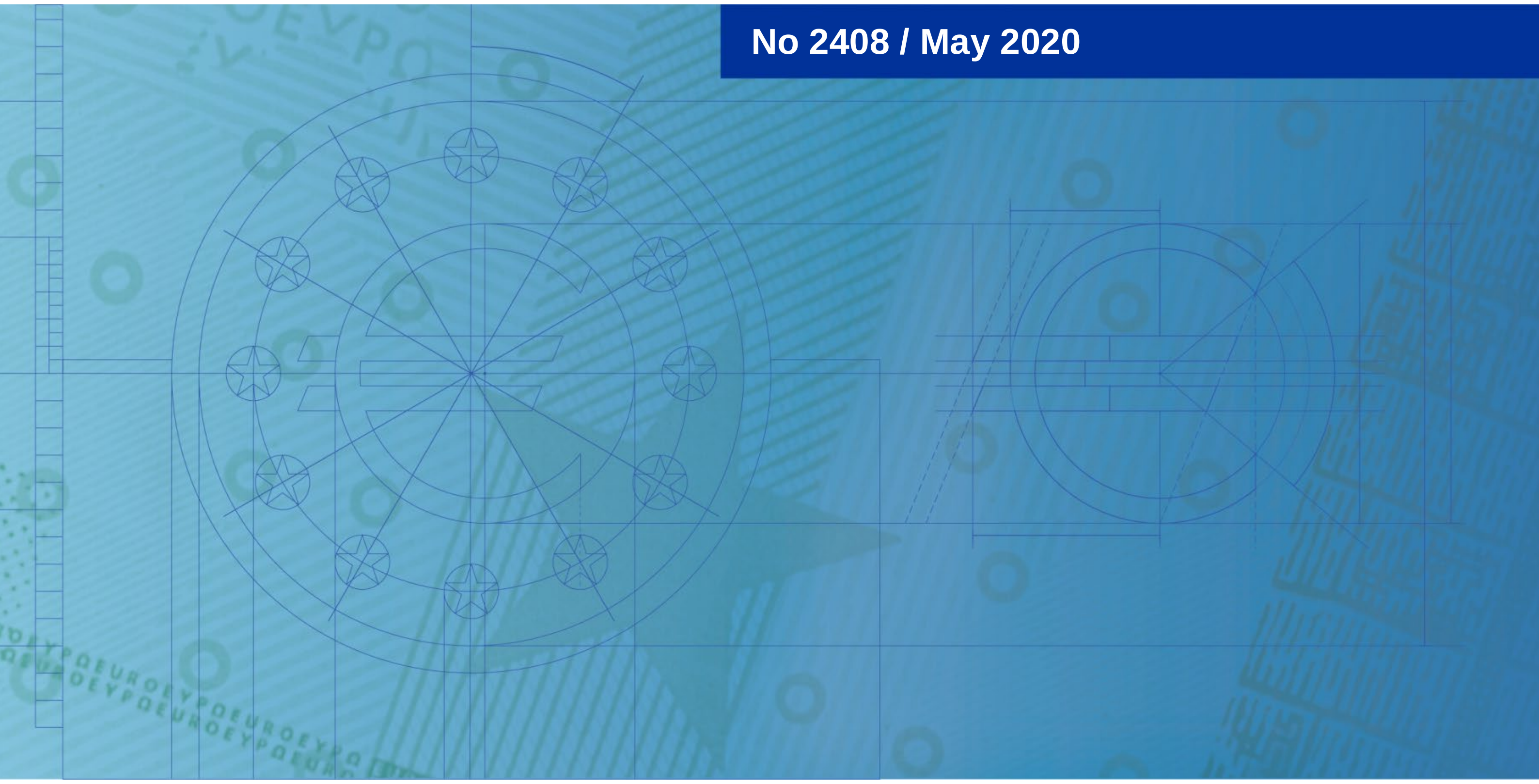
EUROPEAN CENTRAL BANK
EUROSYSTEM

Working Paper Series

Barbara Jarmulska

Random forest versus logit models: which offers better early warning of fiscal stress?

No 2408 / May 2020



Disclaimer: This paper should not be reported as representing the views of the European Central Bank (ECB). The views expressed are those of the authors and do not necessarily reflect those of the ECB.

Abstract

This study seeks to answer whether it is possible to design an early warning system framework that can signal the risk of fiscal stress in the near future, and what shape such a system should take. To do so, multiple models based on econometric logit and the random forest models are designed and compared. Using a dataset of 20 annual frequency variables pertaining to 43 advanced and emerging countries during 1992-2018, the results confirm the possibility of obtaining an effective model, which correctly predicts 70-80% of fiscal stress events and tranquil periods. The random forest-based early warning model outperforms logit models. While the random forest model is commonly understood to provide lower interpretability than logit models do, this study employs tools that can be used to provide useful information for understanding what is behind the black-box. These tools can provide information on the most important explanatory variables and on the shape of the relationship between these variables and the outcome classification. Thus, the study contributes to the discussion on the usefulness of machine learning methods in economics.

JEL Classification: C40, C53, H63, G01

Keywords: early warning system; interpretability of machine learning; predictive performance

Non-technical summary

Early warning models, also known as early warning systems, aim to identify possible processes in an economy that may indicate the build-up of vulnerabilities, with the rationale for using them being the observation that certain variables behave differently in periods preceding a crisis than in other periods. The aim of using early warning models is not to predict a crisis, but rather to signal increased risk of its occurrence in the near term. The justification for using these models is that early identification of pre-stress periods grants more time to implement appropriate measures, which is especially crucial in a globalized world.

The goal of the study is to design an effective early warning model signalling the risk of a fiscal stress event in the near future and simultaneously to shed light on the usefulness of the methods that can be employed to construct such a model. Specifically, I am interested in comparing the prediction accuracy and interpretability of models based on random forest and logistic regression.

The contributions of this work are as follows. First, to the best of my knowledge, there is a lack of a comprehensive research project comparing standard methods (discrete dependent variable models) with a more modern approach based on machine learning methods applied to predict a fiscal stress event. Second, in my opinion, methods stemming from machine learning have not been given sufficient attention to date in economics, even though they could offer great benefits. Against this backdrop, this study makes a relevant contribution to the literature by providing the discussion on the usefulness of machine learning tools in economics, by focusing on a comparison of random forest and logit models in the context of early warning of fiscal stress.

I selected random forest as my method of interest. Its advantages relate to its relative simplicity, as it requires neither many choices on multiple parameters, nor a large database, while at the same time it belongs to the class of robust and effective machine learning tools. Of course, random forest is not the only possible choice that could contribute to the discussion on the usefulness of machine learning approaches in economics; other tools may be equally suitable to this end (e.g., gradient boosting machines, neural networks, and support vector machines).

The effectiveness of random forests proved to be higher than that obtained by logit models, offering an average prediction accuracy of the fiscal stress events and the tranquil events of slightly below 80%, in contrast to 70-75% of logits. Signalling many of fiscal stress episodes related to the sovereign debt crisis in euro area would have been possible, had an early warning model based on random forest been implemented in the past. When focusing on the first year of the stress event only, the prediction accuracy dropped to 64-73% for random forests and 68-72% for logits. While this clearly shows that prediction of the first year of the stress event is more difficult than prediction of ongoing stress, early warning models proposed still offer the prediction accuracy that could be considered useful. This conclusion is underpinned by the underlying reason to use early warning systems, given that the outcome signal should be interpreted as a warning of heightened level of vulnerabilities, and not as a forecast of a crisis. Therefore, it is worth striving to build even an imperfect tool.

The random forest is understood to offer lower interpretability of results than the logit models it outperforms, which represents a relevant limitation for economists. Some of the especially useful features of econometric models are not available when using the random forest; however, alternative sources of similar information are available. Variable importance measures and game theory-derived Shapley values can help to assess which predictors are especially useful for the classification problem, and as such, provide information partially akin to the significance of explanatory variables. Furthermore, partial dependence and accumulated local effects plots of the random forest aid understanding of the impact of a given variable on the outcome obtained, providing information analogous to the coefficients estimated by an econometric model.

1 Introduction

Early warning models, also known as early warning systems, aim to identify possible processes in an economy that may indicate the build-up of vulnerabilities, with the rationale for using them being the observation that certain variables behave differently in periods preceding a crisis than in other periods (Reinhart and Rogoff (2008)). The aim of using early warning models is not to predict a crisis, but rather to signal increased risk of its occurrence in the near term. The justification for using these models is that early identification of pre-stress periods grants more time to implement appropriate measures, which is especially crucial in a globalized world (Hernández de Cos et al. (2014)).

The goal of the study is to design an effective early warning model signalling the risk of a fiscal stress event in the near future and simultaneously to shed light on the usefulness of the methods that can be employed to construct such a model. Specifically, I am interested in comparing the prediction accuracy and interpretability of models based on random forest and logit regression.

The contributions of this work are as follows. First, to the best of my knowledge, there is a lack of a comprehensive research project comparing standard methods (discrete dependent variable models) with a more modern approach based on machine learning methods applied to predict a fiscal stress event. Second, in my opinion, methods stemming from machine learning have not been given sufficient attention to date in economics, even though, as argued by Athey (2019), they could offer great benefits. Against this backdrop, this study makes a relevant contribution to the literature by providing the discussion on the usefulness of machine learning tools in economics, by focusing on a comparison of random forest and logit models in the context of early warning of fiscal stress.

I selected random forest as my method of interest. Its advantages relate to its relative simplicity, as it requires neither many choices on multiple parameters, nor a large database, while at the same time it belongs to the class of robust and effective machine learning tools (Caruana and Niculescu-Mizil (2006)). Of course, random forest is not the only possible choice that could contribute to the discussion on the usefulness of machine learning approaches in economics; other tools may be equally suitable to this end (e.g., gradient boosting machines, neural networks, and support vector machines).

It has been established in the literature that early warning models aimed at signalling fiscal stress work best when they are based on a possibly broad set of variables reflecting the state of the whole economy, and should not be restricted to fiscal variables (Berti et al. (2012); Bruns and Poghosyan (2016); Ciarlone and Trebeschi (2005); Manasse et al. (2003)). Thus, this study uses a broad set of 20 variables. They can be grouped into categories related to labour and financial markets, competitiveness, indebtedness, global environment, as well as institutional, fiscal, and macroeconomic variables. The analysis is performed using a database for the years 1992-2018 for 43 countries, defined by the IMF as 19 emerging countries and 24 advanced countries.

In the empirical analysis, models based on the two employed methods of interest are compared, and their effectiveness in signalling fiscal stress events is checked. The early warning models based on the random forest turn out to be more effective than the logit models. Random forest-based models outperform all versions of logit models, which offer slightly lower prediction accuracy. On the one hand, logit models can be interpreted in the standard econometric way, based on the significance of the variables and the coefficients estimated (i.e., marginal effects), while random forest-based models cannot. On the other hand, machine learning tools, including random forest, can be made more interpretable through the use of tools that provide information partially akin to that of the significance of variables and the estimated coefficients. The empirical analysis provides an example of Breiman's variable importance and Shapley values to provide information on which explanatory variables contributed the most to the outcome obtained, and partial dependence and accumulated local effects plots to check the shape of the relationship between the explanatory variables and the outcome classification, thereby enabling interpretation of the black-box.

The rest of the paper is structured as follows. Section 2 reviews the literature on early warning models. Section 3 explains the methodology, and Section 4 presents the data used. Section 5 summarizes the empirical findings and Section 6 concludes.

2 Literature review

Most of the research on early warning systems is concentrated on signalling currency or banking crises, with relatively few works aimed at signalling fiscal stress periods. In the existing works, standard methods used are a non-parametric signalling approach and discrete dependent variable models, such as logistic regression. The signalling approach is a simple, non-parametric method aimed at optimizing thresholds, above (or below) which each of the variables sends a warning signal. The method has been popularized by a seminal paper on currency crises by Kaminsky et al. (1998) and by subsequent papers on the co-occurrence of currency and banking crises by Kaminsky (1999) and Kaminsky and Reinhart (1999).

Discrete dependent variable models have been used by many researchers and have become a standard tool in the early warning literature, also when applied to the fiscal stress event or debt crisis example. Barrell et al. (2010) used logit models to signal systemic banking crises. Demirgüç-Kunt and Detragiache (2005) used both the signalling approach and discrete dependent variable models to predict banking crises and to analyse their determinants. Kumar et al. (2003) used logit models to signal currency crises, using financial and macroeconomic variables. Manasse et al. (2003) used both logit models and decision trees to forecast debt crises based on a broad set of variables on solvency and liquidity of countries, external factors, and macroeconomic and political conditions. Ciarlone and Trebeschi (2005) designed an early warning model signalling debt crises using a multinomial logit model. Fuertes and Kalotychou (2006, 2007) analysed early warning systems based on logit models and the k-means clustering algorithm. Bussiere and Fratzscher (2006) also used logit models to predict financial crises in developing countries. Lo Duca and Peltonen (2013), using discrete dependent variable models aimed at signalling systemic financial crises, showed that vulnerabilities build up non-linearly. Bruns and Poghosyan (2016) used a logit model aggregation, called extreme bound analysis, to signal fiscal stress events.

Discrete dependent variable models have limitations, including an assumption about the functional form of the relationship between the left-hand side and explanatory variables. Owing to progress in quantitative methods, gradual improvement in data access, and the practical need for effective early warning models, new approaches based on non-parametric, flexible methods inspired by machine learning are becoming more common in the literature. These methods enable consideration of both non-linearity between the dependent and explanatory variables and cross-dependency between the variables. However, research based on these methods, including classification and regression trees (CART), is still relatively less common in the literature than analyses based on the signalling approach or logit models.

Davis and Karim (2008) analysed whether it would have been possible to forecast the sub-prime crisis, which started in 2007, using early warning systems based on logit models and decision trees. Sarlin (2012) used biological sciences-inspired neural networks to design an early warning model aimed at signalling financial crises, and the model outperformed standard statistical methods. Duttagupta and Cashin (2008) used binary decision trees to analyse banking crises and identified variables crucial for the prediction. Fioramanti (2008) showed that early warning systems based on neural networks could yield better results than when standard parametric methods were used. Manasse and Roubini (2009), using decision trees, derived a collection of ”rules of thumb” that help identify the typical characteristics of countries facing a sovereign debt crisis. Demyany and Hasan (2009) comprehensively compared methods used in the early warning literature, using the sub-prime crisis example. Holopainen and Sarlin (2017) conducted a comprehensive comparison of 12 methods that could be used to design early warning models, and compared the predictive accuracy of models aimed at forecasting banking crises in 15 European countries during 1973-2014. The authors concluded that more sophisticated methods, based on machine learning techniques, like k-means clustering and neural networks, were more effective than standard statistical approaches. Alessi and Detken (2018) proposed a random forest-based early warning model to identify excessive credit growth and aggregate leverage.

This study places itself at the crossing of two literature streams: first, on early warning models, and second, on the usefulness of machine learning tools in economics. My method of interest is the random forest, a robust and fairly simple approach, extensively used in other disciplines, but, in my view, relatively underutilised in economic analyses. While the random forest-based approach has been already used in numerous studies on early warning models, this study makes the novel contribution to the literature by applying this approach to fiscal stress events, and by discussing the usefulness of this approach in comparison to logit models in terms of prediction accuracy and interpretability of results.

3 Methodology

3.1 Metrics Used to Assess the Effectiveness of Early Warning Models

The study employs standard measures used in the literature to assess the effectiveness of early warning models, namely, sensitivity, specificity, their average, and the area under receiver operating curve (AUROC). Sensitivity is the relationship between the number of true positives and all stress periods, which is a proportion of actual positives correctly classified. Specificity is the relationship between all true negatives and all tranquil periods, that is, the proportion of tranquil periods correctly classified.

Specificity and sensitivity depend on the threshold chosen to distinguish the observations that are classified as stress events from the observations classified as tranquil events. In other words, the exact cut-off point determines the relative count of type I and type II errors. The measure that does not depend on the specific threshold chosen is the AUROC, reflecting the area under the receiver operating curve (ROC). The ROC shows the relationship between false positives (i.e., 1 minus specificity) and sensitivity for all possible thresholds. Purely random division of the sample into categories results in an AUROC of 0.5 on average, and thus, any AUROC above that value means that the model has value added over a random assignment of classes to observations.

3.2 Logit Models

The first approach used is a discrete dependent variable model, logit, of which two versions are utilised: standard logit model and a logit model with a least absolute shrinkage and selection operator (LASSO) penalisation. LASSO penalisation is one of the methods that performs simultaneous variable selection and regularisation in order to enhance the prediction accuracy and interpretability of the model (Tibshirani (1996)). The difference between the logit with the LASSO penalisation and regular logit is the restriction imposed on estimated coefficients. If the LASSO penalisation is included, the sum of absolute values of estimated coefficients cannot exceed the pre-specified free parameter, t, which effectively determines the amount of penalisation imposed. The LASSO estimates are defined by (Hastie et al. (2012)):

$$\hat{\beta}^{lasso} = \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^{N} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2$$

subject to

$$\sum_{j=1}^{p} |\beta_j| \leq t. \tag{1}$$

Alternatively, the LASSO problem in the Lagrangian form is given as follows:

$$\hat{\beta}^{lasso} = \underset{\beta}{\operatorname{argmin}} \left\{ \frac{1}{2} \sum_{i=1}^{N} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\}. \tag{2}$$

where λ is a penalisation parameter. As follows from equations (1) and (2), there is a one-on-one correspondence between the parameters t and λ. The higher the penalisation (i.e. the higher λ or, equivalently, lower t), the higher the number of variables whose coefficients are shrunk to zero, effectively eliminating them from the model. However, the information content of the variables retained in the model also decreases. λ (or t) should typically be selected in such a way that the resulting model would minimise the out-of-sample error.

In practice, the penalisation parameter is usually obtained empirically using a cross-validation technique (Holopainen and Sarlin (2017)). In the first step of the cross-validation the sample is divided into k subgroups. Next, all of the subgroups with an exception of one are used as a train set (on which the model will be estimated), and the remaining subgroup is used as a test set, utilised to check the out-of-sample performance of the model on unseen data. In this step, a LASSO logit model is estimated for each of the penalisation parameters λ from the chosen interval. Subsequently, a selected metric of a model performance, for example AUROC, is computed using the model estimated using the train set, and the data from the test set. This step is performed k times (the number of train-test sets following from the division of the sample into k subgroups) for each of the penalisation parameters λ from the chosen interval. Hence, also AUROC for each of the penalisation parameters λ from the chosen interval is obtained k times. In the last step the average measures of AUROC obtained for all the penalisation parameters are compared and this λ which is associated with the best AUROC is chosen as a preferred penalisation parameter. In summary, cross-validation, which the present study employs, yields a penalisation parameter that maximises out-of-sample prediction accuracy of the model and limits the overfitting problem (Chetverikov et al. (2016)). Therefore in practice, by shrinking particular parameters to zero, LASSO logit effectively chooses the variables that should be retained in the model.

3.3 Random Forest

The second method used is CART for binary classification and their ensemble into random forests, which is a non-parametric, supervised learning approach stemming from machine learning. The CART method yields new knowledge by building a model that reflects dependencies found directly in the database, with the goal of minimising the loss function. This is achieved based on a set of examples composed of explanatory variables X and the set of corresponding outcomes y (Breiman (2001)).

The CART algorithm is based on a recursive partition, when each non-terminal node is divided into two child nodes. The construction of a tree is based on a set of questions with binary responses: yes or no. The answers determine the questions that follow in consecutive nodes until the terminal node, where the observation is assigned an outcome category. The expected prediction error is:

$$Err(\phi) = E_{X,y}\{L(y, \phi(X))\} \tag{3}$$

where L is the loss function.

Each terminal node w can be understood as a little model defined locally on $X_i \times y$ which assigns the same class $\hat{y}_w$ to all the observations within this node. Hence, for each tree ϕ the minimisation of the overall classification error is strictly equivalent to the minimisation of the error at the level of the terminal nodes. Hence, fitting the best CART implies choosing the best classification $\hat{y}_w$ assigned to each of the terminal nodes w. Error at the level of each node is minimised by forecasting a more likely class.

In practice, the trees are fitted recursively. The sample is split into subsamples in each of the non-terminal nodes, and all the observations in the terminal node are assigned one class. The algorithm chooses the best possible combination of a variable and a threshold that would allow minimising the loss function. The loss function is an increasing function of a number of observations classified incorrectly, and thus, its minimisation results in the best possible prediction accuracy. The Gini index, used as a loss function in CART, is one possible impurity criteria representing the precision of the fit:

$$g(w) = \sum_{k \neq j} p_{wk} p_{wj} = \sum_{k} p_{wk}(1 - p_{wk}) \tag{4}$$

where p_{wj} is the probability distribution of class j in node w.

CART is a non-parametric technique, allowing detection of dependencies between many variables. Its advantage is that single trees are interpretable. Out-of-sample prediction obtained by CART has low bias, but high variance, resulting in low prediction accuracy. To address this problem, trees ensemble processes are applied, as they decrease the variance significantly; however, the cost is an increased bias (Louppe (2014)). The logic of ensemble processes is that they use multiple, decorrelated trees, instead of fitting one tree and using it for out-of-sample classification. Each of the observations, being an entry vector, is assigned a class by each of the trees, and the outcome of the operation of the ensemble process is the numerical count of the classes obtained (voting). Final binary classification requires choosing a threshold, depending on the researcher's or decision-maker's preference.

One of the most popular ensemble methods is the random forest, which this study uses. The trees are randomised in two ways. First, each tree is fitted on the permutated sample of observations. Second, when, in each node of the tree, the algorithm divides the sample based on the best variable and the threshold, it can pick that variable only from the limited set of m explanatory variables randomly drawn in each node.

The forecast error of the random forest depends on two factors. One is the correlation between the trees - the higher the correlation, the higher the prediction error. The second factor is the strength of the single tree in the forest - a tree with low error is said to be strong.

The higher the parameter m, the stronger the trees of the random forest, because they are fit using the best variable in each node chosen from a larger set of explanatory variables. However, increasing parameter m also leads to increased correlation between the trees. Hence, parameter m is the key decision that impacts the outcome random forest. The number of trees in the random forest influences the results to a lesser extent, since increasing the number of trees results in the classification error converging to the value dependent on the characteristics of single trees. This explains why increasing the number of trees in the forest does not lead to overfitting (Breiman (2001)).

Random forest can be used for complex databases, in which explanatory variables affect not only the dependent variable, but also each other. The great advantage of this method is that it is one of the most robust and effective machine learning models; it does not require imposing assumptions on the shape of the relationship between dependent and explanatory variables (Breiman (2001)). However, it is considered to be less interpretable than econometric models, similarly to other black-box machine learning tools.

4 Data

4.1 Fiscal Stress Event Variable

There is no unique quantitative variable for fiscal stress. The difficulty with choosing the exact definition is that fiscal stress is not directly observable - it can be proxied only by using differing criteria. The definition of the explained event has a paramount effect on the characteristics of the final model obtained, and hence, it is crucial to choose the one that reflects the goals and preferences of the researcher to the extent possible.

In line with the convention adopted in the literature, the dependent variable is a binary variable equal to 1 in the case of a fiscal stress event and 0 otherwise. In more recent literature in this field, the dependent variable tends to be defined broadly, reflecting not only outright default or debt restructuring, but also less extreme events. Therefore, following Baldacci et al. (2011), the definition used in the present study is broad, and the focus is on signalling fiscal stress events, in contrast to a narrower event of a fiscal crisis related to outright default or debt restructuring. Fiscal problems can take many forms; in particular, some of the outright defaults can be avoided through timely, targeted responses, like support programs of international institutions.

According to the chosen definition, a fiscal stress event is ongoing when at least one of the following conditions is met:

- Debt default or restructuring is present, computed based on three data sources (Beers and Mavalwalla (2017); Cruces and Trebesch (2013); Trebesch et al. (2012));

- A large support program financed by an international institution is assumed to reflect fiscal problems of a given country, since these programs tend to be a way to avoid risk materialisation in the form of a strong fiscal stress;

- Hyperinflation is defined as exceeding 35% in advanced countries and 500% in emerging countries, where the thresholds are based on Reinhart and Rogoff (2010) and Baldacci et al. (2011);
- Deterioration in access to financial markets is aimed at identifying a situation in which governments are faced with substantial worsening of investor sentiment, defined as a period in which the spreads of a country's bond yield to a benchmark bond yield increase considerably. Following the literature (Knedlik and von Schweinitz (2012)), the spread has to exceed the average spread by more than two standard deviations.

Some studies in the literature argue that periods of ongoing stress except for the first year should be excluded, as these atypical observations may bias the results and cause the prediction power of the early warning model to deteriorate (Bussiere and Fratzscher (2006)). However, other studies show that models based on the whole sample of available observations exhibit higher prediction accuracy than those based only on the subset of the observations (Fuertes and Kalotychou (2006)). Moreover, in the case of analyses based on macroeconomic variables, the size of the database and the number of stress events are important considerations, and researchers have to be careful about disregarding certain available information. In the present study, among 987 observations, 163 (nearly 17%) were stress periods. If the analysis excludes observations of ongoing stress, the share of stress periods in the database drops from 17% to slightly over 4%, or 38 observations. Therefore, in the present study, models are fitted on the whole sample, but the results are presented in two versions: signalling for all the observations; and signalling when periods of ongoing stress are excluded, with the focus on prediction accuracy of the first years of the stress events.

Stress periods are not equally distributed across time and country groups (Figure 1 and A1 in the Appendix). The majority of them, 123, were in developing countries (29% of observations for developing countries). Among observations for advanced countries only 7% were stress periods. In particular, most of the stress periods in advanced countries are related to the fiscal stress in the euro area which started around 2010-2011. It is rare for early warning literature to focus on advanced countries exclusively, but such analyses have been performed - see, for example, Hernández de Cos et al. (2014).

4.2 Explanatory Variables

Given data availability, the database used includes annual frequency data for 43 countries, defined by the IMF as 24 advanced countries and 19 emerging countries, for the years 1992-2018. The database contains 20 explanatory variables that can be classified as follows: macroeconomic and global economy; financial, including private indebtedness; fiscal; competitiveness and domestic demand; and labor market. Macroeconomic variables and those related to global markets are important owing to a possible contagion between countries. The relevant channels here are trade, financial market integration, and dependencies between financial institutions. Variables

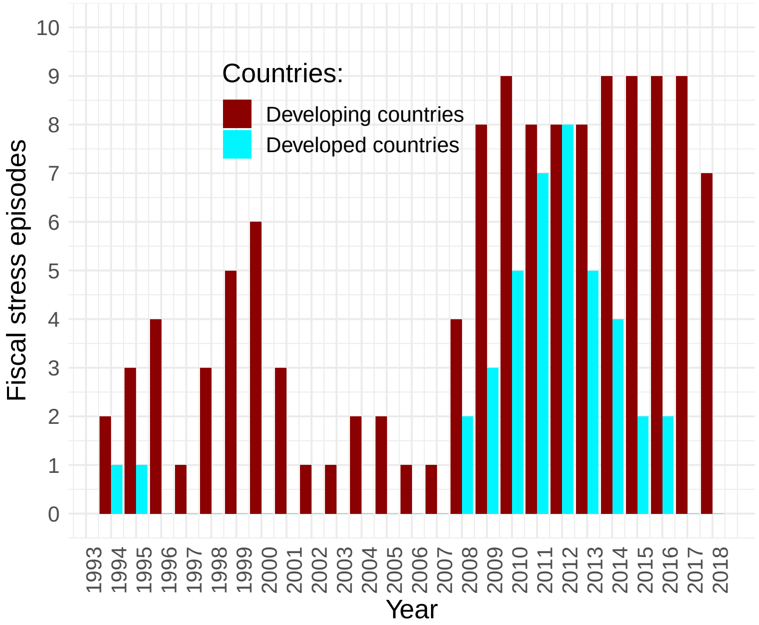


Figure 1: Stress Periods in Analysed Countries

Note: Purple bars represent the number of stress periods in emerging countries, and blue bars represent those in advanced countries.

related to competitiveness and domestic demand can be indicators of fiscal stress, as a decrease in competitiveness may result in vulnerabilities building up in many sectors, which in turn can result in deterioration of the public sector financial position. Financial variables can indicate a build-up of vulnerabilities mainly related to the credit growth channel, as excessive credit dynamics tend to precede the outbreak of crises (Schularick and Taylor (2012)), and the problems of the financial sector can spill over to the public sector through the need to support banks. Variables related to the labour market can be associated with the flexibility of the economy in reacting to shocks, and can impact public finances through automatic stabilisers. Finally, fiscal variables are natural choices for variables signalling fiscal stress. They can indicate problems, because high levels of debt or deficit impact negatively on investor sentiment, which in turn translates into worse financing conditions for the country. Disadvantageous financing conditions can result in difficulties in financing borrowing needs and, as a result, can lead to solvency issues (Zhuang and Dowling (2002)). I model the country-year observations as independent in the classification problem I consider, given a small sample size at my disposal, and following the literature (for example Bluwstein et al. (2020)). Explanatory variables used in the analysis are transformed in a way allowing for cross-country and time comparisons and to reduce non-stationarity. The set of variables considered is shown in Table 1.

The database used to design and test early warning models includes the explanatory variables lagged 2 years with regard to the dependent variable; thus, the dependent variable is used for the years 1994-2018, and the explanatory variables for 1992-2016 (with data for some countries

Table 1: Explanatory Variables

Category	Variable	Source
Macroeconomic and global economy	Interest rates in the US	OECD
	Real GDP in the US, y-o-y	OECD
	Real GDP in China, y-o-y	World Bank
	Oil price, y-o-y	BP p.l.c.
	VIX	CBOE
	Real GDP, y-o-y	World Bank, OECD, IMF WEO
	GDP per capita in PPS	World Bank
Competitiveness and domestic demand	Currency overvaluation	IMF WEO
	Current account balance, % GDP	IMF WEO
	Share in global exports, y-o-y	World Bank, OECD
	Gross fixed capital formation, y-o-y	World Bank, OECD
	CPI	IMF IFS, IMF WEO
	Real consumption, y-o-y	World Bank, OECD
Financial	Nominal USD exchange rate, y-o-y	IMF IFS
	Private credit to GDP, change in p.p.	IMF IFS, World Bank and OECD
Fiscal	General government balance, % GDP	IMF WEO
	General government debt, % GDP	IMF WEO
	Effective interest rate on the g.g. debt	IMF WEO
Labor market	Unemployment rate, change in p.p.	IMF WEO
	Labor productivity, y-o-y	ILO

available slightly later). A lag of 2 years is chosen with the aim of closely mirroring the actual availability of data when assessing the performance of the early warning models. In other words, the data pertaining to year t becomes available with a lag lasting up until the middle of year $t + 1$. Therefore, the signal relating to year $t + 2$ leaves the decision-makers some time for a reaction.

Comparison of the means for observations 2 years before the stress event, and those for 2 years before a tranquil period confirms the intuition that variables tend to behave in a particular way in the period preceding the occurrence of the stress event (Table 2). As the distributions of variables are not normal (as per the Shapiro–Wilk test), the Wilcoxon test with statistical significance at 0.05% is employed to check whether the means are statistically different in subgroups. The Wilcoxon test is a non-parametric alternative to a standard t-student test and can be used to compare two unpaired groups with non-normal distributions. In general, the differences in means before the stress period and before the tranquil period are statistically significantly different and in line with intuition in most of the cases, with the exception of dynamics of GDP in China, dynamics of consumption, dynamics of credit to GDP, dynamics of fixed capital formation, dynamics of export share, and Chicago Board Options Exchange Volatility Index (VIX).

Overall, pairwise correlations between the variables are at low levels in most cases, which means that the variables provide different information that may be potentially useful for the early warning models (Figure 2). Correlations higher than 65% are present in the following cases: dynamics of labour productivity and of GDP (75.6%), dynamics of fixed capital formation and of GDP (68.9%), and currency overvaluation and GDP per capita (66.5%). High pairwise correlation is a problem for econometric models, which means that such variables should not be included together in the specification of the logit model. However, for random forest, high

Table 2: Means of Variables Used in the Analysis

Variables	All periods	Tranquil periods	Stress periods	P-value	Significance
CPI	4.26	3.68	7.18	0.00	yes
GDP dynamics	2.9	3.14	1.71	0.00	yes
China GDP dynamics	9.62	9.63	9.59	0.55	no
US GDP dynamics	2.46	2.58	1.82	0.00	yes
US interest rates	4.27	4.4	3.64	0.00	yes
Oil price dynamics	5.05	5.87	0.89	0.03	yes
Consumption dynamics	2.78	2.94	1.98	0.06	no
FX rate dynamics	1.87	0.69	7.82	0.00	yes
Credit to GDP change	1.42	1.58	0.58	0.40	no
Net lending	-2.47	-2.06	-4.55	0.00	yes
Public debt	58.51	56.69	67.71	0.00	yes
Interest on debt	3.58	3.32	4.92	0.00	yes
Currency overvaluation	-33.76	-31.59	-44.73	0.00	yes
Current account balance	-0.52	0.28	-4.57	0.00	yes
Fixed capital formation dynamics	7.82	8	6.96	0.34	no
Export share dynamics	0.6	0.78	-0.36	0.08	no
Unemployment change	-0.04	-0.14	0.48	0.00	yes
Labor productivity dynamics	1.76	1.9	1.05	0.00	yes
VIX	20.17	20.07	20.7	0.74	no
GDP per capita	26.13	28.08	16.28	0.00	yes

Note: The Wilcoxon test with statistical significance at 0.05% is employed to check whether the means of two unpaired groups with non-normal distribution are statistically significantly different.

pairwise correlations between some explanatory variables are not problematic, and hence, there is no need to exclude particular variables from the analysis.

5 Findings

5.1 Aggregate Forecasting Properties

The empirical analysis is focused on developing a set of early warning systems aimed at signalling a fiscal stress event in $t+2$, checking their out-of-sample performance, and discussing their usefulness and interpretability. Early warning systems based on both logit and the random forest models are implemented using a recursive approach. First, models are based on data until 2006 (containing information about the fiscal stress occurrence up to 2008), second, they use observations up to 2007 (and information about the stress event until 2009), and so on. Using a recursive approach is based on the assumption that if such tools were used in practice, they might be refitted every year, in order to extract maximum information from the data available. In the case of logistic regression, in order to avoid multicollinearity and resulting bias of the parameter estimates, the models are fitted using a database without three variables whose pairwise correlation with other variables exceed 65% (i.e. I excluded dynamics of labour productivity, dynamics of gross fixed capital formation and currency overvaluation). For logits, first, the most general model is fitted. Second, with the support of a Wald test and variance decomposition, consecutive variables are removed and the models with and without each variable are compared using the likelihood ratio test, which aims to determine whether a given variable

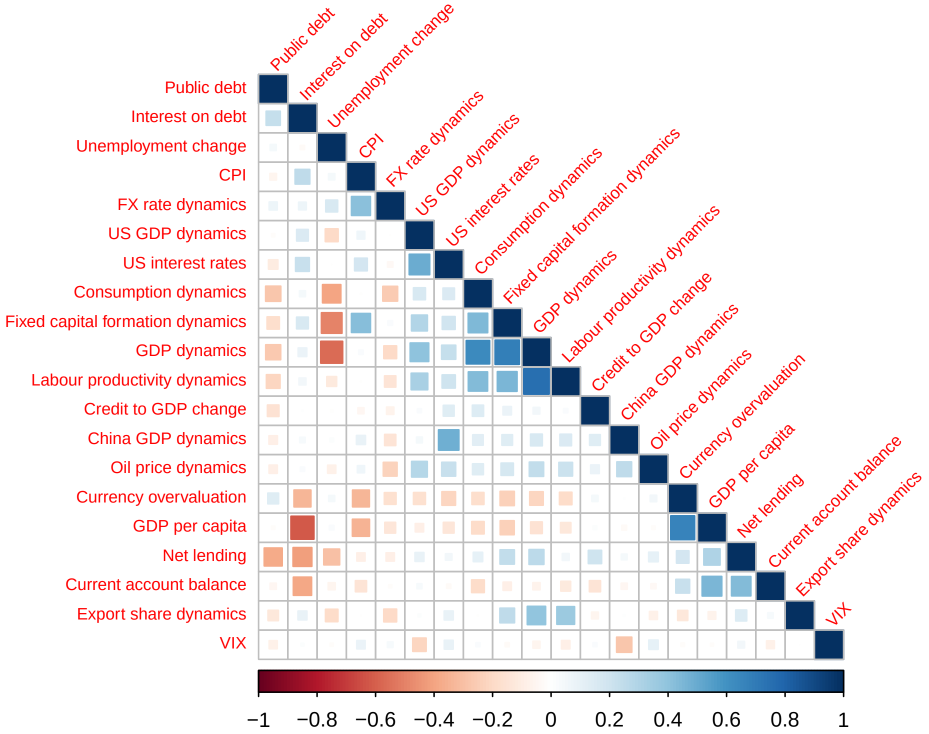


Figure 2: Pairwise Correlations Between the Variables Used in the Analysis

Note: Darker shades of colour mean higher absolute values of correlation between two variables. In particular, blue represents positive correlation, and red negative correlation.

improves the model. Both logit and random forest models are fitted in two versions. In one of them, the GDP per capita accounts for the level of economic development of the country, as historically the likelihood of the fiscal stress event occurrence is lower for more advanced economies than for these less advanced. Next, the GDP per capita is replaced by the binary variable differentiating between advanced and emerging countries, taking the value of 1 when a country is classified as advanced and 0 otherwise. Additionally, in the case of logit models, interactions between the explanatory variables and the advanced/emerging dummy variable are added in order to capture the differences in behaviour of variables between these two groups of countries. Similar approach is not necessary in the case of a random forest model, as this tool accounts for interactions between variables by construction.

Over time, more variables prove to be significant and relevant for explaining the occurrence of the fiscal stress event (Table 3). The coefficient estimates exhibit relative stability. Country-specific variables chosen consistently throughout the period are the current account balance to GDP with a negative coefficient estimate, and CPI with a positive coefficient estimate. GDP per capita is a slow-moving variable which controls for the level of development of a given country. As most of stress events in the past occurred in developing countries, the higher the level of the variable, the lower the likelihood of the fiscal stress event occurrence. Net lending to GDP starts being significant at 5% only since 2010, coinciding with the onset of the fiscal stress in euro area countries. Similarly, change in unemployment rate is significant since

Table 3: Logit Models Fitted Using the Recursive Approach

Variables	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016
Oil price dynamics	-0.45* (0.19)	-0.47* (0.19)	-0.35* (0.17)	-0.54** (0.19)	-0.47** (0.18)	-0.43* (0.18)	-0.43** (0.16)	-0.44** (0.16)	-0.45** (0.15)	-0.29* (0.13)	-0.29* (0.13)
China GDP dynamics	0.42* (0.16)	1.13*** (0.26)	1.08*** (0.23)	0.81*** (0.17)	0.86*** (0.17)	0.85*** (0.16)	0.87*** (0.17)	0.85*** (0.17)	0.86*** (0.17)	0.8*** (0.16)	0.8*** (0.16)
US GDP dynamics	0.3 (0.2)	0.56* (0.28)	0.45 (0.28)	0.01 (0.24)	0.35. (0.21)	0.4. (0.21)	0.4* (0.19)	0.41* (0.18)	0.41* (0.18)	0.24 (0.15)	0.24 (0.15)
VIX		0.55* (0.23)	0.94*** (0.26)	0.58** (0.18)	0.55** (0.17)	0.56*** (0.16)	0.56*** (0.16)	0.52*** (0.16)	0.51*** (0.15)	0.39** (0.14)	0.39** (0.14)
US interest rates		-0.45* (0.23)	-0.43. (0.24)	-0.41. (0.24)	-0.98*** (0.22)	-1.15*** (0.22)	-1.3*** (0.2)	-1.41*** (0.21)	-1.45*** (0.21)	-1.3*** (0.18)	-1.3*** (0.18)
GDP per capita	-1.47*** (0.34)	-1.27*** (0.29)	-1.09*** (0.25)	-0.99*** (0.23)	-0.88*** (0.21)	-0.88*** (0.2)	-0.97*** (0.19)	-1.04*** (0.19)	-1.13*** (0.18)	-1.22*** (0.18)	-1.22*** (0.18)
GDP dynamics	-0.31. (0.16)	-0.23 (0.16)	-0.3* (0.15)	-0.36* (0.15)	-0.12 (0.17)	-0.11 (0.16)	-0.14 (0.16)	-0.17 (0.16)	-0.17 (0.15)	-0.21 (0.15)	-0.21 (0.15)
CPI	0.4** (0.12)	0.5*** (0.13)	0.55*** (0.13)	0.51*** (0.12)	0.5*** (0.12)	0.5*** (0.12)	0.48*** (0.12)	0.47*** (0.11)	0.4*** (0.1)	0.37*** (0.1)	0.37*** (0.1)
Current account balance	-0.44. (0.24)	-0.5** (0.19)	-0.63*** (0.19)	-0.6*** (0.18)	-0.52** (0.16)	-0.51*** (0.15)	-0.59*** (0.15)	-0.66*** (0.15)	-0.68*** (0.14)	-0.69*** (0.14)	-0.69*** (0.14)
Net lending	0.43. (0.22)	0.36. (0.21)	0.14 (0.2)	-0.06 (0.19)	-0.38* (0.16)	-0.45** (0.16)	-0.35* (0.17)	-0.32* (0.16)	-0.27. (0.15)	-0.23 (0.15)	-0.23 (0.15)
Unemployment change					0.26. (0.14)	0.33* (0.14)	0.34* (0.14)	0.36** (0.13)	0.35** (0.13)	0.34** (0.13)	0.34** (0.13)
Public debt							0.17 (0.13)	0.19 (0.13)	0.26* (0.12)	0.27* (0.11)	0.27* (0.11)
(Intercept)	-3.54*** (0.33)	-3.36*** (0.28)	-3.13*** (0.25)	-2.95*** (0.22)	-2.79*** (0.2)	-2.74*** (0.2)	-2.72*** (0.19)	-2.72*** (0.19)	-2.69*** (0.18)	-2.66*** (0.17)	-2.66*** (0.17)
AIC	243.63	286.59	330.59	384.26	441.52	476.99	511.19	539.66	573.00	605.38	605.38
R2	0.26	0.27	0.29	0.28	0.28	0.29	0.31	0.32	0.32	0.32	0.32
ROC	0.86	0.87	0.87	0.86	0.86	0.86	0.87	0.87	0.87	0.87	0.87

Note: Years in the table relate to the cut-off year for the explanatory variables, for example, 2006 means that the latest observations used to fit the model contain explanatory variables until 2006 and the dependent variable until 2008. To avoid multicollinearity and resulting bias of the parameter estimates, the models are fitted using a database without three variables whose pairwise correlation with other variables exceed 65%. GDP per capita accounts for the level of economic development of countries.

2011. In turn, public debt to GDP is retained only since 2014. Dynamics of export share, dynamics of FX rates, change in the ratio of credit-to-GDP and interest on public debt are not retained in any of the models. Additionally, Table A1 in the Appendix provides results of recursive logit model estimation when the continuous GDP per capita is replaced by the dummy for advanced/emerging economies and its interactions with other explanatory variables, accounting for different behaviour of explanatory variables in advanced and emerging economies. Current account balance and change in unemployment are particularly important for advanced economies, similar to net lending, which is however significant only in the period of the sovereign debt crisis in some of the euro area countries. In contrast, CPI and public debt seem to be more important in the case of emerging economies.

Moreover, a number of variables are included to capture the impact of global developments on the likelihood of the fiscal stress event: US and China GDP dynamics, US long-term interest rates, oil price dynamics, and VIX. The lower levels of long-term US interest rates are associated with increased likelihood of the fiscal stress event in the future. In the period analysed, US long-term interest rates were falling until 2012, driven by loosening of US monetary policy, aimed at supporting the economy during the crisis. The parameter estimate for oil price dynamics is negative, that is, oil price increases tend to precede the periods characterised by lower likelihood of the fiscal stress event materialising. By contrast, the increases in VIX, interpreted as proxying investor sentiment, tend to precede the periods of fiscal stress. Variables aiming at capturing the impact of global developments are not significant when interacted with the dummy taking the value of one for advanced economies, with an exception of US interest rates (Table A1).

If early warning models were used in practice and refitted every year, it could have been a pragmatic choice to use an automatic variable selection method to avoid discretion and communication challenges related to changes in the model from period to period. LASSO tends to retain many variables in the model, even those that are not significant, such as export share dynamics, FX rate dynamics, credit to GDP change, and interest on public debt (see Table A2 in the Appendix with GDP per capita and Table A3 with the advanced country dummy). This difference with regard to logit models in Table 3 could be related to the fact that logit LASSO models are fitted using the forecasting properties of the model, as they are implemented using a cross-validation of five folds and a choice of a penalisation parameter is based on the AUROC measure on the out-of-sample data. Hence, it may seem that the model contains unnecessary variables, but if they increase the forecasting properties, they are still retained. This illustrates the difference between the explanatory and forecasting approach to modelling, since different goals result in different approaches, as comprehensively explained by Shmueli (2010). LASSO penalisation shrinks parameter estimates, some of them to zero, effectively eliminating a variable from the model. Logit models used in this study, both with and without LASSO penalisation, are fitted using standardised variables, and hence, the coefficients of the variables within a given model are comparable. The levels of coefficients obtained by the model with LASSO penalisation cannot be compared with the levels of the coefficients (and marginal effects) obtained by the logit models presented in Table 3, as some of the coefficients have been shrunk. However, the signs can be compared, as can whether the variable is retained both in the models in Table 3 and by logit models with LASSO penalisation.

Early warning systems using the random forest models are fitted using the entire set of 20 variables that are considered as potentially useful in signalling the risk of a fiscal stress event, as multicollinearity between the predictors is not a problem for this tool. For each year, a random forest of 10000 trees is used. Parameter m is set at the default proposed value of the floor of the square root of the number of explanatory variables, which amounts to four in this application.

Table 4 shows the average prediction accuracy of the compared approaches of the stress events in the years 2009-2018, using as a metric the threshold-independent AUROC and the threshold-dependent sensitivity, specificity, and their unweighted average. Each of the three methods is used in two variants - when using the dummy for advanced/emerging economies, and when using GDP per capita as a continuous proxy for the level of economic development of a country. Based on the results obtained, the early warning system is more accurate in classifying stress and tranquil periods when it is designed using random forest models than when using logit models. The average prediction accuracy for methods used is 70-75% for logit models and 78-80% for random forest models.

Table 4: Average Prediction Accuracy of Early Warning Models for Years 2009-2018, Using all Observations

	Logit LASSO		Logit manual		Random forest	
	advanced dummy	GDP per capita	advanced dummy	GDP per capita	advanced dummy	GDP per capita
% of correctly classified stress episodes	86.95	73.33	82.36	78.88	89.56	91.18
% of correctly classified tranquil episodes	59.77	67.03	61.65	70.78	65.74	67.95
Average	73.36	70.18	72.01	74.83	77.65	79.56
AUROC	0.83	0.84	0.81	0.84	0.88	0.89

Note: Models fitted in two variants - with the dummy taking the value of one for advanced and zero for emerging economies, and when using GDP per capita as a continuous proxy for the level of economic development of countries. In the case of logit model fitted manually, also interactions between the dummy variable and the other predictors are included.

The early warning models whose results are presented in Table 4 and Figure 3 have different weights assigned to sensitivity and specificity. In the baseline scenario, thresholds are obtained by maximising the weighted sum of sensitivity and specificity of a given classifier, assigning a 50% higher weight to sensitivity than to specificity. Additionally, Tables A4-A6 in the Appendix provide the detailed annual results of variants of early warning models based on logit and the random forest models, with thresholds optimised when weighting sensitivity equal to or more important (50% and 100%) than specificity (following Alessi and Detken (2011)). The comparison of the share of correctly classified stress and tranquil events in these variants underpins the relevance of choosing a weighting scheme and a resultant threshold that reflects preferences of a researcher or a decision-maker to the extent possible (Manasse and Roubini (2009)).

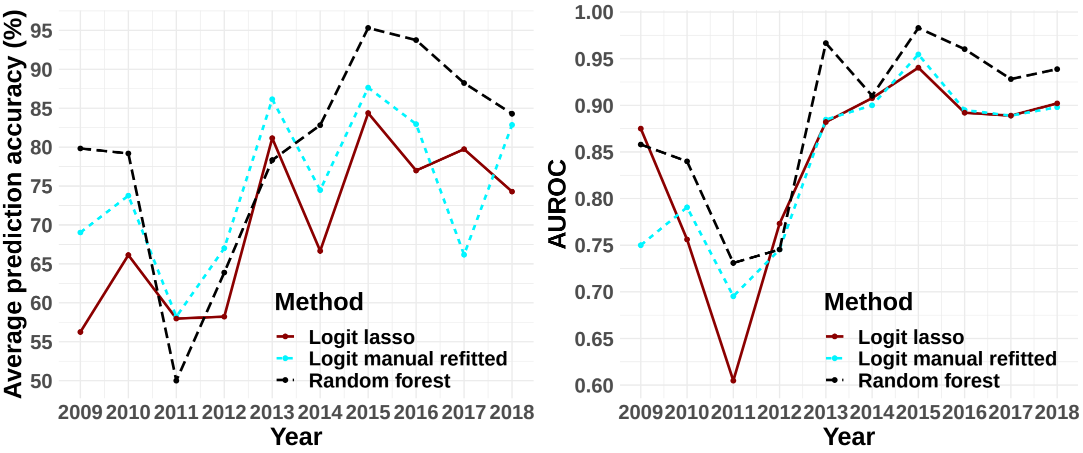


Figure 3: Average Prediction Accuracy and AUROC of Early Warning Models for Year $t + 2$
Note: Blue lines show the average prediction accuracy/AUROC using logit models specified manually, purple lines show the results obtained using models refitted with the LASSO penalisation, and black dashed lines show average prediction accuracy/AUROC for the random forest. GDP per capita accounts for the level of economic development of countries. Average prediction accuracy for years 2009-2018 is presented in Table 4.

A feature worth mentioning is a relative weakness of the models' performance during 2010-2012, the period of the sovereign debt crisis in some euro area countries (Figure 3). Perhaps it should not be surprising that a purely quantitative model is unable to capture events that were largely driven by qualitative factors, like the strength and credibility of fiscal frameworks, investor sentiment, and loss of credibility in view of large revisions of statistical fiscal data reported. Reputation and credibility are very important factors that cannot be directly considered in a model of the presented kind, Japan being a point in case - it is highly indebted, but still positively perceived by investors. Overall, the likelihood of the stress event tends to be lower in economies with strong fiscal institutions than in those with weak institutions, even in a situation of identical macroeconomic and financial parameters. According to Mitchell and Stansel (2016), fiscal stress is positively associated with prior spending growth. Therefore, decision-makers should avoid substantially increasing spending in good times, in order to minimise fiscal stress risk in the future. Consequently, credible fiscal rules can substantially change the perception of a country, as tax and expenditure limits constrain the range of decisions that can be taken by decision-makers (Ostaszewski and Wrzesiński (2018)). An additional factor is that the preparedness of the country for potential difficulties can play a crucial role. For example, sizeable ”rainy day” funds created during economic expansions help to minimise fiscal stress during recessions (Jimenez (2017)).

Next, the prediction accuracy of the first year of the stress period is assessed. All the models used are fitted using all the observations in the sample, but Table 5 reports the prediction accuracy when the observations pertaining to the ongoing stress periods are removed, and only the first year of a given stress period is retained. If there is only one non-stress tranquil period between two stress periods, it is assumed that the same stress period is ongoing, and the tranquil period in between is also removed. This explains why the percentages of correctly classified tranquil periods reported in Tables 4 and 5 differ slightly. Forecasting fiscal stress events is important in all cases, during both stress and tranquil periods. During the stress period, information that the stress may continue is relevant for decision-makers. However, information about a potential shift from a tranquil to a stressed period can be considered especially valuable.

The models based on logistic regressions and random forest classified correctly 62-81% the first years of the periods of fiscal stress (Table 5). Comparing the prediction accuracy of the first year of the stress event with the prediction accuracy obtained for the entire period shows clearly that it is more difficult to forecast only the first year of the stress event than to forecast a continuation of an ongoing stress - even though the current state of the economy (stress or tranquil period) is not included among the predictors. This notwithstanding, the prediction accuracy of the models is relatively high also when focusing on the first stress year, which shows that the methods proposed could be a helpful tool for signalling the risk of a fiscal stress in the near future.

Table 5: Average Prediction Accuracy of Early Warning Models for Years 2009-2018, Observations Pertaining to the Ongoing Stress Period Removed

	Logit LASSO		Logit manual		Random forest	
	advanced dummy	GDP per capita	advanced dummy	GDP per capita	advanced dummy	GDP per capita
% of correctly classified stress episodes	73.08	69.23	80.77	65.38	61.54	76.92
% of correctly classified tranquil episodes	60.26	66.78	62.54	70.68	67.10	69.38
Average	66.67	68.00	71.65	68.03	64.32	73.15
AUROC	0.54	0.52	0.69	0.64	0.69	0.70

Note: Models fitted in two variants - with the dummy taking the value of one for advanced and zero for emerging economies, and when using GDP per capita as a continuous proxy for the level of economic development of countries. In the case of logit model fitted manually, also interactions between the dummy variable and the other predictors are included.

The models proposed are universal across countries. An alternative approach would have been to estimate country-specific models, considering that countries can be structurally different. However, this approach should be employed only when time series are long enough, each of the variables per country exhibits sufficient variability, and each of the countries has at least some stress events included in the database. Since these conditions are not met in the current analysis, using universal models seemed to be a natural choice. This approach is also employed in most studies in the early warning literature (Knedlik and von Schweinitz (2012); Lo Duca and Peltonen

(2011); Manasse and Roubini (2009).

5.2 Forecasting Properties Based on the Sovereign Debt Crisis in the Euro Area

This subsection presents a visualisation of the application of the random forest-based early warning models to countries that were affected by the sovereign debt crisis which started around 2010 in euro area: Cyprus, Greece, Spain, Ireland, Portugal, and Italy. In line with the approach presented in Figure 3 and Table 3, models are fitted using a recursive approach until a certain year, and tested in the following period. The thresholds in the models are universal across countries, but vary over time, as the models are re-estimated adding 1 year of data at a time. When the probability of the stress event exceeds the threshold, the given model classifies an observation as a stress event, and a warning signal is issued.

If an early warning system based on the random forest had been used in practice in the period preceding the sovereign debt crisis in the euro area, its effectiveness would have been as follows. It would have correctly classified the first year of a stress period in Italy, Cyprus, Spain, Greece and Ireland (Figure 4). The model would not have correctly classified the first year of a stress period in Portugal, although the stress probability would have been only slightly below the threshold. While this accounts for an incorrectly classified observation, in practice, this information could have been useful as well. This highlights the importance of choosing the weights of sensitivity and specificity, and corresponding thresholds that reflect the underlying preference, which could in certain cases mean more conservative choices than the baseline presented. While the early warning system based on the random forest model would not have been helpful in predicting the imminent stress period in Portugal, the observations for the following years are classified correctly. In the same period the performance of the logit model would have been less satisfactory, as most of the first years of the fiscal stress periods in countries affected by the sovereign debt crisis would not have been signalled (see the Appendix, Figure A2 for details). However, the first year of the stress period in Portugal would have been classified correctly, so using random forest and logit models in parallel could have sent a complete set of warning signals.

An important aspect is that euro area countries that experienced fiscal stress in the analysed period did not experience fiscal stress in the period used to fit the models (except for Greece in 1994-1995). Hence, early warning models proposed could have signalled many of the stress periods only based on fiscal stress events experienced by other countries, mostly developing countries. This in turn shows that the behaviour of certain variables in the period preceding the stress event tends to be universal across countries, even if country-specific features still play a role.

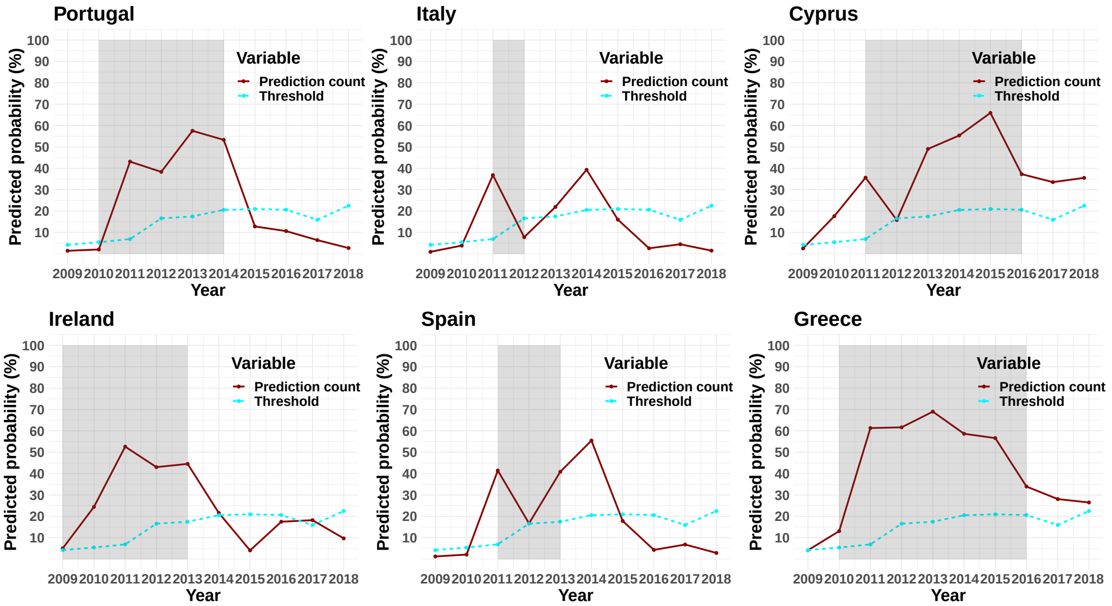


Figure 4: Prediction Accuracy of Early Warning Models Based on Random Forest

Note: Grey areas mean that given periods are classified as stress events. Blue lines show the threshold optimised for a given year and purple lines show the probability of a fiscal stress event, which is the share of decision trees among the 10000 used that classifies a given observation as a stress period. Results presented are based on random forest models using GDP per capita to control for the level of economic development of countries.

5.3 Interpretability of a Random Forest-based Early Warning Model

Based on the results provided, random forest-based early warning models offer higher average prediction accuracy than logit models do. However, apart from the precision of the classification, another important aspect for economists and decision-makers alike is the interpretability of the outcome obtained. When a signal is issued, it could be very relevant information on which variables were crucial for that result. The standard perception is that logit models offer higher interpretability than the random forest models do. It is certainly true that it is easier to communicate and understand logit models. The coefficients estimated allow for computation of marginal effects, and Wald statistics allow assessment of the statistical significance of variables; those features of econometric models are appreciated and very useful. However, there are also advantageous features that help with the interpretability of machine learning models, a case in point being the variables' importance measures (Breiman (2001)), Shapley values (Shapley (1953)), partial dependence plots (Friedman (2001)) and accumulated local effects plots (Apley (2016)). These measures offer valuable information, partially akin to estimated coefficients and the significance of the variables.

A variable's importance (Breiman (2001)) is understood as the average improvement of the model's performance when a given variable is included, compared with the case in which the variable is excluded. In practice, the values of a variable of interest are permutated, which breaks the relationship between that variable and the true class of the dependent variable. A variable's importance can be quantified as the mean increase in the model's accuracy, or the average decrease in impurity, as measured by the Gini index. The interpretation of the measure is straightforward, which is also its great advantage: variables with high importance have a substantial impact on the outcome classification, while variables with low importance contribute little to the outcome and may be omitted from the model without a large impact on the resultant classification. Moreover, by construction, the variable's importance automatically takes into account all interactions with other variables, as replacing the original variable with its permutation breaks not only the link to the dependent variable, but also to other explanatory variables. On the other hand, permutation adds randomness to the measures, but reporting an average of multiple computations of measures of variable's importance addresses this issue. Moreover, the variable's importance can be biased and should be interpreted with caution when some variables are correlated, as when only one of them gets permutated, unrealistic observations are created and used to compute the measure of importance. Moreover, if some variables are correlated, the resultant measure of the importance of these variables will be decreased, as the actual importance will be split between these predictors.

Alternatively, Shapley values (Shapley (1953)) from cooperative game theory can provide information on which variables are driving the prediction (Štrumbelj and Kononenko (2014)). The Shapley value of a predictor is the average marginal contribution of this variable across all possible coalitions of explanatory variables. In practice, this approach relies on decomposing the difference between the predicted value of each observation and the mean predicted value into a sum of contributions of explanatory variables. Hence, a Shapley value for a predictor j is interpreted as a contribution to a prediction of a particular class compared to an average prediction for a dataset. These contributions are Shapley values at observation level, and the means of their absolute values provide information about the average contribution of a given variable to the outcome obtained in different coalitions. The advantage of Shapley values is that they are backed by solid theory. However, their interpretation is not intuitive - the Shapley value of a variable is the contribution of a given variable to the difference between the actual prediction for a given observation and the mean prediction, given the set of variables.

Figure 5 shows the importance of predictors measured by the Breiman's variable importance using Gini index and by the Shapley values, for the random forest models fitted on the entire dataset. The ranking provided by the two approaches is relatively similar. In particular, the same five variables are ranked the highest by two methods (current account balance, CPI, GDP per capita, US interest rate and net lending), albeit in a slightly different order. Moreover, the variables with the highest importance are in line with the significant variables consistently retained in the logit models (Table 3). Breiman's variable importance measures quantified by the Gini index are broadly in line with these measured by the mean increase in the model's accuracy (Figure A3 in the Appendix).

Partial dependence plots (Friedman (2001)) are a very intuitive tool showing the marginal effect that one variable has on the predicted outcome. For classification in which the model provides probabilities (as in random forest), the partial dependence plot displays the impact

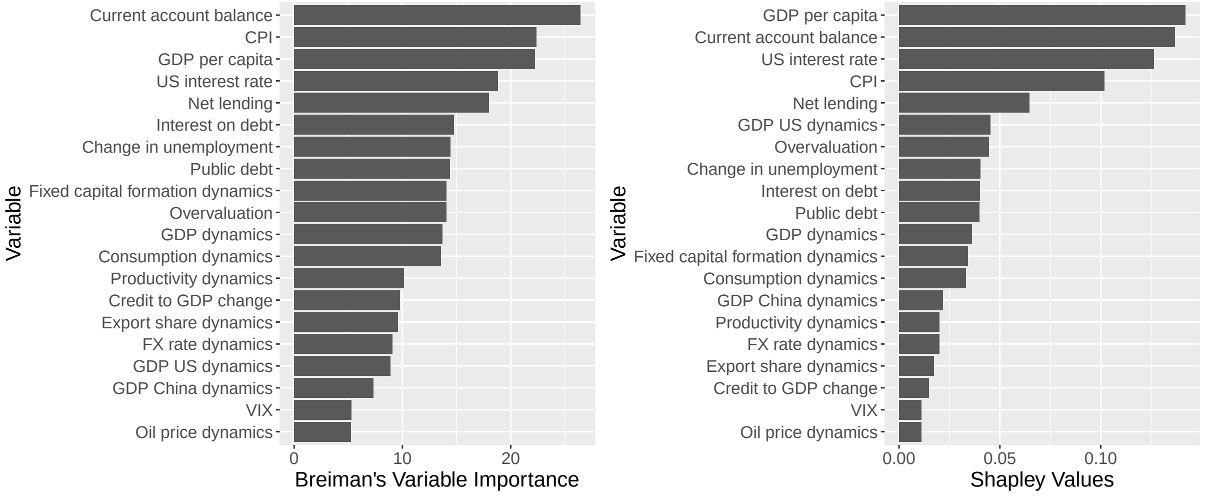
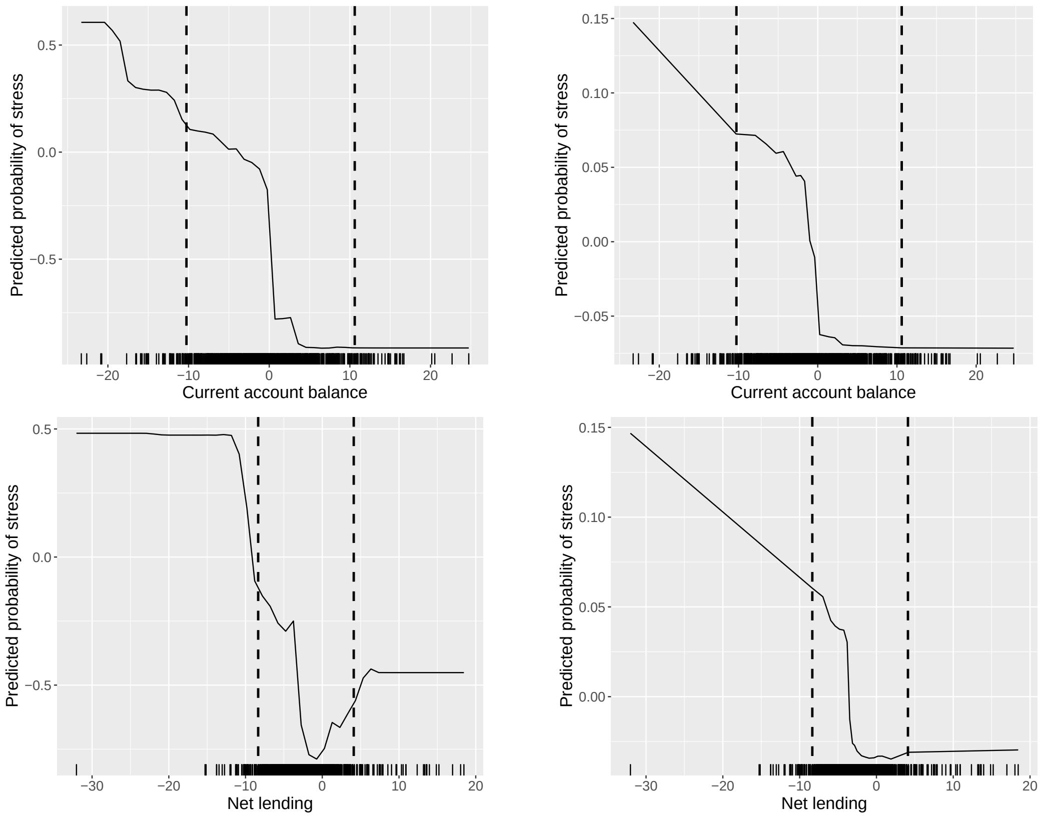


Figure 5: Variable Importance and Shapley Values of Predictors Included in the Random Forest-based Early Warning System

Note: The variable's importance is measured as the average improvement of the random forest model's performance when a given variable is included, compared with the case in which the variable is excluded, here measured by mean decrease in Gini impurity index, average of 100 repetitions (to account for the randomness of the permutation). Shapley values are computed as averages of absolute values of Shapley values of predictors in each observation. Results presented are based on random forest models using GDP per capita to control for the level of economic development of countries, fitted on the entire dataset.

on the probability of a certain class given different values of an independent variable. To this end, the model with averaged values of other independent variables is fitted and in this way, a function that depends only on a given predictor is obtained. Viewing plots of the partial dependence of the random forest on the selected variable can help to qualitatively describe its properties (Hastie et al. (2012)). It can be compared to a graphical representation of linear regression model coefficients that extends to the random forest model. Partial dependence plots for linear econometric models always show a linear relationship, while the actual underlying relationship may not be linear - in the logit model fitted on the entire dataset (presented in Table 3) hardly any predictor exhibits a linear relationship with the logit of the dependent variable (see Figure A6 in the Appendix for the check of the linearity assumption). Partial dependence plots provide a way to visualise non-linearities which can be accounted for by the random forest-based model. However, partial dependence plots, similar to Breiman's variable importance, also suffer from limitations when explanatory variables are correlated, as some of the artificially created observations are unlikely in reality and can bias the results.

Another possibility allowing to understand better the shape of the relationship between a given predictor and the outcome classification are accumulated local effects plots (Apley (2016)). The results provided by this tool are not biased when some variables are correlated, as by design the unrealistic data instances mentioned above are not created. Accumulated local effects plots show differences in predictions driven by changes of a given explanatory variable in a small interval, accumulated over a grid of intervals of this independent variable. Hence, conditional



(a) Partial Dependence Plots

(b) Accumulated Local Effects Plots

Figure 6: Partial Dependence and Accumulated Local Effects Plots

Note: Results presented are based on random forest models using GDP per capita to control for the level of economic development of countries, fitted on the entire dataset.

on a given value of an explanatory variable, accumulated local effects plots provide information on a relative effect of changing the value of this variable. In other words, accumulated local effects plots show how the model predictions change around v, from $v-h$ to $v+h$, providing pure effects of the variable of interest. In contrast, partial dependence plots show what the model predicts on average when a given variable has a value of v in all observations, regardless of whether these artificial observations with the given variable at v make sense or not. Accumulated local effects plots are centred at zero, so the value at each point is the difference compared to the mean prediction.

Figure 6 shows the partial dependence (column (a)) and accumulated local effects plots (column (b)) based on the random forest model fitted on the entire dataset for the example of two variables: the current account balance and net lending (Figures A4 and A5 in the Appendix show the plots for the remaining variables). The shape of the relationship between the explanatory variables and the outcome dependent variable is similar in two approaches compared in columns (a) and (b), in spite of a potential bias of the partial dependence plots driven by the correlation between the variables of interest and other explanatory variables. Also, the direction of the impact of the explanatory variables on the dependent variable that can be read from the plots is broadly in line with the signs obtained in logistic regressions (Tables 3), but in most cases, this impact is far from linear. A surplus on the current account balance decreases the probability of the occurrence of the stress event, while a deficit increases its probability. This is in line with the negative coefficient for the current account balance in the logit model. However, as can be read from the partial dependence and accumulated local effects plots for the current account balance, the relationship is not linear, and the contribution to the positive outcome (a signal of stress) falls abruptly around zero. Another example among crucial variables could be net lending to GDP. While the overall downward-sloping shape is also in line with the negative coefficient estimated by the logit model, the negative contribution to the "no stress" classification seems to be observed for values above few percentage points below zero. The shape of the plot for net lending is however different for the two tools for values above zero, with accumulated local effects plot showing that, compared to the mean prediction, the impact of net lending above few percentage points below zero on the probability of the stress event is negative, and stable regardless of the exact level of the variable - a small deficit or a fiscal surplus. In contrast, in partial dependence plots the values of net lending slightly below zero contribute the strongest to lower probability of the stress event, and higher net lending, also a fiscal surplus, contribute less to this probability. This result is rather counterintuitive, and could be related to the bias of a partial dependence plot driven by the correlation of net lending with other variables, for example public debt and interest on debt.

An interesting aspect is a U-shape of the partial dependence and accumulated local effects plots for most of the country-level cyclical variables, which suggests that both cyclical downswings and upswings contribute to future fiscal stress. These U-shapes of the plots are largely driven by tails of the distributions - for most of the variables, the share of stress periods is higher in regions below the 5th percentile and above the 95th percentile. These parts of the plots are based on few observations, which may be problematic both for partial dependence and the accumulated local effects plots, and hence should be interpreted with caution (see the "rugs" below the plots, which show how many observations of a given explanatory variable were available to inform each part of the plot, and vertical lines at 5th and 95th percentile of a given explanatory country-specific variable). In the regions of cyclical downswing (low GDP dynamics, large increase of unemployment, low dynamics of consumption and fixed capital formation) many of the observations pertain to the period characterised by the sovereign debt crisis in some of the euro area countries (Cyprus, Greece, Ireland, Italy, Latvia, Portugal). In contrast, the observations with stress periods in two years in the regions of cyclical upswings to large extent relate to emerging countries, some of them smaller: Albania, Algeria, Jamaica, El Salvador, Pakistan, or Ukraine. Moreover, this category contains also observations related to the strong cyclical upswing right before the bust, occurring in the period of strong build-up of vulnerabilities in euro area countries (for example Latvia or Iceland in 2006). Finally, some of these observations with both large and small values of explanatory variables are related to Argentina, for which most of the periods (20 out of 24 in the database) are classified as stress periods. Both in the case of the cyclical upswing and downswing the observations driving the U-shapes of the plots are valid and informative, and their possible removal as outliers should be considered very carefully against the backdrop of losing available information. Therefore, given a relatively small database, I decided to retain these observations. Alternatively, if a larger database with more stress periods was available it would be interesting to compare the results and the plots with the variant when only the first year of the stress period was coded as 1 and the periods of ongoing stress were removed. This is however not possible in this study, as my database contains very few stress observations and the results obtained are unstable.

The measures of Breiman's variable importance, Shapley values and the partial dependence and accumulated local effects plots can help understand how the outcome of a random forest model could be interpreted. Despite being heuristics, they provide valuable information partially akin to information available for logit models, and to large extent mitigate the perceived limitation of a lack of interpretability of machine learning tools. Some recent studies in the literature interpret black-box machine learning methods. For example, Bluwstein et al. (2020) use Shapley values to identify the best predictors of a crisis, and a thorough explanation of partial dependence plots is provided by Zhao and Hastie (2019).

6 Conclusions

The study aims to determine the effectiveness and usefulness of various approaches that can be used to design early warning models aimed at signalling fiscal stress. Specifically, this study focuses on stress events, a broad category encompassing fiscal crises related to public debt default or restructuring as well as less extreme events, which could nonetheless indicate difficulty with public debt or deficit financing. Policymakers could benefit from an early warning of risk of a fiscal stress event in the near future. Using this information, they could implement measures aimed at limiting or pre-empting the effects of the problems.

Central to the analysis is a comparison of the effectiveness and usefulness of the random forest with logistic regression. Logit models are standard tools broadly used in the literature on early warning models. Meanwhile, the random forest model is less widespread yet more innovative, and its popularity is increasing, as this approach is very effective for analyses aimed at classifying the outcome to the appropriate category.

In this study, the effectiveness of the random forest models proved to be slightly higher than that obtained by logit models. The former yields an average prediction accuracy of fiscal stress events and tranquil events of slightly below 80%, and the latter 70-75%. If an early warning model based on random forest had been implemented in the past, signalling of many of the fiscal stress episodes related to the sovereign debt crisis in the euro area would have been possible. Focusing on the first year of the stress event only, the average prediction accuracy dropped to 64-73% for the random forest models and 67-72% for the logit models. While this clearly shows that it is more difficult to predict the first year of the stress event than ongoing stress, the proposed early warning models still offer useful prediction accuracy. This conclusion is underpinned by the underlying reason for using early warning systems, given that the outcome signal should be interpreted as a warning of heightened level of vulnerability, and not as a forecast of a crisis. Therefore, it is worth striving to develop even an imperfect tool.

The random forest is understood to offer lower interpretability of results than the logit models, which constitutes a relevant limitation for economists. Some of the especially useful features of econometric models are not available when using the random forest; however, alternative sources of similar information are available. Variable's importance measure proposed by Breiman (2001) and game theory-derived Shapley values (Shapley (1953)) can help to assess which predictors are especially useful for the classification problem. Furthermore, partial dependence (Friedman (2001)) and accumulated local effects plots (Apley (2016)) of the random forest aid understanding of the impact of a given variable on the outcome obtained.

Further research could be pursued as follows. First, it may be valuable to pay more attention to institutional variables. Two countries characterised by a similar set of explanatory variables but differing with regard to institutional strength face distinct likelihood of problems with debt and deficit financing, related to, among other factors, investor sentiment. To this end, an interesting future research direction is the construction of an index utilising information about fiscal rules applying in a particular country, coupled with additional information provided by international organisations.

Second, further discussion on the usefulness of machine learning approaches in economics could benefit from application of other methods to the early warning problem, including, but not limited to, gradient boosting machines, neural networks, and support vector machines. Finally, the present study only briefly touched on the use of various measures enabling interpretation of black-box random forest models in the ongoing discussion of the usefulness of machine learning in economics. However, the field is growing and already provides many more tools that could be explored. In this regard, the bio-, chemical, and medical engineering literature could be inspiring.

References

Alessi, L. and C. Detken (2011). Quasi real time early warning indicators for costly asset price boom/bust cycles: A role for global liquidity. *European Journal of Political Economy 27*(3), 520–533.

Alessi, L. and C. Detken (2018, apr). Identifying excessive credit growth and leverage. *Journal of Financial Stability 35*, 215–225.

Apley, D. W. (2016). Visualizing the effects of predictor variables in black box supervised learning models.

Athey, S. (2019, may). The Impact of Machine Learning on Economics. In *The Economics of Artificial Intelligence: An Agenda*, pp. 507–547. University of Chicago Press.

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

Barrell, R., E. P. Davis, D. Karim, and I. Liadze (2010, sep). Bank regulation, property prices and early warning systems for banking crises in OECD countries. *Journal of Banking & Finance 34*(9), 2255–2264.

Beers, D. and J. Mavalwalla (2017). Database of Sovereign Defaults, 2017. *Ssrn No. 101.*

Berti, K., M. Salto, and M. Lequien (2012). An early-detection index of fiscal stress for EU countries. *European Economy. Economic Papers 475.*

Bluwstein, K., M. Buckmann, A. Joseph, M. Kang, S. Kapadia, and Ö. Simsek (2020). Credit growth, the yield curve and financial crisis prediction: evidence from a machine learning approach. *Bank of England Staff Working Paper 848.*

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

Bruns, M. and T. Poghosyan (2016). Leading Indicators of Fiscal Distress: Evidence from the Extreme Bound Analysis. *IMF Working Paper WP/16/28.*

Bussiere, M. and M. Fratzscher (2006). Towards a new early warning system of financial crises. *Journal of International Money and Finance 25*(6), 953–973.

Caruana, R. and A. Niculescu-Mizil (2006). An Empirical Comparison of Supervised Learning Algorithms. In *Presented at Proc. Int. Conf. Machine Learn., 23rd, Pittsburgh, PA.*

Chetverikov, D., Z. Liao, and V. Chernozhukov (2016, may). On cross-validated Lasso. *Cornell University Library*.

Ciarlone, A. and G. Trebeschi (2005, dec). Designing an early warning system for debt crises. *Emerging Markets Review 6*(4), 376–395.

Cruces, J. J. and C. Trebesch (2013). Sovereign Defaults: The Price of Haircuts.

Davis, E. P. and D. Karim (2008, oct). Could Early Warning Systems Have Helped To Predict the Sub-Prime Crisis? *National Institute Economic Review 206*(1), 35–47.

Demirgüç-Kunt, A. and E. Detragiache (2005). Cross-Country Empirical Studies of Systemic Bank Distress: A Survey. *IMF working paper WP/05/96.*

Demyany, Y. and I. Hasan (2009). Financial crises and bank failures: a review of prediction methods. *Bank of Finland Research Discussion Papers 35.*

Duttagupta, R. and P. Cashin (2008). The Anatomy of Banking Crises. *IMF working paper WP/08/93.*

Fioramanti, M. (2008, jun). Predicting sovereign debt crises using artificial neural networks: A comparative approach. *Journal of Financial Stability 4*(2), 149–164.

Friedman, J. H. (2001). Greedy Function Approximation: A Gradient Boosting Machine. *The Annals of Statistics Vol. 29, N*, 1189–1232.

Fuertes, A.-M. and E. Kalotychou (2006, nov). Early warning systems for sovereign debt crises: The role of heterogeneity. *Computational Statistics & Data Analysis 51*(2), 1420–1441.

Fuertes, A.-M. and E. Kalotychou (2007). Optimal design of early warning systems for sovereign debt crises. *International Journal of Forecasting 23*(1), 85–100.

Hastie, T., R. Tibshirani, and J. Friedman (2012). *The Elements of Statistical Learning.* Springer.

Hernández de Cos, P., G. B. Koester, E. Moral-Benito, and C. Nickel (2014, jul). Signalling Fiscal Stress in the Euro Area - A Country-Specific Early Warning System. *ECB Working Paper No. 1712.*

Holopainen, M. and P. Sarlin (2017, dec). Toward robust early-warning models: a horse race, ensembles and model uncertainty. *Quantitative Finance 17*(12), 1933–1963.

Jimenez, B. S. (2017, jun). Institutional Constraints, Rule-Following, and Circumvention: Tax and Expenditure Limits and the Choice of Fiscal Tools During a Budget Crisis. *Public Budgeting & Finance 37*(2), 5–34.

Kaminsky, G. L. (1999). Currency and Banking Crises : The Early Warnings of Distress. *IMF Working Paper No. 99/178.*

Kaminsky, G. L., S. Lizondo, and C. M. Reinhart (1998). Leading Indicators of Currency Crises. *IMF Staff Papers* (November), 45.

Kaminsky, G. L. and C. M. Reinhart (1999, jun). The Twin Crises: The Causes of Banking and Balance-of-Payments Problems. *American Economic Review 89*(3), 473–500.

Knedlik, T. and G. von Schweinitz (2012, sep). Macroeconomic Imbalances as Indicators for Debt Crises in Europe. *JCMS: Journal of Common Market Studies 50*(5), 726–745.

Kumar, M., U. Moorthy, and W. Perraudin (2003). Predicting emerging market currency crashes. *Journal of Empirical Finance 10*(4), 427–454.

Lo Duca, M. and T. Peltonen (2011). Macrofinancial vulnerabilities and future financial stress: assessing systemic risks and predicting systemic events. In *Macroprudential regulation and policy*, Volume 60, pp. 82–88. Bank for International Settlements.

Lo Duca, M. and T. A. Peltonen (2013). Assessing systemic risks and predicting systemic events. *Journal of Banking & Finance 37*, 2183–2195.

Louppe, G. (2014, jul). Understanding Random Forests: From Theory to Practice. *Cornell University Library.*

Manasse, P. and N. Roubini (2009, jul). "Rules of thumb" for sovereign debt crises. *Journal of International Economics 78*(2), 192–205.

Manasse, P., N. Roubini, and A. Schimmelpfennig (2003). Predicting Sovereign Debt Crises. *IMF Working Paper No. 03/221.*

Mitchell, D. T. and D. Stansel (2016). The Determinants of the Severity of State Fiscal Crises. *Public Budgeting & Finance 36*(4), 50–67.

Ostaszewski, J. and M. Wrzesiński (2018). *Etyka, sprawiedliwość i racjonalność w dorobku nauki o finansach w latach 1918-2018.* Szkoła Główna Handlowa. Oficyna Wydawnicza.

Reinhart, C. and K. Rogoff (2010, mar). From Financial Crash to Debt Crisis. *National Bureau of Economic Research.*

Reinhart, C. M. and K. S. Rogoff (2008). Is the 2007 US Sub-prime Financial Crisis So Different? An International Historical Comparison. *American Economic Review 98*(2), 339–344.

Sarlin, P. (2012, mar). On biologically inspired predictions of the global financial crisis. *Neural Computing and Applications 24*(3-4), 663–673.

Schularick, M. and A. M. Taylor (2012). Credit Booms Gone Bust: Monetary Policy, Leverage Cycles, and Financial Crises, 1870-2008. *American Economic Review 102*(2), 1029–1061.

Shapley, L. S. (1953). A value for n-person games. *Contributions to the Theory of Games 2.28*, 307–317.

Shmueli, G. (2010). To Explain or to Predict? *Statistical Science 25*(3), 289–310.

Štrumbelj, E. and I. Kononenko (2014, dec). Explaining prediction models and individual predictions with feature contributions. *Knowledge and Information Systems 41*(3), 647–665.

Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological) 58*(1), 267–288.

Trebesch, C., M. G. Papaioannou, and U. S. Das (2012). Sovereign Debt Restructurings 1950-2010; Literature Survey, Data, and Stylized Facts. *IMF Working Paper No. 12/203.*

Zhao, Q. and T. Hastie (2019, jun). Causal Interpretations of Black-Box Models. *Journal of Business & Economic Statistics*, 1–19.

Zhuang, J. and J. M. Dowling (2002). Causes of the 1997 Asian Financial Crisis: What Can an Early Warning System Model Tell Us? *ERD WORKING PAPER SERIES NO. 26.*

APPENDIX

IMPLEMENTATIONAL DETAILS

I list the R packages and functions I used to perform the analysis summarised in the paper.

Logistic Regression. I used the *glm* implementation (Fitting Generalized Linear Models) from *stats* package, family=binomial.

Logistic Regression with LASSO Penalisation. I used the *cv.glmnet* implementation from *glmnet* package, with 5 folds for the cross-validation, using parameter *lambda* that gives minimum cross-validated error.

Random Forest. I used the *randomForest* implementation of Breiman's random forest algorithm (based on Breiman and Cutler's original Fortran code) for classification from *randomForest* package, with 10000 trees. Parameter m is set at the default value of the floor of the square root of the number of explanatory variables, which amounts to four in this application.

Breiman's Variable Importance. I used the *importance* component from the object of class *randomForest*, mentioned above.

Shapley Values. I used the *Shapley* implementation from *iml* package, which computes feature contributions for single predictions with the Shapley value. Function *Shapley* uses an object of class *Predictor*, which holds any machine learning model and the data to be used for analysing the model, and which can be created by a function *Predictor* (*iml* package). I report Shapley values per explanatory variable as means of absolute values of Shapley values across all observations.

Partial Dependence Plots. I used the *Partial* implementation from *pdp* package, which computes partial dependence functions for various model fitting objects.

Accumulated Local Effects Plots. I used the *FeatureEffects* implementation from *iml* package, method = "ale". Function *FeatureEffects* computes feature effects for multiple features at once. Function *FeatureEffects* uses an object of class *Predictor*, mentioned above.

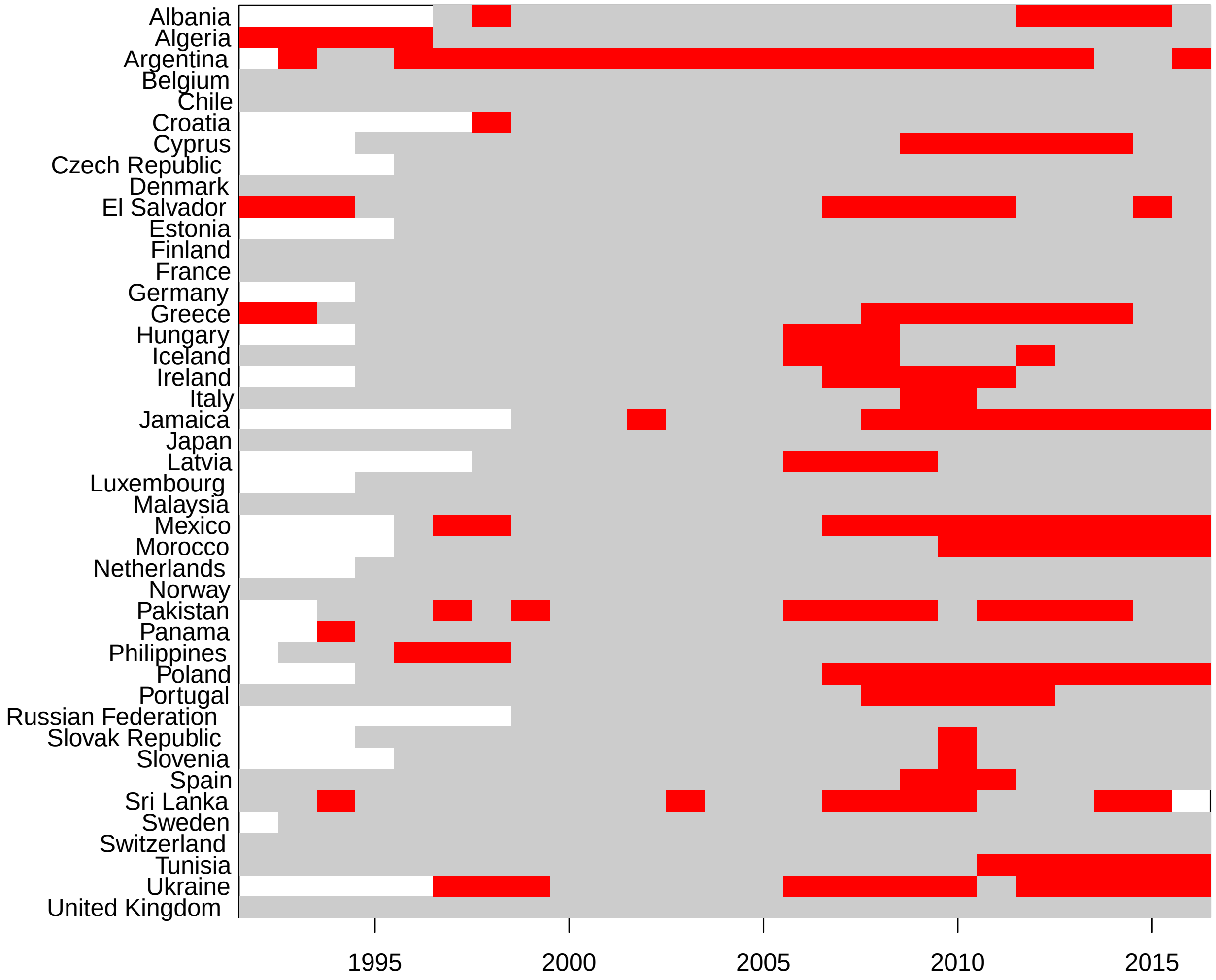


Figure A1: Stress and Non-stress Observations

Note: Red areas depict stress periods, grey areas non-stress periods, white areas missing data.

Table A1: Logit Models Fitted Using the Recursive Approach

Variables	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016
Oil price dynamics	-0.54** (0.19)	-0.44* (0.18)	-0.4* (0.18)	-0.55** (0.2)	-0.51** (0.19)	-0.47* (0.19)	-0.53** (0.17)	-0.54** (0.17)	-0.55*** (0.16)	-0.41** (0.14)	-0.41** (0.14)
US GDP dynamics	0.5* (0.22)	0.48* (0.22)	0.58. (0.3)	0.29 (0.26)	0.45. (0.24)	0.5* (0.23)	0.56* (0.22)	0.54* (0.21)	0.53** (0.2)	0.34* (0.17)	0.34* (0.17)
China GDP dynamics	0.38* (0.19)	0.96*** (0.25)	1.04*** (0.26)	0.81*** (0.2)	0.8*** (0.2)	0.78*** (0.19)	0.79*** (0.19)	0.74*** (0.19)	0.74*** (0.19)	0.67*** (0.19)	0.69*** (0.19)
VIX		0.54* (0.25)	0.97*** (0.28)	0.67** (0.21)	0.59** (0.19)	0.58** (0.19)	0.58** (0.18)	0.5** (0.18)	0.5** (0.17)	0.38* (0.16)	0.38* (0.15)
US interest rates			-0.28 (0.25)	-0.25 (0.26)	-0.58* (0.24)	-0.77** (0.24)	-1.02*** (0.22)	-1.12*** (0.23)	-1.14*** (0.23)	-0.96*** (0.2)	-1.01*** (0.21)
Oil price dynamics*Advanced	-7.96 (11722.02)	22.63 (1304.66)	2.92 (3.45)	2.53 (2.63)	0.62 (1.1)	0.72 (1.14)	0.92 (0.62)	0.89 (0.6)	0.38 (0.49)	0.41 (0.45)	0.43 (0.46)
US GDP dynamics*Advanced	-11.73 (5997.3)	-17.07 (790.01)	-7.58 (5.25)	-4.9 (3.71)	-1.02 (1.21)	-1.14 (1.25)	-1.41. (0.75)	-1.21. (0.71)	-0.49 (0.57)	-0.48 (0.49)	-0.47 (0.5)
China GDP dynamics*Advanced	24.46 (4058.87)	41.3 (2892.28)	1.55 (1.75)	1.12 (1.15)	0.76 (0.58)	0.85 (0.6)	0.86 (0.6)	0.77 (0.56)	0.44 (0.49)	0.54 (0.51)	0.59 (0.52)
VIX*Advanced		-20.97 (1315.6)	-3.57 (2.48)	-1.52 (1.24)	-0.61 (0.62)	-0.61 (0.63)	-0.73 (0.52)	-0.64 (0.51)	-0.29 (0.44)	-0.29 (0.39)	-0.27 (0.39)
US interest rates*Advanced			2.22 (2.52)	1.64 (1.8)	-1.62. (0.89)	-1.66. (0.95)	-1.67* (0.78)	-1.53* (0.77)	-1.18. (0.67)	-1.39* (0.69)	-1.43* (0.72)
Advanced	-96.78 (6056.5)	-86.65 (5544.66)	-10.45* (5.22)	-6.29* (2.64)	-4.62*** (1.08)	-4.85*** (1.1)	-4.83*** (1.03)	-4.71*** (0.96)	-3.99*** (0.71)	-4.14*** (0.71)	-4.19*** (0.7)
GDP dynamics	-0.32* (0.16)	-0.29. (0.16)	-0.34* (0.16)	-0.4* (0.17)	-0.33. (0.18)	-0.3. (0.18)	-0.31. (0.18)	-0.31. (0.17)	-0.27. (0.16)	-0.3. (0.16)	-0.32* (0.16)
CPI	0.25* (0.12)	0.25* (0.12)	0.35** (0.12)	0.36** (0.12)	0.37** (0.12)	0.37** (0.11)	0.34** (0.11)	0.34** (0.11)	0.29** (0.1)	0.25* (0.1)	0.26** (0.1)
Current account balance	-0.12 (0.28)	-0.13 (0.24)	-0.3 (0.23)	-0.29 (0.21)	-0.27 (0.2)	-0.31. (0.19)	-0.33. (0.18)	-0.4* (0.18)	-0.4* (0.17)	-0.43* (0.17)	-0.44** (0.16)
Net lending	0.17 (0.26)	0.03 (0.25)	0.02 (0.25)	-0.06 (0.24)	-0.17 (0.24)	-0.21 (0.23)	-0.08 (0.23)	-0.09 (0.22)	-0.1 (0.22)	-0.02 (0.2)	0.03 (0.2)
Unemployment change					0.03 (0.18)	0.12 (0.17)	0.12 (0.17)	0.12 (0.17)	0.14 (0.16)	0.12 (0.15)	0.11 (0.15)
Public debt							0.34 (0.21)	0.33 (0.2)	0.44* (0.2)	0.52** (0.19)	0.58** (0.19)
GDP dynamics*Advanced	-0.42 (2074.09)	0.37 (1.09)	0.35 (0.58)	0.08 (0.48)	0.83 (0.51)	0.88. (0.51)	0.89. (0.51)	0.79 (0.49)	0.43 (0.45)	0.47 (0.45)	0.49 (0.45)
CPI*Advanced	22.03 (1959.68)	8.24 (5.25)	2.53* (1.17)	0.87 (0.68)	1.16 (0.72)	1.17. (0.71)	1.46* (0.71)	1.24. (0.69)	0.59 (0.63)	0.73 (0.64)	0.73 (0.64)
Current account balance*Advanced	-22.27 (2975.58)	-0.22 (1.24)	-1.54* (0.78)	-1.45* (0.57)	-1.14* (0.46)	-1.23** (0.45)	-1.27** (0.44)	-1.38** (0.44)	-1.46*** (0.42)	-1.41*** (0.42)	-1.4*** (0.41)
Net lending*Advanced	-6.84 (2476.58)	0.43 (1.13)	-0.78 (0.68)	-0.68 (0.48)	-0.94* (0.44)	-1.05* (0.44)	-0.87* (0.43)	-0.65. (0.39)	-0.49 (0.35)	-0.56 (0.34)	-0.62. (0.34)
Unemployment change*Advanced					0.84* (0.38)	0.93* (0.39)	1.02** (0.39)	1.15** (0.39)	0.86* (0.35)	0.88** (0.34)	0.9** (0.33)
Public debt*Advanced							-0.02 (0.36)	0.15 (0.33)	0.03 (0.3)	-0.03 (0.3)	-0.09 (0.3)
(Intercept)	-1.87*** (0.23)	-1.78*** (0.21)	-1.73*** (0.2)	-1.64*** (0.2)	-1.55*** (0.18)	-1.5*** (0.18)	-1.4*** (0.17)	-1.36*** (0.17)	-1.28*** (0.17)	-1.17*** (0.16)	-1.14*** (0.15)
AIC	215.27	254.03	299.6	355.13	408.18	439.93	474.16	503.61	544.54	571.67	590.74
R2	0.4	0.4	0.4	0.38	0.37	0.38	0.39	0.39	0.38	0.38	0.39
ROC	0.92	0.92	0.92	0.91	0.90	0.90	0.91	0.91	0.90	0.90	0.91

Note: Years in the table relate to the cut-off year for the explanatory variables, for example, 2006 means that the latest observations used to fit the model contain explanatory variables until 2006 and the dependent variable until 2008. To avoid multicollinearity and resulting bias of the parameter estimates, the models are fitted using a database without three variables whose pairwise correlation with other variables exceed 65%.

Table A2: Logit Models Fitted Using Automatic Variable Selection LASSO Penalisation Procedure

Variables	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016
Oil price dynamics		-0.10	-0.03	-0.45	-0.43	-0.28	-0.38	-0.15	-0.40	-0.20	-0.14
China GDP dynamics		0.32	0.49	0.77	0.86	0.72	0.89	0.51	0.90	0.76	0.68
VIX			0.42	0.57	0.55	0.43	0.55	0.20	0.52	0.33	0.25
US interest rates				-0.39	-0.96	-0.93	-1.30	-0.95	-1.50	-1.27	-1.19
US GDP dynamics				-0.04	0.30	0.16	0.37	0.00	0.43	0.18	0.08
GDP per capita	-0.30	-0.56	-0.53	-0.88	-0.85	-0.79	-0.89	-0.82	-1.07	-1.10	-1.06
CPI	0.27	0.36	0.45	0.41	0.39	0.40	0.38	0.38	0.31	0.29	0.28
Current account balance		-0.16	-0.39	-0.57	-0.53	-0.50	-0.59	-0.61	-0.68	-0.69	-0.67
Consumption dynamics					0.02		0.12		0.13		
Net lending					-0.28	-0.34	-0.31	-0.25	-0.23	-0.17	-0.14
Unemployment change				0.15	0.28	0.33	0.35	0.30	0.36	0.32	0.29
GDP dynamics			-0.03	-0.22	-0.08		-0.21	-0.08	-0.30	-0.25	-0.24
Public debt				0.12	0.13	0.13	0.14	0.13	0.24	0.22	0.21
Export share dynamics				0.06	0.03		0.05		0.13	0.13	0.12
FX rate dynamics		0.01		0.14	0.22	0.16	0.22	0.11	0.21	0.15	0.16
Credit to GDP dynamics				-0.02	-0.05	-0.01	-0.06		-0.05	-0.06	-0.04
Interest on debt				0.12	0.07	0.04	0.13	0.05	0.14	0.18	0.19
(Intercept)	-2.58	-2.62	-2.56	-2.88	-2.78	-2.61	-2.71	-2.46	-2.71	-2.62	-2.55
ROC	0.83	0.85	0.86	0.87	0.86	0.86	0.87	0.87	0.88	0.87	0.88

Note: Years in the table relate to the cut-off year for the explanatory variables, for example, 2006 means that the latest observations used to fit the model contain explanatory variables until 2006 and the dependent variable until 2008. GDP per capita accounts for the level of economic development of countries.

Table A3: Logit Models Fitted Using Automatic Variable Selection LASSO Penalisation Procedure

Variables	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016
Oil price dynamics		-0.21	-0.07	-0.37	-0.49	-0.30	-0.37	-0.26	-0.44	-0.20	-0.20
China GDP dynamics		0.38	0.49	0.56	0.87	0.70	0.80	0.60	0.89	0.69	0.80
VIX			0.39	0.40	0.56	0.41	0.47	0.29	0.52	0.27	0.32
US interest rates				-0.17	-0.91	-0.83	-1.10	-0.96	-1.37	-1.05	-1.20
US GDP dynamics				-0.12	0.40	0.20	0.35	0.16	0.50	0.17	0.24
Advanced	-0.59	-1.58	-1.44	-1.55	-1.62	-1.46	-1.56	-1.56	-1.77	-1.72	-1.82
CPI	0.17	0.30	0.40	0.39	0.37	0.38	0.37	0.37	0.30	0.27	0.26
Current account balance		-0.34	-0.53	-0.66	-0.67	-0.62	-0.72	-0.75	-0.81	-0.80	-0.81
Consumption dynamics				-0.03	0.01	-0.01	0.04		0.12		0.01
Net lending					-0.33	-0.38	-0.35	-0.31	-0.27	-0.19	-0.16
Unemployment change				0.15	0.33	0.37	0.38	0.36	0.41	0.35	0.35
GDP dynamics			-0.10	-0.20	-0.12	-0.04	-0.18	-0.14	-0.34	-0.27	-0.35
Public debt				0.09	0.14	0.14	0.15	0.17	0.29	0.27	0.29
Export share dynamics					0.03				0.14	0.12	0.18
FX rate dynamics		0.08		0.03	0.17	0.11	0.12	0.07	0.16	0.09	0.14
Credit to GDP dynamics					-0.03		-0.03		-0.05	-0.06	-0.06
Interest on debt							0.05		0.10	0.15	0.20
(Intercept)	-2.21	-1.91	-1.85	-1.82	-1.86	-1.76	-1.74	-1.61	-1.63	-1.52	-1.52
ROC	0.85	0.88	0.89	0.89	0.87	0.88	0.88	0.88	0.89	0.89	0.89

Note: Years in the table relate to the cut-off year for the explanatory variables, for example, 2006 means that the latest observations used to fit the model contain explanatory variables until 2006 and the dependent variable until 2008. Binary dummy taking value of one for advanced and zero for emerging countries accounts for the level of economic development of countries.

Table A4: Average Prediction Accuracy of Early Warning Models for Years 2009–2018 for Logit Models, Different Weights of Sensitivity Relative to Specificity

Weight of sensitivity relative to specificity	Equal weights	50% higher weight for sensitivity	100% higher weight for sensitivity
2009 (AUROC 0.75)			
% of correctly classified stress episodes	81.82	81.82	81.82
% of correctly classified tranquil episodes	59.38	56.25	56.25
Average	70.60	69.03	69.03
2010 (AUROC 0.79)			
% of correctly classified stress episodes	42.86	78.57	78.57
% of correctly classified tranquil episodes	82.76	68.97	68.97
Average	62.81	73.77	73.77
2011 (AUROC 0.70)			
% of correctly classified stress episodes	20.00	20.00	60.00
% of correctly classified tranquil episodes	100.00	96.43	60.71
Average	60.00	58.21	60.36
2012 (AUROC 0.75)			
% of correctly classified stress episodes	31.25	56.25	56.25
% of correctly classified tranquil episodes	85.19	77.78	77.78
Average	58.22	67.01	67.01
2013 (AUROC 0.88)			
% of correctly classified stress episodes	92.31	92.31	92.31
% of correctly classified tranquil episodes	80.00	80.00	80.00
Average	86.15	86.15	86.15
2014 (AUROC 0.90)			
% of correctly classified stress episodes	92.31	92.31	100.00
% of correctly classified tranquil episodes	56.67	56.67	36.67
Average	74.49	74.49	68.33
2015 (AUROC 0.95)			
% of correctly classified stress episodes	90.91	90.91	100.00
% of correctly classified tranquil episodes	84.38	84.38	68.75
Average	87.64	87.64	84.38
2016 (AUROC 0.89)			
% of correctly classified stress episodes	90.91	90.91	90.91
% of correctly classified tranquil episodes	75.00	75.00	75.00
Average	82.95	82.95	82.95
2017 (AUROC 0.89)			
% of correctly classified stress episodes	100.00	100.00	100.00
% of correctly classified tranquil episodes	32.35	32.35	29.41
Average	66.18	66.18	64.71
2018 (AUROC 0.90)			
% of correctly classified stress episodes	85.71	85.71	85.71
% of correctly classified tranquil episodes	80.00	80.00	71.43
Average	82.86	82.86	78.57
2009-2018 (AUROC 0.84)			
% of correctly classified stress episodes	72.81	78.88	84.56
% of correctly classified tranquil episodes	73.57	70.78	62.50
Average	73.19	74.83	73.53

Note: Results presented are based on the better-performing logit model fitted manually presented in Table 4, i.e. including the GDP per capita among the explanatory variables.

Table A5: Average Prediction Accuracy of Early Warning Models for Years 2009–2018 for Random Forest, Different Weights of Sensitivity Relative to Specificity

Weight of sensitivity relative to specificity	Equal weights	50% higher weight for sensitivity	100% higher weight for sensitivity
2009 (AUROC 0.86)			
% of correctly classified stress episodes	45.45	90.91	90.91
% of correctly classified tranquil episodes	90.63	68.75	68.75
Average	68.04	79.83	79.83
2010 (AUROC 0.84)			
% of correctly classified stress episodes	50.00	92.86	92.86
% of correctly classified tranquil episodes	89.66	65.52	65.52
Average	69.83	79.19	79.19
2011 (AUROC 0.73)			
% of correctly classified stress episodes	100.00	100.00	100.00
% of correctly classified tranquil episodes	7.14	0.00	0.00
Average	53.57	50.00	50.00
2012 (AUROC 0.75)			
% of correctly classified stress episodes	50.00	50.00	93.75
% of correctly classified tranquil episodes	77.78	77.78	40.74
Average	63.89	63.89	67.25
2013 (AUROC 0.97)			
% of correctly classified stress episodes	100.00	100.00	100.00
% of correctly classified tranquil episodes	66.67	56.67	56.67
Average	83.33	78.33	78.33
2014 (AUROC 0.91)			
% of correctly classified stress episodes	92.31	92.31	100.00
% of correctly classified tranquil episodes	80.00	73.33	66.67
Average	86.15	82.82	83.33
2015 (AUROC 0.98)			
% of correctly classified stress episodes	100.00	100.00	100.00
% of correctly classified tranquil episodes	90.63	90.63	90.63
Average	95.31	95.31	95.31
2016 (AUROC 0.96)			
% of correctly classified stress episodes	100.00	100.00	100.00
% of correctly classified tranquil episodes	87.50	87.50	81.25
Average	93.75	93.75	90.63
2017 (AUROC 0.93)			
% of correctly classified stress episodes	88.89	100.00	100.00
% of correctly classified tranquil episodes	82.35	76.47	76.47
Average	85.62	88.24	88.24
2018 (AUROC 0.94)			
% of correctly classified stress episodes	85.71	85.71	85.71
% of correctly classified tranquil episodes	82.86	82.86	82.86
Average	84.29	84.29	84.29
2009-2018 (AUROC 0.89)			
% of correctly classified stress episodes	81.24	91.18	96.32
% of correctly classified tranquil episodes	75.52	67.95	62.95
Average	78.38	79.56	79.64

Note: Results presented are based on the better-performing random forest model presented in Table 4, i.e. including the GDP per capita among the explanatory variables.

Table A6: Average Prediction Accuracy of Early Warning Models for Years 2009–2018 for Logit LASSO, Different Weights of Sensitivity Relative to Specificity

Weight of sensitivity relative to specificity	Equal weights	50% higher weight for sensitivity	100% higher weight for sensitivity
	2009 (AUROC 0.88)		
% of correctly classified stress episodes	9.09	100.00	100.00
% of correctly classified tranquil episodes	100.00	12.50	12.50
Average	54.55	56.25	56.25
	2010 (AUROC 0.76)		
% of correctly classified stress episodes	35.71	35.71	92.86
% of correctly classified tranquil episodes	96.55	96.55	44.83
Average	66.13	66.13	68.84
	2011 (AUROC 0.60)		
% of correctly classified stress episodes	0.00	26.67	26.67
% of correctly classified tranquil episodes	100.00	89.29	89.29
Average	50.00	57.98	57.98
	2012 (AUROC 0.77)		
% of correctly classified stress episodes	31.25	31.25	68.75
% of correctly classified tranquil episodes	85.19	85.19	77.78
Average	58.22	58.22	73.26
	2013 (AUROC 0.88)		
% of correctly classified stress episodes	84.62	92.31	100.00
% of correctly classified tranquil episodes	80.00	70.00	26.67
Average	82.31	81.15	63.3
	2014 (AUROC 0.91)		
% of correctly classified stress episodes	100.00	100.00	100.00
% of correctly classified tranquil episodes	33.33	33.33	33.33
Average	66.67	66.67	66.67
	2015 (AUROC 0.94)		
% of correctly classified stress episodes	100.00	100.00	100.00
% of correctly classified tranquil episodes	68.75	68.75	68.75
Average	84.38	84.38	84.38
	2016 (AUROC 0.89)		
% of correctly classified stress episodes	72.73	72.73	72.73
% of correctly classified tranquil episodes	84.38	81.25	81.25
Average	78.55	76.99	76.99
	2017 (AUROC 0.89)		
% of correctly classified stress episodes	88.89	88.89	100.00
% of correctly classified tranquil episodes	70.59	70.59	44.12
Average	79.74	79.74	72.06
	2018 (AUROC 0.90)		
% of correctly classified stress episodes	85.71	85.71	85.71
% of correctly classified tranquil episodes	71.43	62.86	62.86
Average	78.57	74.29	74.29
	2009-2018 (AUROC 0.84)		
% of correctly classified stress episodes	60.80	73.33	84.67
% of correctly classified tranquil episodes	79.02	67.03	54.14
Average	69.91	70.18	69.40

Note: Results presented are based on the better-performing LASSO logit model presented in Table 4, i.e. including the GDP per capita among the explanatory variables.

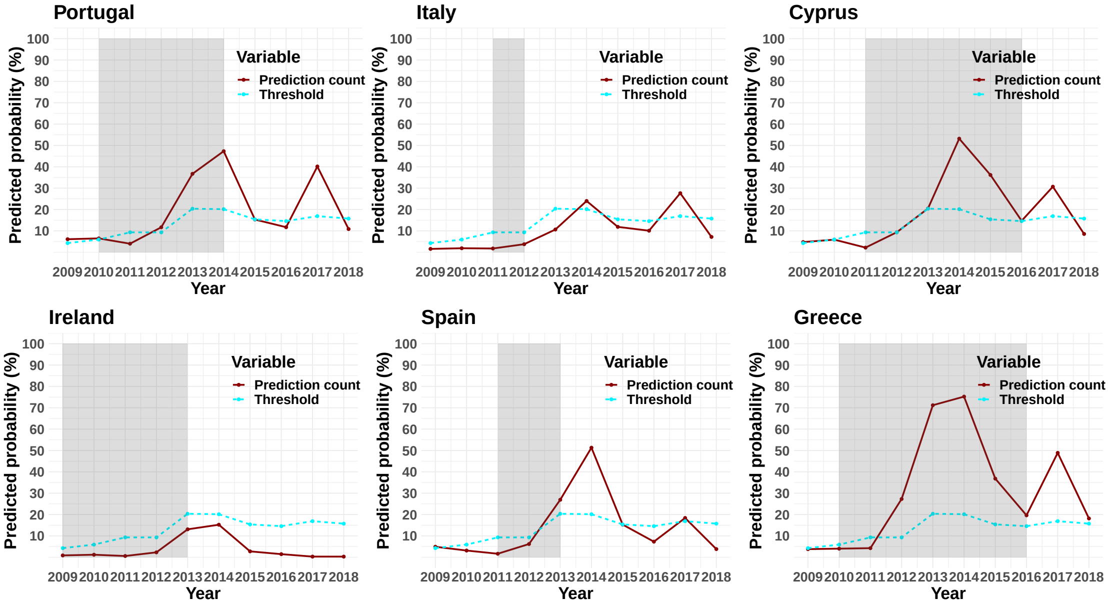


Figure A2: Prediction Accuracy of Early Warning Models Based on Logit Regression

Note: Grey areas mean that given periods are classified as stress events by the definition adopted. Blue lines show the threshold calibrated for a given year, maximising the weighted sum of sensitivity and specificity on the training set. Purple lines show the probability of a fiscal stress event given by the model used. Results presented are based on logit models including the GDP per capita among the explanatory variables.

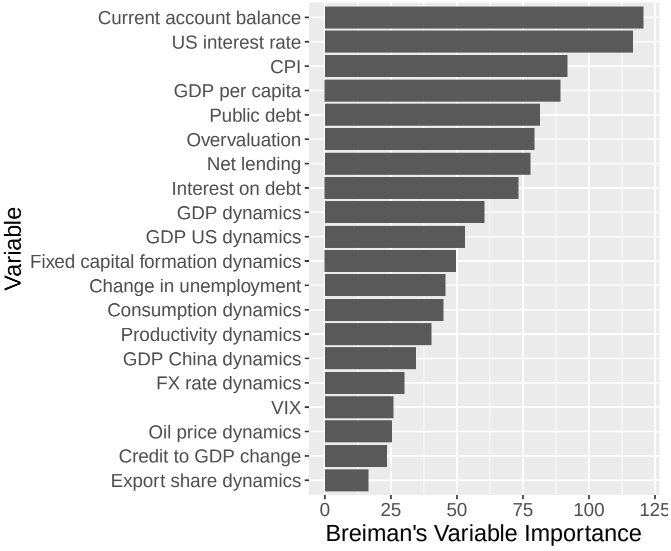


Figure A3: Variable Importance of Predictors Included in the Random Forest-based Early Warning System

Note: The variable's importance is understood as the average deterioration of the random forest model's performance when a given variable is excluded, compared with the case in which the variable is included, here measured by mean decrease in classification accuracy. Results presented are based on random forest models using GDP per capita to control for the level of economic development of countries, fitted on the entire dataset.

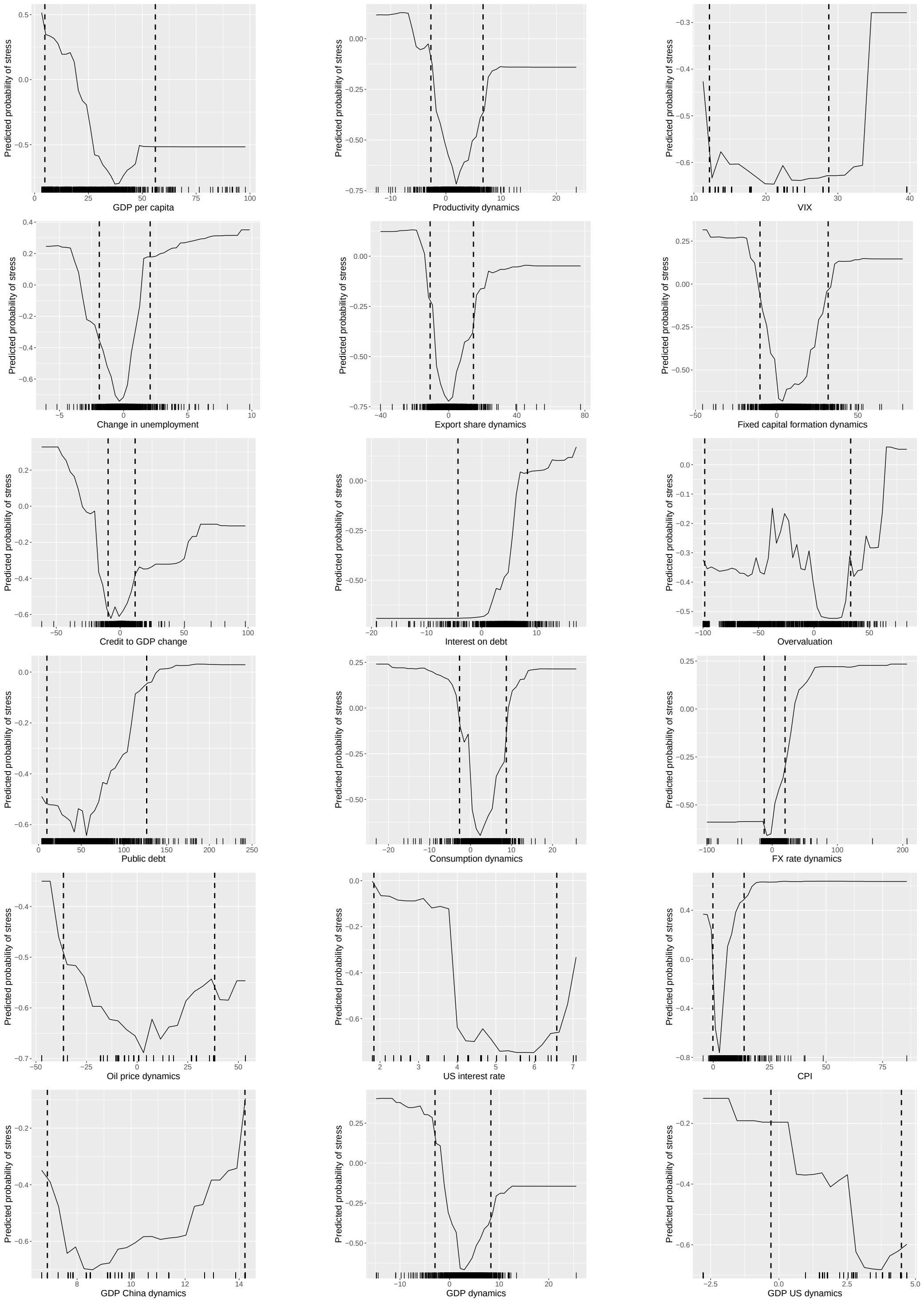


Figure A4: Partial Dependence Plots of Predictors Included in the Random Forest Model

Note: Results presented are based on random forest models using GDP per capita to control for the level of economic development of countries, fitted on the entire dataset.

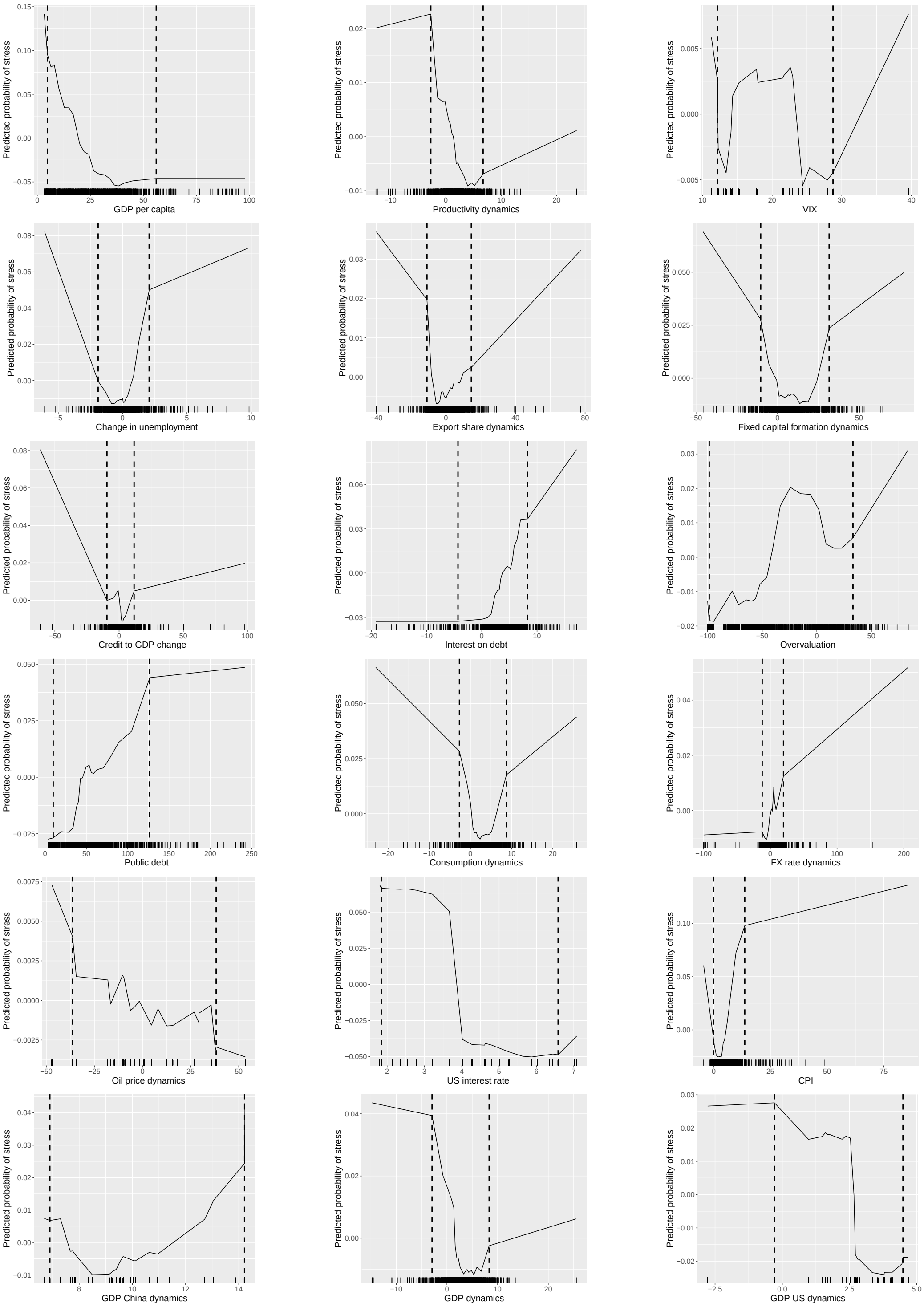


Figure A5: Accumulated Local Effects Plots of Predictors Included in the Random Forest Model

Note: Results presented are based on random forest models using GDP per capita to control for the level of economic development of countries, fitted on the entire dataset.

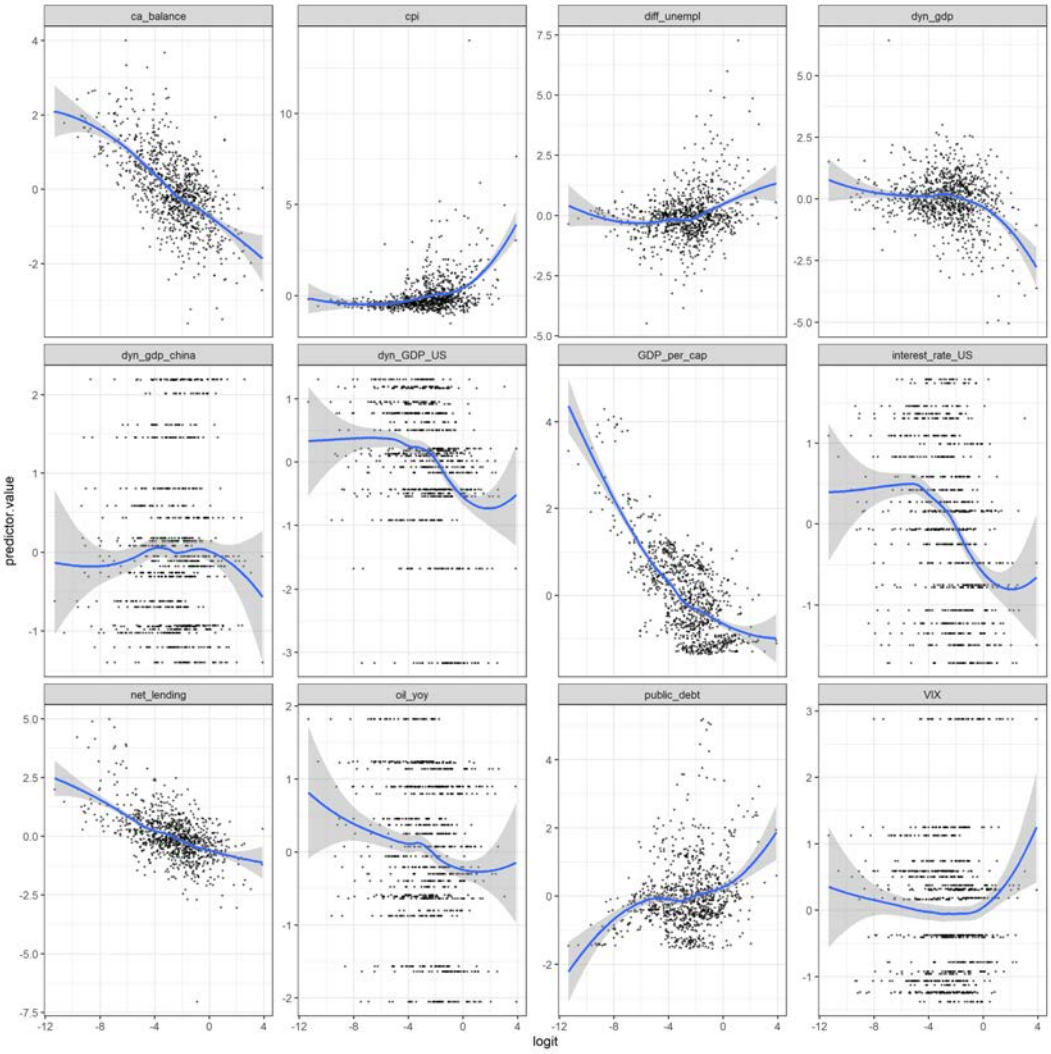


Figure A6: Scatter Plot Between Each Predictor and the Logit Values for the Logit Model
Note: Results presented are based on a logit model using GDP per capita to control for the level of economic development of countries, fitted on the entire dataset.

Acknowledgements

I would like to thank an anonymous referee and the Editorial Board of the ECB Working Paper Series for their very useful comments and suggestions which I incorporated into this paper. The study has also benefited from comments and discussions with seminar participants at the European Central Bank (in Directorate General Economics and at Early Warning Models Workshop organised by the Directorate General Macroprudential Policy and Financial Stability), Warsaw School of Economics (Collegium of Management and Finance), the OECD (Department of Economics), the National Bank of Poland (Research and Financial Innovation Department) and the Ministry of Finance of Poland (Macroeconomic Policy Department). I am very grateful to Janusz Ostaszewski, Joanna Beza-Bojanowska, Gianni de Nicolo, Carsten Detken, Jon Danielsson and Anna Matusik for valuable comments. All views expressed in this paper are mine and do not necessarily correspond to those of the European Central Bank.

Barbara Jarmulska
European Central Bank, Frankfurt am Main, Germany; Warsaw School of Economics; email: barbara.jarmulska@ecb.europa.eu

Postal address 60640 Frankfurt am Main, Germany
Telephone +49 69 1344 0
Website www.ecb.europa.eu

This paper can be downloaded without charge from www.ecb.europa.eu, from the Social Science Research Network electronic library or from RePEc: Research Papers in Economics. Information on all of the papers published in the ECB Working Paper Series can be found on the ECB's website.

PDF ISBN 978-92-899-4051-1 ISSN 1725-2806 doi:10.2866/214327 QB-AR-20-060-EN-N