Credit Growth, the Yield Curve and Financial Crisis Prediction: Evidence from a Machine Learning Approach*

Kristina Bluwstein
Bank of England

Marcus Buckmann
Bank of England

Andreas Joseph
Bank of England

Miao Kang
Bank of England

Sujit Kapadia
European Central Bank

Özgür Şimşek
University of Bath

This draft: January 2, 2020

Abstract

We develop early warning models for financial crisis prediction using machine learning techniques on macrofinancial data for 17 countries over 1870-2016. Machine learning models mostly outperform logistic regression in out-of-sample predictions and forecasting. We identify economic drivers of our machine learning models using a novel framework based on Shapley values, uncovering nonlinear relationships between the predictors and crisis risk. Throughout, the most important predictors are credit growth and the slope of the yield curve, both domestically and globally. A flat or inverted yield curve is of most concern when nominal interest rates are low and credit growth is high.

Keywords: machine learning; financial crises; financial stability; credit growth; yield curve; Shapley values; out-of-sample prediction.

JEL Classification: C40; C53; E44; F30; G01.

*The views expressed in this paper are those of the authors and do not reflect those of the Bank of England or the European Central Bank. We are grateful to David Aikman, David Bholat, Michael Bordo, Ambrogio Cesa-Bianchi, Julia Giese, Klaus Peter Hellwig, Christopher Kurz, Ricardo Reis, Hélène Rey and Rhiannon Sowerbutts for helpful comments and suggestions. We would also like to thank seminar participants of the "Forecasting & Empirical Methods" at the NBER Summer Institute (Cambridge, MA, July 2019), "Forecasting Political and Economic Crises: Economics meets Machine Learning" at the 2019 GSE Summer Forum (Barcelona, June 2019), EEA conference (Manchester, August 2019), "Nontraditional Data, Machine Learning, and Natural Language Processing in Macroeconomics" at the Federal Reserve Board (Washington DC, October 2019), "Big Data and Forecasting Economic developments" at the European Commission (Ispra, May 2019), the ESCoE Conference on Economic Measurement (London, May 2019), "Modelling with Big Data and Machine Learning" (Bank of England, London, November 2018) and the University of Southampton (Southampton, December 2018) for useful comments and discussions.

1 Introduction

Financial crises have huge economic and social costs (Hoggarth et al., 2002; Ollivaud and Turner, 2015; Laeven and Valencia, 2018; Aikman et al., 2018). Spotting their warning signs sufficiently early is therefore of great importance for policy makers. Doing so can facilitate the timely activation of countercyclical macroprudential policies, and reduce the likelihood and severity of financial crises in the face of rising risks (Giese et al., 2013; Cerutti et al., 2017; Akinci and Olmstead-Rumsey, 2018). But identifying a reliable set of early warning predictors is challenging for several reasons. First, there are a relatively limited set of observed crises, which makes robust modelling difficult. Second, crisis indicators often only flash red when it is already too late to intervene. Third, it can be challenging to distil complicated early warning models into simple and transparent indicators that can help guide timely intervention by macroprudential authorities.

This paper uses machine learning models to tackle these issues. Despite the relatively small sample of crisis observations, we show that these models generally outperform a baseline logistic regression in both the out-of-sample prediction and the forecasting of financial crises on a multi-year horizon. Due to their greater flexibility, machine learning models have the advantage that they may uncover important nonlinear relationships and variable interactions which may be difficult to identify using classical techniques. As financial crises are rare and extreme events that are likely to exhibit unknown nonlinear dependencies prior to their crystallisation (Alessi and Detken, 2018), such methods are particularly well suited for developing a reliable early warning system.

To the best of our knowledge, our paper is the first to provide a rigorous inference analysis of *how* black box machine learning models predict financial crises by decomposing their predictions into the contributions of individual variables using the Shapley value framework (Strumbelj and Kononenko, 2010; Joseph, 2019).[1] This approach allows us both to identify the key economic drivers of our models and to test those statistically. It also helps to tackle a key challenge faced by policy makers in using machine learning models to inform their decisions, because it provides narratives which can be used to

[1]Other approaches to decompose the predictions of models into the contributions of individual variables include Ribeiro et al. (2016) and Shrikumar et al. (2017). But only the Shapley value framework has a set of appealing analytical properties (see Lundberg and Lee (2017)) while being applicable to any model family.

justify policy actions which may be partially based on such models. Such economic reasoning is important in reaching a rounded assessment which integrates insights from machine learning models with other models, data, market intelligence, and judgement. And it is essential for transparancy and accountability, given that public policymakers need to explain the rationale for their decisions and cannot simply point to black box models to justify their interventions.

The paper aims to identify robust predictors for financial crises one to two years in advance. This gives time to implement policies that can potentially avert a crisis altogether or dampen its negative consequences. We first estimate a benchmark logistic regression model using the Macrohistory Database by Jordà et al. (2017), which covers macroeconomic and financial variables from 17 advanced economies over more than 140 years and contains a binary financial crisis variable. We then test the out-of-sample performance of a range of machine learning models: decision trees, random forests, extremely randomised trees, support vector machines (SVM), and artificial neural networks. We find that, with the exception of individual decision trees, all machine learning models outperform the logistic regression. Among other findings, the best-performing machine learning model, extremely randomised trees, also correctly predicts the global financial crisis of 2007–2008, giving differentiated signals between countries that reflect different economic realities and outcomes.

Investigating the drivers of our models, we find that credit growth and the slope of the yield curve are the most important predictors for financial crises across a diverse set of models. While the importance of domestic credit growth is well known in the literature (Borio and Lowe (2002); Drehmann et al. (2011); Schularick and Taylor (2012); Aikman et al. (2013); Jordà et al. (2013, 2015b); Giese et al. (2014)), the role of the yield curve has been far less explored. We find that the flatter or more inverted the yield curve is, the higher the chance of a crisis. This could reflect the search for yield and increased risk-taking that can often be observed prior to financial crises. Both credit growth and the yield curve slope are also highly important predictors at the global level, with one major difference. The global slope provides a robust signal over the entire period from the 1870s until the present, while global credit is a key predictor of the global financial crisis of 2007–2008 but less important for the prediction of other crises. Our results also indicate that stock prices, money and the current account have lower overall predictive power when controlling for other factors. House price, by contrast, slightly do improve model performance in the post-1945 period, but not robustly, i.e. this may be an important indicator for some countries at certain times but not throughout the full sample. More generally, we also leverage our long sample to explore how the importance of different variables has varied over time.

We uncover relatively simple and intuitive nonlinear relationships and interactions for our key indicators. For example, crisis probability increases materially at high levels of global credit growth but this variable has nearly no effect at low or medium levels. Similarly, interactions seem to be important—particularly between global and domestic variables. For example, many crises fall into an environment of strong domestic credit growth and a globally flat or inverted yield curve. We also find that a flat or inverted yield curve is more concerning when nominal yields are at low levels.

Our paper develops from the extensive literature on early warning systems for crisis prediction that applies classical regression techniques or classifies leading indicators in a binary way according to whether they correctly signalled crises or generated false alarms (see e.g. Kaminsky and Reinhart (1999); Bussiere and Fratzscher (2006); Drehmann et al. (2011); Frankel and Saravelos (2012); Schularick and Taylor (2012); Drehmann and Juselius (2014); Babecký et al. (2014); Giese et al. (2014)). This literature typically identified domestic private credit or credit-to-GDP growth and indebtedness as key predictors of financial crises, with more recent work (Alessi and Detken, 2011; Duca and Peltonen, 2013; Cesa-Bianchi et al., 2019) also highlighting the importance of global credit growth in predicting crisis after 1970. Our results are in line with these findings.

The domestic yield curve is a well-established leading indicator for economic recessions (Estrella and Hardouvelis, 1991; Wright, 2006; Rudebusch and Williams, 2009). But, only a few studies have have linked it empirically to the risk of financial crises and these studies have not discussed this result in detail or examined the role of the yield curve on the global level (Babecký et al., 2014; Joy et al., 2017; Vermeulen et al., 2015). At the same time, our work is compatible with several theoretical models which investigate the relationships between nominal risk-free returns, risk taking, credit and financial stability (Aikman et al., 2015; Martinez-Miera and Repullo, 2017; Coimbra and Rey, 2017; Korinek and Novak, 2017). These models tend to highlight the importance of credit booms, particularly in a low interest rate environment, counter-cyclical risk premia and search-for-yield behaviour prior to financial crises.

A more recent line of work has started to use machine learning techniques for financial crisis prediction. Several studies apply random forests, a well-established machine learning model that uses decision trees. For example, Alessi and Detken (2018) employ them to predict banking crises in a quarterly dataset spanning 1970–2012 across EU countries, while Joy et al. (2017) use them to predict banking and currency crises in 36 advanced economies between 1970 and 2010 and Ward (2017) uses them to predict financial crises in the long-run Macrohistory Database and two post-1970 datasets.[2] Other machine learning models have also been used to predict financial crises. Adaboost, with decision trees as its base model, was shown to outperform logistic regression in forecasting financial crises in 100 advanced and emerging economies between 1970 and 2017 (Casabianca et al., 2019). Tölö et al. (2019) shows that recurrent neural networks yield better early warning models than both ordinary neural networks and logistic regression in the Macrohistory database. And Fouliard et al. (2019) combine several predictive models, including regression and decision trees, to forecast financial crises in seven countries between 1985 and 2018. While all these of studies find that machine learning methods generally outperform a regression regression approach, Beutel et al. (2018) reach the opposite conclusion. They find that logistic regression consistently outperforms a set of machine learning models in forecasting financial crises based on quarterly post-1970 data.

We contribute to the above literature in four main ways. First, we believe that our study is the first to compare a very diverse set of machine learning models on a long-run dataset of more than 140 years in both out-of-sample cross-validation and forecasting testing. Second, we are the first to tackle the black box critique of machine learning models for crisis prediction by identifying the key economic drivers of our models within a well-defined framework. Third, we uncover novel economic relationships which speak to the drivers of financial crises. In particular, we find that the slope of the yield curve is an important predictor for crises even after controlling for recessions and examine potential

[2] Other examples of tree applications in economics are Manasse and Roubini (2009) and Savona and Vezzoli (2015) for sovereign crises, and Duttagupta and Cashin (2011) for banking crises in emerging and developing countries.

reasons for this in detail. We also identify the importance of yield curves globally for financial crisis risk. Fourth, we identify important interactions and nonlinearities of key variables. We find particularly strong relationships between global factors and domestic indicators like the global slope of the yield curve and domestic credit.

The remainder of the paper is structured as follows. Section 2 describes the dataset and reviews the literature on those variables that we choose as predictors. Section 3 presents the benchmark logistic regression. Section 4 outlines the methodology and provides a brief description of the different machine learning models applied and the Shapley value framework. Section 5 compares the predictive performance of all models. Section 6 investigates the importance of the predictors using Shapley values. Section 7 looks at the robustness of indicators in a macroprudential policy context and investigates in detail the role of the yield curve. Section 8 concludes.

2 Dataset and variable selection

2.1 The financial crisis dataset

Financial crises are rare events. While there are a handful of truly global financial crises such as the Great Depression and the Global Financial Crisis of 2007–08, the majority of crises mostly occur in a single country or a small cluster of countries. Given the infrequency of financial crises, we exploit the longest cross-country dataset available.

The Jordà-Schularick-Taylor Macrohistory Database (Jordà et al., 2017) contains annual macroeconomic and financial measures from 17 developed countries between 1870 and 2016 (see Figure I).[3] For each of the 2499 country-year observations, the dataset contains a binary variable indicating whether (n=90) or not (n=2409) the country suffered from a financial crisis in a particular year. The authors define financial crises as "events during which a country's banking sector experiences bank runs, sharp increases in default rates accompanied by large losses of capital that result in public intervention, bankruptcy, or forced merger of financial institutions." The crisis variable synthesises several previous databases (Bordo et al., 2001; Laeven and Valencia, 2008; Reinhart and

[3]We obtained the third version of this dataset in January 2019 from http://www.macrohistory.net.

Rogoff, 2009; Cecchetti et al., 2009) and has been confirmed by experts for the respective countries (Schularick and Taylor, 2012).

Since we wish to predict crises ahead of time, we set our binary outcome variable to positive values for one and two years before the beginning of the crisis. The actual year of the crisis and the following four years are excluded from the analysis to avoid post-crisis bias (Bussiere and Fratzscher, 2006), which would otherwise mean that years where the economy is healthy are in the same class as post-crisis years, where the economy is recovering and still affected by crisis dynamics. For the same reason, we also exclude all observations between 1933 and 1939, the later years of the Great Depression, which is generally considered to have lasted from 1929 to 1939. (Bernstein, 1987; Gordon and Krenn, 2010). The two world wars (1914–1918, 1939–1945) are also excluded. To ensure full coverage, we also exclude all observations with any missing values of the predictors, which particularly restricts the sample in the 19th century. Figure I summarises these exclusions.

After these exclusions, 1249 observations remain from the original dataset and constitute our baseline dataset. Of these observations, 95 have a positive class value indicating the build-up phase to 49 distinct crises.

2.2 Explanatory variables and related literature

We treat the prediction of crises as a classification problem and model the ⟨country, year⟩ pairs as independent observations. We explore the following predictors in our baseline analysis (see also Table I for a summary): the slope of the yield curve (difference of short and long-term interest rates[4]), credit (loans to the non-financial private sector), stock prices, the debt service ratio (credit × long-term interest rate over GDP), consumption, investment, the current account, public debt, broad money, and CPI. The slope of the yield curve is left in levels, while CPI, stock prices, and real consumption per capita are transformed into percentage growth rates of the given indices. All other variables, i.e. credit, money, public debt, debt servicing, investment, and the current account, are differences of GDP-ratios. Variable transformations address potential issues of compa-

[4] Short-term rates are either risk-free or market-based rates depending on data availability. Our results are robust to the type of short-term rates used. Long-term rates refer to long-term government debt.

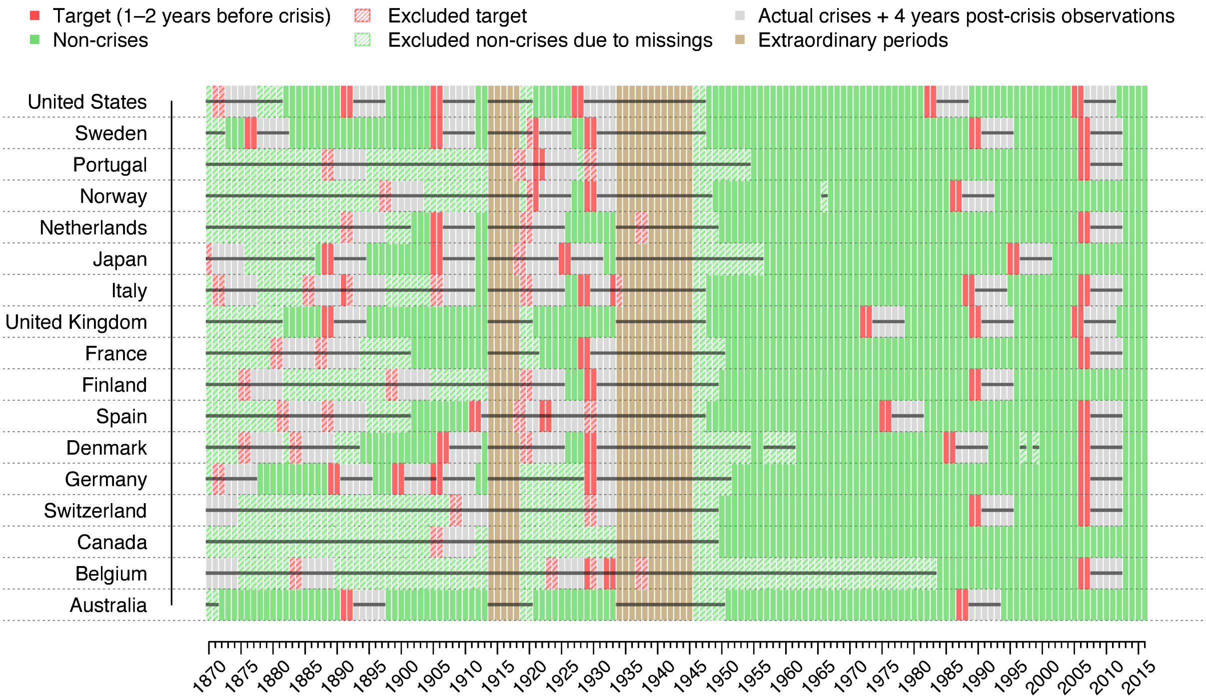

FIGURE I: Observations in the dataset. Green bars show non-crisis observations, red bars show the target 1–2 years before crisis. All excluded observations are highlighted by the thick black lines. We exclude: (i) the actual crisis observations and the following four years (grey); (ii) observations of both world wars and the second half of the great depression (brown); and (iii) observations with missing values of the predictors (shaded green). Shaded red bars show target observations excluded for any of these three reasons.

rability and non-stationarity. In addition to these 10 domestic variables, we define two global variables, namely global credit growth and the global slope of the yield curve. They are computed for a country-year pair $\langle c, y \rangle$ by the mean credit to GDP growth (mean slope of the yield curve) in all countries except c in year y.[5] Finally, Table I also lists a small number of additional variables that we consider in various extensions and robustness checks. In what follows, we briefly discuss the potential economic relevance of each of our predictors with reference to the relevant literature.

Credit growth has been found to be a crucial predictor of financial crises (Borio and Lowe, 2002; Drehmann et al., 2011; Schularick and Taylor, 2012; Aikman et al., 2013). High credit growth often reflects a period of excessive risk taking, which can subsequently

[5]Appendix B.2 discusses the computation of the global variables in detail.

lead to financial instability (Minsky, 1977). Indeed, Aliber and Kindleberger (2015) describes financial crises as "credit booms gone wrong". Financial accelerator effects (Bernanke and Blinder, 1992) can even mean that a rather small credit bubble may be very detrimental if a negative spiral amplifies an initial shock. And, collateral constraints may serve as a further amplifier (Kiyotaki and Moore, 1997; Bernanke et al., 1999).

Beyond domestic credit growth, several studies have identified the importance of global credit growth. Financial crises often occur on an international scale and may reflect global financial cycles (Rey, 2015), or be driven by cross-country spillovers rather than only domestic imbalances. For example, Cesa-Bianchi et al. (2019) find an increasing correlation of credit growth across countries over time and show that global credit growth is an even stronger predictor for financial crises than domestic credit. Similarly, Alessi and Detken (2011) and Duca and Peltonen (2013) show that the global credit gap is an effective early warning signal.

Rising asset prices—including equity and house prices—are also often associated with pre-crisis periods (Aliber and Kindleberger, 2015; Reinhart and Rogoff, 2008). In particular, rapid rises of asset prices could indicate the formation of a bubble.

The slope of the yield curve, i.e. the difference between the long and short-term interest rate, is often seen as a strong predictor of an impending economic recession (Estrella and Hardouvelis, 1991; Wright, 2006), especially of a longer horizon of 12–18 months (Rudebusch and Williams, 2009; Liu and Moench, 2016; Croushore and Marsten, 2016). But while some early warning models for financial crises have identified the slope of the yield curve as an important predictor of financial crises (Babecký et al., 2014; Joy et al., 2017; Vermeulen et al., 2015), they have not explored the drivers of its predictive power in detail.

The yield curve reflects expectations of the future path of short-term interest rates, as well as a risk premium (i.e. the term premium) for holding an asset for a longer duration. In normal times, the slope is positive, which means that long-term interest rates are higher than short-term rates. But there are two distinct reasons why a flat or negative sloping yield curve might be predictive of financial crises, separate from the possible signal on the macroeconomic outlook.

First, for a given macroeconomic environment, a flatter yield curve tends to be associated with lower net interest margins and weaker banking sector profitability (Adrian et al., 2010; Borio et al., 2017). This may potentially directly affect the resilience of the banking sector. It could also lead to a contraction in credit supply with implications for real economic activity. If these effects are severe enough, the slope of the yield curve might be a useful predictor for financial crises.

Second, a flat or inverted yield curve may often be associated with low term premia. In such an environment, investors might have to search for riskier investment, rather than longer maturity, to achieve higher absolute returns, and they may not be properly compensated for their increased risk exposure. For example, Coleman et al. (2008) find that house prices in the United States rose with the flattening of the yield curve prior to the global crisis of 2007–2008. They suggest (p. 286), that "the hunger for spread during this period of a flat yield curve could have been fuelling sub-prime and other alternative mortgage activity". Such a system-wide build-up of under-priced risk leaves the financial system highly exposed to a sharp correction which may result in a crisis. In this regard, the *levels* of short and long-term interest rates may also be important. For instance, low nominal interest rates may also drive excessive risk taking in the financial system as banks and other intermediaries search for yield (Taylor, 2009; Adrian and Shin, 2010; Borio and Zhu, 2012). While we asses the importance of the level of short and long-term interest rates in more detail later in the paper, we do not incorporate them into the baseline model to avoid a direct linear dependency of the yield curve slope and the interest rates.

Beyond the domestic slope, we also test a global slope indicator. Several studies have shown strong dependencies of interest rates across countries (Frankel et al., 2004; Obstfeld et al., 2005) and have found a systematic global factor of the yield curve (Diebold et al., 2008; Abbritti et al., 2018). At a global level, a flattening of the yield curve could point towards a global economic slowdown, which could be a likely trigger for existing financial vulnerabilities. It could also precipitate weaker profitability for banks operating globally. Or, it could be associated with collectively underestimated risk premia and/or search for yield behaviour in line with the views of shared narratives in global financial markets (Shiller, 2017; Gennaioli and Shleifer, 2018).

The debt service ratio has also been identified as a good early warning indicator (Drehmann and Juselius, 2014). It measures interest payments relative to income. This can provide a gauge of how overextended borrowers are: the higher the debt service ratio, the more vulnerable borrowers are to falls in their incomes or increases in the interest rate. Overextension in borrowing could result in an increased rate of defaults, a loss in consumption smoothing capabilities, or a lack of new investment. The downside of our simplistic debt service ratio measure (credit × long-term interest rate over GDP), which is driven by data availability, is that it does not capture short-term lending rates, capital repayments, or the maturity structure of the debt, all of which may also be important.

We also explore the potential role of the current account. Current account imbalances have often been found to be a strong driver of crises due to capital inflows pushing down interest rates and thus encouraging excessive risk-taking behaviour potentially financed by flightly funding (Reinhart and Rogoff, 2008; Bernanke, 2009; King, 2010). To account for crises which could be caused by fiscal vulnerabilities we include public debt. Finally, we also control for general macroeconomic conditions which could trigger financial crises by including real consumption per capita, investment, the consumer price index (CPI), and money supply.

To obtain a quick sense of the potential importance of these explanatory variables, Table I also compares their mean values shortly before the crises and during normal economic conditions. A t-test confirms that there are significant differences in nearly all of the variables.

Variable	Transformation	Crisis build-up Mean	Crisis build-up SD	Non-crises Mean	Non-crises SD	Difference of mean
Baseline experiment						
Yield curve slope	level	-0.34	1.59	0.85	1.75	-1.19 ***
Credit	2-year difference of GDP-ratio × 100	6.59	7.92	2.07	5.46	4.52 ***
CPI	2-year growth rate of index	4.60	8.12	7.76	9.89	-3.16 ***
Debt service ratio	2-year difference of GDP-ratio × 100	0.55	1.07	-0.02	1.08	0.57 ***
Consumption	2-year growth rate of index	3.27	5.23	4.74	4.77	-1.47 ***
Investment	2-year difference of GDP-ratio × 100	1.06	3.13	0.21	2.30	0.86 **
Public debt	2-year difference of GDP-ratio × 100	-0.37	8.84	-0.36	8.04	-0.02
Broad money	2-year difference of GDP-ratio × 100	2.67	4.76	1.05	4.62	1.62 ***
Stock market	2-year growth rate of index	18.16	28.38	19.46	41.25	-1.30
Current account	2-year difference of GDP-ratio × 100	-0.59	2.875	0.06	2.70	-0.64 **
Global† yield curve slope	level	0.18	0.80	0.90	0.84	-0.72 ***
Global† credit	2-year difference of GDP-ratio × 100	4.18	3.21	2.20	2.33	1.98 ***
Additional variables used in robustness checks						
Domestic nominal short-term rate	level	6.14	3.51	5.19	3.69	0.95 **
Domestic nominal long-term rate	level	5.80	3.15	6.05	3.43	-0.25
Household loans††	2-year difference of GDP-ratio × 100	3.69	4.96	1.62	3.29	2.07 ***
Business loans††	2-year difference of GDP-ratio × 100	4.45	5.62	0.45	4.05	4.00 ***
House price (index)†††	2-year growth rate of index	14.95	19.15	13.77	18.58	1.19

Table I: Overview and descriptive statistics of the variables for observations one and two years before a crises (build-up) and non-crises observations. A *t*-test is used to determine whether the difference in the mean is statistically significant with *p< 0.1; **p<0.05; ***p<0.01. †: Mean of all other countries; ††: Based on a subset of 901 observations (52 with positive crisis outcome); †††: Based on 1081 observations (83 with positive crisis outcome). The statistics of the remaining variables are based on our baseline dataset with 1249 observations. In robustness checks, we also test other variable transformations than those specified here.

3 The benchmark logistic regression model

For our first exercise, we fit a simple logistic regression model to our dataset. To better compare the predictive power of the individual variables, we standardise them in this and all following regression analyses such that they have a mean of 0 and a standard deviation of 1.[6]

We include all 12 baseline variables in Table I, focussing particularly on domestic and global credit growth and the yield curve slopes. The first model in Table II shows that domestic credit growth is an important predictor for financial crises even after controlling for all covariates apart from global credit and yield curve slopes. This is in line with the literature—for example Schularick and Taylor (2012) found that a 2-year lag of credit growth is highly predictive with a standardised regression coefficient of 0.50.

The second specification adds global credit to the model. We find that this variable obtains a higher weight than domestic credit, in line with Cesa-Bianchi et al. (2019).[7]

Next, we add the slope of the yield curve. Its weight is negative, indicating that a negative (or small positive) slope corresponds to a higher estimated probability of crisis. Adding the global slope in Model 4, the weight of the domestic slope decreases but both remain important and highly significant as do the credit variables.[8] But the significance of CPI and the debt service ratio both drop out when adding the global slope to the model. Likelihood ratio tests confirm that each increment from model 1 to 4 improves the goodness of fit of the models significantly ($p < 0.001$).

[6] This is equivalent to the standardisation of the regression coefficients suggested by Agresti (1996) and recommended over other approaches by Menard (2004).

[7] These two variables have a correlation of 0.25 and an analysis of multicollinearity across all variables does not indicate problematic levels.

[8] Collinearity between both yield curve variables again does not indicate problematic levels.

	(1)	(2)	(3)	(4) Baseline specification
Domestic credit	0.420***	0.360***	0.362***	0.426***
	(0.127)	(0.128)	(0.135)	(0.137)
Global credit		0.560***	0.668***	0.668***
		(0.117)	(0.126)	(0.127)
Domestic slope			−0.786***	−0.581***
			(0.131)	(0.144)
Global slope				−0.613***
				(0.151)
CPI	−0.509***	−0.561***	−0.414**	−0.238
	(0.157)	(0.163)	(0.167)	(0.170)
Broad money	0.124	0.136	−0.016	0.036
	(0.138)	(0.145)	(0.154)	(0.155)
Stock market	0.080	0.071	−0.093	−0.126
	(0.148)	(0.153)	(0.158)	(0.167)
Consumption	−0.469***	−0.448***	−0.484***	−0.418***
	(0.130)	(0.131)	(0.136)	(0.139)
Public debt	−0.044	−0.084	−0.055	−0.026
	(0.132)	(0.139)	(0.134)	(0.134)
Investment	0.322***	0.306**	0.379***	0.316**
	(0.121)	(0.123)	(0.131)	(0.131)
Current account	−0.166	−0.140	−0.083	−0.084
	(0.126)	(0.130)	(0.131)	(0.133)
Debt service ratio	0.615***	0.528***	0.355**	0.158
	(0.150)	(0.159)	(0.166)	(0.168)
Observations	1,249	1,249	1,249	1,249
Log Likelihood	-287.997	-272.134	-257.605	-248.885
Akaike Inf. Crit.	595.994	566.268	539.211	523.769
Area under the curve	0.756	0.785	0.836	0.852

TABLE II: Logistic regression models fitted to all data points. The outcome variable is our crisis indicator, which is set to positive one and two years before an actual crisis. The standard errors of the regression weights are shown in parentheses. Significance levels: *p<0.1; **p<0.05; ***p<0.01.

4 Machine learning methodology

The benchmark regression (4) in Table II is easy to interpret and has low computational costs. This approach, however, has some shortcomings for an exercise like ours. In particular, logistic regression does not automatically account for nonlinearities and interactions, which are both likely to be relevant prior to financial crises. For example, Cesa-Bianchi et al. (2019) find a significant quadratic association between global credit and financial crises, while Alessi and Detken (2018) observe a significant interaction between domestic and global credit growth. To account for nonlinearities and interactions in a logistic regression, the modeller explicitly needs to add polynomial or interaction terms to the model. Choosing the right terms is challenging; choosing many terms is problematic because it reduces the stability of the model and the statistical power of finding an effect. We address this shortcoming by using machine learning models that are capable of learning nonlinearities and interactions from the data without the need to specify them explicitly. The remainder of this section explains our methodology in detail.

Theoretical (Wolpert et al., 1997) and empirical (Fernández-Delgado et al., 2014) evidence suggests that different machine learning models models work well for different prediction problems. As it is challenging to deduce a priori from the characteristics of the data which model will perform well on a problem, we employ a range of diverse machine learning models, as summarised in Section 4.1.

Fitting a model to the data does not tell us how well it fares in prediction, as (in-sample) fitting accuracy is in most cases higher than (out-of-sample) prediction accuracy. This is true for linear regression but the discrepancy is often more pronounced for flexible machine learning models which may fit perfectly to data even though they may perform poorly in out-of-sample predictions. This situation corresponds to overfitting the data and to avoid it we adapt an experimental procedure which allows for extensive out-of-sample tests (Section 4.2).

Finally, Section 4.3 introduces the novel framework based on Shapley values which aim to address the black box critique of machine learning models and identify the contributions of individual predictors. This section also explains *Shapley regressions* (Joseph, 2019) through which we are able to determine whether or not a predictor makes a statistically significant contribution to the accuracy of the model.

4.1 Machine learning models

Let f be a prediction model $\hat{y} = f(\mathbf{X})$, where $\mathbf{X}_{n \times k}$ is the predictor matrix containing n observations and k variables and $\hat{y} \in [0, 1]$ is the predicted probability of a crisis. The observed class label for each observation is denoted by $y \in \{0, 1\}$, where 1 marks the pre-crisis target years, one and two years before an actual crisis in our baseline approach. It is referred to as the positive class and 0 indicates no crisis and is referred to as the negative class.

We compare a diverse set of machine learning classification algorithms ranging from simpler, more interpretable models such as decision trees to more complex approaches such as random forests and neural networks. In what follows, we only provide a high level, non-technical explanation of the models we use. A description of the implementation details is found in Appendix A and several excellent textbooks provide a more comprehensive overview of the different machine learning techniques (Bishop, 2006; Friedman et al., 2008).

Decision trees

A decision tree is an interpretable model that successively splits the data into subsets by testing a single predictor at each node (e.g. *Credit growth* > 1%). Starting at the root node of the tree, all observations are divided into two child nodes, one for which the condition in the node is true and one for which it is false. This process is recursively repeated in the respective child nodes.

Decision trees are very flexible models. However, the bigger a tree grows, the less likely it will generalise well to out-of-sample data. Big trees tend to fit to the specific noise of a data sample and therefore perform substantially worse on a new set of observations drawn from the same population. This phenomenon is usually referred to as *overfitting*. There exists a plethora of *pruning* techniques to reduce overfitting by controlling the size of decision trees (Rokach and Maimon, 2005). We use the C5.0 algorithm (Quinlan, 1993; Kuhn et al., 2014), which uses a statistical heuristic to control the complexity of the tree. Nevertheless, decision trees often have limited predictive power compared to more complex methods such as random forests, especially when the dataset is small. Fundamentally, this is related to high variance in the (relatively small) leaf nodes. Ensemble methods, such as random forests or bootstrapped model averages, avoid this problem by reducing this variance, i.e. they lead to a better bias-variance trade-off (see also Chakraborty and Joseph (2017)).

Random forests

A random forest (Breiman, 2001) is a collection of many, often hundreds, of decision trees. By averaging the predictions of the trees, random forests usually suffer less from overfitting than any individual tree by reducing the overall variance of predictions from the model. Each tree overfits differently and averaging their predictions cancels out these noisy components and increases the ability to predict on unseen data. This only works if the trees are sufficiently different to each other; similar trees fit to the noise in similar ways. To diversify the collection of trees, the random forest algorithm uses two techniques: First each tree is trained on a different subset of the data, which is drawn with replacement from all observations.[9] Second, the algorithm does not choose the best of all possible splits but randomly samples m candidates from the k predictors, optimises the split for each of them and then chooses the best split from this subset. In a forest, each individual tree predicts either the positive or negative class for an observation. The mean prediction across all trees gives an estimate of the probability that an instance belongs to the positive class.

A random forest often performs substantially better than individual decision trees and many other machine learning algorithms. Indeed, in a large-scale empirical comparison of 179 classification algorithms conducted on a diverse set of 121 real world datasets, it was the best performing algorithm on average (Fernández-Delgado et al., 2014). But, this comes at the cost of interpretability. As it aggregates the predictions of hundreds of trees, a random forest is not decomposable into a simple set of rules.

[9] This approach is referred to as *bagging* (short for *bootstrap aggregating*) in the machine learning literature and is a general technique to improve the stability of prediction models (Breiman, 1996).

Extremely randomised trees

Extremely randomised trees (Geurts et al., 2006) are similar to random forests but tend to produce predictions that are more continuous as a function of the predictors. They achieve that by creating more diverse trees. The method differs in two aspects from random forests. First, each tree is trained on the complete training data and not on a resampled subset of the data. Second, the splitting process in each tree is more random. For each of the m candidate predictors that are randomly sampled, a split is not optimised but made completely at random across the range of the values of the indicator. Of these random splits, the best one is used in the tree. In what follows, we refer to this method as *extreme trees*.[10]

Support vector machines

A support vector machine (SVM) is similar to a logistic regression as it learns a linear function of the inputs. However, these inputs are previously transformed by using a nonlinear kernel function, allowing them to model nonlinear classification problems. Hereby, kernels efficiently transform the data into a higher linear dimensional space in which the SVM then learns to separate the positive from the negative class. In the study by Fernández-Delgado et al. (2014), SVMs were, on average, the second best algorithm. However, by using a nonlinear kernel, the SVM becomes a black box because each prediction is not easily attributable to individual predictors as it is for a regression or a simple decision tree. A popular kernel, which we also use in the following analyses, is the radial basis function (Gaussian kernel, Vert et al. (2004)). We do not train a single SVM but average the predictions of 25 SVM models that are trained on the same training set (see Appendix A).

[10] We also tested gradient boosting, which has been successfully employed in other economic prediction problems such as predicting recessions (Ng, 2014; Döpke et al., 2017) and bankruptcy (Carmona et al., 2019; Zięba et al., 2016). In our experiments, gradient boosting performed better than logistic regression but fell behind the other decision tree ensembles, random forests and extreme trees, so we do not report its results in what follows.

Artificial neural networks

Artificial neural networks have been the most researched machine learning technique in recent years. They have achieved landmark successes in classification problems such as face (Schroff et al., 2015) and speech recognition (Amodei et al., 2016), though these and other prominent applications of neural networks use very large datasets.

A neural network consists of an input layer that represents the values of the predictors, at least one hidden layer, and an output layer. The inputs are passed from one layer to the next and are finally integrated as a prediction in the output layer. Note that, without a hidden layer, a neural network is a linear function of the input layer, such as a linear regression.

The nodes in a hidden layer are connected to the previous and subsequent layer by weights. A node in a hidden layer computes the weighted sum of all its inputs and transforms it using an activation function (e.g. a logistic function) before this output is passed to the next layer. Given a dataset with k predictors and a network with a single hidden layer containing m nodes, $k \times m$ weights are needed to fully wire the input layer to the hidden layer, and m weights are needed to connect the hidden layer with the output layer, which only contains a single node in a binary classification task.

A neural network has hyperparameters that control the structure of the model such as the number of hidden layers and nodes or the activation function. The high number of parameters and hyperparameters, and a network's sensitivity to these, makes learning a predictive network with an appropriate architecture challenging, especially when the available data are small. We do not train a single neural network but average the predictions of 25 models that are trained on different samples drawn with replacement from the training set (see Appendix A).

4.2 Experimental procedure

In our main analysis, we use cross-validation to evaluate the out-of-sample predictive performance of our models. This entails calibrating the models exploiting all available data on random 80% selections of the these data, called the training set, and evaluating them on the remaining 20% of observations, the test set. However, in Section 5.3 we also show results for a forecasting approach. Using an expanding window, this requires predictions to be made based only upon data available until the given point in time, so that any out-of-sample test is later in time.

One may argue that cross-validation is not an adequate empirical test for an early warning model. By randomly sampling training and test sets from the whole sample, we predict crises in the past with data from the future. Therefore, our baseline performance estimates do not reflect the real-time performance of an early warning model that is predicting crises in the future. Despite this, there are good reasons for using cross-validation instead of forecasting crises with past data only. For example, previous research suggests that the global financial crisis is qualitatively different from other crises in that global credit plays a crucial role. In a forecasting experiment, we cannot reliably test a model that learned from the global financial crisis, as our dataset does not contain many observations after that crisis. More generally, due to the limited number of crises in the data, a comparison of the forecasting performance across models suffers from low statistical power.

In our cross-validation set-up, all included observations between 1870 and 2016 are randomly assigned to one of five different subsets that we refer to as *folds*. Each fold is used once as a test set in which the performance of the prediction models is evaluated, while the remaining folds are used for training the models. In this way, each training set contains 80% of the observations, while the remaining 20% constitute the corresponding test set. To obtain stable results, the random assignment of folds is repeated 100 times.

Recall that each crisis observation in the raw data is recoded to two positive class labels: one and two years before the actual crisis. As these observations are highly similar, we always assign them to the same fold. This avoids an overly optimistic out-of-sample performance estimate whereby, for example, training on observations two years prior to a given crisis is likely to imply strong out-of-sample performance on the observations one year before the same crisis. Appendix B.1 examines this bias and other ways of cross-validation in detail.

Some of the machine learning methods require learning hyperparameters (see Appendix A). Hyperparameters control the flexibility of the model, such as the number of layers or nodes in a neural network. These parameters cannot simply be optimised in the training set because the most flexible model structure would always obtain the best fit. Instead, the hyperparameters need to be evaluated on out-of-sample data. To achieve that, we employ *nested cross-validation*: within each training set S of the 5-fold cross-validation procedure, we apply 5-fold cross-validation to assess the performance of all possible combinations of hyperparameters. The parameter combination that obtains the best performance in this 5-fold cross-validation is then used to train a model on the complete training set S.

4.3 Shapley values

The machine learning models described above are non-parametric and error consistent (Stone, 1977; Joseph, 2019), which means that they approximate any sufficiently well-behaved function arbitrarily well when provided with enough training data. But their high flexibility typically makes them difficult to interpret. In particular, it is hard to ascertain which specific variables drive model predictions and through what functional relationship they are important.

We address this issue by adopting the *Shapley additive explanations* framework (Strumbelj and Kononenko, 2010; Lundberg and Lee, 2017). It uses the concept of Shapley values (Shapley, 1953; Young, 1985) from cooperative game theory. In that context, Shapley values are used to calculate the payoff distribution across a group of players. Analogously, we use them to calculate the 'payoff' for including different predictors in the models. More precisely, the predicted crisis probability for each individual observation is decomposed into a sum of contributions from each predictor, namely its *Shapley values*. This enables us to understand which variables are driving each prediction. As such, we can determine whether or not variables like credit growth, the yield curve slope, the current account, and asset prices have large predictive value in our machine learning models.

Corresponding to the predictor matrix $\mathbf{X}_{n\times k}$ described in Section 4.1, we define the Shapley value matrix as $\mathbf{\Phi}_{n\times k}$ and ϕ_{ij} as the Shapley value of observation i and predictor j. The predicted value of observation i is decomposed into the sum of the Shapley values $\hat{y}_i = \sum_{j=1}^{k} \phi_{ij} + c$, where c is the base value that is set to the mean predicted value in the training set.

How are the Shapley values calculated? For a linear regression model, the Shapley value of predictor j is simply the product of its regression coefficient w_j and the difference between the predictor value X_{ij} and its mean, i.e. $\phi_{ij} = w_j(X_{ij} - \mathbb{E}_i[X_{ij}])$. Computing Shapley values for a more general machine learning model is computationally more complex.

An intuitive way to understand the computation of Shapley values is to consider the problem that motivated their invention. In a cooperative game, the individual contribution within a coalition of players is not directly observable but the payoff generated by the the group as a whole is. To determine the contribution of player j, coalitions can be formed sequentially and j's contribution can be measured by her marginal contribution when entering a coalition, which also depends on the other players in the group. Imagine player j joins a coalition in which player k has similar skills. In this case, j's contribution is smaller than if she had joined the group when k was absent. Therefore, all possible coalitions of players need to be evaluated to make a precise statement of j's contribution to the payoff.

More formally, let N be the set of all players in the game, and $f(S)$ be the payoff of a coalition S. Then the Shapley value for player j is computed by:

$$\phi_j = \sum_{S \subseteq N \setminus j} \frac{|S|!(|N| - |S| - 1)!}{|N|!} [f(S \cup \{j\}) - f(S)]. \tag{1}$$

In our case, we make the analogy between the payoff and the predicted probability estimated by the model for a particular observation i, i.e. $f(X_i) = \sum_{j=1}^{k} \phi_{ij}(X_i) + c$ (Strumbelj and Kononenko, 2010). The set of players N correspond to the predictors used in the model. It follows that the computation of the Shapley values has to be done for each individual observation for which we want to explain the predicted value. To compute the exact Shapley value of variable j for observation i, one has to compute how much variable j adds to the predictive value $(f_i(S \cup \{j\}) - f_i(S))$ in all possible subsets of the other variables $(S \subseteq N \setminus j)$. Contrary to a cooperative game, predictors not in S cannot be left out as this would not allow a (machine learning) model to produce predictions. Instead, these predictors are integrated out using all observed values in the training set. For a given set S and observation i, the values of i of variables not contained in S are replaced and averaged over according to (1).[11]

Within the Shapley value framework, we can also measure how much interactions of variables contribute to the predictions (Fujimoto et al., 2006; Lundberg et al., 2018). To compute the *Shapley interaction* $\phi^{\mathbf{i}}_{r\times s}$ of variables r and s we use

$$\phi^{\mathbf{i}}_{r\times s} = \sum_{S\subseteq N\setminus\{r,s\}} \frac{|S|!(|N|-|S|-2)!}{2(|N|-1)!} \nabla rs(S), \tag{2}$$

where $r \neq s$ and $\nabla_{rs}(S)$ is the contribution of the interaction without the contribution of the predictors r and s on their own:

$$\nabla_{rs} = f(S\cup\{r,s\}) - f(S\cup\{r\}) - f(S\cup\{s\}) + f(S) \tag{3}$$

To estimate the Shapley values and interactions, we again use 5-fold cross-validation. The models are learned in the training set and the Shapley values are computed for the objects in the test set. We repeat the cross-validation procedure 100 times to obtain stable estimates.

Shapley regressions

Shapley values measure how much individual variables drive predictions of a model, independent of the overall accuracy of the model. In other words, taken in isolation, Shapley values do not show how reliably the variables actually predict the true outcome, which is a question of statistical inference.

To judge the economic and statistical significance of predictors, we use *Shapley regressions* (Joseph, 2019). To the best of our knowledge, there is no other statistical framework that estimates the significance of individual predictors or general hypothesis tests on non-parametric models. In our context, the Shapley regression framework achieves this by regressing the crisis indicator y on the Shapley values $\mathbf{\Phi}_{n\times k}$ using a logistic regression.

[11] We use the `shap` Python package (Lundberg, 2018) to estimate the Shapley values efficiently. Lundberg and Lee (2017) provide a detailed explanation of how Shapley values are computed in the context of explaining predictions of machine learning models. A tacit assumption behind the above calculation is variable independence which cannot be accounted for using non-tree models (Lundberg et al., 2018). However, the robustness of variable importances measured by Shapley values across all models, especially for dominant predictors, suggests that any contemporaneous dependences between variables can be neglected in the current application.

That is, the nonlinear and unobservable function of the predictors in a black box machine learning model is transformed via Shapley values into an additive parametric space which makes the estimation of p-values a simple regression exercise.[12] In line with the Shapley values for a linear regression model, the surrogate Shapley regression has the appealing property that if the estimated model is a linear function of the predictors, the Shapley regression will reproduce the linear model. We include all 100 individual Shapley estimates for each observation from different bootstraps in the regression to account for variability across replications. We estimate clustered standard errors on the country-year level to account for dependences between observations in our experimental setting.

5 Machine learning prediction performance

5.1 Model comparison

We now evaluate the predictive performance of the different machine learning models in the cross-validation exercise and also compare them to a logistic regression approach. All of our models aim to predict the occurrence of a financial crisis. Therefore, we can evaluate their performance in the *Receiver Operating Characteristic* (ROC) space. Here, the vertical axis shows the true positive rate, also known as the hit rate, which is defined as the proportion of positive instances (crises) correctly identified as such. The horizontal axis shows the false positive rate, also known as the false alarm rate, which is defined as the proportion of negative instances (non-crises) incorrectly identified as positive (crisis). The perfect model would obtain a hit rate of 1 and a false alarm rate of 0. In practice, a higher hit rate comes at the cost of a higher false alarm rate.

The trade-off between the hit rate and the false alarm rate can be controlled by setting different thresholds on the probabilities predicted by a model to trigger an alarm. The overall performance of a model in the ROC space can be summarised by the *Area Under the Curve* (AUC). The main advantage of ROC analysis is that it does not force the modeller to specify the relative costs of the two types of classification errors (failing to

[12]Note that this regression on the regressors generated by the Shapley values is performed on the test data making the machine learning model and Shapley regressoin coefficient estimation independent. See Pagan (1984) and Joseph (2019) for discussions of validity conditions of inference which are fulfilled in the current setting.

predict a crisis when there is one and predicting a crisis when there is none), which is often a non-trivial endeavour and usually depends on the context in which the model is applied.

5.1.1 Baseline analysis

Figure II shows how all models compare in out-of-sample prediction in ROC space when using the baseline explanatory variables from Table I. For the logistic regression it is to note that this is a different more demanding prediction exercise than the in-sample fitting exercise summarised in Section 3. So the corresponding AUC is lower here than in model (4) in Table II.[13] It is immediately clear that machine learning approaches have potential value relative to standard regression methods when seeking to predict financial crises. Under the AUC metric, four out of our five machine learning algorithms perform better than the logistic regression with extreme trees being the most accurate, followed by random forests. To quantify this difference, extreme trees only generate 19% false alarms when calibrated to achieve a hit rate of 80% compared to 32% for the logistic regression.[14] Only the decision tree performs worst. This is not surprising, as individual decision trees tend to overfit and produce unreliable probability estimates if the training data is small (Perlich et al., 2003).

5.1.2 Robustness checks

To show that the relative model ranking is robust, Table III reports robustness checks that test different sets of predictors and transformations with all experiments repeated 100 times using 5-fold cross validation. We did not test SVMs and neural networks across all of these combinations because of the generally weaker predictive performance in the baseline and the extensive computational time involved.

Adding new variables or changing the transformation may lead to a change in the

[13] We also tested regularised logistic regression with ridge, lasso (Ng, 2004), and elastic net (Zou and Hastie, 2005) penalties to reduce overfitting. However, none of the regularisations improved the out-of-sample performance so we do not report these results in what follows.

[14] Note that this difference cannot be read of exactly from Figure II, as the curves are generated by averaging ROCs over cross-validation folds while the precise 80% threshold is calculated across all 100 test repetitions. Basing the curve on the latter would suggest an overly good performance of the decision tree model.

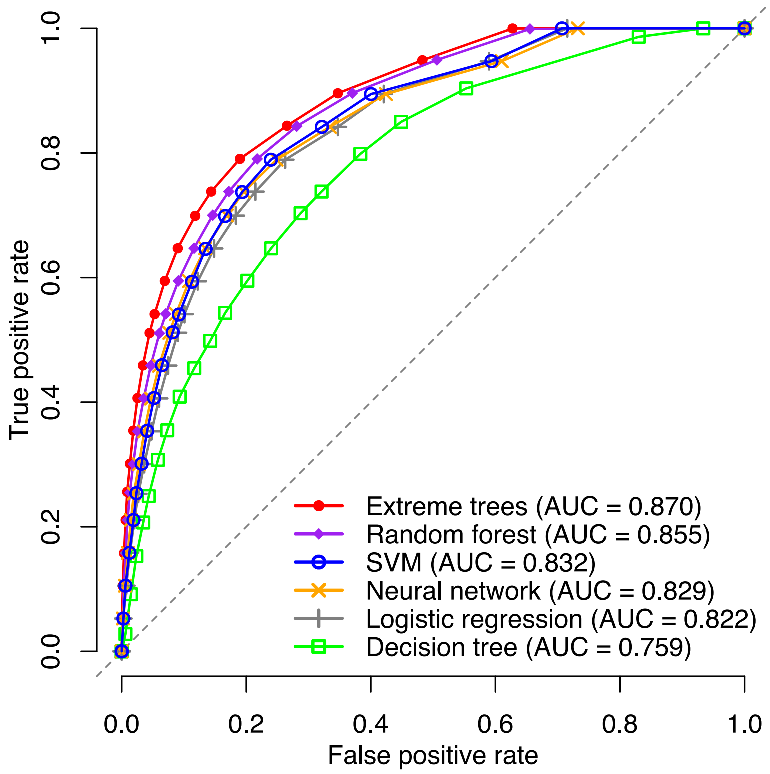

FIGURE II: ROC curves for baseline models.

number of observations due to missing values. To provide a fair comparison, we need to retrain the baseline models on exactly the same sets of observations as the robustness check is trained on. In Table III, the retrained baseline models are marked with an asterisk. If the pool of observations does not change in a robustness check with respect to the baseline, the results can be directly compared with the baseline in the first row.

Variable transformations. To ensure that scaling most variables by GDP ratios is indeed superior, we performed additional analyses using both growth rates and filtered data to detrend the data. For the growth rate experiment, the slope of the yield curve is left in levels, the current account is scaled by GDP (as it contains positive and negative values) and all other variables are transformed into 2-year percentage growth rates. Another de-trending method used to identify the gap between the long-term trend of a variable and the observed change, is the regression filter proposed by Hamilton (2018).We set the parameter $h = 2$, and regress on the four most recent values. The filter is applied to consumption and to the following variables after scaling them by GDP: do-

Experiments	Crises	Extreme trees	Random forest	Logistic regression	Decision tree
Unit	#	AUC			
Baseline	95	0.87	0.86	0.82	0.76
ALTERNATIVE TRANSFORMATIONS					
Growth rates only	95	0.86	0.85	0.81	0.76
Hamilton filter	89	0.86	0.85	0.81	0.76
*	89	0.87	0.85	0.82	0.76
ALTERNATIVE VARIABLE SETS					
Nominal interest rates (alt.)	95	0.87	0.85	0.82	0.76
Real interest rates (alt.)	95	0.86	0.85	0.82	0.76
1-yr change nom. s.t. rate (add.)	95	0.86	0.85	0.82	0.76
2-yr change nom. s.t. rate (add.)	95	0.87	0.85	0.82	0.76
Loans by sector (alt.)	52	0.82	0.83	0.83	0.76
*	52	0.83	0.83	0.83	0.79
House prices (add.)	83	0.88	0.87	0.82	0.75
*	83	0.87	0.86	0.82	0.75
TRANSFORMATION HORIZONS					
1 year	95	0.85	0.84	0.82	0.76
*	95	0.87	0.85	0.82	0.76
3 years	92	0.86	0.85	0.80	0.74
*	92	0.87	0.85	0.82	0.76
4 years	90	0.87	0.86	0.80	0.74
*	90	0.87	0.85	0.82	0.77
5 years	89	0.86	0.85	0.80	0.75
*	89	0.87	0.85	0.82	0.76

TABLE III: Results of the robustness checks for different model specifications. Asterisks indicate the retrained baseline experiment on exactly the same observations as the respective robustness check.

mestic credit, global credit, money, public debt, debt servicing, investment, and current account. Across all models, using changes scaled by GDP (our baseline) leads to more accurate or as accurate predictions when compared to using growth rates or the Hamilton filter.[15]

Additional variables. Next, we investigate how the performance changes if we use

[15] We did not test GDP ratios where variables where given as an index (consumer prices, consumption and stock prices). Growth rates gave the best test results here. We also did not mix transformations within variable sets but tested transformations against each other whenever possible.

alternative or additional variables. Replacing the yield curve slope by the nominal or real long and short-term interest rates does not improve performance, though we examine the interplay between the slope and the level of interest rates more carefully in Section 7.

We also added the change in the nominal short-term rate over one and two years as a possible proxy for monetary policy actions. But again, there was no improvement in the predictions of the various models.

Replacing total loans by household and business loans does not improve the performance of any model, either. Adding house prices does increase the performance of extreme trees and random forests by one and two percentage points, respectively. This is in line with the observation that credit booms after 1945 are often strongly driven by increases in mortgage debt and that rapid house price appreciations are indicative of future financial crises (Jordà et al., 2015a). However, we find that house prices do not obtain a significant weight in a logistic regression when controlling for our covariates, including domestic credit growth even if we calibrate the model only on observations after 1945. The inclusion of house prices also reduces the crisis sample. Together, these findings lead to the decision to exclude them from the baseline model, while acknowledging they may have useful value as a supplementary indicator.

Horizon of growth rates. The horizon of the growth rates and changes scaled by GDP are set to 1, 3, 4, and 5 years for all respective variables rather than the 2 years in our baseline. There are only small differences between horizons but the baseline almost always produced the most accurate results for all prediction models.

Cross-validation procedure. In Appendix B, we compare four different types of cross-validation. Our results are stable across these different approaches.

Overall, extreme trees and random forests perform best in each of these additional experiments. This confirms their value in this prediction problem and justifies our focus on extreme trees in the more detailed analyses which follows.

5.2 Exploring the best predictive model: extreme trees

To enhance understanding of our results, we examine the best performing model—extremely randomised trees—in more detail. We average the out-of-sample predictions across the 100 replications and pick one plausible working point on the ROC curve, namely that with a hit rate of 80%. The corresponding threshold at which the model identifies a crisis is a predicted probability of 9.6%. This setting results in a false alarm rate of 19%.

Using this threshold, Figure III depicts correctly identified crises (green circles), missed crises (red triangles), false alarms (grey triangles), and the predicted probability of crisis (black line) for all observations in our sample. To improve legibility, the most prevalent outcome by far, true negatives (correctly identified non-crisis), are only shown in light green in the pie charts to the right which depict the overall distribution of all four outcomes for each country.

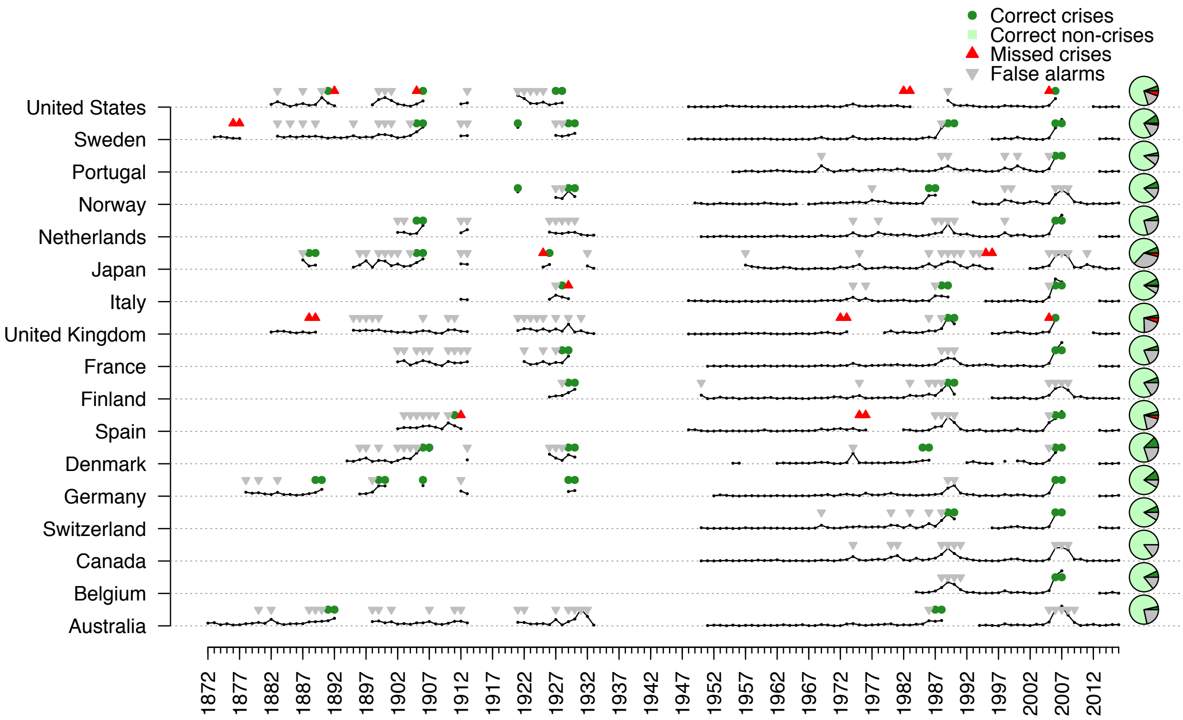

FIGURE III: Crisis probability estimated by extremely randomised trees (black line) and the classification (crisis vs. non-crisis) when imposing a probability threshold of 9.6% to achieve a hit rate of 80%.

The model fully misses only six out of 49 distinct crisis events: Sweden (1878), United Kingdom (1890, 1974), Spain (1977), United States (1984), and Japan (1997). For another six crises, the model only misses either the first or second year ahead of the actual crisis but not both of these observations. All of the missed crisis episodes can be related either to unclear crisis timing or aspects not fully captured in our models, including vulnerabilities to concentrated exposures overseas as discussed below.

After a significant expansion of the Swedish railway industry in the early 1870s, Sweden's economy declined causing substantial losses in the financial sector which was heavily exposed to the railway industry. The late 19th century crisis in the UK was known as the Barings crisis or the Panic of 1890. It was precipitated by poor investments in Argentina that led to the bankruptcy of Barings Bank (Mitchener and Weidenmier, 2008).

The secondary banking crisis in 1974 in the UK (Reid, 1982) was linked to house prices which are not included in our baseline model but do appear to have some value as an indicator (see Section 5.1).

According to Betrán and Pons (2013), the crisis in Spain in 1977 was caused by the global oil price shock and the government's interventions to dampen its effects, which delayed firm consolidation and increased public spending. Another reason was the rapid institutional change in the financial sector after the end of the Franco regime (Martin-Acena, 2014), all factors not captured by our predictive variables.

The 1984 financial crisis in the US reflects the saving and loan crisis caused both by the global recession in the early 1980s (Reinhart and Rogoff, 2008) and loans by some US banks heavily exposed to Latin America. The particular financial problems of the saving and loan associations that led to the failure of more than 1,000 of them are not reflected in our data. In particular, after a substantial increase of the discount rate, the interest rates of existing long-term loans issued by the saving and loan associations were lower than the rates at which they could borrow. The associations were deregulated and had to make risky investments to remain profitable. Finally, the missed crisis in Japan in 1997 is partly linked to the Asian financial crisis of 1997, to which Japan was exposed (Wade, 1998). But as reflected in 'false alarms', our model predicted a high probability of crisis for much of the late 1980s and the beginning of the *Lost decade* in the early 1990s. This is line with other research that identified 1992 as the start of the financial crisis in Japan (Reinhart and Rogoff, 2009; Bordo et al., 2001).

It is also worth noting that the relatively high proportion of false alarms is somewhat misleading for several reasons. First, through the choice of the targeted 80% hit rate, the model is, by construction, fairly risk averse and calibrated to ensure fewer crises are missed at the cost of a higher false alarm rate. Second, several of the the false alarms occur more than two years in advance of an actual crisis and so still provide a useful very early warning signal. Third, the false positives cluster around periods when other countries experience financial crises, which indicate periods of elevated global risks. Linked to this, the model does not account for any policy measures that might have mitigated a crisis. In particular, the model might have correctly detected an impending crisis or elevated (global) risks which did not hit particular countries because of mitigating policy actions. In these cases, even the false positives might provide useful information for policy makers by indicating when vulnerabilities are building up.

Figure III also shows that the number of false alarms is substantially higher before World War 2. Given substantial changes to the global economy over time, a general model covering more than 140 years may not predict consistently well over the full sample period. As we have fewer observations before WW2 than after, the earlier period also has less weight when training the model, which means that it is geared to perform better on more recent observations. Finally, the quality of the earlier data is likely to be lower than that of more recent data. Section 7.2 discusses the robustness of the model across time in more detail.

5.3 Forecasting experiment

All results shown so far are based on cross-validation. If we want to employ a model to predict future crises in a strict forecasting sense, all observations in the training set must be from earlier years than the observations in the test set. This simulates how early warning models are actually used in practice. So we now implement a recursive forecasting experiment, where we use all observations up to year $t-2$ to train the models and test them on observations of year t, where $1946 \leq t \leq 2016$.[16]

In this way, the models learn from training samples with very different proportions of crises at different points in time. For example, after the global financial crisis, the proportion of crises in the training data is substantially higher than before that crisis.

[16] Note that we do not use observations at $t-1$ to make a prediction at time t. As in the cross-validation experiment, whereby we avoid positively biased performance estimates that may occur if one observation of a crisis (two years before an actual crisis) is in the training set and the other observation of that crisis (one year before a crisis) is in the test set.

	Forecasting period		
	1946–2016	1946–2003	2004–2016
Neural network	0.833	0.770	0.872
Extreme trees	0.813	0.748	0.870
SVM	0.808	0.700	0.911
Random forest	0.792	0.735	0.846
Logistic regression	0.789	0.704	0.867
Decision tree	0.788	0.727	0.867

TABLE IV: Forecasting performance (AUC) on all observations after 1945 and those before and after 2004.

As the predicted probability, and therefore the AUC estimate, is highly sensitive to the proportion of crises in the training set, for comparability we resampled all training sets such that they contain the same number of crisis and non-crisis observations.[17]

Table IV compares the forecasting performance of the models. It shows the AUC on all observations between 1946–2016, and for the period before and after 2004. The logistic regression again performs relatively poorly. Across the entire forecasting period, the best model is the neural network, followed by extreme trees and the SVM. But the test set is small and all performance differences in all three periods are insignificant at the 5% level according to a DeLong test (DeLong et al., 1988). We report more detailed results for extreme trees below to remain consistent with the previous cross-validation exercise.

Figure IV shows the forecasting performance of extreme trees at a hit rate of 80%. Compared to the cross-validation results in Figure III, there are substantially more false alarms. However, the pie charts show that most predictions are still correct. The forecasting model is able to correctly forecast the global financial crisis as well as a string of crises in the early 90s with the Japanese crisis now also correctly signalled. Furthermore, the pattern of missed crisis observations in Figure IV (red triangles) is almost identical to that of Figure III, indicating that models in the cross-validation and forecasting

[17]For all algorithms, we apply two techniques of resampling: upsampling and downsampling. Let n_+ and n_- be the number of crisis and non-crisis observations in the training set, respectively. Using upsampling, we increase the number of crisis observations by drawing n_+ observations with replacement. Using downsampling, we decrease the number of non-crisis observations by sampling (without replacement) from n_- non-crisis observations. To obtain stable results we repeat the resampling and model estimation 50 times and average the predictions across the iterations. For each model, we do only report the maximum performance obtained by using up- or downsampling.

approaches see similar signals.

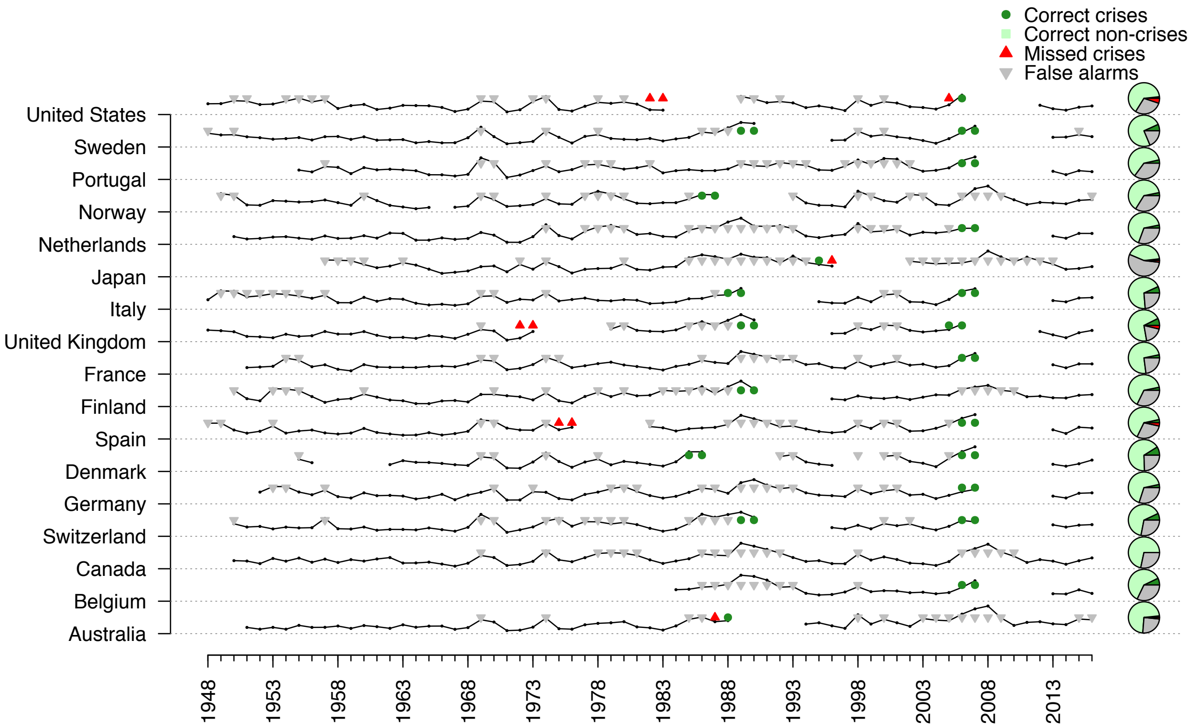

FIGURE IV: Forecasting performance of extreme trees over 1946–2016.

6 Inference on machine learning models

6.1 Shapley decomposition: variable importance

To assess the importance of the individual predictors across all observations, we compute mean absolute Shapley values for all predictors. We refer to this measure as the *predictive share* of a variable and show it in Figure V for all predictors in the baseline approach across different models. The variables are ordered by decreasing predictive share for extreme trees.

The two variables with the largest predictive shares are the global yield curve slope and global credit growth. Both are consistently ranked as the top two across the five models. The domestic yield curve slope and domestic credit follow after that, again with a high degree of consistency across different models. CPI, the debt servicing ratio, consumption and investment come next but often with significant variabilities across models. This ranking of the variables closely matches the strength of the predictors in the in-sample logistic regression (Table II). It strengthens the view that the key variables of credit and the slope of the yield curve, both domestically and globally, are robust indicators for predicting financial crises.

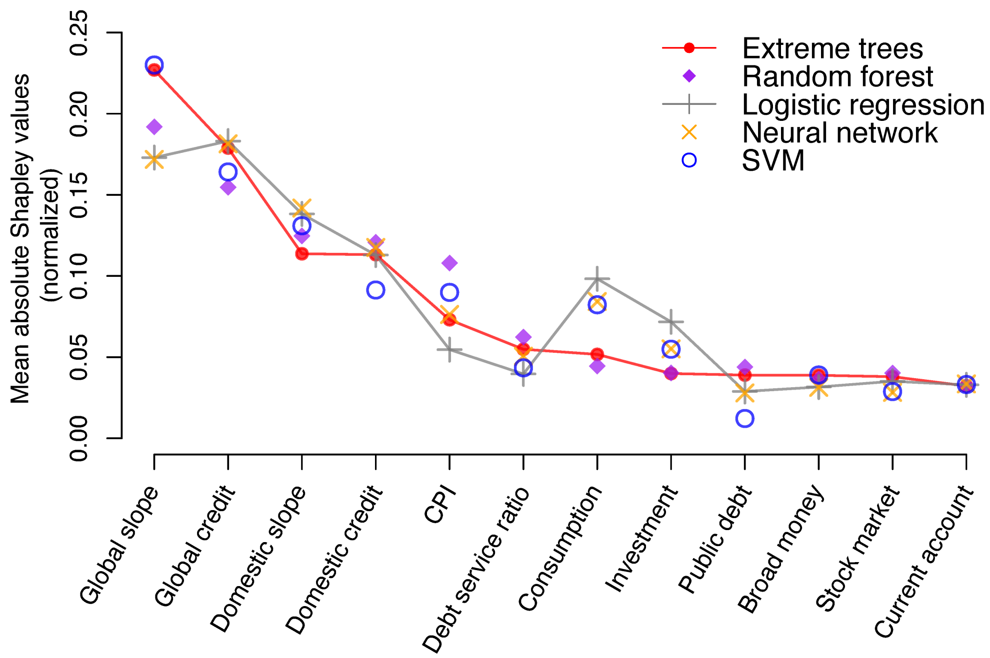

FIGURE V: Mean absolute Shapley values of individual variables across different models.

To illustrate the potential value of the Shapley decomposition in interpreting the predictions of our machine learning models, Figure VI shows the decomposition over time for the United States (US), Sweden, and Spain. This is again based on the predictions of extreme trees, our baseline machine learning model. To retain legibility, only the Shapley values of the yield curve slope and credit growth (both domestic and global) are shown in different colours; the remaining predictors are summed up in grey bars. All Shapley values and the mean predicted value in the training set (black dashed horizontal line) add up to the predicted value, shown by the black circle. The red dotted line shows the threshold corresponding to a 80% hit rate above which the model predicts a crisis. Vertical red bars represent our target one or two years ahead of a crisis, grey bars the actual beginning of the crisis.

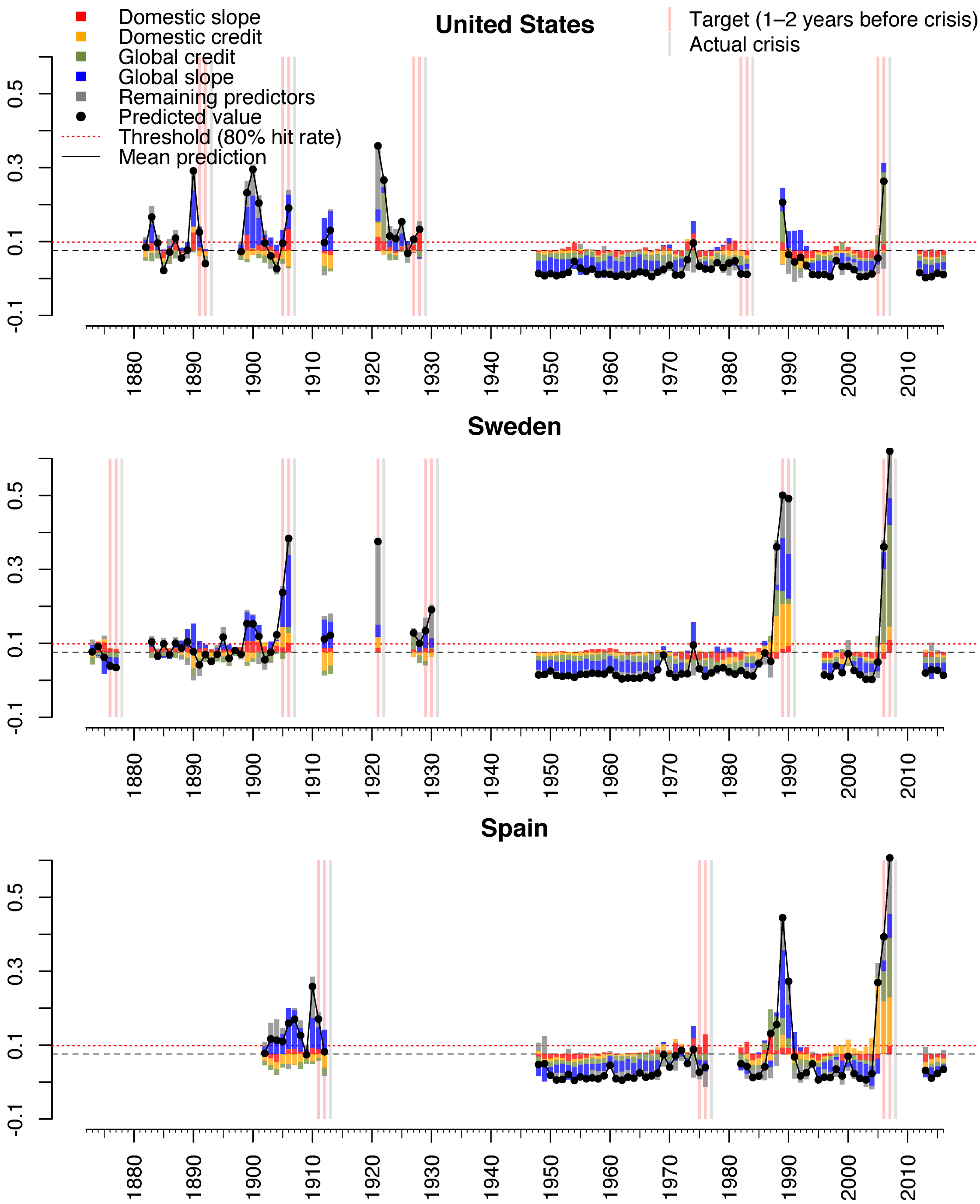

FIGURE VI: Shapley values as a function of time for the United Sates (top), Sweden (middle) and Spain (bottom).

Model performance varies across countries. Generally predictions are more noisy in the pre-WW2 period. Extreme trees correctly predict most crises for the US. Early in the sample, the global yield curve slope appears to play a strong role in crisis prediction, though there is also a substantial number of false positives. The only crisis fully missed is the Savings and Loans crisis in the 1980s. In Sweden, the model performs very well overall with the global yield curve again being important early in the sample. The Nordic financial crisis in 1991, which hit Sweden and Finland most severely (Jonung et al., 2009), is collectively predicted by several factors with domestic credit playing a strong role. By contrast, the global financial crisis is mainly predicted by global factors, especially global credit growth.

While global credit is the dominating factor predicting the global financial crisis in the US and Sweden, the prediction for Spain is strongly driven by domestic credit as well. This reflects the Spanish housing bubble prior to the crisis (Gentier, 2012). The high false alarms in the late 1980s may be associated wit the severe recession affecting many developed countries at the time, even though this did not translate into a financial crisis in Spain.

Notable events not driven by the leading four indicators are the crises in the US and Sweden in the early 1920s, when the former experienced a strong recession linked to the global flu pandemic and the latter to a (correctly signalled) banking crisis. The predictions are strongly driven by falls in consumption in both cases and also by a fall in prices in the Swedish case, though domestic credit also played a role.

Overall, the Shapley approach allows us to explain the predictions of our benchmark model well with clear attribution to economic and financial conditions surrounding individual events. As such, it can considerably alleviate the black box critique of machine learning models.

6.2 Shapley regressions: variable significance

We now use Shapley regressions to determine the statistical significance of the predictors in our machine learning models. The crisis indicator is regressed on the Shapley values, which can be interpreted as an additive feature transformation. The left half of Table V shows the output from this exercise for the extreme trees model. The normalised mean absolute Shapley values (corresponding to the read line in Figure V) are displayed in the *share* column. The coefficients represent the effects of the Shapley values for a one standard deviation change of Shapley values on the predicted log-odds of crisis ($\log \frac{\hat{y}}{1-\hat{y}}$). It is important to note that the sign of the coefficients does not indicate the sign of the association between the predictors and the probability of crisis, which is separately captured in the direction column taken from the baseline logistic regression in Table II. Rather, the coefficients are expected to be positive because higher Shapley values should reflect an increase in the predicted probability of the positive (crisis) class. Negative coefficients are statistically insignificant, which indicates that the machine learning models made little use of the information in that variable.

Consistent with our previous results, global and domestic credit and yield curve slopes obtain the highest coefficients and lowest p-values. Consumption, investment and changes in stock market indices are also significant. This means that, despite the small magnitude of their signals in terms of predictive shares, their values are significantly aligned with the crisis indicator and so they provide a useful supplementary indicator. By contrast, variables like the debt servicing ratio, public debt and the current account balance have some predictive weight but their signals cannot be differentiated from the null, i.e. there is no clear alignment with actual crises. The same is true for house price growth which is not included in the baseline presented here.

It is also informative to compare the Shapley coefficients to the coefficients of our standard logistic regression. The right hand side of Table V reproduces our baseline regression (Model 4 in Table II). This model also gives highest weight to the yield curve slope and credit but does also produce some results which are qualitatively different for the remaining predictors. For instance, stock market prices and CPI are significant predictors for extreme trees but not in the logistic regression. This generally richer signal from machine learning models is partly expected given their greater flexibility.

6.3 Nonlinearities in the importance of variables

Using Shapley values, we can also depict nonlinearities in the importance of different variables as captured by the machine learning models. Figure VII plots the Shapley values of the key predictors as a function of the actual input values. Each circle shows one observation with the crisis observations being highlighted in red. A Shapley value

	Shapley regression					Logistic regression		
	Direction	Share	Coefficient	α-level	p	Coefficients	α-level	p
Global slope	−	0.23	0.55	***	0.000	-0.61	***	0.000
Global credit	+	0.18	0.33	***	0.000	0.67	***	0.000
Domestic slope	−	0.11	0.37	***	0.000	-0.58	***	0.000
Domestic credit	+	0.11	0.34	***	0.000	0.43	***	0.002
CPI	−	0.07	0.28	***	0.002	-0.24		0.160
Debt service ratio	+	0.05	0.06		0.236	0.16		0.347
Consumption	−	0.05	0.17	**	0.029	-0.42	***	0.003
Investment	+	0.04	0.18	***	0.005	0.32	**	0.016
Public debt	−	0.04	-0.04		0.313	-0.03		0.845
Broad money	+	0.04	-0.11		0.095	0.04		0.817
Stock market	−	0.04	0.16	**	0.020	-0.13		0.451
Current account	−	0.03	-0.05		0.282	-0.08		0.525

TABLE V: Left side: Shapley regression. Direction of alignment between predictor and crisis outcome (same as sign of logistic regression), coefficients, p-values against the null hypothesis (positive coefficients only) and predictive share of variable. Right side: Coefficients and p-values of a logistic regression from Model (4) in Table II. Significance levels: *p<0.1; **p<0.05; ***p<0.01.

greater than zero indicates an increase in the predicted probability of a crisis relative to the model mean, while the opposite holds for negative values.

To test the importance of nonlinearities, we fit linear (black line) and cubic polynomial (blue line) regressions to the input-Shapley value relations. The goodness-of-fit in terms of R^2 is substantially better for nonlinear relationships, particularly global credit. The nonlinear relationships are also intuitive. A severe flattening or inversion of the yield curve is associated with a more than proportional increase in the probability of crisis, as is higher global and domestic credit growth. By contrast, when credit growth is muted or the yield curve is strongly upwards sloping, changes in these variables make relatively limited difference to the predicted crisis probability. Together, these results highlight that financial systems are particularly susceptible to a crisis when some variables are in the risky tails of their distributions.

Does modelling nonlinearities improve the predictive power of all predictors? To answer this question, we conduct a simple test by regressing the crisis outcome on each indicator independently, once on its actual values (reflecting a linear model) and once on

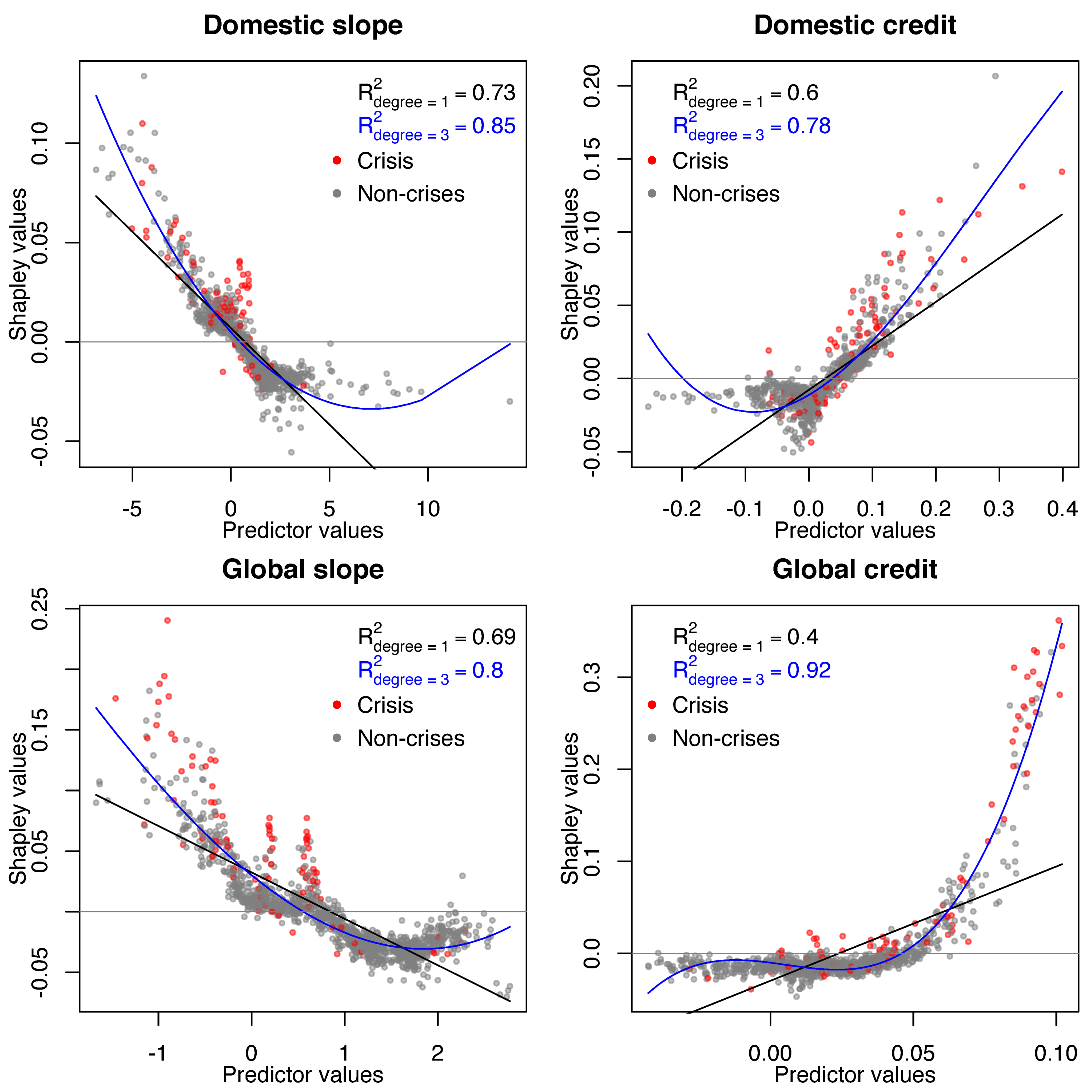

FIGURE VII: Indicator values plotted against Shapley values for each observation on the four most predictive indicators. Crisis observations are highlighted in red.

the Shapley values of extreme trees.[18] Table VI compares the AUC and the goodness-of-fit (log-likelihood) of the two models for each variable in the baseline approach. For our key indicators, the goodness-of-fit and the AUC score are significantly better when we regress on Shapley values. Together with the observations in Figure VII, this highlights the value of machine learning approaches in capturing nonlinearities that are likely to be both statistically and economically meaningful when predicting financial crisis.

[18]It should be noted that the Shapley values of each predictor do not show the effect independent of the other predictors but rather include all interactions with the other predictors.

	AUC		Log likelihood	
	Linear	Shapley	Linear	Shapley
Global slope	0.736	0.812***	-305	-273***
Global credit	0.667	0.720*	-309	-297***
Domestic slope	0.710	0.765***	-316	-307***
Domestic credit	0.686	0.677	-312	-305**
CPI	0.597	0.699***	-331	-316***
Debt service ratio	0.651	0.722***	-322	-311***
Investment	0.608	0.643	-330	-323
Stock market	0.470	0.628**	-336	-301***
Consumption	0.558	0.626**	-332	-322**
Broad money	0.606	0.594	-331	-332
Current account	0.572	0.594	-334	-335
Public debt	0.528	0.493	-336	-335

TABLE VI: AUC and log likelihood of univariate logistic regressions. The crisis outcome is regressed on the values of the variables (Linear) and on the Shapley values extracted from extreme trees. We use the DeLong test (DeLong et al., 1988) and Vuong's closeness test (Vuong, 1989) to evaluate whether the difference in AUC and log-likelihood are significant. Significance levels: *p<0.1; **p<0.05; ***p<0.01.

7 Robust indicators for financial crisis prediction

Using the models and techniques developed, we examine our main findings in greater detail. We focus particularly on the economic interpretation of our yield curve results, interactions across variables and the robustness of different variables across time.

7.1 The role of the yield curve

7.1.1 The yield curve and recessions

In the literature, the slope of the yield curve has often been seen as a harbinger of recessions (Estrella and Hardouvelis, 1991; Rudebusch and Williams, 2009). Financial crises and recessions are correlated events that regularly co-occur. To ensure that we are not just predicting recessions but indeed financial crises, we control for recessions when testing the predictive power of the slope. We follow Baker et al. (2018), who use the same annual dataset, and define a recession as a period where real GDP declines after it has increased in the previous year.

We differentiate between two types of crises: (1) crises that were preceded by a recession one or two years ahead or co-occur with a recession in the same year; (2) crises that do not co-occur or were not preceded by a recession but may be followed by one.

We re-estimate our baseline logistic regression Model 4 for both types of crises. Concretely, we conduct the regressions on the subset of crises of the respective type and all non-crisis observations. The results of this exercise are summarised in Table VII.

The domestic slope of the yield curve loses part of its predictive power when a crisis co-occurs or is preceded by a recession (Model 11). This is not surprising as the onset of a recession typically causes the yield curve to steepen. This leads to a lower predicted crisis probability even though a crisis subsequently occurs. By contrast, the domestic yield curve is particularly informative in the *absence* of a recession (Model 12). Moreover, the significance of the global slope in both cases highlights the importance of global interest rate conditions when evaluating the likelihood of a crisis. Together, these results strongly suggest that the yield curve slope can help to predict financial crises over and above the value it may have in predicting recessions. We examine why this may be the case in the next subsection.

7.1.2 The slope of the yield curve and the level of interest rates

While credit growth is an established predictor for financial crises in the literature, the role of the yield curve remains relatively underexplored. To further analyse the potential economic relevance of the yield curve in financial crisis prediction, we investigate its components, i.e. the short and long-term nominal interest rates using logistic regression in in-sample fitting. To increase the statistical power, we exclude the global slope from the regression analyses but include all other covariates.

Table VIII presents the results with Model 5 using only the slope and Models 6 and 7 respectively showing how predictive the domestic nominal short-term and long-term rates are. The short-term rate is a significant predictor, while the long rate is not. Model 8 uses both interest rates. Compared to using the short-term rate alone, the goodness of fit improves significantly. Model 8 implicitly learns a function of the interest rates that closely mimics the slope. In particular, let l and s be the long and short-term rate, respectively. Then, the model learns $1.641s - 1.367l = -1.367(l - 1.2s)$. This model is not significantly better ($p = 0.44$) than Model 5, which only uses the slope.

	(11) Crises and recession co-occur (n = 62)	(12) Crises and recession do not co-occur (n = 33)
Domestic slope	−0.369*	−0.793***
	(0.193)	(0.197)
Global slope	−0.664***	−0.476**
	(0.196)	(0.221)
+ Covariates	Global credit, domestic credit, CPI, debt service ratio, broad money, public debt investment, current account	
Observations	1,216	1,187
Log Likelihood	-167.607	-121.615
Akaike Inf. Crit.	361.214	269.230

TABLE VII: Logistic regression fitting financial crises for two subsets of crises. Model 11: Crises that co-occur with a recession in the same year or are preceded by recession 1–2 years ahead. Model 12: Remaining crises, which do not follow or co-occur with a recession. Significance levels: *p<0.1; **p<0.05; ***p<0.01.

Together with the machine learning robustness checks previously presented in Table III, this analysis confirms that the yield curve slope is of particular interest rather than just the level of short or long-term interest rates. But as discussed in Section 2, under certain theoretical mechanisms, a flat or inverted yield curve may be of greater concern when nominal yields are low. Models 9 and 10 test this hypothesis. They establish a statistically significant between the yield curve slope and both the short and long-term rate, with the former showing a stronger interaction effect. Using real interest rates rather than nominal rates (lower part of Table VIII), Models 6–8 do not qualitatively change. However, the significance of the interaction of the slope with the interest rates disappears (Models (16) and (17)).

Figure VIII illustrates these interactions. It shows the predicted probability of crisis as a function of the domestic slope (horizontal axis), when the nominal short-term rate (left panel) and long-term rate (right panel) is at its mean and one standard deviation

	(5)	(6)	(7)	(8)	(9)	(10)
Domestic slope	−0.786***				−1.105***	−0.826***
	(0.131)				(0.206)	(0.130)
Domestic nominal short rate		0.698***		1.641***	0.405*	
		(0.167)		(0.272)	(0.220)	
Domestic nominal long rate			0.044	−1.367***		0.226
			(0.182)	(0.313)		(0.192)
Domestic slope × nominal short rate					0.482***	
					(0.147)	
Domestic slope × nominal long rate						0.186**
						(0.076)
+ Covariates	as in Table VII					
Observations	1,249	1,249	1,249	1,249	1,249	1,249
Log Likelihood	-257.605	-267.668	-276.245	-257.305	-251.219	-255.670
Akaike Inf. Crit.	539.211	559.336	576.489	540.610	530.438	539.339

		(13)	(14)	(15)	(16)	(17)
Domestic real short-term rate		0.552***		1.835***	0.307	
		(0.151)		(0.303)	(0.189)	
Domestic real long-term rate			0.097	−1.607***		0.179
			(0.156)	(0.330)		(0.186)
Domestic slope × real short rate					0.156	
					(0.110)	
Domestic slope × real long rate						0.011
						(0.100)
+ Covariates	as in Table VII					
Observations	1,249	1,249	1,249	1,249	1,249	1,249
Log Likelihood	-257.605	-269.507	-276.084	-257.062	-256.039	-257.056
Akaike Inf. Crit.	539.211	563.014	576.168	540.123	540.077	542.112

TABLE VIII: Logistic regression model including domestic nominal (upper part) and real (lower part) short and long-term interest rates. Significance levels: *p<0.1; **p<0.05; ***p<0.01.

above or below it. All other predictors are held constant at their mean value. It is evident that when the yield curve is inverted, the predicted probability of crisis is higher when the level of interest rates is low (red line). These effects are stronger for nominal than for real interest rates in line with the finding presented in Table VIII.

We also test whether our best performing machine learning model (extreme trees) exploits these interactions in our out-of-sample experiments, which have the advantage that they do not explicitly pre-specify any interactions. We use the baseline set of variables, excluding the global slope and add nominal short and long-term interest rates, respectively. The interactions of both rates with the domestic slope obtain significant coefficients ($p < 0.01$) in a Shapley regression.

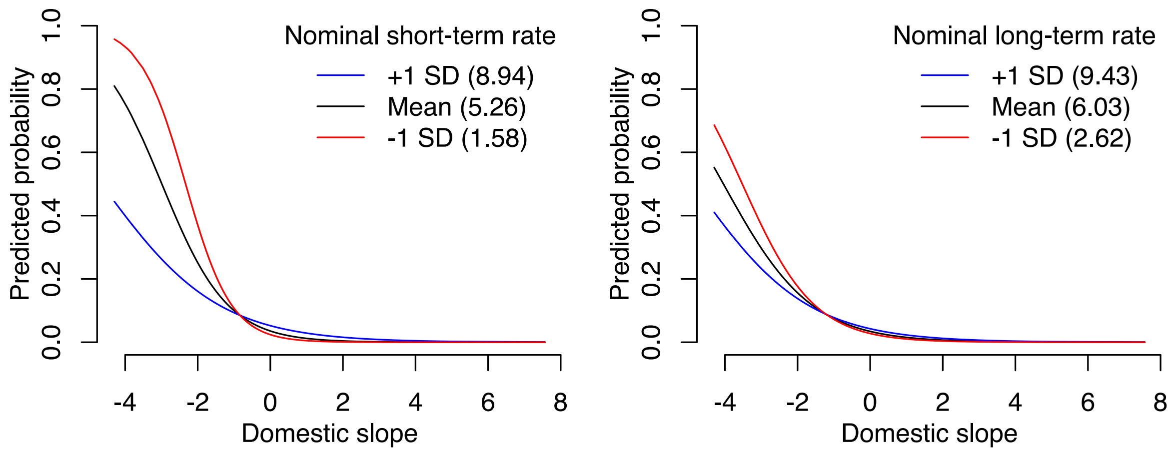

FIGURE VIII: Interaction effects in Models 9 and 10. The plot depicts the effect of the slope on the predicted probability of crisis at three different levels of the short-term rate and long-term rate (mean, ± 1 standard deviation). All remaining predictors are held constant at their mean value.

Taken together, these results imply that a flat or inverted yield curve is of greater concern when nominal yields are low. If the yield curve slope only affected crisis probabilities via its effect on net interest margins, the interaction with the level of nominal interest rates should not be that important except at the effective lower bound which is not relevant in most of our sample. So these results point to the potential importance of a search-for-yield channel when the yield curve is flat or inverted, with financial market players taking on more risk to boost nominal returns when term premia and nominal yields are both low.

7.1.3 Robustness of the global yield curve

A natural question is whether the importance of the global yield curve proxies a particular country. We address this by replacing the global slope variable with the domestic slopes of all individual countries and then compare model performances and the importance of variables for our best performing model (extreme trees).[19] The predictive performance of extreme trees consistently deteriorates using individual country slopes. Using a combination of the UK (pre-WW2) and the US slope (post-WW2) also leads to inferior results. Therefore, we conclude that we are truly picking up global financial conditions with our global yield curve variable rather than simply reflecting the conditions in, for example, a dominant country in the global financial system.

7.2 The importance of variables across time

The financial and economic system has changed substantially over the period covered in our dataset. We therefore expect that the prediction of crises is also subject to changes over time. As we are interested in how well the predictors differentiate between crisis and non-crisis observations, we compute the *Shapley difference*, i.e. the mean Shapley value of crisis observations subtracted by the mean Shapely values of non-crisis observations. Figure IX shows the Shapley differences for specific time periods in the data (i.e. pre and post-WW2, crises in the 1990s, and the global financial crisis).

Before World War 2, the global slope and domestic credit mainly differentiates crisis from non-crisis observations. During the series of financial crises that occurred in the 1990s, domestic credit, the global slope, and the debt service ratio are key predictors. During the global financial crisis, and only then, by far the most important predictor is global credit. This may be partly driven by financial globalisation which has magnified the importance of international credit growth (Cesa-Bianchi et al., 2019). For example, Germany and Switzerland experienced negative domestic credit to GDP growth before the global financial crisis. Nevertheless, both countries experienced a financial crisis because

[19] We changed the cross-validation procedure to avoid the machine learning models overfitting because the value of the slope is the same for all individual countries in a specific year. We assigned all observations of the same year to the same fold and also required that the two observations before a crisis (positive outcome) are in the same fold (see Appendix B.1)

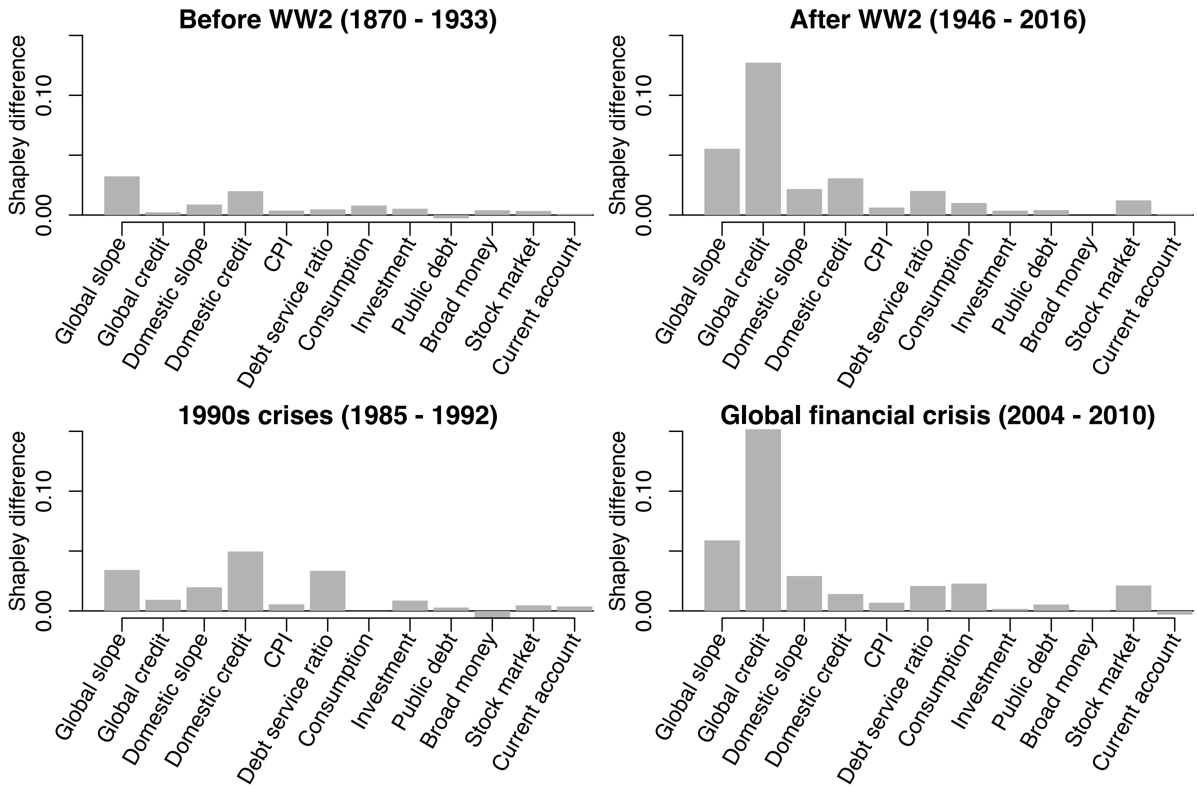

FIGURE IX: Mean difference of Shapley values (crisis minus non-crisis observations) for different periods

their banking sectors were highly exposed to global risks.

The results from Figure IX are based on a single model covering the whole sample period. It may, however, be the case that relations between variables changed fundamentally such that a single, albeit flexible model, cannot adequately differentiate between different regimes. To test this, we fit extreme trees independently to pre and post-WW2 samples and repeat the above exercise. Reassuringly, all main findings hold. The only main change is an increased importance of the debt servicing ratio in the pre-WW2 sample. We present these results in Appendix B (Figure B.III).

The analysis highlights that the global slope of the yield curve is a key predictor across the whole period covered by our dataset. This might be explained by two regimes. First, the Gold Standard and then pegged exchange rates established a close connection of macroeconomic policies across countries (Obstfeld et al., 2005). Later, the globalisation of the world economy and financial markets, especially a greater global bond market integration (Diebold et al., 2008) may have cemented the importance of the global yield curve.

7.3 Global-domestic interactions

The Shapley values shown in Figure V show the total effect of a variable on model predictions. But they do not tell us how much of the effect can be attributed to that variable alone and how much to the interaction with other variables. But the Shapley value framework also allows us to measure explicitly how much a particular interaction drives a prediction (Equations 2 and 3).

We investigate Shapley interactions in the extreme trees model. We focus on pairwise interactions of our two global variables for credit growth and the slope of the yield curve with the remaining variables that have a major predictive share and a significant main effect according to Table V. The reasons for this choice are two-fold. First, global variables are overall the most important predictors, and, second, interactions involving only domestic factors turn out to be generally weaker than interactions with global factors.

		Shapley regression					**Logistic regression**		
Variable 1	Variable 2	Direction	Share	Coefficient	α	p	Coefficients	α	p
Global slope	Domestic credit	−	0.07	0.24	***	0.004	-0.36	**	0.018
Global slope	Domestic slope	−	0.04	0.09		0.169	-0.06		0.700
Global slope	Investment	−	0.04	0.41	***	0.000	-0.46	***	0.002
Global slope	CPI	+	0.04	0.25	***	0.003	0.36	**	0.015
Global slope	Consumption	+	0.03	0.13	*	0.058	0.03		0.848
Global slope	Stock market	−	0.03	0.09		0.185	-0.25		0.167
Global credit	Domestic credit	−	0.03	0.11	*	0.083	-0.07		0.522
Global credit	Domestic slope	−	0.03	0.18	**	0.027	0.06		0.705
Global credit	Investment	+	0.02	0.14	**	0.036	-0.07		0.576
Global credit	CPI	+	0.04	0.39	***	0.001	0.53	***	0.007
Global credit	Consumption	−	0.03	0.23	***	0.002	-0.18		0.209
Global credit	Stock market	+	0.03	0.23	**	0.014	0.38	**	0.014

TABLE IX: Shapley regression and logistic regression on variable interactions. Each row is based on a different regression including the respective interaction and the main effects of the 12 predictors. To estimate the direction of an interaction, we regress the crisis outcome on the respective input variables and their interaction. Significance levels: *p<0.1; **p<0.05; ***p<0.01.

Table IX shows the summary statistics for the interaction terms in Shapley regressions.

Each interaction is tested in an individual Shapley regression that controls for the main effects of all 12 predictors but does not contain the other interactions to avoid collinearity issues. We again include the corresponding results for the logistic regression for reference.

Interactions with the global yield curve slope are particularly strong. Especially, a low or inverted global slope is of particular concern when coupled with high growth in domestic credit and investment. The predictive power of the former surpasses that of most individual variables by share and significance when compared to Table V.

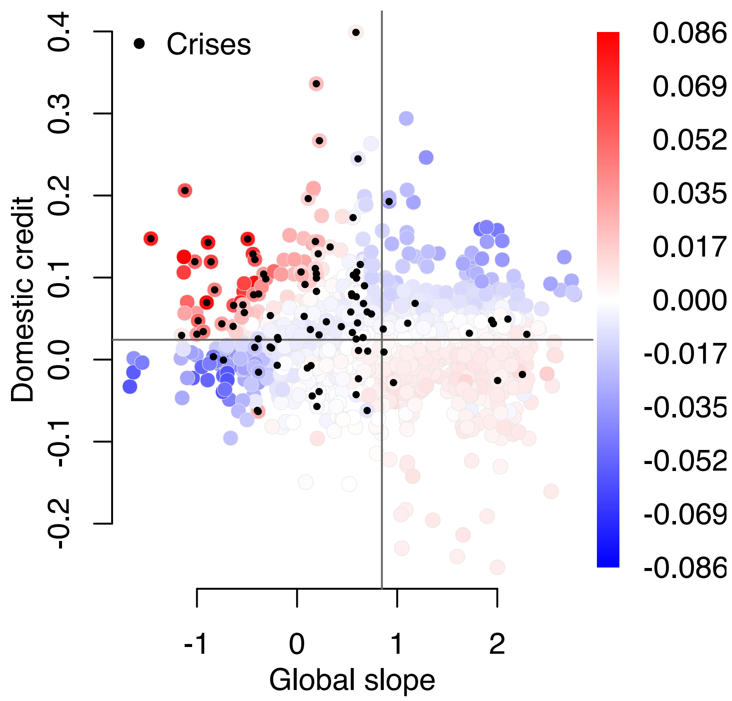

FIGURE X: Shapley interactions between domestic credit and the global slope of the yield curve. The scatter plot shows all observations as a function of their values on two predictors. The colour of the observations denotes the value of the Shapley interaction, with darker red indicating a higher predicted probability of crisis. Crisis observations are highlighted with black circles. The quadrants are defined by the mean value of each variable.

To illustrate this further, Figure X shows the Shapley interactions of the global slope of the yield curve and domestic credit for extreme trees. The values of the input variables are shown on the horizontal and vertical axis. The grey lines represent the means of variables, indicating four quadrants of low/high value combinations. The value of the Shapley interaction is shown by the colour with darker red indicating a higher probability of crisis. Indeed, most crises (black circles) fall into the upper left quadrant of high credit growth and a globally flat/inverted yield curve, corroborating the findings from Table IX. Overall, this analysis points towards the potential importance of the international yield curve environment in amplifying domestic exuberance.

8 Conclusion

This paper shows that machine learning models outperform logistic regression in predicting financial crises on a macroeconomic dataset covering 17 countries between 1870 and 2016 in both out-of-sample cross-validation and recursive forecasting. The consistently most accurate models are decision-tree based ensembles, such as extremely randomised trees and random forests. The gains in predictive accuracy justify the use of initially more opaque machine learning models. To understand their predictions, we apply a novel Shapley value framework which allows us to examine the contributions of individual predictors economically and statistically.

All models consistently identify similar predictors for financial crises, although there are some variations across time reflecting changes in the nature of the global monetary and financial system over the past 150 years. These key early warning signs include: (i) prolonged high growth in domestic credit relative to GDP; (ii) a flat or inverted yield curve especially when nominal yields are low, and (iii) a shared global narrative in both of these dimensions as indicated by the importance of global variables. While the crucial role of credit is an established result in the literature, the predictive power of the yield curve has obtained far less attention as an early warning indicator. Indeed, the global yield curve slope is a robust crisis indicator throughout the sample period. This contrasts with global credit, which proves a strong indicator only for the recent global financial crisis.

We also inspect nonlinearities and interactions identified by the machine learning models. Global credit shows a particularly strong nonlinearity—only very high global credit growth beyond a certain point influences the prediction of the models. Interactions are particularly strong between global and domestic indicators. For instance, a globally flat or inverted yield curve coupled with strong domestic credit growth may highlight a significant crisis risk. Overall, our findings suggest a combination of low risk perception, search-for-yield behaviour and strong credit growth in the years preceding a crisis.

Our results help policy makers to identify the risk of financial crises in advance and potentially act on these signals. The ability to do so is crucial given the enormous economic, political, and social consequences that financial crises entail. With more accurate predictive models and reliable indicators complementing softer information and judgement, policy makers can pre-emptively adjust macroprudential measures such as countercyclical capital buffers (BCBS, 2010). Such action may help to avoid or at least reduce the consequences of financial crises.

More generally, our results highlight the potential value of machine learning models for broader economic policy making in two key dimensions. First, our approach illustrates how machine learning techniques can uncover important nonlinearities and interactions which facilitate superior out-of-sample prediction and forecasting even in situations characterised by relatively small datasets with limited observations of the event of interest. Second, the novel Shapley value approach demonstrates how the black box concern linked to the practical policy application of machine learning models may be overcome. In particular, by providing a mechanism to identify the key economic drivers of the predictions generated by such models, it allows insights from machine learning models to be integrated into a broader decision making framework while preserving the transparency and accountability of any resulting public policy decision.

A Machine learning models implementation

This Appendix describes the implementation and the parameter settings of the machine learning algorithms. If a parameter is not specified in the following, we used its default value.

Logistic regression. We used the `SGDCClassifier` implementation from the Python package `sklearn` with *penalty* = None and *loss* = log. We also tried regularised logistic regression (Lasso, Elastic-net) but did not observe an improvement in performance.

Random forest. We used the `RandomForestClassifier` implementation from the Python package `sklearn` with *n_ estimators*= 1000 and used the default values of the other hyperparameters as random forests are known to be rather insensitive to the choice of hyperparameters. Nevertheless, we also tested a version of random forest for which we searched for *max_features* $\in \{1, 2, ..10\}$ and *max_ depth* $\in \{2, 3, 4, 5, 7, 10, 12, 15, 20\}$ using 5-fold cross-validation in the training set. It did not improve the performance.

Extremely randomised trees. We used the `ExtraTreesClassifier` implementation from the Python package `sklearn` with *n_ estimators=1000* and used the default values of the other hyperparameters. We also tested a version for which we searched for hyperparamters *max_features* $\in \{1, 2, ..10\}$ and *max_ depth* $\in \{2, 3, 4, 5, 7, 10, 12, 15, 20\}$ using cross-validation in the training set but it did not improve the performance.

Support vector machine. We used the `SVC` implementation from the Python package `sklearn` and searched for hyperparameters $C \in \{2^{-5+15\times\frac{0}{9}}, 2^{-5+15\times\frac{1}{9}}, ..., 2^{-5+15\times\frac{9}{9}}\}$ and *gamma* $\in \{2^{-10+13\times\frac{0}{9}}, 2^{-10+13\times\frac{1}{9}}, ..., 2^{-10+13\times\frac{9}{9}}\}$ using cross-validation in the training set. We trained 25 SVMs in each training sample. For each model, we upsample the crisis observations, i.e. we randomly draw with replacement as many crisis observations as there are non-crisis observations in the training set. The hyperparameter search was conduced for each model independently. The final prediction is the mean predicted value across all models.

Neural network. We used the `MLPClassifier` implementation from the Python package `sklearn` with *solver*=lbfgs and searched for hyperparameters *alpha*=$\in \{10^{-3+6\times\frac{0}{9}}, 2^{-3+6\times\frac{1}{9}}, ..., 2^{-3+6\times\frac{9}{9}}\}$, *activation* $\in$ {tanh, relu}, and *hidden_ layer_ sizes* $\in \{n/3, n/2, n, (n, n/2), (n, n), (2n, n), (2n, 2n)\}$, where n is the number of predictors. Numbers all rounded to the nearest integer. We trained 25 neural networks on bootstrapped samples of each training set. The hyperparameter search was conduced for each model independently. The final prediction is the mean predicted value across all models.

Decision Tree C5.0 We used the `C5.0` implementation from the R package `C50` with *trials*= 1, *noGlobalPruning* = False, and *minCases*= 1. We weight the observations such that both classes contribute equally to the training set. The objects in the positive class (N_+) were weighted by $0.5/\frac{N_+}{N_+ + N_-}$ and the objects in the negative class (N_-) by $0.5/(1 - \frac{N_+}{N_+ + N_-})$.

CART We used the `rpart` implementation from the R package `rpart` with *maxdepth*= 10 and cross-validated the complexity parameter. We weight the objects such that both contribute equally to the training set. We do not report CART in the paper because it performed less well to the other decision tree algorithm C5.0.

Gradient boosting We used the `XGBClassifier` implementation from the Python package `xgboost` and searched for hyperparameters *learning_ rate* $\in \{0.01, 0.05, 0.1, 0.15, 0.2\}$, *min_ child_ weight* $\in \{1, 5, 10\}$, and n_estimators $\in \{50, 100, 250, 500\}$. We upsampled the crisis observations, i.e. we randomly draw with replacement as many crisis observations as there are non-crisis observations in the training set. We do not report Gradient Boosting in the paper because it performed less well to the other tree ensembles random forest and extremely randomised trees.

B Results Appendix

B.1 Four types of cross-validation

In our main experiment, we use 5-fold cross-validation to estimate the out-of-sample performance of the prediction models. Different constraints can be imposed when assigning the observations to the folds. Here, we investigate whether these constraints materially change our results, both in terms of predictive performance and in terms of variable importance. We test four types of cross-validation. First, in unconstrained cross-validation, country-year pairs are randomly assigned to the five folds. Second, we impose the constraint that the two observations of the same crisis (two years before the actual crisis observation) are assigned to the same fold. This type of cross-validation is the approach reported in the main body of the paper. Third, we assign all observations of the same year to the same fold. Fourth, we combine the two constraints and require that observations of the same year and crisis are in the same fold.

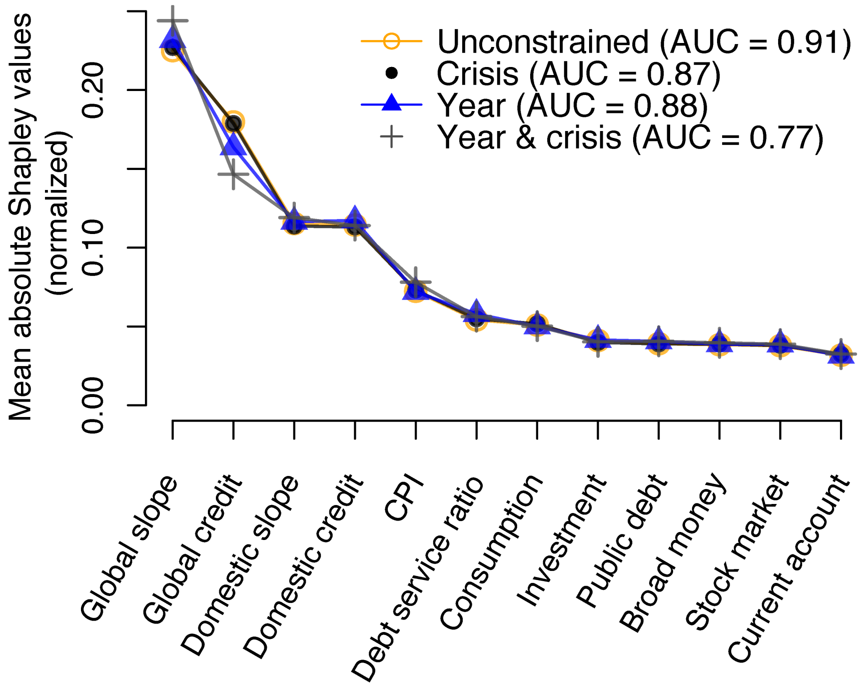

FIGURE B.I: Shapley values for extreme trees for the four different cross-validation experiments. "Crisis" corresponds to the baseline approach presented in the main part of the paper.

In the empirical test of these four types of cross-validation, we use the variables and transformations of our baseline experiment and report the performance of the most accurate model, extreme trees. The unconstrained cross-validation achieves the highest performance (AUC = 0.91), followed by the constrained procedure that assigns all observation of the same year to the same fold (AUC of 0.88). The baseline procedure achieves an AUC of 0.87. The strictest constraint of assigning both the year and crisis to the same fold reduces the AUC to 0.77. This pronounced decline in performance is mostly driven by the reduced accuracy on the global financial crisis. In 15 of the 17 countries, the onset of the crisis was in 2008. With the constrained cross-validation, the observations of the two years before 2008 are either all in the training set or in the test set. In the former case, the importance of global credit is learned but is not very useful for the prediction in the test set. In the latter case, the importance of global credit cannot be learned from the training data and therefore the prediction on the observations of the global financial crisis in the test set is not very accurate.

Figure B.I confirms this explanation. It shows the mean absolute Shapley values for the four types of cross-validation. Generally, they all show highly similar patterns. However, the global credit variable is a less important predictor for the constrained cross-validation with the year plus crisis constraint. But, in all four types of cross-validation, extreme trees still outperform logistic regression, by at least 4 percentage points in AUC.

B.2 Global variables

The most straightforward operationalisation of global credit to GDP growth is the mean credit to GDP growth across all countries in a particular year. Similarly, the global slope could be measured as the mean slope of the yield curve across all countries. However, this implementation is problematic, as it creates a data leakage between training and test sets.

For example, assume that half of the 2008 observations are in the training and set and the other half in the test set. As most countries experienced a crisis in 2008, a flexible machine learning model learns to associate the exact value of the global variable in that year with a high probability of crisis. It implicitly learns the year, instead of learning a trend from the values of the variable. To confirm that, we trained extreme trees on each of the global variables separately. We randomly shuffled the actual values of the global variables across years and just made sure that all observations of the same year had the same value. The out-of-sample AUC was 0.82 for both global variables. By implicitly learning an association between year and country, without any actual information about the level of global credit, or global slope, we obtain a very high predictive performance.

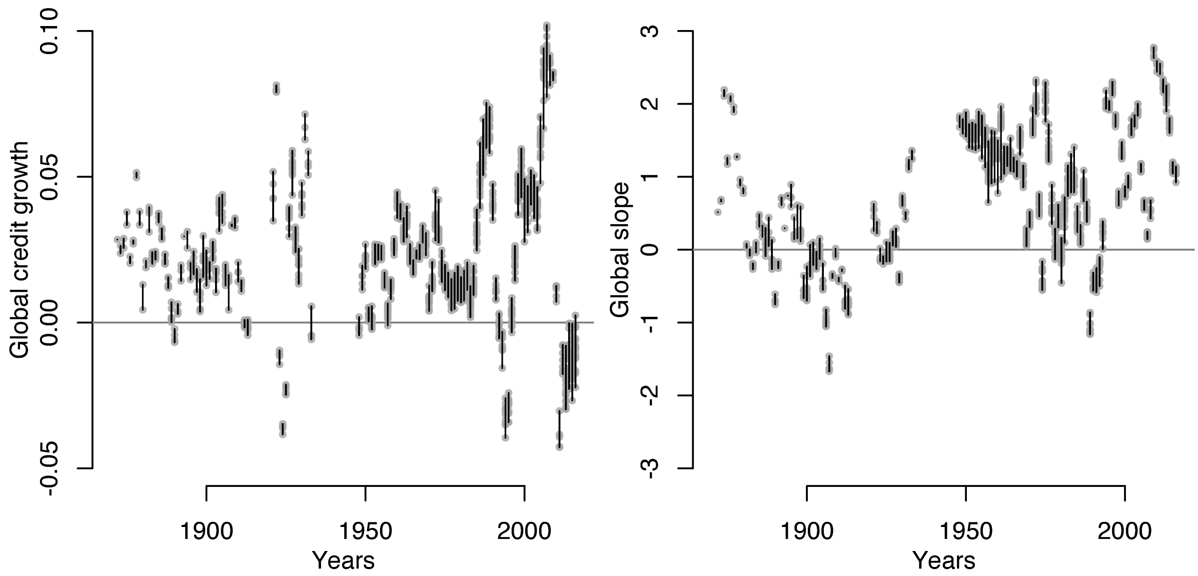

FIGURE B.II: Depiction of the global variables. The grey circles show the actual values, the vertical lines show the range of values in each year.

To avoid this effect, we defined the global variable for country c in year y as the average value of the domestic variables in year y for all countries except c. Several checks confirm that this operationalisation of the global variables is not prone to the same problem as the the simple average across all countries and that our cross-validation results are therefore not positively biased.

First, Figure B.II shows our global variables (circles). The range of the values overlaps between years such that the model cannot infer the year from the variable. Second, we used our global variables as the only predictor in the cross-validation experiment. Now, extreme trees obtained an AUC of only 0.58 and 0.62 for global credit and slope, which confirms that our variable does not directly map to years. Third, the Shapley analysis in Figure VII depicts that extreme trees learn a smooth monotonic association between the actual value of the global variables and the probability of financial crises rather than a direct mapping of values to probability of crisis. Fourth, the constrained cross-validation (Figure B.I) and the forecasting experiment both confirm the crucial role of the global variables. In these experiments, an implicit learning of the year can be ruled out as observations of the same year are constrained to be all in the training or test set but not distributed among them.

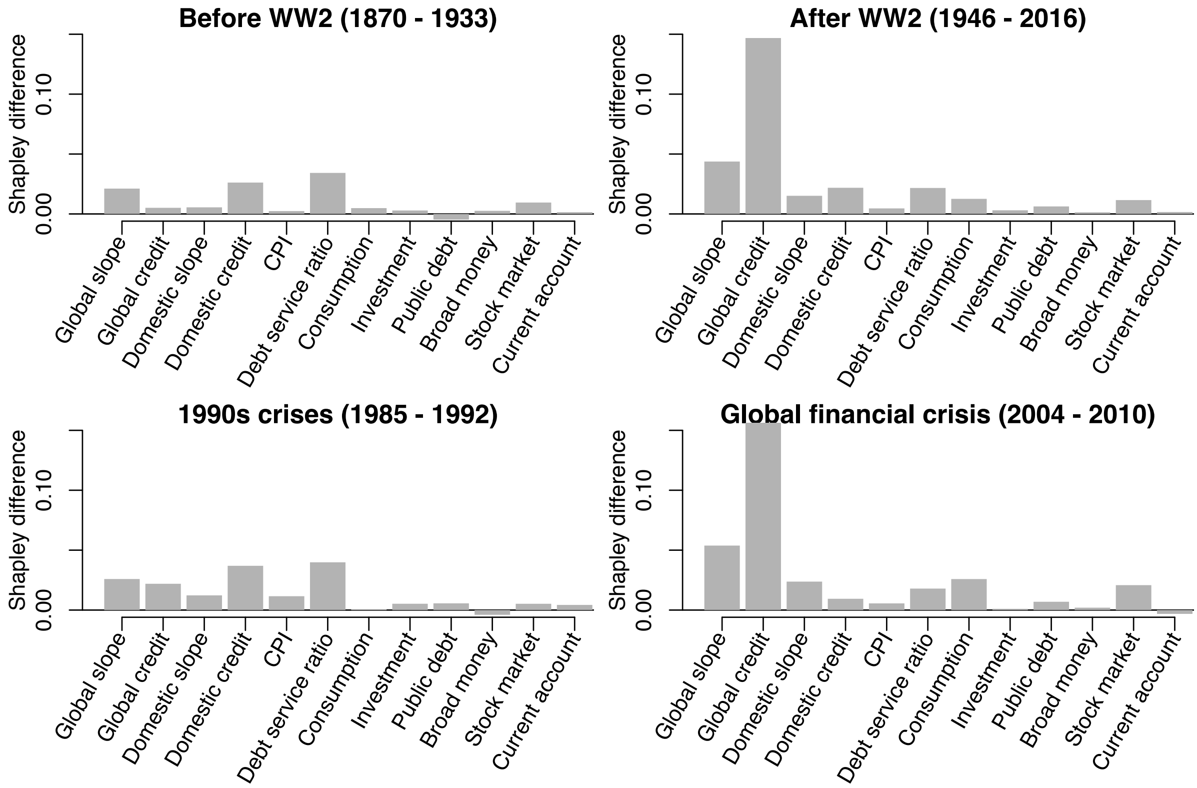

Figure B.III: Mean difference of Shapley values (crisis - non-crisis observations) for different periods. The top left plot is based on an extreme trees model trained on pre-WW2 observations, only. The other plots are based on an extreme trees model based on post-WW2 observations, only. Note that this figure is qualitatively very similar to Figure IX in the main text which is based on a single model trained on the whole sample.

References

Abbritti, Mirko, Salvatore Dell'Erba, Antonio Moreno, Sergio Sola et al. (2018) "Global factors in the term structure of interest rates", *International Journal of Central Banking*, Vol. 14, No. 2, pp. 301–340.

Adrian, Tobias and Hyun Song Shin (2010) "Financial intermediaries and monetary economics", in *Handbook of monetary economics*, Vol. 3: Elsevier, pp. 601–650.

Adrian, Tobias, Arturo Estrella, and Hyun Song Shin (2010) "Monetary cycles, financial cycles, and the business cycle", Staff Reports 421, Federal Reserve Bank of New York.

Agresti, Alan (1996) *An introduction to categorical data analysis*, Chap. Building and applying logistic regression models, pp. 137–172, Hoboken, New Jersey: John Wiley.

Aikman, David, Andrew G. Haldane, and Benjamin D. Nelson (2013) "Curbing the Credit Cycle", *The Economic Journal*, Vol. 125, No. 585, pp. 1072–1109.

Aikman, David, Andrew Haldane, Marc Hinterschweiger, and Sujit Kapadia (2018) "Rethinking financial stability", Bank of England working papers 712, Bank of England.

Aikman, David, Benjamin Nelson, and Misa Tanaka (2015) "Reputation, risk-taking, and macroprudential policy", *Journal of Banking & Finance*, Vol. 50, pp. 428 – 439.

Akinci, Ozge and Jane Olmstead-Rumsey (2018) "How effective are macroprudential policies? an empirical investigation", *Journal of Financial Intermediation*, Vol. 33, pp. 33–57.

Alessi, Lucia and Carsten 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*, Vol. 27, No. 3, pp. 520–533.

——— (2018) "Identifying excessive credit growth and leverage", *Journal of Financial Stability*, Vol. 35, pp. 215–225, April.

Aliber, Robert Z. and Charles P. Kindleberger (2015) *Manias, Panics, and Crashes: A History of Financial Crises, Seventh Edition*, Basingstoke, Hampshire New York: Palgrave Macmillan, 7th edition.

Amodei, Dario, Sundaram Ananthanarayanan, Rishita Anubhai, Jingliang Bai, Eric Battenberg, Carl Case, Jared Casper, Bryan Catanzaro, Qiang Cheng, Guoliang Chen et al. (2016) "Deep speech 2: End-to-end speech recognition in English and Mandarin", in *International Conference on Machine Learning*, pp. 173–182.

Babeckỳ, Jan, Tomas Havranek, Jakub Mateju, Marek Rusnák, Katerina Smidkova, and Borek Vasicek (2014) "Banking, debt, and currency crises in developed countries: Stylized facts and early warning indicators", *Journal of Financial Stability*, Vol. 15, pp. 1–17.

Baker, Sarah S., J. David López-Salido, and Edward Nelson (2018) "The money view versus the credit view", *Board of Governors of the Federal Reserve System. Finance and Economics Discussion Series*, Vol. 042.

BCBS (2010) *Guidance for national authorities operating the countercyclical capital buffer*, Basel Committee on Banking Supervision, Bank for International Settlements.

Bernanke, Ben S. (2009) *Asia and the Global Financial Crisis*, Speeach at the Federal Reserve Bank of San Francisco's Conference on Asia and the Global Financial Crisis, Santa Barbara, California.

Bernanke, Ben S. and Alan S. Blinder (1992) "The federal funds rate and the channels of monetary transmission", *American Economic Review*, Vol. 82, No. 4, pp. 901–921.

Bernanke, Ben S., Mark Gertler, and Simon Gilchrist (1999) *The financial accelerator in a quantitative business cycle framework*, Chap. Chapter 21, pp. 1341–1393: Elsevier.

Bernstein, Michael A (1987) *The Great Depression: delayed recovery and economic change in America, 1929-1939*: Cambridge University Press.

Betrán, Concha and María A Pons (2013) "Understanding Spanish financial crises, 1850-2000: What determined their severity?", in *European Historical Society Conference*, Vol. 6.

Beutel, Johannes, Sophia List, and Gregor von Schweinitz (2018) "An evaluation of early warning models for systemic banking crises: Does machine learning improve predictions?", *Deutsche Bundesbank Discussion Paper Series*, No. 48.

Bishop, Christopher M (2006) *Pattern recognition and machine learning*, New York: Springer.

Bordo, Michael, Barry Eichengreen, Daniela Klingebiel, and Maria Soledad Martinez-Peria (2001) "Is the crisis problem growing more severe?", *Economic Policy*, Vol. 16, No. 32, pp. 52–82.

Borio, Claudio and Philip Lowe (2002) "Asset prices, financial and monetary stability: exploring the nexus", BIS Working Papers 114, Bank for International Settlements.

Borio, Claudio and Haibin Zhu (2012) "Capital regulation, risk-taking and monetary policy: A missing link in the transmission mechanism?", *Journal of Financial Stability*, Vol. 8, No. 4, pp. 236–251.

Borio, Claudio, Leonardo Gambacorta, and Boris Hofmann (2017) "The influence of monetary policy on bank profitability", *International Finance*, Vol. 20, No. 1, pp. 48–63.

Breiman, Leo (1996) "Bagging predictors", *Machine Learning*, Vol. 24, No. 2, pp. 123–140.

——— (2001) "Random forests", *Machine Learning*, Vol. 45, No. 1, pp. 5–32.

Bussiere, Matthieu and Marcel Fratzscher (2006) "Towards a new early warning system of financial crises", *Journal of International Money and Finance*, Vol. 25, No. 6, pp. 953–973.

Carmona, Pedro, Francisco Climent, and Alexandre Momparler (2019) "Predicting failure in the US banking sector: An extreme gradient boosting approach", *International Review of Economics & Finance*, Vol. 61, pp. 304–323.

Casabianca, Elizabeth Jane, Michele Catalano, Lorenzo Forni, Elena Giarda, Simone Passeri et al. (2019) "An early warning system for banking crises: From regression-based analysis to machine learning techniques", Marco Fanno Working Papers 235, Dipartimento di Scienze Economiche "Marco Fanno".

Cecchetti, Stephen G, Marion Kohler, and Christian Upper (2009) "Financial crises and economic activity", NBER Working Papers 15379, National Bureau of Economic Research.

Cerutti, Eugenio, Stijn Claessens, and Luc Laeven (2017) "The use and effectiveness of macroprudential policies: New evidence", *Journal of Financial Stability*, Vol. 28, pp. 203–224.

Cesa-Bianchi, Ambrogio, Fernando Eguren Martin, and Gregory Thwaites (2019) "Foreign booms, domestic busts: The global dimension of banking crises", *Journal of Financial Intermediation*, Vol. 37, pp. 58–74.

Chakraborty, Chiranjit and Andreas Joseph (2017) "Machine learning at central banks", Staff Working Paper No. 674, Bank of England.

Coimbra, Nuno and Hélène Rey (2017) "Financial cycles with heterogeneous intermediaries", NBER Working Papers 23245, National Bureau of Economic Research, Inc.

Coleman, Major IV, Michael LaCour-Little, and Kerry D Vandell (2008) "Subprime lending and the housing bubble: Tail wags dog?", *Journal of Housing Economics*, Vol. 17, No. 4, pp. 272–290.

Croushore, Dean and Katherine Marsten (2016) "Reassessing the relative power of the yield spread in forecasting recessions", *Journal of Applied Econometrics*, Vol. 31, No. 6, pp. 1183–1191.

DeLong, Elizabeth R, David M DeLong, and Daniel L Clarke-Pearson (1988) "Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach", *Biometrics*, Vol. 44, No. 3, pp. 837–845.

Diebold, Francis X, Canlin Li, and Vivian Z Yue (2008) "Global yield curve dynamics and interactions: a dynamic Nelson–Siegel approach", *Journal of Econometrics*, Vol. 146, No. 2, pp. 351–363.

Döpke, Jörg, Ulrich Fritsche, and Christian Pierdzioch (2017) "Predicting recessions with boosted regression trees", *International Journal of Forecasting*, Vol. 33, No. 4, pp. 745–759.

Drehmann, Mathias and Mikael Juselius (2014) "Evaluating early warning indicators of banking crises: Satisfying policy requirements", *International Journal of Forecasting*, Vol. 30, No. 3, pp. 759–780.

Drehmann, Mathias, Claudio Borio, and Kostas Tsatsaronis (2011) "Anchoring countercyclical capital buffers: The role of credit aggregates", *International Journal of Central Banking*, Vol. 7, No. 4, pp. 189–240.

Duca, Marco Lo and Tuomas A Peltonen (2013) "Assessing systemic risks and predicting systemic events", *Journal of Banking & Finance*, Vol. 37, No. 7, pp. 2183–2195.

Duttagupta, Rupa and Paul Cashin (2011) "Anatomy of banking crises in developing and emerging market countries", *Journal of International Money and Finance*, Vol. 30, No. 2, pp. 354–376, March.

Estrella, Arturo and Gikas A Hardouvelis (1991) "The term structure as a predictor of real economic activity", *The Journal of Finance*, Vol. 46, No. 2, pp. 555–576.

Fernández-Delgado, Manuel, Eva Cernadas, Senén Barro, and Dinani Amorim (2014) "Do we need hundreds of classifiers to solve real world classification problems?", *The Journal of Machine Learning Research*, Vol. 15, No. 1, pp. 3133–3181.

Fouliard, Jérémy, Michael Howell, and Hélène Rey (2019) "Answering the Queen: Machine learning and financial crises", Working paper.

Frankel, Jeffrey and George Saravelos (2012) "Can leading indicators assess country vulnerability? evidence from the 2008–09 global financial crisis", *Journal of International Economics*, Vol. 87, No. 2, pp. 216–231.

Frankel, Jeffrey, Sergio L Schmukler, and Luis Serven (2004) "Global transmission of interest rates: monetary independence and currency regime", *Journal of International Money and Finance*, Vol. 23, No. 5, pp. 701–733.

Friedman, Jerome, Trevor Hastie, and Robert Tibshirani (2008) *The Elements of Statistical Learning*, New York: Springer Series in Statistics, 2nd edition.

Fujimoto, Katsushige, Ivan Kojadinovic, and Jean-Luc Marichal (2006) "Axiomatic characterizations of probabilistic and cardinal-probabilistic interaction indices", *Games and Economic Behavior*, Vol. 55, No. 1, pp. 72–99.

Gennaioli, Nicola and Andrei Shleifer (2018) *A Crisis of Beliefs: Investor Psychology and Financial Fragility*, Princeton, New Jersey: Princeton University Press.

Gentier, Antoine (2012) "Spanish banks and the housing crisis: Worse than the subprime crisis?", *International Journal of Business*, Vol. 17, No. 4.

Geurts, Pierre, Damien Ernst, and Louis Wehenkel (2006) "Extremely randomized trees", *Machine Learning*, Vol. 63, No. 1, pp. 3–42.

Giese, Julia, Henrik Andersen, Oliver Bush, Christian Castro, Marc Farag, and Sujit Kapadia (2014) "The credit-to-GDP gap and complementarity indicators for macroprudential policy: Evidence from the UK", *International Journal of Finance & Economics*, Vol. 19, No. 1, pp. 25–47.

Giese, Julia, Benjamin Nelson, Misa Tanaka, and Nikola Tarashev (2013) "How could macroprudential policy affect financial system resilience and credit? lessons from the literature", *Bank of England Financial Stability Papers*, No. 21.

Gordon, Robert J and Robert Krenn (2010) "The end of the great depression 1939-41: Policy contributions and fiscal multipliers", NBER Working Papers 16380, National Bureau of Economic Research.

Hamilton, James D (2018) "Why you should never use the Hodrick-Prescott filter", *Review of Economics and Statistics*, Vol. 100, No. 5, pp. 831–843.

Hoggarth, Glenn, Ricardo Reis, and Victoria Saporta (2002) "Costs of banking system instability: Some empirical evidence", *Journal of Banking & Finance*, Vol. 26, No. 5, pp. 825 – 855.

Jonung, Lars, Jaakko Kiander, and Pentti Vartia (2009) *The Great Financial Crisis in Finland and Sweden: The Nordic Experience of Financial Liberalization*: Edward Elgar Publishers Cheltenham, UK.

Jordà, Òscar, Moritz Schularick, and Alan M. Taylor (2013) "When Credit Bites Back", *Journal of Money, Credit and Banking*, Vol. 45, No. s2, pp. 3–28.

Jordà, Òscar, Moritz Schularick, and Alan M Taylor (2015a) "Betting the house", *Journal of International Economics*, Vol. 96, pp. 2–18.

Jordà, Òscar, Moritz Schularick, and Alan M. Taylor (2015b) "Leveraged bubbles", *Journal of Monetary Economics*, Vol. 96, pp. 1–20.

Jordà, Òscar, Moritz Schularick, and Alan M Taylor (2017) "Macrofinancial history and the new business cycle facts", *NBER Macroeconomics Annual*, Vol. 31, No. 1, pp. 213–263.

Joseph, Andreas (2019) "Shapley regressions: A universal framework for statistical inference on machine learning models", Staff Working Paper Series 784, Bank of England.

Joy, Mark, Marek Rusnák, Kateřina Šmídková, and Bořek Vašíček (2017) "Banking and currency crises: Differential diagnostics for developed countries", *International Journal of Finance & Economics*, Vol. 22, No. 1, pp. 44–67.

Kaminsky, Graciela and Carmen Reinhart (1999) "The twin crises: The causes of banking and balance-of-payments problems", *American Economic Review*, Vol. 89, No. 3, pp. 473–500, June.

King, Mervyn (2010) "Speech at the University of Exeter (no title)", Bank of England.

Kiyotaki, Nobuhiro and John Moore (1997) "Credit cycles", *Journal of Political Economy*, Vol. 105, No. 2, pp. 211–248.

Korinek, Anton and Martin Novak (2017) "Risk-taking dynamics and financial stability", Working paper.

Kuhn, Max, Steve Weston, and Nathan Coulter. C code for C5.0 by R. Quinlan (2014) *C50: C5.0 Decision Trees and Rule-Based Models*. R package version 0.1.0-21.

Laeven, Mr Luc and Fabian Valencia (2008) *Systemic banking crises: a new database*, No. 8-224: International Monetary Fund.

Laeven, Luc and Fabian Valencia (2018) "Systemic banking crises revisited", IMF Working Papers 18/206, International Monetary Fund.

Liu, Weiling and Emanuel Moench (2016) "What predicts US recessions?", *International Journal of Forecasting*, Vol. 32, No. 4, pp. 1138–1150.

Lundberg, Scott M (2018) "Shap (shapleyy additive explanations)", https://github.com/slundberg/shap/.

Lundberg, Scott M. and Su-In Lee (2017) "A unified approach to interpreting model predictions", in *Advances in Neural Information Processing Systems*, pp. 4765–4774.

Lundberg, Scott M, Gabriel G Erion, and Su-In Lee (2018) "Consistent individualized feature attribution for tree ensembles", *arXiv preprint arXiv:1802.03888*.

Manasse, Paolo and Nouriel Roubini (2009) ""Rules of thumb" for sovereign debt crises", *Journal of International Economics*, Vol. 78, No. 2, pp. 192–205, July.

Martin-Acena, Pablo (2014) "The savings banks crises in Spain: When and how?", Documentos de Trabajo (DT-AEHE) 1404, Asociación Española de Historia Económica.

Martinez-Miera, David and Rafael Repullo (2017) “Search for yield”, *Econometrica*, Vol. 85, No. 2, pp. 351–378.

Menard, Scott (2004) “Six approaches to calculating standardized logistic regression coefficients”, *The American Statistician*, Vol. 58, No. 3, pp. 218–223.

Minsky, Hyman P. (1977) “The Financial Instability Hypothesis: An Interpretation of Keynes and an Alternative to“Standard” Theory”, *Challenge*, Vol. 20, No. 1, pp. 20–27.

Mitchener, Kris James and Marc D Weidenmier (2008) “The baring crisis and the great Latin American meltdown of the 1890s”, *The Journal of Economic History*, Vol. 68, No. 2, pp. 462–500.

Ng, Andrew Y (2004) “Feature selection, L1 vs. L2 regularization, and rotational invariance”, in *Proceedings of the twenty-first international conference on Machine learning*.

Ng, Serena (2014) “Boosting recessions”, *Canadian Journal of Economics*, Vol. 47, No. 1.

Obstfeld, Maurice, Jay C Shambaugh, and Alan M Taylor (2005) “The trilemma in history: tradeoffs among exchange rates, monetary policies, and capital mobility”, *Review of Economics and Statistics*, Vol. 87, No. 3, pp. 423–438.

Ollivaud, Patrice and David Turner (2015) “The effect of the global financial crisis on OECD potential output”, *OECD Journal: Economic Studies*, Vol. 2014, No. 1, pp. 41–60.

Pagan, Adrian (1984) “Econometric issues in the analysis of regressions with generated regressors”, *International Economic Review*, Vol. 25, No. 1, pp. 221–47.

Perlich, Claudia, Foster Provost, and Jeffrey S Simonoff (2003) “Tree induction vs. logistic regression: A learning-curve analysis”, *Journal of Machine Learning Research*, Vol. 4, No. Jun, pp. 211–255.

Quinlan, J Ross (1993) *C4. 5: programs for machine learning*, San Mateo, California: Morgan Kaufmann Publishers.

Reid, Margaret (1982) *The secondary banking crisis, 1973–75: its causes and course*, London, UK: Springer.

Reinhart, Carmen M and Kenneth S Rogoff (2008) “Is the 2007 US sub-prime financial crisis so different? An international historical comparison”, *American Economic Review*, Vol. 98, No. 2, pp. 339–44.

Reinhart, Carmen M. and Kenneth S. Rogoff (2009) *This Time Is Different: Eight Centuries of Financial Folly*, Princeton, New Jersey: Princeton University Press.

Rey, Hélène (2015) “Dilemma not trilemma: The global financial cycle and monetary policy independence”, Working Paper 21162, National Bureau of Economic Research.

Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin (2016) “Why should I trust you?: Explaining the predictions of any classifier”, in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1135–1144.

Rokach, Lior and Oded Maimon (2005) "Top-down induction of decision trees classifiers-a survey", *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, Vol. 35, No. 4, pp. 476–487.

Rudebusch, Glenn D and John C Williams (2009) "Forecasting recessions: the puzzle of the enduring power of the yield curve", *Journal of Business & Economic Statistics*, Vol. 27, No. 4, pp. 492–503.

Savona, Roberto and Marika Vezzoli (2015) "Fitting and Forecasting Sovereign Defaults using Multiple Risk Signals", *Oxford Bulletin of Economics and Statistics*, Vol. 77, No. 1, pp. 66–92.

Schroff, Florian, Dmitry Kalenichenko, and James Philbin (2015) "Facenet: A unified embedding for face recognition and clustering", in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 815–823.

Schularick, Moritz and Alan M Taylor (2012) "Credit booms gone bust: Monetary policy, leverage cycles, and financial crises, 1870-2008", *American Economic Review*, Vol. 102, No. 2, pp. 1029–1061.

Shapley, Lloyd S (1953) "A value for n-person games", *Contributions to the Theory of Games*, Vol. 2, No. 28, pp. 307–317.

Shiller, Robert (2017) "Narrative Economics", *American Economic Review*, Vol. 107, No. 4, pp. 967–1004.

Shrikumar, Avanti, Peyton Greenside, and Anshul Kundaje (2017) "Learning important features through propagating activation differences", *arXiv preprint arXiv:1704.02685*.

Stone, Charles (1977) "Consistent nonparametric regression", *The Annals of Statistics*, Vol. 5, No. 4, pp. 595–620, 07.

Strumbelj, Erik and Igor Kononenko (2010) "An efficient explanation of individual classifications using game theory", *Journal of Machine Learning Research*, Vol. 11, pp. 1–18.

Taylor, John B (2009) "The financial crisis and the policy responses: An empirical analysis of what went wrong", NBER Working Papers 14631, National Bureau of Economic Research.

Tölö, Eero et al. (2019) "Predicting systemic financial crises with recurrent neural networks", research discussion papers, Bank of Finland.

Vermeulen, Robert, Marco Hoeberichts, Bořek Vašíček, Diana Žigraiová, Kateřina Šmídková, and Jakob de Haan (2015) "Financial stress indices and financial crises", *Open Economies Review*, Vol. 26, No. 3, pp. 383–406.

Vert, Jean-Philippe, Koji Tsuda, and Bernhard Schölkopf (2004) "A primer on kernel methods", *Kernel Methods in Computational Biology*, Vol. 47, pp. 35–70.

Vuong, Quang H (1989) "Likelihood ratio tests for model selection and non-nested hypotheses", *Econometrica*, pp. 307–333.

Wade, Robert (1998) "The Asian debt-and-development crisis of 1997-?: Causes and consequences", *World Development*, Vol. 26, No. 8, pp. 1535–1553.

Ward, Felix (2017) "Spotting the danger zone: Forecasting financial crises with classification tree ensembles and many predictors", *Journal of Applied Econometrics*, Vol. 32, No. 2, pp. 359–378.

Wolpert, David H, William G Macready et al. (1997) "No free lunch theorems for optimization", *IEEE transactions on evolutionary computation*, Vol. 1, No. 1, pp. 67–82.

Wright, Jonathan H (2006) "The yield curve and predicting recessions", *Federal Reserve Board: Finance and Economics Discussion Series.*

Young, Peyton (1985) "Monotonic solutions of cooperative games", *International Journal of Game Theory*, Vol. 14, pp. 65–72.

Zięba, Maciej, Sebastian K Tomczak, and Jakub M Tomczak (2016) "Ensemble boosted trees with synthetic features generation in application to bankruptcy prediction", *Expert Systems with Applications*, Vol. 58, pp. 93–101.

Zou, Hui and Trevor Hastie (2005) "Regularization and variable selection via the elastic net", *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, Vol. 67, No. 2, pp. 301–320.