzer, G., Kapadia, S., Katsikopoulos, K., Kothiyal, A., ... & Neumann, T. (2021). Taking uncertainty seriously: simplicity versus complexity in financial regulation. *Industrial and Corporate Change, 30*(2), 317-345.

Alfaro, L., & Kanczuk, F. (2019). *Undisclosed debt sustainability* (No. w26347). National Bureau of Economic Research.

Badia, M., Medas, P., Gupta, P., & Xian, Y. (2022). Debt is not free. *Journal of International Money and Finance*.

Bailey, M. A., Strezhnev, A., & Voeten, E. (2017). Estimating Dynamic State Preferences from United Nations Voting Data. *The Journal of Conflict Resolution, 61*(2), 430–456.

Bird, G., & Rowlands, D. (2001). IMF lending: how is it affected by economic, political and institutional factors. *The Journal of Political Reform*, 243-2790.

Bird, G., & Rowlands, D. (2008). A disaggregated empirical analysis of the determinants of IMF arrangements: Does one model fit all? *Journal of International Development*.

Breiman, L. (2001). Random forests. *Machine learning, 45*, 5-32.

Burman, P., Chow, E., & Nolan, D. (1994). A cross-validatory method for dependent data. *Biometrika, 81*(2), 351-358.

Cerovic, S., Gerling, K., Hodge, A., & Medas, P. (2018). Predicting Fiscal Crises. *IMF Working Paper*.

Cerutti, E. (2007). IMF Drawing Programs: Participation Determinants and Forecasting. *IMF Working Paper*.

Chen, T., & Guestrin, C. (2016, August). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining* (pp. 785-794).

Elekdağ, S. (2006). How Does the Global Economic Environment Influence the Demand for IMD Resources? *IMF Working Paper*.

Fouliard, J., Howell, M., & Rey, H. (2021). Answering the Queen: Machine Learning and Financial Crises. *NBER Working Paper*.

Hacibedel, M. B., & Qu, R. (2022). *Understanding and predicting systemic corporate distress: a machine-learning approach*. International Monetary Fund.

He, H., & Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Transactions on knowledge and data engineering, 21*(9), 1263-1284.

Hellwig, P. (2021). Predicting Fiscal Crises: A Machine Learning Approach. *IMF Working Paper*.

Hills, T., Nguyen, H., & Sab, R. (2021). Determinants of Pre-Pandemic Demand for the IMF's Concessional Financing. *IMF Working Paper*.

IMF. (2021). How to Assess Country Risk: The Vulnerability Exercise Approach Using Machine Learning. *IMF Technical Notes and Manuals* No.2021/003.

IMF. (2022). Review of the Adequacy of the Fund's Precautionary Balances.

Iseringhausen, M., Nkusu, M., & Wiranto, W. (2019). Repeated Use of IMF-Supported Programs: Determinants and Forecasting. *IMF Working Paper*.

Jarmulska, B. (2022). Random forest versus logit models: Which offers better early warning of fiscal distress? *Journal of Forecasting*.

Knight, M., & Santaella, J. A. (1997). Economic determinants of IMF fiancial arrangements. *Journal of Development Economics*, 405-436.

Lundberg, S. M., Erion, G. G., & Lee, S. I. (2018). Consistent individualized feature attribution for tree ensembles. *arXiv preprint arXiv*:1802.03888.

Lutz, F., Zessner-Spitzenberg, L. (2023): Sudden Stops and Reserve Accumulation in the Presence of International Liquidity Risk. Journal of International Economics. 141(103729).

Maeder, N., Poulain, J.-G., & Reynaud, J. (2019). Assessing IMF Lending: A Model of Sample Selection. *IMF Working Paper*.

Malladi, R. K. (2022). Application of Supervised Machine Learning Techniques to Forecast the COVID-19 U.S. Recession and Stock Market Crash. *Computational Economics*.

McGettigan, D., & Reynaud, J. (2017). IMF Lending in an Interconnected World. *IMF Working Paper*.

Poulain, J.-G., & Reynaud, J. (2017). IMF Lending in an Interconnected World. *IMF Working Paper*.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *nature, 323*(6088), 533-536.

Samistas, A., Kampouris, E., & Kenourgios, D. (2020). Machine learning as an early warning system to predict financial crisis. *International Review of Financial Analysis*.

Scheubel, B., Stracca, L. (2019): What do we know about the global financial safety net? A new comprehensive data set. *Journal of International Money and Finance*, 99, 102058.

Shapley, L. S. (1953). A value for n-person games. *Princeton University Press*.

Trudel, R. (2005). Effects of Exchange Rate Regime on IMF Program Participation. *Review of Policy Research*.

Vreeland, J. R. (2004). Institutional Determinants of IMF Arrangements. *UCLA: International Institute.*

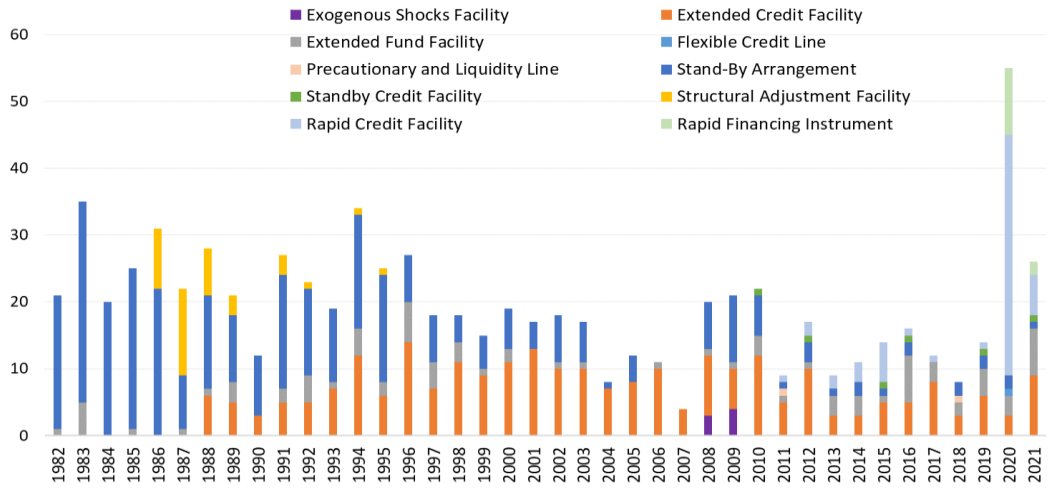
Weisfeld, H., de Carvalho Filho, I., Comelli, F., Giri, R., Hellwig, K. P., Huang, C., ... & Presbitero, A. (2020). Predicting Macroeconomic and Macrofinancial Stress in Low-Income Countries. *IMF Working Paper.*

Annex I. Data and Data Processing

Table I.1: Base Variables

Category	Predictors - Base Variable Names	Source
External	Current Account	WEO Live
	Trade Openness	WEO Live
	Nominal Exchange Rate	WEO Live
	Real Exchange Rate	WEO Live
	PPP Exchange Rate	WEO Live
	External Debt	External Wealth of Nations Database
	Gross Reserves	WEO Live
	Reserves, % of ARA metric	WEO Live
	Terms of Trade	WEO Live
	Net Other Investment	WEO Live
	Net Portfolio Investment	WEO Live
	Net FDI	WEO Live
	Remittances	WEO Live
	Net non-FDI liability inflow	WEO Live
	Trading Partner Growth	WEO Live
Fiscal	Primary Balance	WEO Live
	Overall Balance	WEO Live
	General government revenue	WEO Live
	Fiscal interest expense / Fiscal revenue (%)	WEO Live
	Public Debt to Revenue	WEO Live
	Public Debt Service to Exports	WEO Live
	Public Debt Service to Revenue	WEO Live
	Total Debt Service Paid	WEO Live
	External Public Debt Amortization	WEO Live
	Public External Debt, in USD billion	WEO Live
	Central government debt	IMF GDD
	Interest rate on government debt minus growth	WEO Live
Financial	Broad Money	WEO Live
	External Bank Liabilities	WEO Live
	Loan to deposit ratio	IFS
	Capital Adequacy Ratio	IFS
	Financial Inclusion	SPRAIMU
Global	CBOE Volatility Index	FRED
	Dollar Appreciation	IFS
	Federal Funds Effective Rate	FRED
	10-Year U.S. Treasury Securities Yield	FRED
	TED Spread	FRED
	J.P.Morgan EMBI	Bloomberg
Real	PPP GDP	WEO Live
	PPP income per capita	WEO Live
	5-year growth deviation	WEO Live
	Inflation	WEO Live
	Food Price Inflation	WEO Live
	Oil Price Inflation	WEO Live
	HP Output Gap	WEO Live
Structural	Cost of natural disaster hazards	EM-DAT and WEO Live
	Ideal Point Distance to US	Bailey, Strezhnev, and Voeten (2021)
	Population, log, 1-year growth rate	WEO Live
	Bureaucracy Score (ICRG)	ICRG
	Corruption Score (ICRG)	ICRG
	Government Score (ICRG)	ICRG
	Polity Score (ICRG)	ICRG
Financial-Real	Corporate Debt Sub-Investment Grade	Vulnerability Exercise Securities Database, based o
	Total Credit	SPRAIMU
	Credit Gap	SPRAIMU
	Private Credit	WDI
	Real Short-term Deposit Rate	WEO Live
	House Price Acceleration	RES Real Estate Template
	Total Debt	WEO Live, WDI and EWN
Dummy	Membership in an RFA	Scheubel & Straccia (2019)
	SWAP Agreement Membership	Scheubel & Straccia (2019)
	Bankin Crisis	Laeven & Valencia
	Currency Crisis	Laeven & Valencia
	Sovereign Crisis	Laeven & Valencia
	Twin Crisis	Laeven & Valencia
	Triple Crisis	Laeven & Valencia
	Fund Accounts, Overdue Obligations	IFS
	Post-arrangement dummy	IMF, Financial Data Query Tool
	GRA arrangement	IMF, Financial Data Query Tool
	PRGT arrangement	IMF, Financial Data Query Tool
	Precautionary arrangement	IMF, Financial Data Query Tool
	SBA arrangement, GRA	IMF, Financial Data Query Tool
	EFF arrangement, GRA	IMF, Financial Data Query Tool
	FCL arrangement, GRA	IMF, Financial Data Query Tool
	PLL arrangement, GRA	IMF, Financial Data Query Tool
	ESF arrangement, PRGT	IMF, Financial Data Query Tool
	ECF arrangement, PRGT	IMF, Financial Data Query Tool
	SCF arrangement, PRGT	IMF, Financial Data Query Tool
	SAF arrangement, PRGT	IMF, Financial Data Query Tool
	RCF loan, PRGT	IMF, Financial Data Query Tool
	RFI loan, GRA	IMF, Financial Data Query Tool
	Exchange Rate Peg	Ilzetzki Reinhart Rogoff (2017)
	Net Debitor	Dummy
	Net Creditor	Dummy
	Fuel Exporter	Dummy
Group	AE	IMF WEO
	MAE	IMF WEO
	OAE	IMF WEO
	EMDE	IMF WEO
	EDA	IMF WEO
	EDE	IMF WEO
	LAC	IMF WEO
	MECA	IMF WEO
	SSA	IMF WEO
	EU27	EU
	LIDC	IMF WEO
	PRGT	IMF WEO
	PRGT & SDS	IMF WEO
	UN	UNCTAD
ID	Country	IMF WEO
	IFS	IMF WEO
	ISO	IMF WEO

Figure I.1: Number of Approved IMF Arrangements, 1982–2021



Missing Value Imputation

Figure I.2 summarizes the distribution of selected features with different shares of missing values. Each sub-graph compares the distribution lines of the feature pre- and post-imputation. While the mean and median imputed series show a sharp increase in the density at the country group specific means and medians (see, e.g., capital adequacy ratio), the general shape of the KNN imputed series remains closer to the base series. A Kolmogorov-Smirnov test (shown in Figure I.3), however, rejects the null of statistically similar distributions for most variables. For KNN imputed variables, all except 3 variables are significantly different from the non-imputed distribution (out of 480 variables in total). For mean (median) imputed variables, all except 40 (41) variables are significantly different from the non-imputed distribution.

Figure I.2: Missing Data Imputation and Distribution

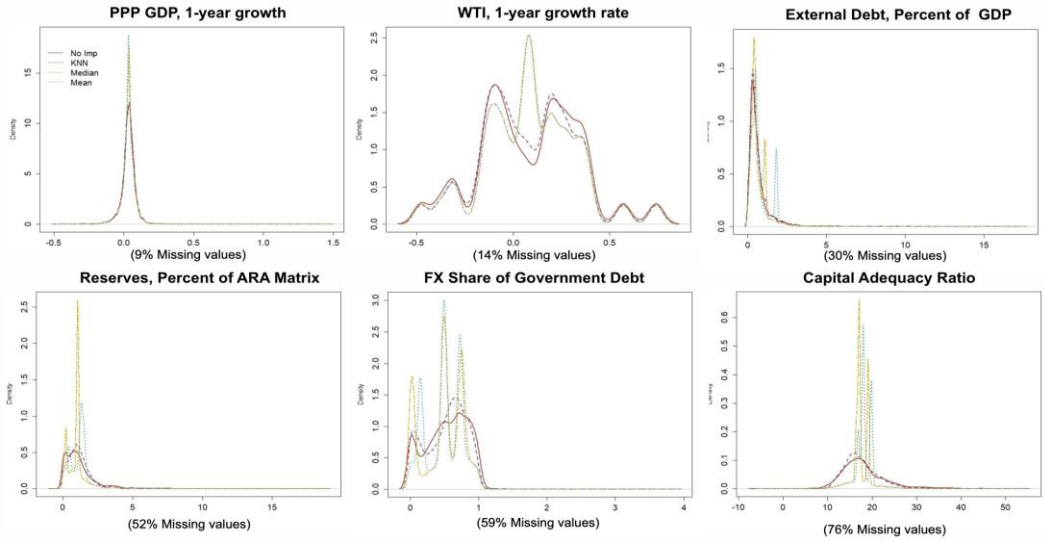
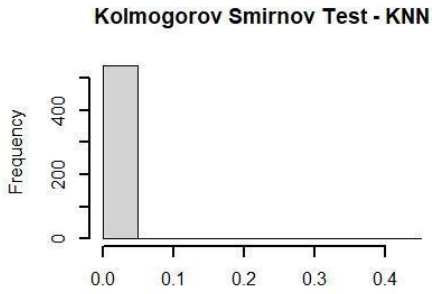
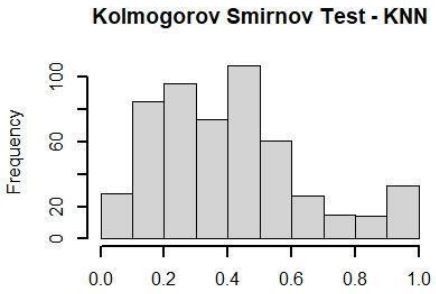
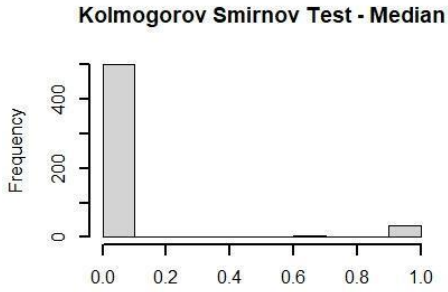
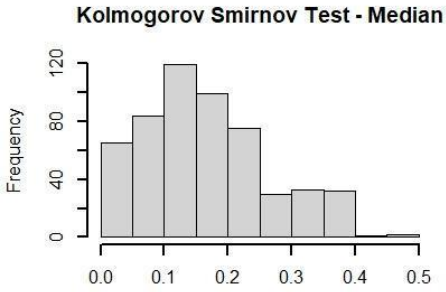


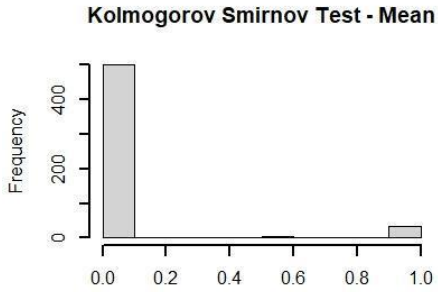
Figure I.3: Kolmogorov-Smirnov Test Results

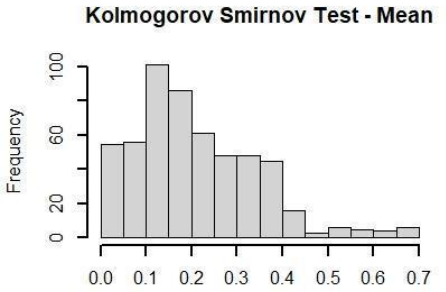

Kolmogorov Smirnov Test - KNN
Frequency
0
200
400
0.0
0.1
0.2
0.3
0.4


Kolmogorov Smirnov Test - KNN
Frequency
0
20
60
100
0.0
0.2
0.4
0.6
0.8
1.0


Kolmogorov Smirnov Test - Median
Frequency
0
200
400
0.0
0.2
0.4
0.6
0.8
1.0


Kolmogorov Smirnov Test - Median
Frequency
0
40
80
120
0.0
0.1
0.2
0.3
0.4
0.5


Kolmogorov Smirnov Test - Mean
Frequency
0
200
400
0.0
0.2
0.4
0.6
0.8
1.0


Kolmogorov Smirnov Test - Mean
Frequency
0
20
60
100
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7

Annex II. Model specifications

Table II.1: Model Specification

Model	Hyperparameters setting
Random Forest	n_estimator=500, max_depth=20, max_features="sqrt', class_weight="balance"
Extra Tree	N_estimator =xx, max_depth=xx, max_feature ='sqrt', class_weight='balanced'
XGBoost	n_estimator=500, reg_lambda=0.01, learning_rate=0.01
AdaBoost	n_estimator = 500, learning_rate=0.1
RUSBoost	n_estimator = 250, learning_rate=0.1, replacement=True
RNN	Layers=1, Nodes=64
KNN	k=9, weights=distance, algorithm=ball_tree / brute / kd_tree
SVM	Gamma=0.01, C=0.1, kernel='rbf'
Logistic	Penalty=none, C=100, solver='saga'
Logistic Lasso	Penalty=l1, C=10, solver='saga'
Logistic Ridge	Penalty=l2, C=1, solver='saga'
Logistic Elastic Net	Penalty=elasticnet, C=100, solver='saga', l1_ratio=0.5

Table II.1 describes the hyperparameters chosen by each model. In ensemble models such as Random Forest and Extra Trees, the number of trees (n_estimators) influences model robustness, with higher values improving performance but increasing computational cost, and lower values potentially leading to underfitting. Tree depth (max_depth) and the number of features considered for splits (max_features) are key in managing overfitting and enhancing model generalization. Class weights can be adjusted for imbalanced datasets. In boosting models like AdaBoost, RUSBoost, and XGBoost, the number of trees and learning rate dictate model adaptability, with higher learning rates potentially leading to overfitting. XGBoost also employs a regularization term (reg_lambda) to mitigate overfitting. Recurrent Neural Networks (RNNs) depend on the number of layers and nodes to capture complex patterns, with too many leading to overfitting and too few causing underfitting. In K-Nearest Neighbors (KNN), the number of neighbors (k) affects model sensitivity, with a smaller *k* potentially leading to overfitting and a larger *k* smoothing the decision boundary but possibly causing underfitting. Kernelized Support Vector Machine (SVM) relies on the Gamma parameter to adjust the model complexity. Finally, in linear models such as Logistic Regression) regularization strength (C) and regularizer type (Penalty) are crucial.

Annex III. Extended Model Results

Figure III.1: Range of Validation AUC Scores by Imputation, Sampling Method, Variable Set and Model

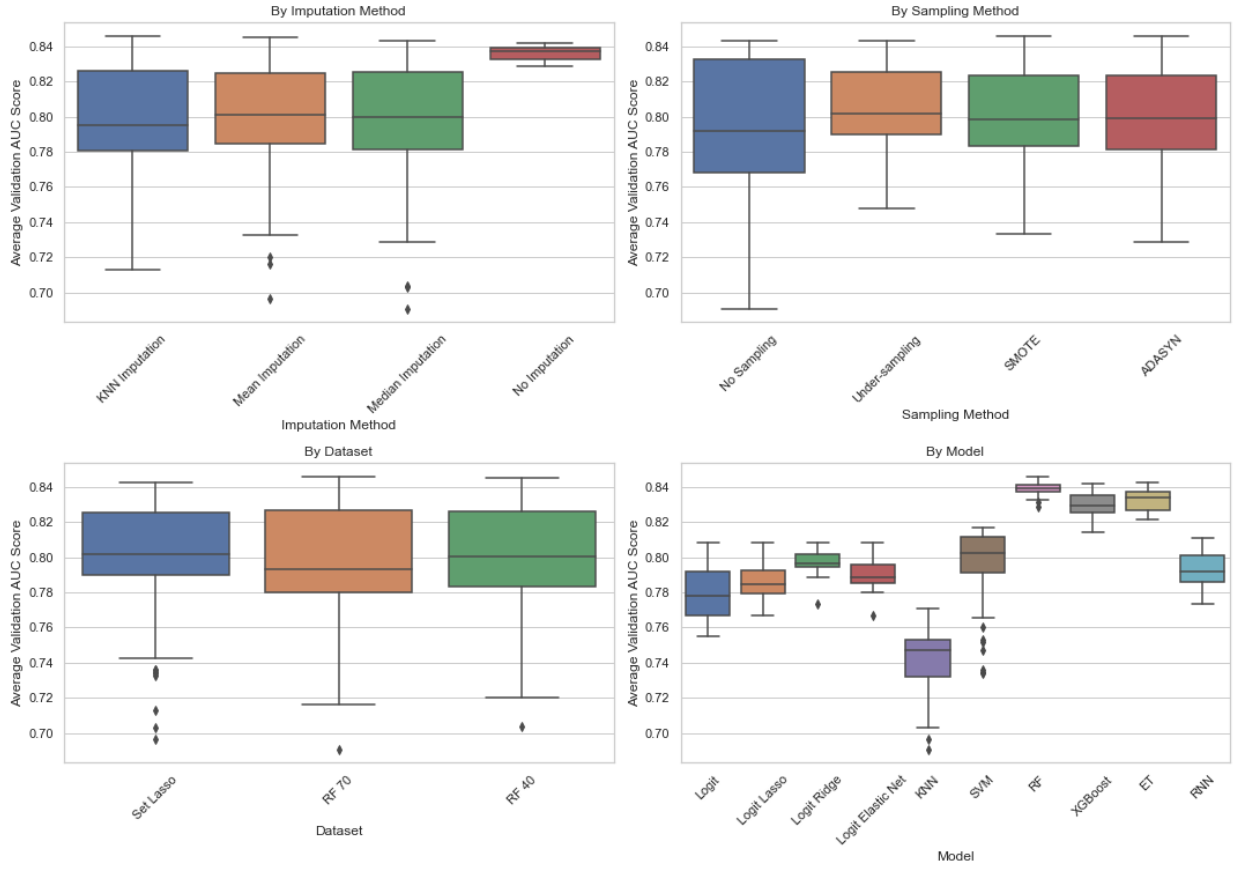


Figure III.1 above compares AUC scores across models (only the best performing ones under each setting), various imputation methods, sampling techniques, and feature sets. While initial observations may suggest a uniform distribution of performance across imputation methods, feature set and sampling methods, a more nuanced examination—controlling for the type of model used—reveals substantial differences in how each model responds to these methods. Following 3 graphs show similar performance indicators for specific models. Specifically, in comparing the performance of Random Forest (RF), Recurrent Neural Network (RNN), and Logistic Regression (Logit) models' it's evident that each model achieves its highest scores under different settings. The Random Forest model attains its peak performance with "SMOTE" sampling, particularly when combined with "KNN" imputation. The Recurrent Neural Network model, on the other hand, shows a slight preference for "SMOTE" and "Under" sampling in achieving its top AUC scores. It also performs exceptionally well with "Median" imputation. Logistic Regression performs well under "Under" sampling, and it tends to perform best with "Mean" imputation, which might be due to the simplicity and effectiveness of this imputation technique for linear models.

In summary, while each model has its unique strengths and preferences, there is a clear indication that the choice of sampling method and imputation technique can significantly influence the model's performance. Achieving the

highest AUC scores often requires a combination of balanced datasets (through methods like "SMOTE", "ADASYN", or "Under" sampling) and appropriate imputation techniques ("KNN" for RF, "Median" for RNN, and "Mean" for Logit). This analysis underscores the necessity of tailoring preprocessing strategies to the specific characteristics and needs of the model at hand.

Figure III.2: Range of Validation AUC Scores – Logistic Regression

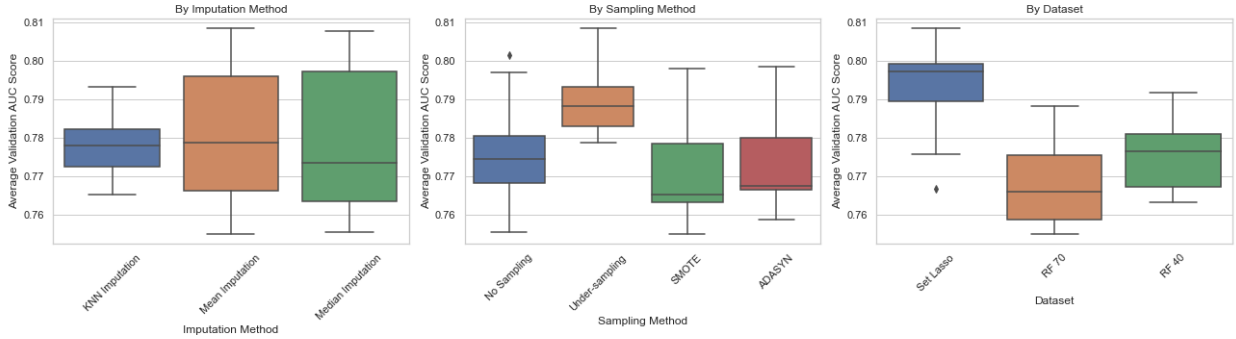


Figure III.3: Range of Validation AUC Scores – Recurrent Neural Network (RNN)

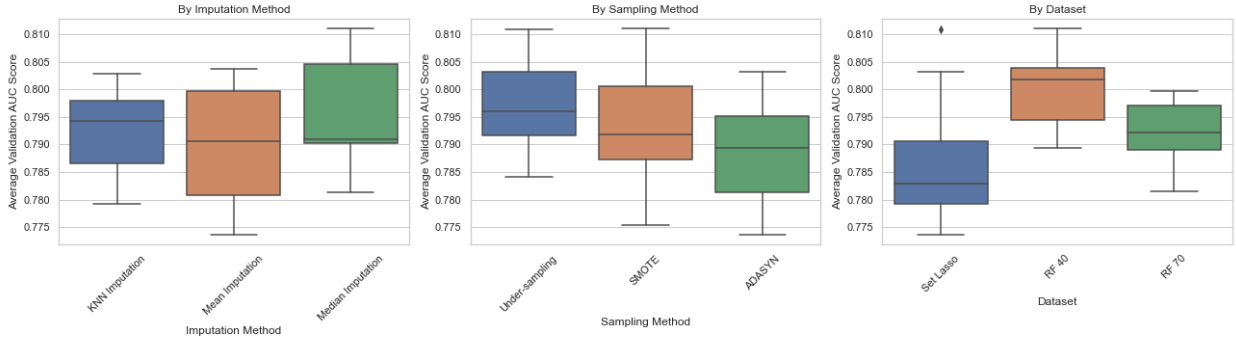
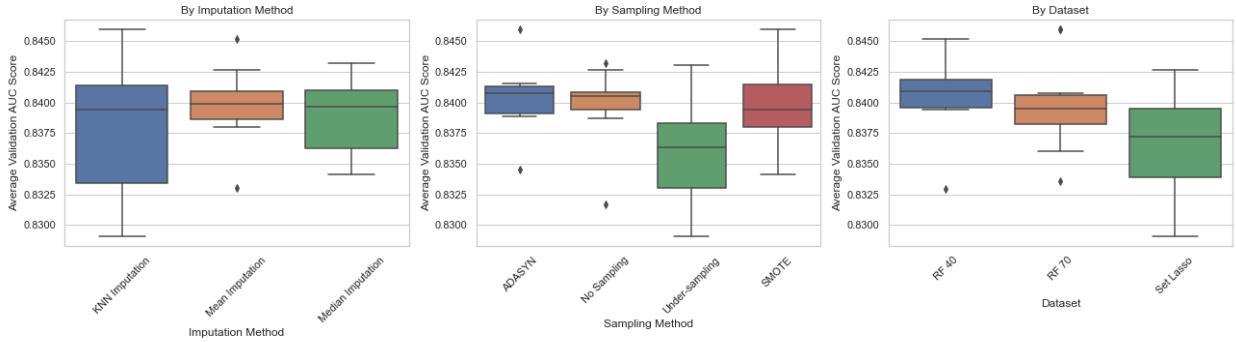


Figure III.4: Range of Validation AUC Scores – Random Forest (RF)



Annex IV. Predicting the Size of IMF Arrangements

Table IV.1: Predicting the Size of IMF Arrangements

VARIABLES	RF (1982-2017)	RF (1990-2017)	RF PLUS (1982-2017)	RF PLUS (1990-2017)
EMDE Dummy	0.2	0.236	0.167	0.247
	(0.2)	(0.327)	(0.241)	(0.326)
Dollar Appreciation (Real Effective Exchange Rate, based on CPI)	-0.1**	-0.335***	-0.152**	-0.448***
	(0.0)	(0.123)	(0.0701)	(0.145)
Membership in RFA	0.1	0.179	0.0356	0.0748
	(0.1)	(0.144)	(0.102)	(0.148)
(Ideal Point Estimate US – ideal point estimate country x)*(-1)	0.1	-0.0583	0.00565	-0.0878
	(0.1)	(0.147)	(0.115)	(0.147)
(Ideal Point Estimate US – ideal point estimate country x)*(-1), lag	-0.2	0.0844	-0.0392	0.133
	(0.1)	(0.152)	(0.114)	(0.155)
Current Account, in USD billion, percent of GDP (in USD)	-0.1	-0.0874	-0.0956	-0.0404
	(0.1)	(0.0925)	(0.0679)	(0.0929)
Current Account, in USD billion, percent of GDP (in USD), lag	0.0	0.00712	0.0624	0.0213
	(0.1)	(0.0750)	(0.0596)	(0.0751)
Exchange Rate, national currency units per U.S. dollar, period average, 1-year growth rate	-0.1	0.0353	-0.0524	0.00930
	(0.1)	(0.248)	(0.0749)	(0.240)
External debt, in USD billion, percent of exports	0.2	0.157	0.0984	0.0539
	(0.1)	(0.141)	(0.102)	(0.140)
External debt, in USD billion, percent of exports, lag	0.1	0.0680	0.0410	0.0488
	(0.1)	(0.101)	(0.0740)	(0.0996)
Gross Reserves, in USD billions, percent of GDP (USD)	0.1	0.00719	0.134	0.00881
	(0.1)	(0.175)	(0.139)	(0.179)
Gross Reserves, in USD billions, percent of imports	0.0	0.0892	0.0399	0.0942
	(0.1)	(0.120)	(0.0922)	(0.121)
FX share of general government debt	-0.1*	-0.163*	-0.132*	-0.146
	(0.1)	(0.0925)	(0.0707)	(0.0921)
FX share of general government debt, lag	-0.1	-0.141**	-0.0629	-0.108
	(0.1)	(0.0702)	(0.0564)	(0.0686)
Hard Peg Indicator (=1 if fine exchange rate regime =1 or =2; =0 otherwise)	0.1	0.00497	0.0596	-0.00596
	(0.1)	(0.108)	(0.0752)	(0.107)
Capital Outflow restrictions	0.0	0.0988	0.0208	0.0860
	(0.0)	(0.0636)	(0.0461)	(0.0630)
Capital Inflows restrictions, lag	0.1	0.0365	0.0672	0.0389
	(0.0)	(0.0674)	(0.0487)	(0.0661)
Trading Partner Growth	-0.0	0.0298	-0.00642	-0.00323
	(0.0)	(0.0520)	(0.0386)	(0.0576)
Overall balance, in USD billion, percent of GDP (in USD)	-0.0	-0.00118	-0.00798	-0.00364
	(0.0)	(0.0190)	(0.0165)	(0.0191)
General government revenue, in national currency, percent of GDP	-0.0	0.0404	-0.0508	-0.0362
	(0.1)	(0.139)	(0.104)	(0.145)
General government revenue, in national currency, percent of GDP, lag	-0.1	-0.0903	-0.0162	0.0115
	(0.1)	(0.0858)	(0.0764)	(0.0977)
Fiscal interest expenses / Fiscal revenue (%)	0.1*	0.0690	0.00452	-0.00672
	(0.0)	(0.0507)	(0.0511)	(0.0714)
Public debt to revenue	0.1	0.104	0.0646	0.0414
	(0.1)	(0.200)	(0.0599)	(0.205)
Public debt to revenue, lag	0.0	0.00458	0.0191	0.0187
	(0.0)	(0.0626)	(0.0237)	(0.0646)
Public external debt, in USD billion, percent of exports, lag	-0.1	-0.0467	-0.0791*	-0.0624
	(0.0)	(0.0586)	(0.0423)	(0.0589)
External debt, in USD billion, percent of exports	-0.1	-0.181	-0.0705	-0.0337
	(0.1)	(0.192)	(0.134)	(0.192)
External debt, in USD billion, percent of exports, lag	-0.1	-0.00195	-0.0146	0.0280
	(0.1)	(0.152)	(0.112)	(0.151)
Loan to deposit ratio	0.1	0.108*	0.0733*	0.0959*
	(0.0)	(0.0577)	(0.0433)	(0.0570)
Financial Inclusion - Access	0.0	0.00824	0.0230	-0.0553
	(0.2)	(0.295)	(0.239)	(0.291)
Financial Inclusion – Access, lag	0.3	0.425	0.250	0.371
	(0.2)	(0.298)	(0.244)	(0.295)
Total credit, in USD billion, percent of GDP (in USD)	0.1	0.0616	0.0773	0.0752
	(0.1)	(0.0784)	(0.0722)	(0.0855)
Private credit, in USD billion, 5-year growth ratee	0.0*	0.0316*	0.0322**	0.0301*
	(0.0)	(0.0174)	(0.0154)	(0.0170)
Private credit, in USD billion, percent of GDP (in USD)	-0.0	-0.179	-0.0979	-0.224

	(0.2)	(0.252)	(0.155)	(0.249)
Private credit, in USD billion, percent of GDP (in USD), lag	0.1	0.284	0.0975	0.172
	(0.1)	(0.234)	(0.148)	(0.233)
PPP income per capita, relative to the U.S.	1.0***	0.780*	0.870**	0.637
	(0.4)	(0.407)	(0.355)	(0.402)
PPP income per capita, relative to the U.S., lag	-0.9**	-0.726*	-0.864**	-0.600
	(0.4)	(0.422)	(0.356)	(0.420)
Consumer Prices, period average, 1-year growth rate	0.0	-0.0904	0.0415	-0.0387
	(0.1)	(0.144)	(0.0623)	(0.143)
Bureaucracy quality (ICRG)	-0.0	0.0293	0.0181	0.0820
	(0.0)	(0.0730)	(0.0491)	(0.0718)
Institutional Quality (ICRG)	-0.1	0.0285	-0.0459	0.00689
	(0.0)	(0.0643)	(0.0430)	(0.0630)
Dollar Appreciation (Real Effective Exchange Rate, based on Consumer Price Index, Index), 5-year growth rate, lag	-0.0	0.118	-0.0127	0.0990
	(0.1)	(0.116)	(0.0576)	(0.127)
Remittances, in USD billion, 1-year growth rate	-0.0	-0.0200	-0.0131	-0.0430
	(0.1)	(0.0563)	(0.0500)	(0.0555)
Total credit outstanding in percent of quota			0.0942***	0.0934**
			(0.0305)	(0.0382)
Reserves as percent of the ARA metric			-0.100*	-0.0524
			(0.0514)	(0.0640)
Capital Adequacy Ratio			0.00872	-0.0754
			(0.0435)	(0.0593)
Access to swap line			1.066***	0.707**
			(0.240)	(0.286)
Net non-FDI liability inflows, lag			0.160*	0.286**
			(0.0947)	(0.138)
Federal Funds Rate, Percent, Monthly, Not Seasonally Adjusted			0.247**	0.469***
			(0.102)	(0.144)
Market Yield on U.S. Treasury Securities at 10-Year Constant Maturity			-0.188	-0.558***
			(0.117)	(0.180)
Terms of Trade			-0.0270	-0.0587
			(0.0595)	(0.0807)
Public debt service to revenue			0.0584	0.110
			(0.0727)	(0.102)
Foreign liability to domestic credit, lag			0.106	0.0899
			(0.0809)	(0.100)
Short-term deposit rate			-0.0158	-0.0299
			(0.0220)	(0.0436)
External bank liabilities, in USD, percent of GDP (in USD)			0.627*	0.599
			(0.320)	(0.460)
Food price inflation			-0.0169	0.00263
			(0.0402)	(0.0505)
Oil price inflation			-0.0322	-0.0639
			(0.0406)	(0.0619)
CBOE Volatility Index: VOX, Index, Daily, Not Seasonally Adjusted, lag			0.0369	0.0756
			(0.0401)	(0.0535)
Constant	4.6***	4.538***	4.569***	4.450***
	(0.2)	(0.284)	(0.224)	(0.287)
Observations	585	388	585	388
R-squared	0.3	0.401	0.372	0.468
Adjusted R-squared	.25	.33	.30	.38
Root MSE	.76	.81	.73	.78
Root MSE OOS	1.01	0.96	1.04	0.90
Sample Period	1982 - 2017	1990 - 2017	1982 - 2017	1990 - 2017

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

PUBLICATIONS

Predicting IMF-Supported Programs: A Machine Learning Approach

Working Paper No. WP/2024/054