INFER
INTERNATIONAL NETWORK FOR ECONOMIC RESEARCH

INFER WORKING PAPER SERIES

INTERNATIONAL NETWORK FOR ECONOMIC RESEARCH

No. 6 | 2025

Do International Reserve Holdings Still Predict Economic Crises? Insights from Recent Machine Learning Techniques

GIANNNIKIS Nikolaos *(Democritus University of Thrace)*
GOGAS Periklis *(Democritus University of Thrace)*
PAPADIMITRIOU Theophilos *(Democritus University of Thrace)*
SAADAOUI Jamel *(University Paris 8)*
SOFIANOS Emmanouil *(University of Strasbourg)*

Website:
https://infer-research.eu/

Contact:
publications@infer.info

Do International Reserve Holdings Still Predict Economic Crises? Insights from Recent Machine Learning Techniques[1]

Nikolaos Giannakis[a], Periklis Gogas[a], Theophilos Papadimitriou[a], Jamel Saadaoui[b], Emmanouil Sofianos[c]

[a] *Department of Economics, Democritus University of Thrace, 69100 Komotini, Greece*
[b] *University of Paris 8, LED, 93200, Saint-Denis, France*
[c] *University of Strasbourg, University of Lorraine, BETA, CNRS, 67000, Strasbourg, France*

This version: Monday, May 26, 2025

Abstract

This study aims to predict currency, banking, and debt crises using a dataset of 184 crisis events and 2896 non-crisis cases from 79 countries (1970-2017). We tested eight machine learning methods: Logistic Regression, KNN, SVM, Random Forest, Balanced Random Forest, Balanced Bagging Classifier, Easy Ensemble Classifier, and Gradient Boosted Trees. The Balanced Random Forest had the best performance with a 72.91% balanced accuracy, predicting 149 out of 184 crises accurately. To address machine learning's black-box issue, we used Variable Importance Measure (VIM) and Partial Dependence Plots (PDP). International reserve holdings, inflation rate, and current account balance were key predictors. Depleting international reserves at varying inflation levels signals impending crises, supporting the buffer effects of international reserves.

Keywords: Currency Crises; Banking Crises; Debt Crises; International Reserve Holdings; Inflation; Machine Learning; Forecasting.

JEL: C53; H68; F37; F47; G2; G01.

[1] Email addresses: papadimi@econ.duth.gr (T.P.); pgkogkas@econ.duth.gr (P.G.); jamelsaadaoui@gmail.com (J.S.); sofianos@unistra.fr (E.S.); ngiannak@econ.duth.gr (N.G.).
* Correspondence: Jamel Saadaoui, jamelsaadaoui@gmail.com.

1. Introduction

The use of machine learning (ML) methods in forecasting economic crises has gained significant attention recently (Beutel et al., 2019, Chen et al., 2023, Liu et al., 2022, Reimann, 2024, Samitas et al., 2020). Traditional econometric models, such as logit regressions, have long been employed to predict banking crises, but recent advancements in computational techniques have prompted researchers to explore the potential use of ML approaches. Although ML methods offer advantages in handling complex interactions and non-linear relationships, the "black box" limitation makes them unpopular in many scientific areas, including economics. To overcome this weakness, researchers are working on a new area called XAI (eXplainable Artificial Intelligence), or XML (eXplainable Machine Learning), that attempts to reveal as many of the relationships hidden inside the black box as possible.

Beutel et al., (2019) provide a comprehensive comparison between conventional logit models and various ML approaches in predicting systemic banking crises across advanced economies. Their findings indicate that while ML methods often achieve high in-sample accuracy, they tend to underperform in recursive out-of-sample forecasts, largely due to issues of overfitting and instability. In contrast, the logit model consistently yields lower prediction errors and identifies key economic indicators (such as credit expansions, asset price booms, and external imbalances) as significant predictors of crises. Their study suggests that despite ML's theoretical advantages, traditional econometric models still effectively utilize available information, making them more reliable in real-world forecasting applications. However, they highlight that certain methodologies may hold promises for improving ML-based crisis prediction in the future.

Kauko (2014) presents a broad survey of empirical literature on early warning indicators of banking crises, tracing the evolution of predictive models from descriptive analyses to cross-national panel studies. His research underscores the recurring boom-bust cycles that lead to banking crises, where international borrowing fuels domestic lending, leading to asset price bubbles and current account imbalances. The study highlights that both developed and developing economies are susceptible to these financial vulnerabilities. This aligns with the broader consensus that systemic banking crises often follow periods of excessive credit growth and external imbalances, reinforcing the need for robust forecasting models. These insights provide a crucial foundation for integrating ML techniques with traditional economic theories to enhance predictive accuracy and stability.

What are the predictors of economic and financial crises? This question is one of the most important for policymakers around the world. In fact, identifying these drivers may be of utmost importance to prevent the occurrence of such crises.

In the literature, several macroeconomic fundamentals and financial variables have been identified as harbingers of financial crises. Frankel and Saravelos (2012) explore the early warning signals explaining the cross-country incidence during the Global Financial Crisis (GFC). They found that the level of international reserves in 2007 explains how badly countries have been hit by the GFC, using several dependent variables capturing economic and financial performance and resilience

(Table 1). Their results are in line with the pre-GFC research Obstfeld et al. (2009 and 2010) and Rose and Spiegel (2009 and 2010). Along with other predictors like lower past credit growth, larger current account balances, and saving rates, lower external and short-term debt are associated with a lower incidence of crisis. However, the holding of international reserves is the most robust determinant of a lower incidence of crises. This remarkable result is robust to multivariate analysis and to different crisis definitions.

Table 1. Summary of pre-2008 early warning indicators.

Leading indicator	KLR (1998)	Hawkins and Klau (2000)	Abiad (2003)	Others	Total
Reserves	14	18	13	5	50
Real exchange rate	12	22	11	3	48
GDP	6	15	1	3	25
Credit	5	8	6	3	22
Current account	4	10	6	2	22
Money supply	2	16	1	0	19
Exports or imports	2	9	4	2	17
Inflation	5	7	1	2	15
Equity returns	1	8	3	1	13
Real interest rate	2	8	2	1	13
Debt composition	4	4	2	0	10
Budget balance	3	5	1	0	9
Terms of trade	2	6	1	0	9
Contagion	1	5	0	0	6
Political/legal	3	2	1	0	6
Capital flows	3	0	0	0	3
External debt	0	1	1	1	3
Number of studies	28	28	20	7	83

Source: reproduced from Frankel and Saravelos (2012).

Among them, the holding of international reserves has been explored by several papers (Aizenman and Riera-Crichton, 2008; Ahmed et al., 2023, Aizenman et al., 2024, Coulibaly et al., 2024). The stabilizing properties of international reserves may be related to two main channels, namely, the balance sheet channel and the intervention channel. The accumulation of an important stock of international reserves may help countries cope with financial stress by intervening directly on financial markets (i.e., the intervention channel). Besides, it can also help to stabilize the exchange rate through the balance sheet channel. The size of international reserves has been demonstrated to possess stabilization properties, which are anchored by expectations.

Besides, when the stock of international reserves falls below a threshold, net new capital inflows may abruptly end, leading to debt rollover problems and ‘flight to safety’, as documented by Dominguez et al. (2012) during the Global Financial Crisis (GFC). These capital outflows can accelerate the depletion rate of international reserves. Thus, adequate management of international reserve holdings during tranquil times and selling them during times of ‘flight to safety’ may contribute to reducing the probability of a full-blown crisis, especially in the corporate sector, as

explained in Aizenman and Jinjarak (2020). This was a major concern for Korea during the East Asian crisis and for many emerging countries during the GFC (Dominguez et al., 2012). For all these reasons, the *ex-ante* stock of international reserves should be an important predictor of economic crises.

In this paper, we employ several machine learning methods to forecast the currency, banking and debt crises for 79 countries spanning the period from 1970 to 2017. We also train a simpler Logistic Regression model to serve as a benchmark. The dataset, we use, is composed of 42 independent macroeconomic and financial variables. In a second part, we employ the Variable Importance Measure (VIM) and the Partial Dependence Plot (PDP) techniques to uncover helpful insights about the interconnections between the variables and the possible triggers of crises.

The paper is structured as follows, in section 2 we describe the methodologies, and the dataset used. In section 3 we present the results, while section 4 concludes the paper.

2. Methodology and Data

2.1. Machine Learning

Machine learning was conceived in the 1950s, as the foundational learning component of artificial intelligence systems. Since it evolved both as a part of AI and independently. Machine learning involves the development of systems that can analyze data, recognize patterns, and make decisions with little to no human input, without being explicitly programmed for these tasks.

Historically, machine learning has relied on large datasets (Gogas and Papadimitriou, 2021). Therefore, machine learning in economics has mainly been applied to financial data, which is the only subfield of economics with large datasets, mainly due to the availability of high-frequency data sampling.

In machine learning, the dataset is split into the training and the testing (out-of-sample) part, the former is used to train the model (identify the hyperparameters of the optimal model), and the latter is utilized to evaluate the generalization ability of the trained model to unknown data.

2.2. SVM

Support Vector Machines (SVM) are a set of methods based on the idea that the optimal binary classification is identified in the linear classifier maximizing the distance between the two classes. The initial concept was further developed to perform multiclass classification and regression.

Let us consider a dataset of N m-sized vectors $\mathbf{x}_i \in \mathbb{R}^m$ (i=1, 2…, N). These vectors are labelled according to the class they belong $y_i \in \{-1, +1\}$.

If the two classes are linearly separable, we define a boundary as:

$$f(\mathbf{x}_i) = \mathbf{w}^T \mathbf{x}_i - b = 0. \quad (1)$$

Subject to:

$$\mathbf{w}^T\mathbf{x}_i - b > 0 \; \forall i: y_i = +1,$$
$$\mathbf{w}^T\mathbf{x}_i - b < 0 \; \forall i: y_i = -1, \tag{2}$$

w is the parameter vector, and b is the bias (Figure 1). So $y_i f(\mathbf{x}_i) > 0, \forall i$.

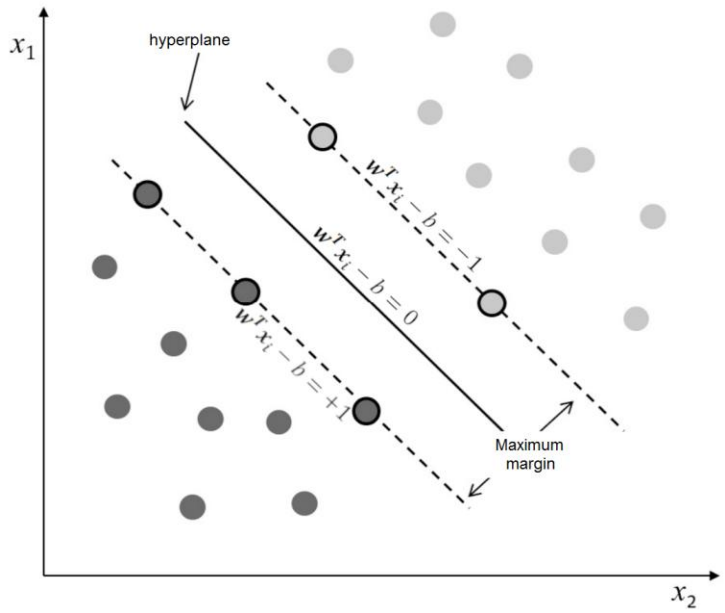

Figure 1. Support vectors and optimal hyperplane identification. The SVs (represented with the pronounced black contour) define the margins that are represented with the dashed lines and outline the separating hyperplane represented by the continuous line.

In scenarios where data is linearly separable, the boundary (represented as a line in two dimensions, a plane in three, and a hyperplane in higher-dimensional spaces) acts as the decision tool that assigns each data point to a class. A limited number of data points, known as support vectors (SVs), determine the exact position of this boundary. As illustrated in Figure 1, support vectors are marked with bold contours. The dashed lines indicate the margins, which are parallel to the optimal separator and are defined by the support vectors, while the separator itself is shown as a solid line. The width of the margin, reflects the separation between the two classes. The objective of an SVM is to find the linear boundary that maximizes this margin.

However, real-world datasets often contain noise and outliers, making them unsuitable for strict linear separation. To address such challenges, Cortes and Vapnik (1995) proposed the introduction of non-negative slack variables along with a parameter C, which allows a controlled level of misclassification.

When the data cannot be separated linearly, SVMs are extended using kernel: functions mapping the original data into a higher-dimensional feature space where linear separation becomes possible. Figure 2 demonstrates this: on the left, red and blue circles represent a dataset that cannot be linearly divided in two-dimensional space. After applying a kernel function, the same data is projected into a three-dimensional feature space (right), where a linear decision boundary can be identified.

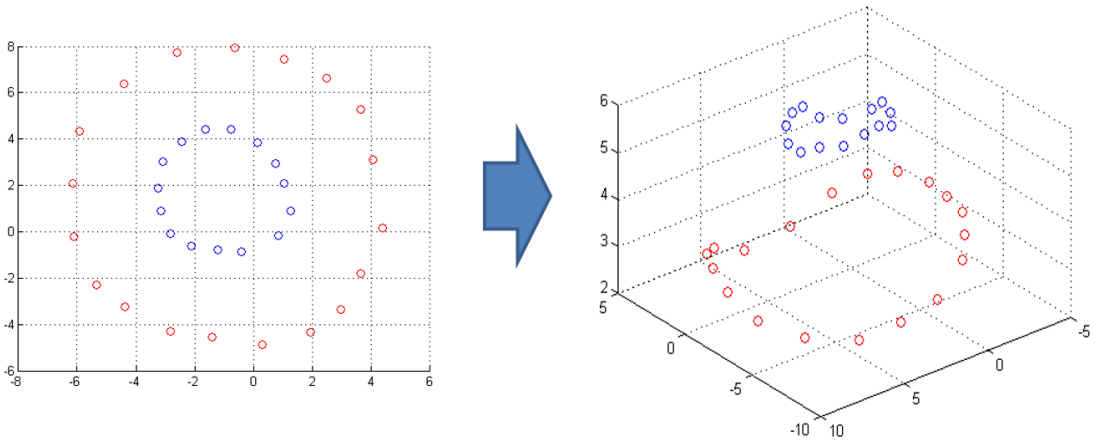

Figure 2. The left side illustrates a two-class scenario in the original data space where the classes are not linearly separable, while the right side shows the corresponding feature space after projection, where the classes become linearly separable.

This study investigates three types of kernel functions: the linear kernel, the radial basis function kernel (RBF), and the polynomial kernel. The linear kernel identifies a separating hyperplane within the original data space, whereas both the RBF and polynomial kernels transform the input data into a higher-dimensional space. A more detailed analysis can be found in Gogas e.a. (2019).

2.3. Decision trees

Decision trees are a type of supervised machine learning algorithm that can be used for classification or regression. They are structured like flowcharts with top-down nodes and branches (Figure 3). The algorithm recursively partitions the data based on the most informative feature (Gogas et al., 2022). Each node represents a splitting criterion, and each branch represents the corresponding outcome. The top node is the root node, representing the complete dataset, while the others are called decision nodes. The leaves, also known as terminal nodes or leaf nodes, represent the final outcomes of the decision-making process, in this case, a class for the classification procedure.

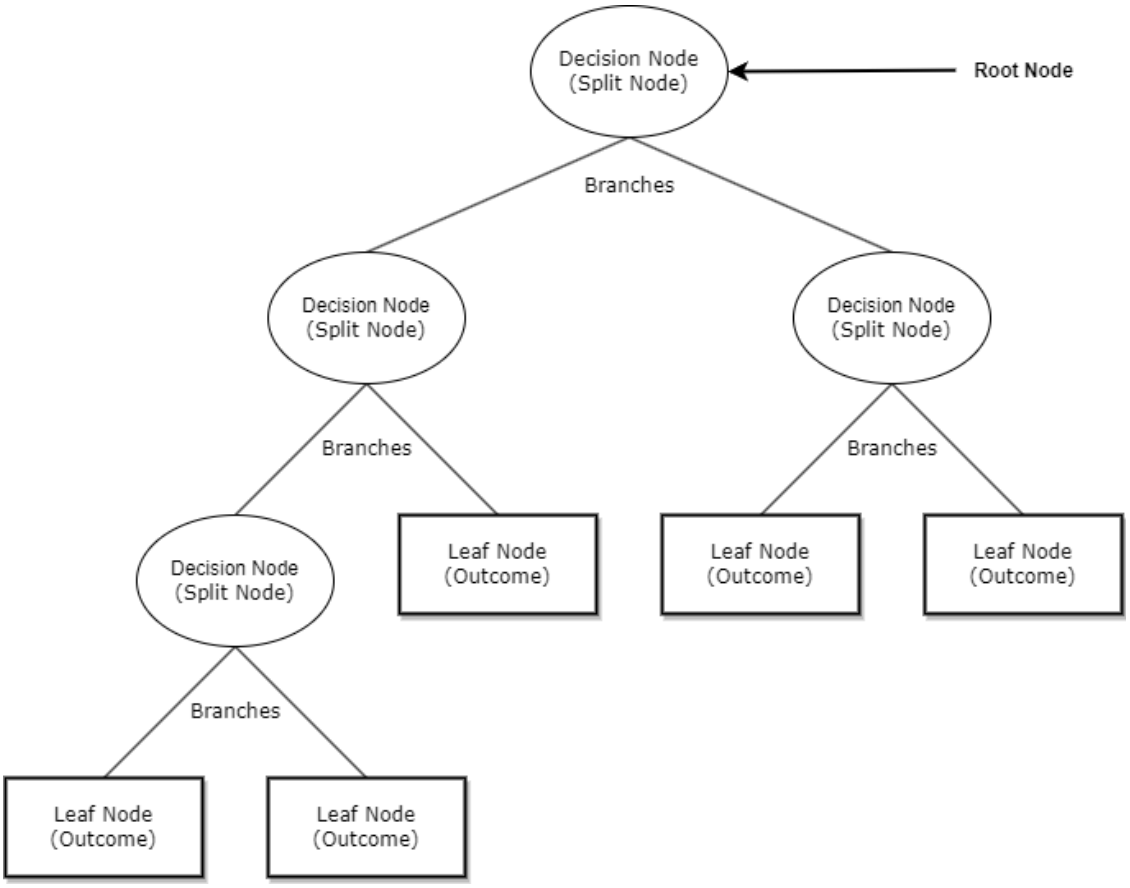

Figure 3: Example of a decision tree.

2.4. Balanced Random Forest

Decision trees have the advantages that they usually perform well with the training data and are easy to interpret. Nonetheless, their main drawback is that their generalization ability is low, and they perform poorly in out-of-sample data – the definition of overfitting. Using machine learning terminology, they tend to have a low bias but a high variance. A common approach to address this issue involves the use of ensemble methods. These methods synthesize many weak predictions from several models into a strong and accurate one. The two main ensemble methods are bagging[2] (Breiman, 1996) and boosting (Freund and Schapire, 1996).

Bagging trains the weak learners independently and in parallel for determining the model's average score. In regression tasks, the prediction is made by calculating the average value of the target variable from the data points within every leaf node. The most common example of bagging is the random forest algorithm (Breiman, 2001). For each tree, a separate subsample is randomly chosen with replacement (bootstrapping). Additionally, a random subset of the initial independent variables is selected for training each tree.

[2] BAGGing (Bootstrapped AGGregating).

Similar to most of the classification algorithms, random forests might result in bad performance when dealing with imbalanced data. The algorithm is constructed to minimize the overall error rate, so it focuses more on predicting the majority class, resulting in poor accuracy in the minority class. To tackle this issue, the balanced random forest algorithm was used (Chen and Liaw, 2004). As Ling and Li in 1998 and Drummond and Holte in 2003 shown, a tree classifier might have better performance based on a metric when making the classes equal either by under sampling or by over sampling within a limit. In imbalanced data, the random forest classifier through the bootstrap technique might create a sample that contains a few or no observations of the minority class. A way of addressing this issue is to use a stratified bootstrap. Balanced random forest will use a stratified bootstrap sample having equal number of samples of each class and apply a modified random forest algorithm: at each tree node it searches a set of randomly selected variables instead of all variables for the optimal split (Chen and Liaw, 2004).

2.5. Easy Ensemble Classifier

Easy Ensemble classifier is being used to tackle imbalanced classification problems (Liu et al., 2009). It is an ensemble of AdaBoost classifiers trained on different balanced bootstrap samples. The balancing between the samples is achieved by employing the random under-sampling technique. In other words, the classification algorithm, randomly samples cases from the majority class to obtain a balanced set and trains a boosting model based on the AdaBoost learner. This process is being repeated multiple times and the final outcome is identified as the most popular class.

2.6. Gradient Boosted Trees

Gradient boosted trees classifier is an ensemble classifier combining decision trees, to create a strong predictive model. The algorithm, in each iteration, marks the misclassified cases, and feeds them to the next model so that they can be correctly classified. The final prediction is a weighted aggregation of all outputs. The gradient boosted trees algorithm is one of the most accurate classification models; it can capture complex patterns in data while preventing overfitting. However, it is computationally expensive due to the sequential nature of the algorithm, especially for large datasets (Friedman, 2001).

2.7. Variable Importance Measure

In tree-based models, the Variable Importance Measure (VIM) can be computed to rank the contribution of each variable to the model performance. At each node of a decision tree, the algorithm determines which feature to split the data based on maximizing class separation or reducing impurity. One of the criteria used for this decision is the Gini index. In an ensemble tree model, the Gini variable importance is calculated by summing the Gini Index reductions caused by each feature over all the trees. Therefore, the more a feature reduces the Gini index when used in a

split across all trees, the more essential it is considered to be. This technique is also referred to as Mean Decrease in Impurity (MDI).

2.8. Overfitting

Overfitting is a common issue in data science and machine learning, where a model learns the training data capturing not only the underlying patterns but also the noise and random fluctuations specific to that dataset, resulting in poor performance when applied to new, unseen data. To assess how well our models generalize beyond the training set, we divided the dataset into two subsets: a training set used for model training and a test set (out-of-sample) used to evaluate the forecasting performance. To further mitigate overfitting and rigorously examine the generalization capability of our models, we employed the cross-validation technique, as illustrated in Figure 4.

Cross-validation is utilized during the training process to determine the model's optimal hyperparameters and prevent overfitting. It involves randomly dividing the training dataset into *k* equal-sized folds (subsets) for testing and training the model *k* times with different data parts. Each iteration uses a different fold for testing, while the remaining *k*-1 folds are used for training. The performance of each set of tested hyperparameters is calculated as the mean performance over all the test *k* folds (Figure 4).

In our tests, to mitigate the potential high variance associated with the small sample size of class 1, we employed a repeated stratified k-fold cross-validation approach. This method ensures a stable and reliable estimation by reducing variability across different training and validation splits (Kim, 2009). Repeated cross-validation is considered a variant of cross-validation technique which involves repeating the process of k-fold cross-validation a certain number of times, reporting the mean metric result of all test folds from all repeats. This mean metric result is expected to be a more accurate estimate of the true unknown mean performance of the model compared to training the model with k-fold cross validation only once because in each iteration, the selection of data for each fold is different. We used a 5-repeated 5-fold stratified configuration.

The dataset was split into stratified train, test and out-of-sample sets for training and evaluation. The above process was repeated 5 times and the cross-validation test-set (in-sample) and the out-of-sample set evaluation results from each model were averaged. Each data point was tested multiple times across different validation sets. In total, 184 crisis instances were evaluated throughout the process.

2.9. Recursive Feature Elimination

In addition, Recursive Feature Elimination was used to include only relevant features (variables) that contribute to the prediction problem. Recursive Feature Elimination is a feature selection technique that recursively removes the least important features iteratively. In the case of the Random Forest classifier, the algorithm ranks the features using the Gini importance (their contribution to the model’s performance). In every step, the least important feature is being removed. The process is repeated until the stopping criterion is met. In each iteration the model

performance is evaluated and at the end, from the resulted performances, the optimal feature selection is defined (Guyon et al., 2002). The feature selection technique has been integrated with a 5-fold cross-validation in order to evaluate the predictive ability of the feature subsets.

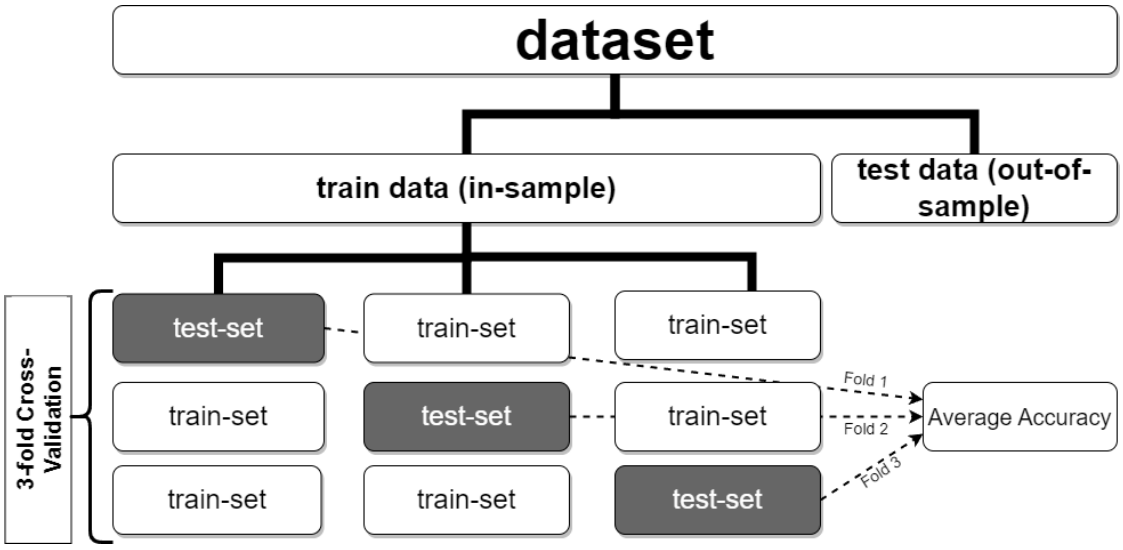

Figure 4: A visual representation of a 3-fold cross-validation process. In each iteration corresponding to a specific set of hyperparameters, one fold is designated as the test set, while the remaining folds are used to train the model. The model's performance is then averaged across the k-test folds to evaluate its in-sample accuracy. The results of this process guide the selection of the final model, which is then tested on the out-of-sample data that was not used during training. (Gogas et al., 2019).

There are several forecasting metrics that can be used to assess the forecasting performance of the models. The metric that we use in this paper is the balanced accuracy, which is commonly used when dealing with imbalanced data. It is defined as:

$$Balanced\ Accuracy = \frac{sensitivity + specificity}{2} = \frac{1}{2}\left(\frac{TP}{TP + FN}\frac{TN}{TN + FP}\right) \quad (4)$$

True Positives (TP) are the outcomes that are correctly predicted as positives while False Positives (FP) are the outcomes inaccurately predicted as positives. Accordingly, True negative (TN) represents the number of correctly classified negative samples and False negative (FN) denotes the number of samples incorrectly classified as negative. Sensitivity, also known as recall or true positive rate, measures the proportion of True Positives out of all the positive cases (True Positives + False Negatives). Specificity, also known as true negative rate, measures the proportion of True Negatives out of all the negative cases (True Negatives + False Positives).

2.10. The dataset

The dataset consists of 79 countries (Table A1 in the appendix details the country list) and 40 variables (Table 1) spanning from 1970 to 2017 in annual frequency. After deleting the data entries with missing values, the final dataset has 3080 observations, 184 crises and 2896 normal cases. The dataset was split 5 times into in-sample and out-of-sample sets based on the ratio 70% - 30% using stratified sampling. This approach is used to identify the model and it is hyperparameters that produce the best results across the dataset.

Table 1: List of explanatory variables

Symbols	Description
current	Current account balance as percent of GDP
crisis	Banking or Debt crisis
deflator	Inflation, GDP deflator (annual %)
eap	Dummy for East Asian and Pacific countries
eca	Dummy for European and Central Asia countries
emg	Dummy for Emerging countries
emg2	Dummy for Emerging countries (alternative grouping)
emp	Number of persons engaged (millions)
eu	Dummy for the EU members
euro	Dummy for the eurozone members
fc	Dummy for Financial centers
fdi_inflow	Foreign direct investment, net inflows (% of GDP)
fixed	Dummy for less flexible exchange rate regimes
flexible	Dummy for more flexible exchange rate regimes
full	Full sample dummy
hc	Human capital index
idc	Dummy for Industrialized countries
infl	Inflation, consumer prices (annual %)
ka_open	Chinn-Ito index, normalized
kaopen	Chinn-Ito index
lac	Dummy for Latin American countries
ldc	Dummy for Less developed countries
mena	Dummy for the Middle East and North Africa group
na	Dummy for North America
nfa	Net Foreign asset
oil	Dummy for Oil exporters
opn	Trade (% of GDP)
rel_inc	Relative income in US, Penn World Tables (PWT)
res	International Reserve-to-GDP
ryg	GDP growth (annual %)
rypc_g	GDP per capita growth (annual %)
rypc_ppp	Real per capita income in PPP, PWT
sa	Dummy for the South Asia group
share_y	Share of income in World, PWT
s_rgdpe	Expenditure-side real GDP at chained PPPs (in mil. 2005US$)
ssa	Dummy for Subsaharan Africa country group
tot	Net barter terms of trade index (2000 = 100)
we	Dummy for the Western Europe country group

3. Empirical Results

3.1. Forecasting Results

As a first step in the empirical part of the study, we performed a feature selection procedure, namely the Recursive Feature Elimination algorithm, using a Random Forest model as the classifier. Then, we trained several machine learning models, which included the: Logistic Regression (Logit), K-Nearest Neighbors (KNN), Support Vector Machines (SVM), Random Forests, Balanced Random Forests, Balanced Bagging Classifier, Easy Ensemble Classification (with ADABoost) and Gradient Boosted Trees (GBT). Since the dataset is heavily imbalanced (94.1% non-crises and 5.9% crises), we used several relevant weighting techniques: regular weights, balanced weights and under-sampling techniques in ratios of 1:1, 1:2 and 1:3[3]. Most of the models and ensemble techniques produced unsatisfactory results except for the Balanced Random Forest, the SVM coupled with the RBF kernel and the Gradient Boosted Trees. The results of the 10 best models[4] (in terms of OOS Balanced accuracy) are presented in Table 2.

Table 2: The 10 best models with out-of-sample (OOS) Balanced Accuracy, in-sample Balanced Accuracy and true positive (crises correctly predicted) for the out-of-sample part.

#	Model	OOS Balanced Accuracy	In-sample Balanced Accuracy	True Positive Rate OOS
1	Balanced Random Forest	72.91%	66.14%	81.00%
2	Balanced Random Forest under sampling proportions 1:3	70.33%	67.00%	73.00%
3	Random Forest balanced weights under sampling proportions 1:1	70.24%	67.73%	76.00%
4	Random Forest regular weights under sampling proportions 1:1	70.16%	67.15%	76.00%
5	Balanced Random Forest under sampling proportions 1:1	69.96%	66.43%	77.00%
6	Balanced Random Forest under sampling proportions 1:2	69.75%	66.20%	75.00%
7	Gradient Boosted Trees	69.60%	66.49%	76.00%
8	Easy Ensemble – AdaBoost	67.81%	63.20%	75.00%
9	SVM (RBF) balanced weights	66.15%	63.73%	72.00%
10	Logistic Regression balanced weights under sampling proportions 1:2	65.53%	63.78%	74.00%

[3] Hasanin and Khoshgoftaar (2018) mention under-sampling the majority class to a 1:1 proportion might improve the model performance, besides, it might also lead to information loss, where valuable data might be removed, causing the dataset to be less representative of the overall population. That's why we don't limit the experiment in 1:1 ratio and go up to 1:3 aiming to find a balance between improved model performance and retaining enough information from the dataset.

[4] A complete list with the results from all models can be found in Table A2 in Appendix.

The optimal model as we can see in Table 2 is the Balanced Random Forest with 120 estimators (trees), a maximum depth of 11 nodes per tree, and splitting the subsets using the Gini index. The feature selection technique was applied on the training set. From Table 2, we can see that the Balanced Random Forest model achieves an average in-sample balanced accuracy of 66.14% and an averaged out-of-sample balanced accuracy of 72.91%. The model reached an 81% accuracy in out-of-sample forecasting the crises (Table 2, true positive rate OOS and Figure 5); it correctly predicted 149 out of 184 crises.

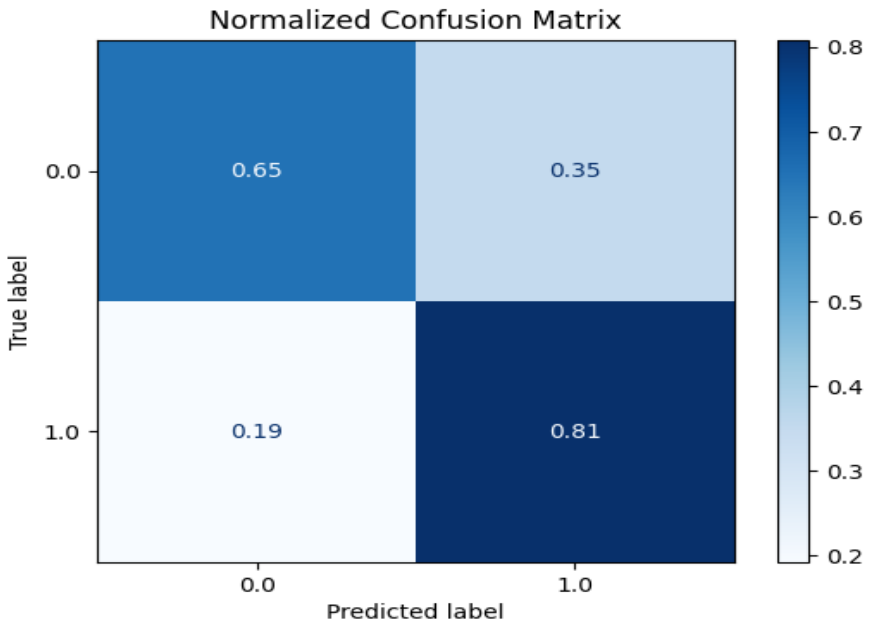

Figure 5: Normalized (in %) Confusion Matrix for the optimal Balanced Random Forest model. A label=1 indicates a crisis and 0 a non-crisis.

3.2. Variable Importance Measure - VIM

We used the Gini Variable Importance Measure to identify and rank the most important variables of the optimal model, the Balanced Random Forest. In Figure 6, we present the ranking of the ten most important variables according to the Gini VIM.

According to Figure 6, the international reserves play a crucial role in forecasting economic crises. This result aligns with the theoretical and empirical literature, which highlights the protective function of reserves in mitigating financial instability.

Hernandez (2018) provides a robust theoretical framework explaining how international reserves reduce the probability of debt crises. His model incorporates both fundamental and market-sentiment-driven crises, demonstrating that reserves serve as a buffer against self-fulfilling rollover crises and external defaults. By ensuring that sovereigns can service their debt obligations even in the absence of new borrowing, reserves shrink the set of states in which a crisis equilibrium can occur. Consequently, a higher level of reserves reduces sovereign spreads, thereby improving

market confidence and access to debt markets. Our findings corroborate this mechanism, as higher reserves are associated with lower crisis probabilities in our predictive model.

The discussion on reserve adequacy indicators by Aydın and Tunç (2022) provides additional insight into the evolving role of international reserves in forecasting crises. Their study suggests that traditional reserve adequacy indicators, such as the short-term debt-to-reserves ratio, have lost their predictive power in recent decades due to the increased global capital mobility. Instead, broader measures, including reserves in months of imports and the broad money-to-reserves ratio, remain effective predictors of currency crises. This shift aligns with our findings, which indicate that international reserves remain a strong predictor of crises.

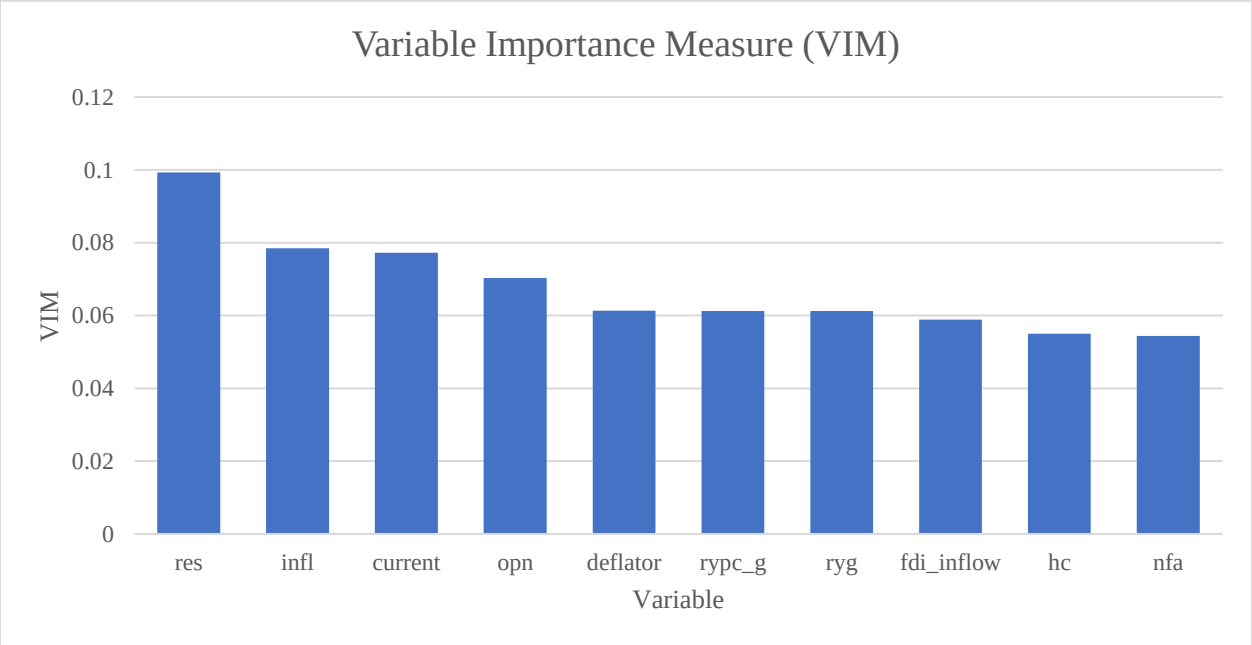

Figure 6: Variable Importance Measure (VIM) results for the optimal Balanced Random Forest model.

3.3. Partial Dependence Plots - PDPs

To explore the relationships between the target variable and the most informative predictors, we employed the Partial Dependence Plots (PDP) methodology. A Partial Dependence Plot reveals the marginal effect of the features on the predicted outcome of a machine learning model (Friedman, 2001).

A Partial Dependence Plot (PDP) shows how the predicted variable changes as a specific independent variable fluctuates while keeping all other variables constant (sensitivity analysis). It provides insights into the relationship between a single variable and the model's predictions. It is a useful tool as it enhances the interpretability of the trained machine learning model in several ways, providing empirical information on:

a) The importance of a specific variable: a PDP with a close to zero slope indicates an independent variable that has no or negligible impact on the target variable.

b) As machine learning models are usually not linear, the linearity (linear PDP) or non-linearity (non-linear PDP) of the effect of an independent variable on the target variable.
c) The monotonicity of the relationship between the specific independent variable and the target variable as we can identify positively or negatively definite monotonic relationships.
d) Monotonically non-decreasing or non-increasing relationships when the independent variable has for some value ranges a slope close to zero and for other ranges a positive or a negative effect respectively.
e) Finally, even non-monotonic, complex, relationships between the independent variables x^j and the target variable *y*, where for some range(s) of a variable x^j the relationship may be positively monotonic and in other range(s) negatively monotonic or independent, i.e. x^j does not affect *y*.

Under this setting, the PDPs provide information a) on the linear and/or nonlinear relationship between the features of the trained model and the target variable and b) how this relationship may differ for various ranges of values of the independent variables.

In our case, where we deal with a binary classification problem, the PDP displays on the horizontal axis the range of values of the selected variable x^j and on the vertical axis the probability of class 1, i.e. a crisis. According to our empirical results and specifically the VIM that provides the most influential features, two of them can be associated with policy decisions: a) the reserves-to-GDP ratio and b)the inflation. Thus, below, we present the PDPs for these two variables. Figure 8 depicts the PDP for the International Reserves-to-GDP ratio, while Figure 9 shows the PDP for the inflation.

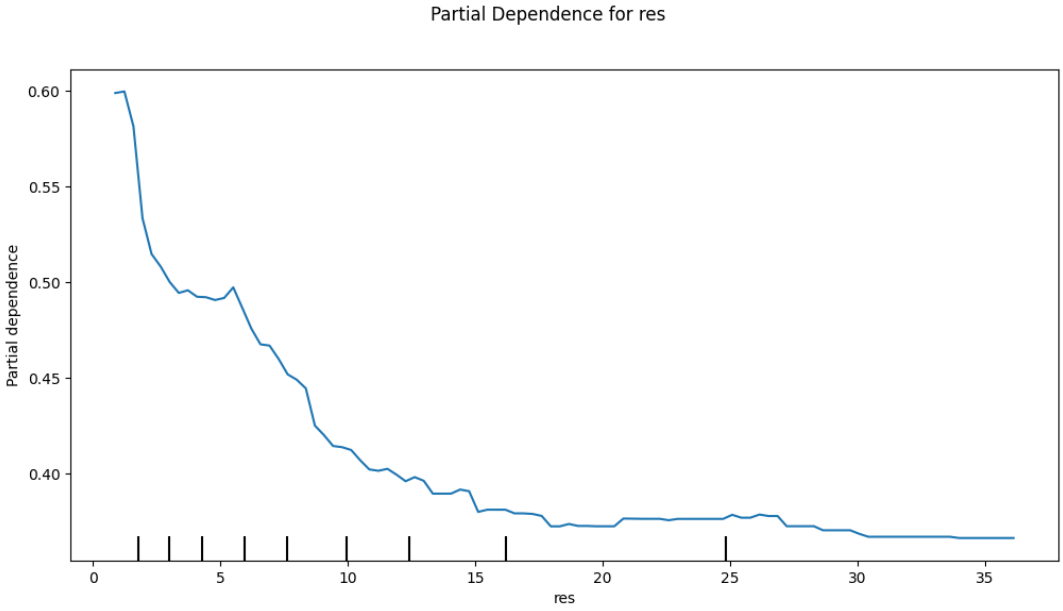

Figure 7: Partial Dependence Plot for International Reserves-to-GDP ratio

From Figure 7, it is evident that there is a monotonically mostly decreasing relationship between the International Reserves-to-GDP ratio and the probability for a crisis to occur. In more detail, we can see that a) The probability of a crisis to occur decreases sharply (from 60% to less than 40%) when the International Reserves-to-GDP ratio increases from 0% to 15%, b) after that, for International Reserves-to-GDP values higher than 15%, the probability of a crisis is almost constant at less than 40%. Thus, in terms of shielding the economy from a possible crisis, international reserves play a crucial role for values between 0% and 15%. Reserves exceeding 15% are redundant in terms of reducing the probability of a crisis. These findings corroborate the results of Benigno, Fornaro, and Wolf (2022). They link reserve accumulation to economic growth and financial crisis prevention. Their model shows that reserve accumulation induces real exchange rate depreciation, fostering growth in the tradable sector. Furthermore, reserves provide liquidity during financial crises, amplifying their positive role for economic stability. This perspective suggests that reserves not only function as a crisis prevention tool but also contribute to general macroeconomic resilience and growth. Our empirical evidence supports this view, as countries with higher reserve levels exhibit a significantly lower probability of a crisis, likely due to their enhanced ability to absorb external shocks, but only up to the the 15% level. An increase in international reserves above 15% is not associated with a further reduction in the probability of a crisis.

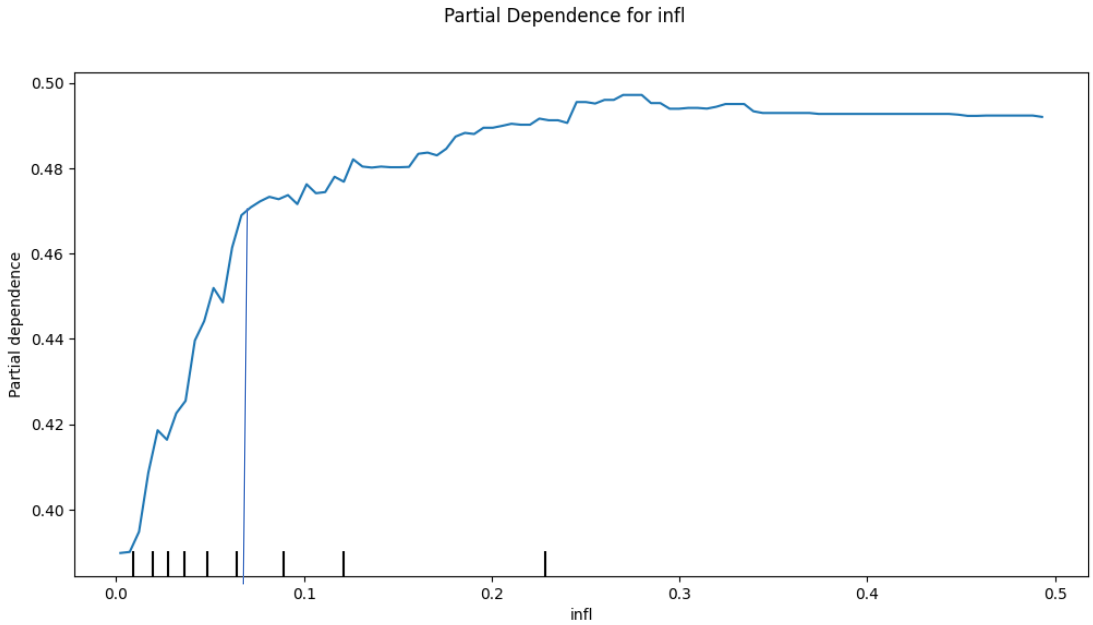

Figure 8: Partial Dependence Plot for Inflation

In Figure 8, we present the PDP for the inflation. We can observe that: a) there is a monotonic non-decreasing relationship between the inflation rate and the probability of a crisis to occur as expected, b) the slope of the PDP is not constant; there is a steep increase in the probability of a crisis to occur from less than 40% to approximately 47% when the inflation rate increases from 0% to approximately 7%, and this slope decreases for inflation rates from 7% to 25% where the

probability of a crisis increases slightly from 47% to approximately 49%, and finally c) an inflation rate greater than 25%, does not seem to increase any further the probability of a crisis. For that range it remains constant at 49%.

In Figure 9, we plot the 2-variable PDP, including both policy variables, the International Reserves-to-GDP ratio and the Inflation rate. The 2-variable PDP is created in a similar manner as the PDPs described above. The only difference is that in the 2-variable PDP, the two variables are iteratively assigned values based on a grid that includes all sample instances of both variables. In Figure 9, the horizontal axis represents the sample values for the inflation rate, while the vertical axis represents the sample values of the International Reserves-to-GDP ratio. The probability of a crisis to occur is represented by the isoquant ranges that are marked with different colors. In general, as we move from the top-left corner to the bottom right, the probability of a crisis increases from 46% to 69%.

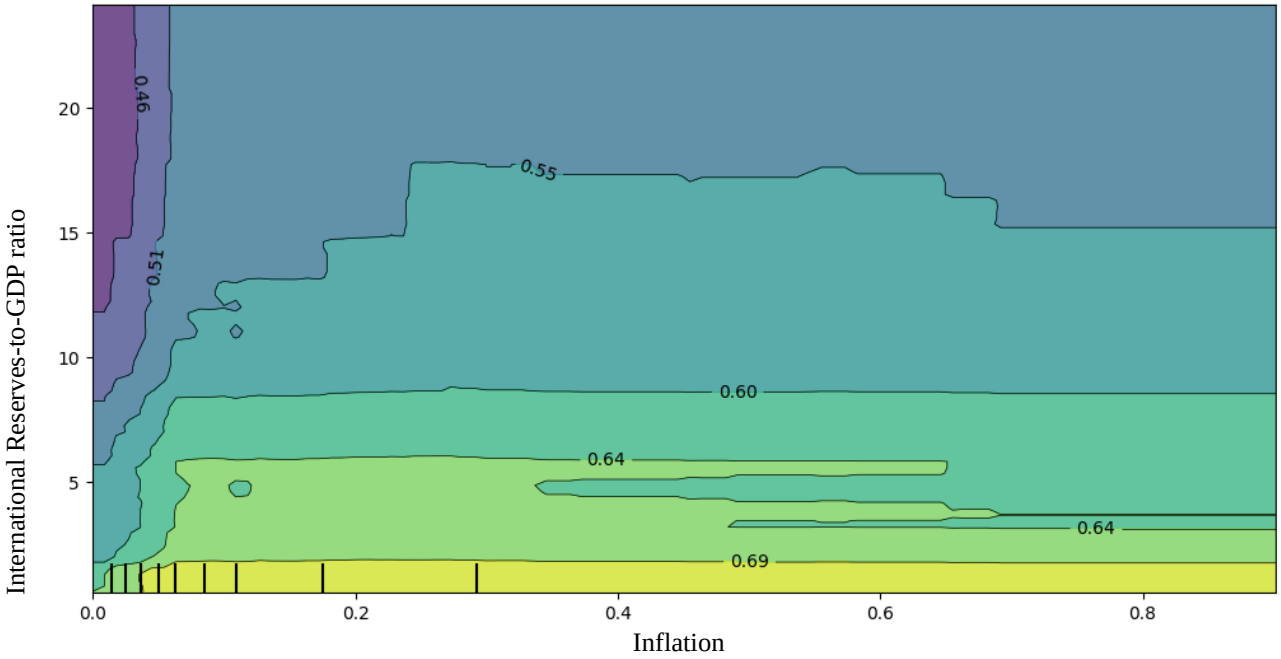

Figure 9: Partial Dependence Plot for Inflation and International Reserves-to-GDP ratio.

There are some very interesting remarks to make from Figure 9. First, as expected, a low inflation rate combined with high reserves is associated with the minimum probability of a crisis. For reserves above 17% of the GDP, any inflation rate above 5% does not affect the probability of the occurrence of a crisis; it is relatively stable at a maximum of 55%. When the reserves fall below 17%, the probability of a crisis is between 55% and 60% for almost all inflation rates. For reserves less than 8%, the probability of a crisis is between 60% and 70%. In general, an inflation rate above 5% gives rise to an increased probability of a crisis for all values of the reserves. The probability of a crisis is minimized at 46% for inflation rates less than 2%-3% and reserves above 12%.

These results highlight the important role of the reserves as a percentage of the GDP. This is somewhat expected as banking, debt and, by definition, currency crises have a significant negative impact on the exchange rate, i.e. the exchange rate depreciates. Thus, the availability of reserves provides the necessary liquidity to the central bank to support the currency in the international markets for as long as it is necessary to mitigate and absorb violent and unpredictable depreciations that may severely impact the national economy. Thus, exports and imports can be stabilized to the pre-crisis levels. Additionally, in economies where the banking sector is heavily exposed to foreign exchange risk, the inability to manage the exchange rate successfully can give rise to the probability of a banking crisis as well, as a result of a currency crisis.

On the other hand, high inflation rates can lead to instability, increasing the risk for domestic and foreign investors and trade partners. For the economies with a substantial portion of debt denominated in a foreign currency, a high inflation rate increases the real value of the debt burden in terms of the domestic currency, as the latter depreciates in response to the high inflation rate. Even in countries where the debt is mostly denominated in the domestic currency, excessive inflation rates can destabilize the economy.

Summarizing, the results point to the fact that the central bank plays a critical role in terms of economic crises. Maintaining an efficient level of reserves and implementing a prudent monetary policy that has inflation under control can significantly reduce the probability of crises.

4. Conclusion

After the global financial crisis, there has been a new wave of research that explores the significance of a stable financial system in ensuring macroeconomic stability. In response to this crisis, novel early warning models for economic crises have been developed and are currently being employed by central banks, to monitor the stability of the financial system and to guide macroprudential policy.

In this research, we aim to forecast the currency, banking and debt crises for 79 countries spanning the period from 1970 to 2017. We are using 40 macroeconomic and financial determinants to train and test multiple machine learning methodologies (KNN, SVM, Random Forest, Balanced Random Forest, Balanced Bagging Classifier, Easy Ensemble Classifier and Gradient Boosted Trees) and compare them to a simpler logistic regression model. Our findings demonstrate the effectiveness of machine learning methods in forecasting financial crises, with the Balanced Random Forest emerging as the best-performing model. Achieving an out-of-sample balanced accuracy of 72.91% and correctly predicting 149 out of 184 crises, this model outperforms traditional approaches in crisis forecasting. These results highlight the potential of ML techniques in early warning systems (EWS), offering valuable tools for policymakers and financial authorities to mitigate risks and prevent economic instability.

Furthermore, our analysis of variable importance reinforces the critical role of international reserves in reducing the likelihood of financial crises. Alongside inflation and current account

balances, reserve holdings stand out as key predictors, lending empirical support to the notion that higher reserves serve as a buffer against economic shocks. These findings align with existing literature on the protective effects of reserves and contribute to the ongoing discourse on financial stability and crisis prevention.

Acknowledgements: Emmanouil Sofianos would like to acknowledge that this work is part of the Interdisciplinary Thematic Institute MAKErS of the ITI 2021-2028 program of the University of Strasbourg, CNRS and INSERM. It has received financial support from the IdEx Unistra (ANR-10-IDEX-0002), and from the Programme Investissement d'Avenir as part of the SFRI-STRAT'US project(s) (ANR-20-SFRI-0012).

Declaration of interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

Abiad, M. A. (2003). Early warning systems: A survey and a regime-switching approach. International Monetary Fund.

Aizenman, J., Ho, S. H., Huynh, L. D. T., Saadaoui, J., & Uddin, G. S. (2024a). Real exchange rate and international reserves in the era of financial integration. Journal of International Money and Finance, 141, 103014.

Aizenman, J., & Jinjarak, Y. (2020). Hoarding for stormy days—Test of international reserves adjustment providing financial buffer stock services. Review of International Economics, 28(3), 656-675.

Aizenman, J., Park, D., Qureshi, I. A., Saadaoui, J., & Uddin, G. S. (2024b). The performance of EMEs during the Fed's easing and tightening cycles: a cross-country resilience analysis. Journal of International Money and Finance, 148, 103169.

Aizenman, J., & Riera-Crichton, D. (2008). Real exchange rate and international reserves in an era of growing financial and trade integration. The Review of Economics and Statistics, 90(4), 812-815.

Ahmed, R., Aizenman, J., Saadaoui, J., & Uddin, G. S. (2023). On the effectiveness of foreign exchange reserves during the 2021-22 US monetary tightening cycle. Economics Letters, 233, 111367.

Aydın, S., & Tunç, C. (2024). A farewell to Greenspan-Guidotti rule: has the role of short-term debt as a reserve adequacy indicator disappeared? Applied Economics Letters, forthcoming.

Benigno, G., Fornaro, L., & Wolf, M. (2022). Reserve accumulation, growth and financial crises. Journal of International Economics, 139, 103660.

Beutel, J., List, S. & von Schweinitz, G. (2019). Does machine learning help us predict banking crises? Journal of Financial Stability, 45, p. 100693.

Breiman, L. (1996). Bagging predictors. Machine Learning, 24(2), pp. 123– 140.

Breiman, L. (2001) Random forests. Machine Learning, 45 pp.5-32

Chen, A. Liaw (2004). Using Random Forest to Learn Imbalanced Data. University of California at Berkeley, Berkeley, California.

Chen, M. et al. (2023). Identifying financial crises using machine learning on Textual Data. Journal of Risk and Financial Management, 16(3), p. 161.

Cortes, C., Vapnik, V. (1995). Support-vector networks. Machine Learning, 20, pp. 273–297.

Coulibaly, I., Gnimassoun, B., Mighri, H., & Saadaoui, J. (2024). International reserves, currency depreciation and public debt: New evidence of buffer effects in Africa. Emerging Markets Review, 60, 101130.

Dominguez, K. M., Hashimoto, Y., & Ito, T. (2012). International reserves and the global financial crisis. Journal of International Economics, 88(2), 388-406.

Drummond, C. & Holte, R. (2003). C4.5, class imbalance, and cost sensitivity: why undersampling beats over-sampling. In Proceedings of the ICML-2003 Workshop: Learning with Imbalanced Data Sets II

Frankel, J. and Saravelos, G. (2012). Can leading indicators assess country vulnerability? evidence from the 2008–09 global financial crisis. Journal of International Economics, 87(2), pp. 216–231.

Frankel, J., & Saravelos, G. (2012). Can leading indicators assess country vulnerability? Evidence from the 2008–09 global financial crisis. Journal of International Economics, 87(2), pp. 216-231.

Freund, Y., & Schapire, R.E. (1996). Experiments with a New Boosting Algorithm. International Conference on Machine Learning.

Friedman, Jerome H. (2001). Greedy function approximation: A gradient boosting machine. Annals of statistics, 29(5), pp. 1189-1232.

Gogas, P., & Papadimitriou, T. (2021). Machine learning in economics and finance. Computational Economics, 57, pp. 1-4.

Gogas, P., Papadimitriou, T. and Sofianos, E. (2019). Money neutrality, monetary aggregates and machine learning, Algorithms, 12(7), p. 137.

Gogas, P., Papadimitriou, T. and Sofianos, E. (2022). Forecasting unemployment in the Euro area with machine learning. Journal of Forecasting, 41(3), pp. 551–566.

Guyon, I., Weston, J., Barnhill, S., & Vapnik, V. (2002). Gene selection for cancer classification using support vector machines. Machine Learning, 46(1-3), pp. 389-422.

Hasanin, T., & Khoshgoftaar, T.M. (2018). The Effects of Random Undersampling with Simulated Class Imbalance for Big Data. 2018 IEEE International Conference on Information Reuse and Integration (IRI), pp. 70-79.

Hernandez, J. (2018). How international reserves reduce the probability of debt crises. Inter-American Development Bank.

Hawkins, J., & Klau, M. (2000). Measuring potential vulnerabilities in emerging market economies. BIS Working Papers 91, Bank for International Settlements.

Kauko, K. (2014) How to foresee banking crises? A survey of the empirical literature. Economic Systems, 38(3), pp. 289–308.

Kim, J. (2009). Estimating classification error rate: Repeated cross-validation, repeated hold-out and bootstrap. Computational Statistics and Data Analysis, 53, pp. 3735-3745.

Kaminsky, G., Lizondo, S., & Reinhart, C. M. (1998). Leading indicators of currency crises. Staff Papers, 45(1), 1-48.

Ling, C. & Li, C. (1998). Data mining for direct marketing problems and solutions. In Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining, New York

Liu, L., Chen, C. and Wang, B. (2022). Predicting financial crises with machine learning methods. Journal of Forecasting, 41(5), pp. 871–910.

Liu, X., Wu, J., & Zhou, Z. (2009). Exploratory Undersampling for Class-Imbalance Learning. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), 39, 539-550.

Obstfeld, M., Shambaugh, J. C., & Taylor, A. M. (2009). Financial instability, reserves, and central bank swap lines in the panic of 2008. American economic review, 99(2), 480-486.

Obstfeld, M., Shambaugh, J. C., & Taylor, A. M. (2010). Financial stability, the trilemma, and international reserves. American Economic Journal: Macroeconomics, 2(2), 57-94.

Rose, A. K., & Spiegel, M. M. (2009). Predicting Crises, Part II: Did Anything Matter (to Everybody)? FRBSF Economic Letter, 30.

Rose, A. K., & Spiegel, M. M. (2010). Cross-country causes and consequences of the 2008 crisis: International linkages and American exposure. Pacific Economic Review, 15(3), 340-363.

Reimann, C. (2024) ‘Predicting financial crises: An evaluation of machine learning algorithms and model explainability for early warning systems’, Review of Evolutionary Political Economy, 5(1), pp. 51–83.

Samitas, A., Kampouris, E. and Kenourgios, D. (2020). Machine learning as an early warning system to predict financial crisis. International Review of Financial Analysis, 71, p. 101507.

Appendix

Table A1: countries involved in this study

Country	
Algeria	Myanmar
Argentina	Nepal
Australia	New Zealand
Austria	Nicaragua
Belgium	Nigeria
Bolivia	Norway
Brazil	Pakistan
Burundi	Panama
Cameroon	Paraguay
Canada	Peru
Central African Rep.	Philippines
Chile	Portugal
Colombia	Rwanda
Congo, Republic of	Saudi Arabia
Costa Rica	Sierra Leone
Cyprus	Singapore
Denmark	South Africa
Dominican Republic	Spain
Ecuador	Sri Lanka
Egypt	Sweden
El Salvador	Syrian Arab Republic
Ethiopia	Tanzania
Finland	Thailand
France	Tunisia
Gabon	Uganda
Germany	United Kingdom
Ghana	United States
Greece	Uruguay
Guatemala	Venezuela, Rep. Bol.
Guyana	Zambia
Honduras	
Hong Kong	
Iceland	
India	
Indonesia	
Ireland	
Israel	
Italy	
Jamaica	
Japan	
Jordan	
Kenya	
Korea	
Kuwait	
Madagascar	
Malaysia	
Mauritania	
Mexico	
Morocco	

Table A2: Results for all models

#	Model	OOS Balanced Accuracy	In-sample Balanced Accuracy	True Positive Rate OOS
1	Balanced Random Forest	72.91%	66.14%	81.00%
2	Balanced Random Forest under sampling proportions 1:3	70.33%	67.00%	73.00%
3	Random Forest balanced weights under sampling proportions 1:1	70.24%	67.73%	76.00%
4	Random Forest regular weights under sampling proportions 1:1	70.16%	67.15%	76.00%
5	Balanced Random Forest under sampling proportions 1:1	69.96%	66.43%	77.00%
6	Balanced Random Forest under sampling proportions 1:2	69.75%	66.20%	75.00%
7	Gradient Boosted Trees	69.60%	66.49%	76.00%
8	Easy Ensemble – AdaBoost	67.81%	63.20%	75.00%
9	SVM (RBF) balanced weights	66.15%	63.73%	72.00%
10	Logistic Regression balanced weights under sampling proportions 1:2	65.53%	63.78%	74.00%
11	SVM (RBF) regular under sampling proportions 1:1	65.35%	64.35%	68.00%
12	SVM (RBF) balanced weights under sampling proportions 1:1	65.35%	64.25%	68.00%
13	Balanced Bagging Logistic Regression	65.31%	63.51%	65.00%
14	Logistic Regression balanced weights under sampling proportions 1:1	65.20%	66.49%	71.00%
15	Logistic Regression regular weights under sampling proportions 1:1	65.20%	66.59%	71.00%
16	Logistic Regression balanced weights	64.94%	62.07%	67.00%
17	SVM (RBF) balanced weights under sampling proportions 1:3	64.91%	63.41%	65.00%
18	SVM (RBF) balanced weights under sampling proportions 1:2	64.90%	62.81%	64.00%
19	KNN under sampling proportions 1:1	64.70%	64.81%	74.00%
20	Logistic Regression balanced weights under sampling proportions 1:3	63.97%	64.29%	68.00%
21	SVM (RBF) regular weights under sampling proportions 1:2	62.02%	62.12%	43.00%
22	Logistic Regression regular weights under sampling proportions 1:2	61.97%	63.23%	43.00%
23	Random Forest balanced weights under sampling proportions 1:2	61.37%	61.30%	41.00%
24	Random Forest regular weights under sampling proportions 1:2	60.90%	60.52%	36.00%
25	SVM (Polynomial) regular weights under sampling proportions 1:1	60.76%	59.24%	56.00%
26	SVM (Polynomial) balanced weights under sampling proportions 1:1	60.76%	59.54%	56.00%

#	Model	OOS Balanced Accuracy	In-sample Balanced Accuracy	True Positive Rate OOS
27	SVM (Linear) regular weights under sampling proportions 1:1	60.66%	59.12%	54.00%
28	SVM (Linear) balanced weights under sampling proportions 1:1	60.66%	59.24%	54.00%
29	SVM (Polynomial) balanced weights under sampling proportions 1:2	59.89%	59.75%	48.00%
30	SVM (Polynomial) regular weights under sampling proportions 1:2	59.89%	59.75%	48.00%
31	SVM (Linear) regular weights under sampling proportions 1:2	59.78%	59.46%	48.00%
32	SVM (Linear) balanced weights under sampling proportions 1:2	59.65%	59.66%	48.00%
33	SVM (RBF) regular under sampling proportions 1:3	59.77%	59.10%	33.00%
34	Random Forest balanced weights under sampling proportions 1:3	59.69%	59.89%	31.00%
35	Random Forest regular weights under sampling proportions 1:3	57.50%	57.04%	22.00%
36	SVM (Polynomial) balanced weights under sampling proportions 1:3	56.98%	56.85%	40.00%
37	SVM (Polynomial) regular weights under sampling proportions 1:3	56.66%	54.65%	38.00%
38	SVM (Linear) balanced weights under sampling proportions 1:3	56.66%	56.85%	38.00%
39	SVM (Linear) regular weights under sampling proportions 1:3	56.66%	56.85%	38.00%
40	SVM (Polynomial) balanced weights	56.36%	54.56%	24.00%
41	SVM (Linear) balanced weights	56.36%	54.56%	24.00%
42	Logistic Regression regular weights under sampling proportions 1:3	55.29%	56.40%	17.00%
43	Random Forest balanced weights	54.88%	51.78%	14.00%
44	SVM (RBF) regular weights	52.57%	51.84%	09.00%
45	SVM (Polynomial) regular weights	50.80%	50.67%	04.00%
46	SVM (Linear) regular weights	50.60%	50.46%	04.00%
47	Random Forest regular weights	50.38%	49.99%	01.00%
48	KNN	50.00%	50.00%	0%
49	KNN under sampling proportions 1:3	50.00%	50.00%	0%
50	Logistic Regression regular weights	50.00%	50.00%	0%
51	KNN under sampling proportions 1:2	49.97%	50.00%	0%
52	Balanced Bagging KNN	49.34%	54.05%	49.00%