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: June 18, 2019

Abstract

We develop accurate early warning models for financial crises using machine learning techniques and identify the key drivers of financial crises using a macroeconomic and financial time series data set for 17 countries between 1870–2016. First, most machine learning models outperform the logistic regression benchmark in out-of-sample prediction and forecasting experiments. Second, we investigate our models using the Shapley value framework. Our models suggest that the most important predictors are the slope of the yield curve and credit growth, across both the global and domestic dimension. These results hold for the logistic regression, as well as for the more complicated and nonlinear machine learning models, while the latter provide richer information about the relation between predictors and the likelihood of a financial crisis.

Keywords: machine learning; financial crises; early Warning system; credit growth; yield curve; random forest; Shapley values.

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 Michael Bordo, Hélène Rey, Ambrogio Cesa-Bianchi, Julia Giese, as well as participants of the workshop “Forecasting Political and Economic Crises: Economics meets Machine Learning” at the 2019 GSE Summer Forum, the ESCoE Conference on Economic Measurement 2019, the CORMISIS Centre for Operational Research, Management Science and Information System at the University of Southampton, workshop on economic and monetary policy at Danmarks Nationlbank, the conference on “Modelling with Big Data and Machine Learning” (2018) at the Bank of England and the workshop “Big Data and Forecasting Economic developments” (2019) at the European Commission for useful comments and discussions.

1 Introduction

Given the huge economic and social costs of financial crisis (Laeven and Valencia, 2018), understanding the early warning signs is of great importance for economic policy makers. However, identifying a reliable set of early warning predictors is challenging because of several reasons. First, there are only limited observations of past crises, which makes robust modelling challenging. Second, crisis indicators often only flash red when it is already too late to intervene. Third, it is challenging to distill complicated early warning models into simple and transparent pointers that can allow policy makers to respond quickly.

In this paper, we use machine learning to address these issues. We find that credit to GDP growth, as well as the slope of the yield curve, both domestically and globally, are key predictors for financial crises across a diverse set of prediction models. While the former is a standard finding in the literature, the role of the yield curve has been far less explored within a crisis prediction context. We observe significant nonlinearities and interactions. For example, a (globally) flat or inverted yield curve paired with high credit growth and falling or flat consumption substantially increases the probability of a crisis two years ahead. The importance of the predictors changes across time and countries. We find that the global financial crisis in 2008 was the first major crisis mostly driven by global credit growth, while the string of financial crises in the early 1990s was mostly due to domestic credit growth. On the other hand, the global yield curve we introduce provides a fairly robust crisis signal across time.

To the best of our knowledge, this 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; Lundberg and Lee, 2017; Joseph, 2019).

We employ the Macrohistory Database by Jordà et al. (2017) which covers macroeconomic and financial variables from 17 advanced economies over 140 years and contains a binary crisis variable. We first estimate a benchmark logistic regression model. As logistic regressions assume linearity in the predictors, we apply machine learning models (e.g. decision trees, random forests, extremely randomised trees, support vector machines

(SVM), and artificial neural networks) to allow for more flexibility such as nonlinear relationships and interactions terms of *a priori* unknown form. As financial crises are rare events and likely to exhibit nonlinear dependencies during their build-up, non-linear methods are particularly suited to construct a reliable early warning system. Almost all machine learning models outperform the basic regression model in extensive out-of-sample prediction tests.

Our work is closely related to the literature on early warning systems (Kaminsky et al., 1998; Bussiere and Fratzscher, 2006; Frankel and Saravelos, 2012; Babecký et al., 2014) and more recent work on applying machine learning techniques to financial crisis prediction (Ward, 2017; Joy et al., 2017; Alessi and Detken, 2018; Beutel et al., 2018).

For example, Alessi and Detken (2018) use random forests to construct an early warning system to predict banking crises. Similar to our results, they find that (a) credit growth is an important indicator for crises, and (b) accounting for global factors helps to improve the predictive performance of early warning systems. Beutel et al. (2018) use a similar dataset but obtain different results. They show that logistic regression consistently outperforms all machine learning competitors, including random forest, and SVMs.[1] Our paper differs from their contributions in a number of ways: (1) Neither paper provides a rigorous analysis of the models' nonlinearities nor measures the contributions of individual variables and interactions to the predictions. (2) Methodologically, Alessi and Detken (2018) use a cross-validation approach to compare the out-of-sample performance of the models, while Beutel et al. (2018) focus on recursive forecasting. We compare the models using both types of methodologies. (3) In terms of data, both these studies use quarterly data starting in 1970 for the European Union[2], while we use annual data that covers a longer time span and contains more crises our models can learn from.

in line with out findings, domestic private credit growth or credit to GDP growth have been found to be key predictors in many studies that use conventional logistic regression modelling (Borio and Lowe, 2002; Drehmann et al., 2011; Schularick and Taylor, 2012; Aikman et al., 2013). Also, confirming our results, global credit growth has been found

[1] Other examples of tree applications in economics are Manasse and Roubini (2009) and Savona and Vezzoli (2015) for sovereign crises or Duttagupta and Cashin (2011) for banking crises in emerging and developing countries. (Ward, 2017) uses random forests on similar datasets as we do but without investigating the drivers of that model nor other possible models.

[2] Beutel et al. (2018) added USA and Japan to the dataset.

to be important (Alessi and Detken, 2011; Duca and Peltonen, 2013) and in some cases even more so than domestic credit (Cesa-Bianchi et al., 2018). Unlike these previous studies, we also test and find that the domestic and global slope of the yield curve is a crucial and robust predictor. The flatter/more inverted the yield curve is the higher the chances of a financial crises in the upcoming two years. While the yield curve is a well-established indicator for economic recessions (Estrella and Hardouvelis, 1991), it is fairly new within the financial crises context. It could indicate the search for yield and increased risk-taking that can be observed in the run-up to a financial crises.

The paper is structured as follows. Section 2 describes the dataset and our selection of variables and transformations. Section 3 outlines the methodology and provides a brief description of the machine learning models and the Shapley value framework. Section 4 presents the main results comparing the out-of-sample performance of all models, investigating the importance of the variables across time and space and providing an analysis of nonlinearities and interactions. Section 5 conducts a recursive forecasting experiment to predict crises. This simulates how the models would have performed at different points in time looking forward. Section 6 concludes.

2 Dataset and Variable Selection

Financial crises are rare events. While there are a handful of truly global financial crises (e.g. 1890, 1907, 1921, 1930/31, and the 2007/08 global financial crisis), the majority of crises mostly occur in a single country or a small cluster of countries. Given that we observe financial crises so infrequently, we use the longest cross-country dataset available.

2.1 The dataset

The Jordà-Schularick-Taylor Macrohistory Database (Jordà et al., 2017)[3] contains annual macroeconomic and financial measures of 17 developed countries[4] between 1870 and 2016. For each of the 2499 observations, the dataset contains a binary variable indi-

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

[4] The countries covered are the United States, Canada, Australia, Belgium, Denmark, Finland, France, Germany, Italy, Japan, the Netherlands, Norway, Portugal, Spain, Sweden, Switzerland, and the United Kingdom.

cating whether a country in a particular year suffered from a financial crisis (n=90) or not (n=2409). The authors define a financial crisis 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 is an integration across several previous databases (Bordo et al., 2001; Laeven and Valencia, 2008; Reinhart and Rogoff, 2009; Cecchetti et al., 2009) and has been confirmed by experts for the respective countries (Schularick and Taylor, 2012).

In order to capture the emergence of a crisis 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 five years are deleted from the data to avoid the post-crisis bias (Bussiere and Fratzscher, 2006).[5] Additionally, the two world wars (1914–1918, 1939–1945) and the years 1933–1938 after the Great Depression are excluded from the data. To ensure full coverage, we also delete all observations with any missing values on the indicators, which especially applies to many observations in the 19th century. Appendix C.2 provides more details on the data sample.

2.2 Variable Selection and Transformation

We use a set of 12 predictors, which can roughly be divided into four categories: financial, fiscal, macro, and external factors. Financial channels are likely to play an important role in capturing the build-up of a financial crisis. We try to capture specific financial channels by including the following variables: total loans to the non-financial private sector (credit), broad money, stock prices, debt service ratio (i.e. debt × long-term interest rate over GDP), and the slope of the yield curve (i.e. long-term interest rate − short-term interest rate).

Credit growth or credit to GDP 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). It can endanger financial stability by amplifying small shocks

[5] If we do not delete the years after a crisis emerged, episodes where the economy is sustainable and healthy are in the same class as post-crisis episodes, where the economy is still affected by the crisis and in a recovering process before it returns to a sustainable level.

via the financial accelerator effect (Bernanke and Blinder, 1992) or by causing financial fragility induced by collateral constraints (Bernanke et al., 1999). Also, credit growth can become an issue in itself by generating financial instability through endogenous credit cycles, as stated by Minsky (1977) or Kindleberger (1978) who coined the term that financial crises are “credit booms gone wrong”.

Additional to domestic credit growth, global credit growth across countries has been identified as an important predictor. It accounts for the fact that financial crises often occur on an international scale and might be the result of cross-country spillovers rather than domestic imbalances. Cesa-Bianchi et al. (2018) investigate the role of global credit growth for the emergence of financial crises. The authors find an increasing correlation of credit growth between 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 for periods of asset price booms and financial distress. Similar to Cesa-Bianchi et al. (2018), we define the global credit for a country-year pair $\langle c, y \rangle$ as the mean credit to GDP growth in all countries except c in year y.[6]

Related to credit growth, money growth can cause an increase in liquidity and lower market rates which could lead to an increase in spending (Friedman, 1970). One might hypothesise that, analogous to the credit view above, a certain event makes agents worry about the value of their money which may result in banks runs, bank insolvencies and further down the line, a financial crisis. Schularick and Taylor (2012) found that money growth is good proxy for credit before the second world war but is not a good predictor for crises after the war. Baker et al. (2018) does not see theoretical nor empirical reasons for strong money growth to be an indicator for financial crises. Rather the authors stress that recessions are often preceded by weak money growth.

The slope of the yield curve, i.e. the term spread, is often seen as a strong predictor of an impending economic recession (Estrella and Hardouvelis, 1991), especially on a longer horizon of 12–18 months (Liu and Moench, 2016), and outperforms professional forecasters from finance, banking, and other fields (Rudebusch and Williams, 2009; Croushore

[6] Using a global variable might bias the results in prediction. Appendix C.2 discusses this potential problem in detail and gives evidence that the bias is, if existent, limited in our experiments.

and Marsten, 2016). In normal times the slope is positive, which means that long-term rates are higher than short-term rates. Investors incorporate expectations of the future path of short term rates, as well as a risk premium (i.e. the term premium) for holding an asset for a longer duration. An inverted yield curve, for which the short rate is larger than the long term rate, would therefore be a signal of worsening expectations about the future macroeconomic path of the economy (Zaloom, 2009). The slope differential is relevant for the balance sheet of banks and other financial intermediaries (Adrian et al., 2010; Borio et al., 2017). A flattening of the yield curve reduces net interest margins and thus the profitability in the financial sector which can lead to a contraction in credit supply and thus affect real economic activity. If this effect is severe enough, the term spread might also prove a useful predictor for financial crises. Coleman et al. (2008) find that the house prices in the United States rose with the flattening of the yield curve in the build-up of the global crisis of 2008 and suggests that "the hunger for spread during this period of a flat yield curve could have been fuelling sub-prime and other alternative mortgage activity", thus highlighting the role of search for yield behaviour in financial markets. While a few empirical studies on early warning models for financial crises have identified the slope of the yield curve as a crucial predictor (Babeckỳ et al., 2014; Joy et al., 2017; Vermeulen et al., 2015), they have not discussed the rational behind its predictive power in detail. Additional to the domestic slope, we also used a global slope indicator. Several studies have shown strong dependencies of interest rates (Frankel et al., 2004; Obstfeld et al., 2005) between countries and have found a systematic global factor of the the yield curve (Diebold et al., 2008; Abbritti et al., 2013). Analogous to global credit, we compute the global slope for a country-year pair $\langle c, y \rangle$ as the mean slope in all countries except c in year y.

Another popular indicator are stock and housing prices, or asset prices in general (Aliber and Kindleberger, 2015; Reinhart and Rogoff, 2008). Rapid rises of asset prices could indicate the formation of a bubble. The eventual burst of that bubble would then affect the economy via a reduction in investment, and a reduction in the balance sheet and liquidity of households.

The debt service ratio has also been identified as a good early warning indicator (Drehmann and Juselius, 2012). It measures interest payment relative to income which can serve as a useful approximation 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, and a lack of new investments. The downside of our simplistic debt service ratio measure (debt × long-term interest rate over GDP) is that it does not capture lending rates, which are important in the run-up to crisis (Drehmann and Juselius, 2012), capital repayments, and maturity structure of the debt.

To account for crises which could be caused by fiscal problems we include public debt. We also control for general macroeconomic conditions, which could trigger a financial crisis by including real consumption per capita, investment, and the consumer price index (CPI).

Finally, we investigate the role of external factors via the current account. Current account imbalances have often been found to be a strong cause of crises, especially for developing countries due to capital flows pushing down the interest rates and thus encouraging risk-taking behaviour in less well regulated financial systems (Bernanke, 2009; King, 2010).

We treat the prediction problem of identifying crises as a classification problem and model the ⟨country, year⟩ pairs as independent observations. As the observations are naturally not independent from each other, we require variable transformation to reduce non-stationarity and ensure comparability of the different measures across countries. The slope of the yield curve is left in levels, while CPI, stock prices, and real consumption per capita are transformed into percentage 2-year growth rates. All other variables, i.e. credit, money, public debt, debt servicing, investment, and current account, are scaled by GDP and expressed in 2-year changes.[7]

After excluding observations with missing values, 1249 observations remain of which 95 have a positive value on our crisis outcome. These observations constitute our baseline dataset.

A very basic test of how useful individual variables are for an early warning system

[7] Section 4.2.2 shows some of the experiments we conducted in order to select the variables and transformations that we use in the main analysis.

	Crisis build-up		Non-crises		
	Mean	SD	Mean	SD	Difference
Global slope	0.18	0.80	0.90	0.84	-0.72***
Global credit	4.18	3.21	2.20	2.33	1.98 ***
Domestic slope	-0.34	1.59	0.85	1.75	-1.19 ***
Nominal short rate	6.14	3.51	5.19	3.69	0.95 **
Nominal long rate	5.80	3.15	6.05	3.43	-0.25
Domestic credit	6.59	7.92	2.07	5.46	4.52 ***
CPI	4.60	8.12	7.76	9.89	-3.16 ***
Debt service ratio	0.55	1.07	-0.02	1.08	0.57 ***
Consumption	3.27	5.23	4.74	4.77	-1.47 ***
Investment	1.06	3.13	0.21	2.30	0.86 **
Public debt	-0.37	8.84	-0.36	8.04	-0.02
Broad money	2.67	4.76	1.05	4.62	1.62 ***
Stock market	18.16	28.38	19.46	41.25	-1.30
Current account	-0.59	2.875	0.06	2.70	-0.64 **
Household loans†	3.69	4.96	1.62	3.29	2.07 ***
Business loans†	4.45	5.62	0.45	4.05	4.00 ***
Mortgage loans†	3.90	5.48	1.52	3.38	2.38 ***
House prices††	14.95	19.15	13.77	18.58	1.19

Table I: Descriptive statistics of the variables for observations one and two years before a crises and non-crises observation. The unit of measurement is percentage points for the interest rates, the slope and the growth in the indices CPI, stock market, investment, and house prices. For the remaining variables, it is the ratio change × 100. 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. The statistics of the variables flagged with † are based on a subset of 901 observations (52 with positive crisis outcome) and those flagged with †† on 1081 observations (83 with positive crisis outcome) because of additional missing values. The statistics of the remaining variables are based on our baseline dataset with 1249 observations.

is to compare the values of the variables shortly before the crises and during normal economic conditions. Table I show the mean values of the variables on observations with positive (one and two years before the crisis) and negative values on our crisis indicator. A t-test confirms that there are significant differences on nearly all of the variables. The differences are often pronounced, such as more than one percentage point difference in domestic slope.[8]

[8] The unit of measurement is percentage points for the interest rates, the slope, and the growth in the indices CPI, stock market, investment, and house prices. For the remaining variables, it is the ratio change × 100. The statistics of the variables flagged with † are based on a subset of 901 observations (52 with positive crisis outcome) and those flagged with †† on 1081 observations (83 with positive crisis outcome) because of additional missing values. The statistics of the remaining variables are based on

3 Methodology

Logistic regression is arguably the most popular prediction method for a classification problem like ours. It is easy to interpret, provides a statistical test of its parameters and has low computational costs. Accordingly, we estimate a logistic regression model on our complete data set as a baseline model. This approach, however, has substantial shortcomings for our prediction problem. Logistic regression does not automatically account for nonlinearities and interactions, which we both expect to be relevant during the build-up of a crisis. For example, when predicting financial crises, Cesa-Bianchi et al. (2018) find a significant nonlinear association between global credit and financial crises and Alessi and Detken (2018) observed a significant interaction between domestic and global credit growth. To account for nonlinearities and interactions in logistic regression, the modeller explicitly needs to add polynomial or interactions term 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.

Also, fitting the model to the data does not tell us how well it fares in prediction. The more parameters the model has, the more likely it will overfit to the data and predicts less well on unseen instances.

We address these shortcomings by investigating a set of machine learning models that automatically learn nonlinearities and interactions from the data without the need to specify them explicitly. To evaluate their predictive performance, we conduct extensive out-of-sample tests. First, we provide the relevant notation and terminology for our classification problem and give a brief description of each model before we explain the experimental out-of-sample procedure in detail. We then describe the Shapley framework which enables us to address the black box critique of machine learning models and actually decompose our results into the contributions of individual predictors. Using Shapley regression (Joseph, 2019), we are also able to determine whether a predictor makes a statistically significant contribution to the accuracy of the model.

our baseline dataset with 1249 observations.

3.1 Machine Learning Models

Let f be a prediction model that predicts the probability of a financial crisis $\hat{y} = f(\mathbf{X})$, where $\mathbf{X}_{n \times k}$ is the predictor matrix containing n observations and k variables. The observed class label is denoted by $y \in \{0, 1\}^n$, where 1 indicates a crisis and is referred to as the positive class and 0 indicates no crisis and is referred to as the negative class. The predicted value of a model is the predicted probability of a crisis and is denoted by $\hat{y} \in [0, 1]^n$. In our out-sample experiments, the models are learned on a subset of the data, the *training set* and are evaluated on a different set of observations not overlapping with the training set, the *test set*.

We compare a diverse set of classification algorithm ranging from simple interpretable models such as decision trees and logistic regression to black box models such as random forest and neural networks. In the following we provide a high level explanation of the algorithms. A description of the implementation details is found in Appendix B.

Decision trees

A decision tree is an interpretable model that successively splits the data into subsets by testing a single predictor in 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, 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 the decision trees (Rokach and Maimon, 2005). Nevertheless, decision trees often have limited predictive power compared to more complex methods such as random forest, especially when the data set is small.

We test the C5.0 algorithm (Quinlan, 1993; Kuhn et al., 2014) which uses Shannon entropy as an impurity measure and a statistical heuristic to control the complexity of the tree.

Random forests

A random forest (Breiman, 2001) is a very popular general-purpose classification algorithm. In a large-scale empirical comparison of 179 classification algorithms tested across a diverse set of 121 real world data sets, it was the best performing algorithm, on average (Fernández-Delgado et al., 2014). A random forest is a collection of many, often hundreds of decision trees. By averaging the predictions of the trees, random forests usually experience less overfitting than an individual tree. Each tree overfits differently and averaging their predictions cancels out these noisy components and increases the ability to generalise on unseen data. This only works if the trees are sufficiently different from 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, drawn with the replacement from the training set.[9] Second, the random forest 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 is a probability estimate for how likely one objects belongs to one class or the other.

Random forests often perform substantially better than individual decision trees, but this comes at the costs of interpretability. Aggregating the predictions of hundreds of trees, a random forest is not decomposable into a simple set of rules as it is the case for individual trees.

Extremely randomised trees

Extremely randomised trees (Geurts et al., 2006) are similar to random forest but tend to produce a smoother classification function by introducing two major changes to the random forest algorithm. First, each trees 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 made completely at random across the range of the values of the indicator. Of these

[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).

random splits, the best one is used in the tree. In the following we refer to this method as *extreme trees*.

Support vector machines

A support vector machine (SVM, Cortes and Vapnik (1995)) uses a hyperplane to separates the positive from the negative class. To avoid overfitting, it chooses the hyperplane that maximises the distance to the nearest data points. Similar to random forests, SVMs are very popular general purpose classification algorithms. In the study by Fernández-Delgado et al. (2014), SVMs were on average the second best algorithm across 121 data sets. The reason for their popularity and predictive strength is their capability to model nonlinear classification problems in an efficient way by using a kernel: The data enters the optimisation as an inner product of all pairs of observations and is implicitly transformed into a higher dimensional nonlinear space. A popular kernel, which we also use in the following analyses, is the radial basis function that transforms the data into infinite dimensions. Using a nonlinear kernel, the model becomes a black box as each prediction is not easily attributable to individual predictors as it is for a regression or simple decision tree.

To obtain more stable predictions, we do not train a single SVM but average the predictions of 25 models that are trained onttt. (see Appendix B).

Artificial neural networks

Artificial neural networks have been the most researched machine learning technique in recent years and achieved landmark successes in classification problems such as face (Schroff et al., 2015) and speech recognition (Amodei et al., 2016).

A neural network consists of an input layer that represents the values of the predictors, at least one hidden layer[10] and an output layer. The inputs are passed from layer to layer as activations and are finally integrated as a prediction in the output layer.

The nodes in a hidden layer are connected to the previous and subsequent layer by weights. For instance, given a data set with N predictors and a network with a single hidden layer containing M nodes, $N \times M$ weights are needed to fully wire the input layer

[10] Without a hidden layer, a neural network is a linear function of 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 networks has hyperparameters that control the structure of the model such as the number of hidden layers and nodes or the activation function that transforms the activations in a node before it is passed to the next layer. The high number of parameters and hyperparameters and the networks' sensitivity to these makes learning a predictive and network with an appropriate architecture challenging, especially when the available data is small. To obtain more stable predictions, we do not train a single SVM but average the predictions of 25 models that are trained onttt. (see Appendix B).

3.2 Experimental procedure

As opposed to logistic regression, nonlinear machine learning models such as neural networks and random forest are flexible enough to obtain perfect in-sample accuracy on the data on which they were trained. This does not mean they successfully learn the data generating process, rather they perfectly fit to the unpredictable noise.

To meaningfully evaluate the performance of the prediction models, they need to be tested out-of-sample. In our main experiment, we use cross-validation to evaluate the predictive performance of the models.[11] All observations between 1870 and 2016 are randomly assigned to one of five different subsets that we refer to as *folds*. Each fold is once used as a test set in which the performance of the prediction models is evaluated, while the remaining folds are used for training the models.

Recall that each crisis observation in the raw data was 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 performance estimate.[12] To obtain stable results, the random assignment of folds is repeated 100 times.

Some of the machine learning methods require to learn hyperparameters (see Appendix B). 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

[11] We will also show results for a forecasting approach in Section 5.

[12] Appendix C.1 examines this bias and other ways of cross-validation in detail.

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. This procedure is computationally costly.[13]

3.3 Shapley framework

Machine learning models, like random forests, are non-parametric and error consistent (Stone, 1977; Joseph, 2019). That is, they approximate any sufficiently well-behaved function arbitrary close provided enough training data. Their high flexibility makes them black box models: It is not straightforward to map model inputs to specific functional dependencies of the model. In other words, the model does not tell us which variables are driving its predictions in which way.

We address this problem by using the *Shapley additive explanations* framework (Lundberg and Lee, 2017; Strumbelj and Kononenko, 2010). It uses the concept of Shapley values (Shapley, 1953; Young, 1985) from cooperative Game theory. Originally, Shapley values are used to calculate the payoff distribution across a group of players. Analogously, we use them to calculate the payoff distribution across predictors. More precisely, the predicted value of each individual observation is decomposed into a sum of contributions from each variable, namely its *Shapley values*. This enables us to make informative statements about which variable is driving each prediction so that we can contribute in the current debate on whether variables like credit growth, current account, and asset prices have large predictive value or not.[14]

[13] For example, a SVM has two hyperparameters C and γ, for which we try 10 different values each. This results in 100 combinations of parameter values and 500 models that are tested in the *inner* 5-fold cross-validation in the training set just to obtain a single performance estimate in one of the five folds of the *outer* cross-validation. A less rigorous approach to determine the hyperparameters only once by doing cross-validation on the complete data set. This is problematic because the test data has already leaked into process of training the models which leads to optimistically biased performance estimates (Cawley and Talbot, 2010).

[14] There do exist other approaches to decompose the predictions of models into the contributions of individual variables Ribeiro et al. (2016); Shrikumar et al. (2017), however only the Shapley framework

Corresponding to the predictor matrix $\mathbf{X}_{n\times k}$ described in Section 3.1, 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 the model 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 usually set to the mean predicted value in the training set.

How are the Shapley values calculated? For a linear model, the Shapley value of feature j is simply the product of the regression coefficient w_j, multiplied by the predictor value X_{ij} relative to the mean predictor value, $\mathbb{E}[x_j]$: $w_j(X_{ij} - \mathbb{E}[x_j])$. Computing the Shapley values for a general model is computationally more challenging.

An intuitive way to understand the computation of Shapley values is to regard the problem that motivated their invention. Shapley values were first introduced in game theory to determine how much a particular player contributes in a corporative game of a group of players. The individual contribution is not directly observable but only the payoff generated by the group as a whole is. To determine the contribution of player j, groups can be formed sequentially and j's contribution is measured by the marginal contribution entering a group. A player's contribution depends on the other players in a group. Imagine player j joins a group in which player k has similar skills. In this case j's contribution is smaller than if he 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. This requirement makes the computation of Shapley values computationally expensive. 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, the payoff f_i is the predicted probability estimated by the model for a particular observation i. 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, one has to compute how much feature j adds to

has both a set of appealing analytical properties Lundberg and Lee (2017) and is applicable to any model family.

the predictive value $(f_i(S \cup \{j\} - f_i(S))$ in all possible subsets of the other variables $(S \subseteq N \setminus j)$. Predictors not in S cannot be left out or set to missing values as this would not allow the machine learning models to produce predictions. Instead, these predictors are integrated out by replacing them with all observed values in the training set.[15].

As in the horse race experiment, we use 5-fold cross-validation, construct the models in the training set and compute the Shapley values for the objects in the test set. We repeated the cross-validation procedure 100 times to obtain stable estimates.

Shapley regression

Shapley values measure how much variables drive the predictions of the model, independent of the accuracy of the model. In other words, Shapley values do not show how well the variables actually predict the true data.

To judge the economic and statistical significance of the predictors we use *Shapley regressions* (Joseph, 2019). To the best of our knowledge, there does not exist another statistical framework that estimates the significance of predictors, or hypothesis testing more generally, for nonlinear models. The Shapley regression framework achieves this by regressing the crisis indicator y on the Shapley values $\mathbf{\Phi}$. In this way, the nonlinear and unobservable function of the predictors in a non-parametric black box model is transformed into a parametric space which makes the estimation of p-values a simple logistic regression exercise. The surrogate Shapley regression has the appealing property that if the model is a linear function of the predictors, the Shapley regression will reproduce the linear model.[16]

	(1)	(2)	(3)	(4) Baseline model
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

*p<0.1; **p<0.05; ***p<0.01

Table II: Logistic regression models fitted to all data points. The outcome variable is our crisis indicator, which is set positive one and two year before the actual crisis.

4 Results

4.1 The benchmark logistic regression model

Our first analysis shows the logistic regression model fitted to all observations in the data.[17] We include the variables we discussed in Section 2 highlighting in particular credit growth and the slope of the yield curve both on the domestic and global level. The first model in Table II shows that domestic credit growth is an important predictor for financial crises even after controlling for all covariates. This is in line with the literature, e.g. 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, which obtains a higher weight than domestic credit. This confirms the study by Cesa-Bianchi et al. (2018) who also find global credit to be more predictive of financial crises than domestic credit[18].

Next, we add the slope of the yield curve. Its weight is negative, indicating that a small or negative slope (long rate smaller than short rate) corresponds to a higher predicted probability of crisis. The slope has a higher standardised weight than credit but does not reduce the effect of the other variables substantially. Adding the global slope in Model 4, the weight of the domestic slope decreases but both remain important predictors[19]. Notably, the weight of CPI declines strongly after 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 ($p < 0.001$).

While credit growth is an established predictor for financial crises in the literature, the slope of the yield curve has only been tangentially discussed (Babeckỳ et al., 2014;

[15] We use the `shap` Python package (Lundberg, 2018) to estimate the Shapley values. The study by Lundberg and Lee (2017) provides a detailed explanation of how Shapley values are computed in the context of explaining predictions of machine learning models.

[16] We include all 100 individual Shapley estimates for each observation in the regression to account for their variability across replications. As observations are not independent, we estimate clustered standard errors (Rogers, 1993) such that each country-year pair is assigned to its own cluster.

[17] To better compare the predictive power of the individual variables, we standardise all variables in this and all following regression analyses such that they have a mean of 0 and a standard deviation of 1. This is equivalent to a standardisation of the regression coefficients suggested by Agresti (1996) and recommended over other approaches by Menard (2004).

[18] Both variables have a medium correlation of 0.25 and an analysis of multicollinearity across all variables does not indicate problematic levels.

[19] Collinearity between both variables does again not indicate problematic levels.

Table III: Results

	(5)	(6)	(7)	(8)
Nominal short-term rate	0.698*** (0.167)		1.641*** (0.272)	0.405* (0.220)
Nominal long term rate		0.044 (0.182)	−1.367*** (0.313)	
Domestic slope				−1.105*** (0.206)
Domestic slope × nominal short term rate				0.482*** (0.147)
+ Covariates	Domestic credit, global credit, CPI, debt service ratio, broad money, stock market, consumption, public debt, investment, current account,			
Observations	1,249	1,249	1,249	1,249
Log Likelihood	-267.668	-276.245	-257.305	-251.219
Akaike Inf. Crit.	559.336	576.489	540.610	530.438
Note:				*p<0.1; **p<0.05; ***p<0.01

Table IV: Logistic regression model of the nominal short- and long-term interest rates and the other variables, except domestic and global slope of the yield curve.

Joy et al., 2017). To better understand its role as an indicator for crises we investigate its components, i.e. the short and long-term nominal interest rates in Table IV. Model 5 and 6 respectively show how predictive nominal short and long-term rates are on their own, after controlling for all other covariates, except domestic and global slope. The short-term rate is a significant predictor, while the long rate is not. Model 7 uses both interest rates. Compared to using the short-term rate alone, the goodness of fit improved significantly ($p < 0.001$).Interestingly, the model implicitly learns a function of the interest rates that closely mimics the slope. Let l and s be the long and short-term rate, respectively, the model learns $1.641s - 1.367l = -1.367(l - 1.2s)$. This model is not significantly better ($p = 0.4383$) than the model which only uses the slope (Model 4). This analysis confirms the predictive power of the slope. We conjecture that a flat or negative slope is a stronger signal for a crisis if the short-term rate is slow. Model 8 test this hypothesis and establishes a statistical significant interaction of the two interest rates. This model has a significantly better fit than Model 4 ($p < 0.001$).

Figure I illustrates the effect of the interaction in Model 8. It shows the predicted probability of crisis as a function of the domestic slope (horizontal axis), when the nominal short-term rate is at the mean and one standard deviation above or below it. All other predictors are held constant at their mean value. Given a low slope, the predicted probability of crisis is substantially higher when the short-term interest rate is low (red line).

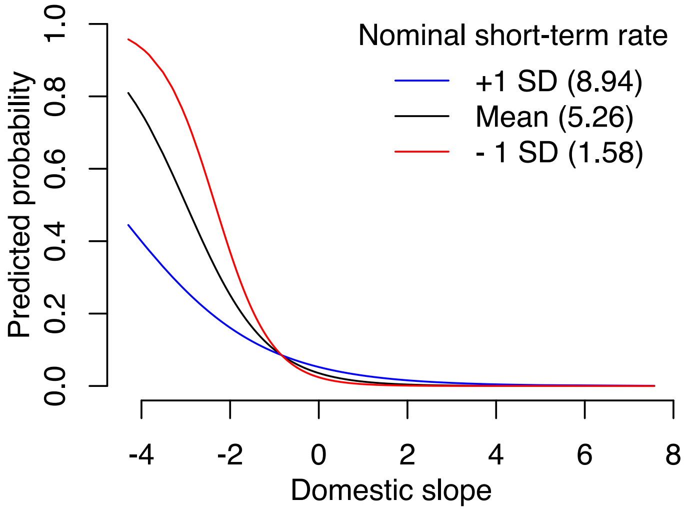

Figure I: Interaction effect in Model 8. The plot depicts the effect of the slope on the predicted probability of crisis at three different levels of the short-term rate (mean, ± 1 standard deviation). All remaining predictors are held constant at their mean value.

Using real rates, the significance of the interaction of the slope with the short-term rate disappears (see Appendix C.3). This is in line with the observation in Model 4 that the significance of CPI drops when controlling for the slope. Models 5–7 do not qualitatively change when using real instead of nominal interest rates.

The interaction of the nominal short rate and the slope did not significantly improve the performance of the logistic regression in the out-of-sample prediction experiments. We therefore focus on the slope alone in the following horse race, as in Model 4 which will be our benchmark.

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

We differentiate between three types of crises: (1) crisis that were preceded by a recession one or two years ahead, (2) crisis that co-occur with a recession but were not preceded by one, and (3) crises that do not co-occur or were preceded by a recession but may be followed by one.

	(9)	(10)	(11)
	Crises preceded by recession ($n = 24$)	Crises that co-occur with recession ($n = 38$)	Remaining crises ($n = 33$)
Domestic slope	-0.283	-0.482^{*}	-0.793^{***}
	(0.272)	(0.258)	(0.197)
Global slope	-0.681^{**}	-0.791^{***}	-0.476^{**}
	(0.312)	(0.251)	(0.221)
+ Covariates	Global credit, domestic credit, CPI, debt service ratio, broad money stock market, consumption, public debt, investment, current account		
Observations	1,178	1,192	1,187
Log Likelihood	-74.933	-109.632	-121.615
Akaike Inf. Crit.	175.866	245.263	269.230

*p<0.1; **p<0.05; ***p<0.01

Table V: Logistic regression predicting financial crises that are preceded by a recession (Model 9), co-occur with a recession (Model 10), or neither of the two (Model 11).

We re-estimate our baseline Model 4 for each of the three 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 V.

The domestic slope of the yield curve loses its predictive power when a crisis is preceded by a recession (Model 9). The is not surprising as a recession causes the yield curve to steepen on average. In line with the findings from Figure I, this leads to a lower predicted probability of crisis despite a crisis occurring.[20] On the contrary, the domestic yield curve is particularly informative in the *absence* of a recession (Model 11). Moreover, the significance of the global slope in all three cases highlights the importance of international market conditions when evaluation the likelihood of a crisis.

Our baseline logistic regression shown in Table II is a useful start for an early warning system. However, as discussed before, it does not account for nonlinearities. We address these shortcomings by investigating a set of machine learning models which automatically learn nonlinearities and interactions from the data without the need to specify an explicit functional form.[21] We evaluate their predictive performance and conduct extensive out-of-sample tests and variable inference analyses.

4.2 Machine learning model performance

All prediction models we investigate estimate the probability that a crisis occurs. Therefore, we can evaluate their performance in the *Receiver Operating Characteristic* (ROC) space. The ROC space depicts the the positive rate (i.e. proportion of crises correctly identified as such) on the vertical axis and the false positive rate (proportion of non-crises incorrectly identified as crises) on the horizontal axis. The perfect model would obtain a hit rate of 1 and a false alarm rate of 0. However, generally, a higher hit rate comes at the cost of a higher false alarm rate.

The trade-off between a high hit rate and a low false alarm rate can be controlled by setting different thresholds on the probabilities predicted by the model. The main advantage of the ROC analysis is that it does not force the modeller to specify the relative costs of the two types of classification errors (missing to predict a crisis when there is one and predicting a crisis when there is none), which is often, as in our case, a non-trivial endeavour and usually depends on the actual application of the model. Figure II shows how the models compare in out-of-sample prediction in the ROC space. Random

[20] The average steeping of the yield curve in the one and two years ahead of a crisis preceded by a recession is about 0.3%. This accounts for the majority of the differences in domestic slopes relative to crisis not associated with a recession.

[21] On a technical level, this is related to the consistency property of non-parametric models (Stone, 1977; Joseph, 2019).

forest and extreme trees are consistently more accurate than their competitors, as their curves are consistently above those of the other models.

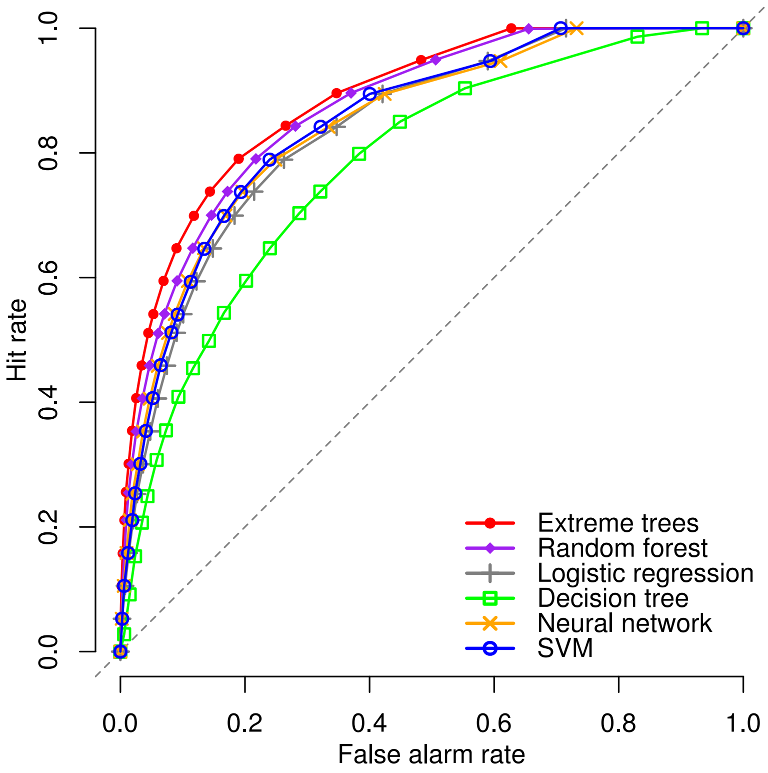

FIGURE II: ROC curves

The performance in the ROC space can be summarised by the *Area Under the Curve* (AUC). This also has an intuitive interpretation besides the geometrical definition. It is the probability that the model assigns a higher predicted value to a randomly drawn objects from the positive crisis class than to a randomly drawn object from the negative non-crisis class.

Table VI shows the mean AUC of the models in the cross-validation experiments. Standard errors are omitted but are consistently below 0.001. The best performing model in terms of AUC is, as implied by the ROC curves, extreme trees, followed by random forest, and support vector machines. 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).

Compared to a probabilistic model like a logistic regression, machine learning models

	AUC	True positives (correctly predicted crises)	False positives (false alarms)	True negatives (correctly predicted non-crises)	False negatives (missed crises)
Extreme trees	0.870	76	219	935	19
Random forest	0.855	76	286	868	19
SVM	0.832	76	320	834	19
Neural network	0.829	76	279	875	19
Logistic regression	0.822	76	367	787	19
Decision tree	0.759	76	244	910	19

Table VI: First column: Mean AUC in the cross-validation experiment across all iterations 100 iterations. The standard error is not shown but is consistently below 0.001. Remaining columns: Mean predictions across the 100 replications at a hit rate of 0.8 as shown in Figure III.

such as random forest, often do not accurately estimate the true posterior probabilities (Niculescu-Mizil and Caruana, 2005).[22] Therefore, the reader should not interpret the predicted values as actual probabilities of crises. Rather, the ranking of the predicted values is meaningful and is reflected in the ROC analysis and the AUC, which are insensitive to the actual predicted probabilities but only considers their order.

4.2.1 Best Model for Crisis Prediction: Extremely Randomised Trees

To convey how well our best model, extreme trees, makes predictions on the individual observations, 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 probability threshold at which the model identifies a crisis is a predicted probability of 9.6%, resulting in a false alarm rate of 19%. Extreme trees reduce the false alarm rate by 40% compared to the logistic benchmark which has a false alarm rate of 32%.[23]

Figure III depicts hits (correctly identified crises using green points), false negatives (missed crises with red triangles), and false positives (false alarms with grey triangles) as

[22] In the machine learning literature there exist methods such as Platt scaling (Platt et al., 1999) and isotonic regression (Zadrozny and Elkan, 2002) to calibrate the probability estimates.

[23] Note that the decision tree performs considerably better than logistic regression in this comparison with a false alarm rate of 21% despite performing worse in the ROC space. The reason for this is that averaging the tree predictions across the 100 replications (and therefore 100 different trees) makes the predictions more robust—similar to the predictions of the tree ensembles random forest and extreme trees.

well as the predicted probability of crisis (black line) for all observations in our sample. To improve legibility, the by far most prevalent outcome, true negatives (correctly identified non-crisis in light green), are only shown in the pie charts at the right that depicts the overall distribution of all four outcomes in each country.

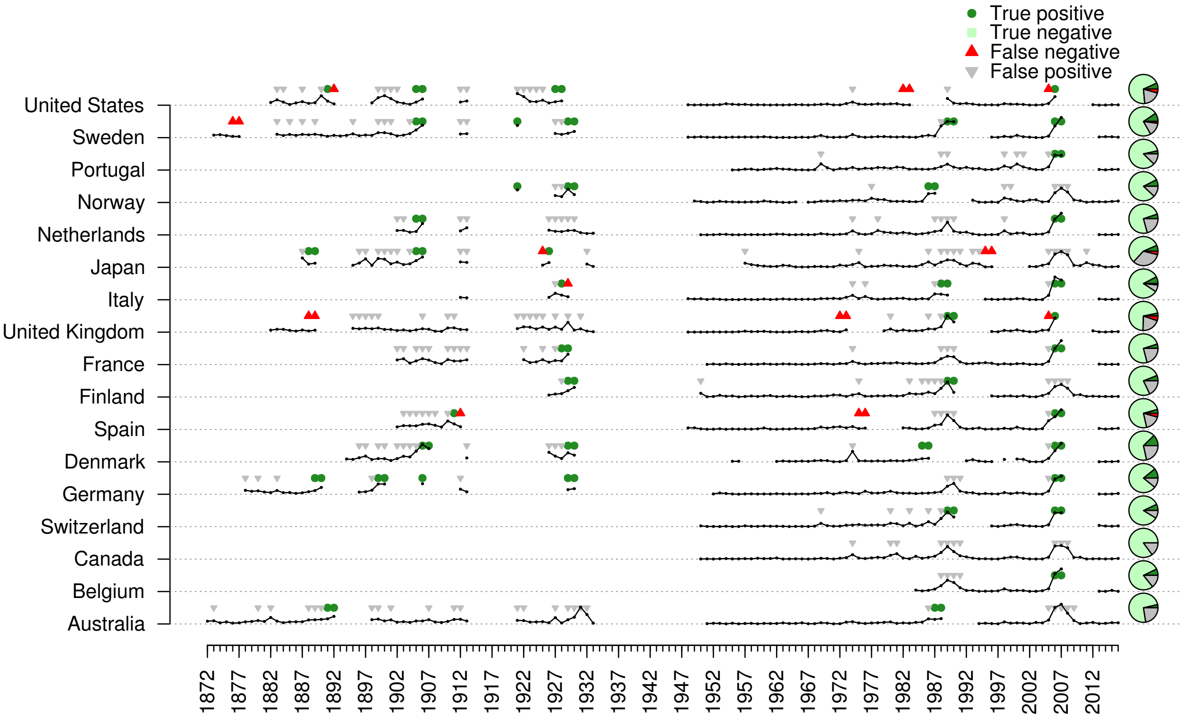

Figure III: Predicted probability of extremely randomised trees for all observations and the classification (crisis vs. non-crisis) according to a prediction threshold of 0.096.

The model only fully misses six out of 48 distinct crisis events: United States (1984), Sweden (1878), Japan (1997), United Kingdom (1890, 1972) and Spain (1977). For another six crisis, the model only missed either one or two years ahead of the actual crisis but not both of these observations. All of the missed crisis episodes can be related to aspects not or only partially captured in our models.

The crisis Spain in 1977 was caused by the global oil price shock and the government's interventions to damp its effects which delayed firm consolidation and increased public spendings (Betrán and Pons, 2013). The missed crisis in Japan is linked to the Asian Financial crisis of 1997, to which Japan was exposed to (Wade, 1998). After an enormous expansion of the Swedish railway industry in the early 1870s, Sweden's economy declined and caused substantial losses in the financial sector which had heavily invested in the railway industry. The crisis in the late 19th century in the UK was known as the Baring crisis or the Panic of 1890 which was due to poor investments in Argentina that led to bankruptcy of Baring bank (Mitchener and Weidenmier, 2008). The same poor investments in Argentina also caused the Financial Panic in the United States (correctly identified in our model) three years later. The Secondary Banking crisis in the UK (1972) was linked to house prices which are not included in our baseline model (see section on robustness checks below).

The high proportion of false alarms is misleading. Firstly, the model is by construction more risk averse and calibrated to ensure less crises are missed at the cost of a higher false alarm rate. In this case, the model predicted a crisis in 19% of the observations where actually no crisis has occurred. Secondly, the false predictions often cluster around periods which see a crisis a couple years later and therefore do provide a useful warning signal. Thirdly, the false positives cluster around periods when other countries have experienced financial crises which again can be seen as a valuable warning signal in line with the importance of global variables we document. Lastly, the model is not taking into account any policy actions that might have mitigated a crisis, so that the model might have correctly detected an impeding crisis that was successfully averted. In either case, even the false positives might provide useful information for policy makers by indicating when vulnerabilities have been building up.

We should also say that our model seems to do better for some countries, e.g. Germany and Switzerland, and worse with others, e.g. Japan, as indicated by the pie charts on the right. Including more region specific variables might help to improve performance for those specific countries.

The figure shows that the number of false alarms is substantially higher before than after World War 2 with respective false positive rates of 52% and 10%. The global economy underwent substantial changes across time and a general model covering more than 140 years may not predict well given time-specific idiosyncrasies. As we have fewer observations before WW2 than after, the earlier period gets less weight when training the model which means that it is geared to perform better on more recent observations.

Experiments	**Crises**	**Extreme trees**	**Random forest**	**Logistic regression**	**Decision tree**
Baseline	95	0.87	0.86	0.82	0.76
Testing transformations					
Growth rates only	95	0.86	0.85	0.81	0.76
Hamilton filter	89	0.86	0.84	0.82	0.76
*	87	0.89	0.85	0.81	0.76
Adding variables					
Nominal rates	95	0.87	0.85	0.82	0.76
Real rates	95	0.86	0.85	0.82	0.76
Loans by sector	52	0.82	0.83	0.83	0.76
*	52	0.83	0.83	0.83	0.79
House prices	83	0.88	0.87	0.82	0.75
*	83	0.87	0.85	0.82	0.76
Changing the horizon					
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	0.76
5 years	89	0.86	0.85	0.80	0.75
*	89	0.87	0.85	0.82	0.76

Table VII: Results of the robustness checks. Asterisks indicate the retrained baseline experiment on exactly the same observations as the respective robustness check.

4.2.2 Robustness checks

We empirically tested different sets of variables and transformations before we chose the experiment on which the main results are based. To show that the relative model ranking is robust to the data specification we use, we report these robustness checks in Table III. We did not test SVMs and neural networks because of the extensive computation time. The *baseline* is the main experiment using the variables and transformations as described in section 2 for which we have a total of 95 crises observations.

Adding new variables or changing the transformation may lead to a loss of observations due to missing values. To provide a fair comparison, we need to retrain the baseline model on exactly the same observations as the robustness check is trained on. The retrained baseline models are marked with an asterisk.[24] All experiments are repeated 100 times using 5-fold cross-validation.

Variable transformations. To ensure that our chosen variable transformation of using GDP ratios is indeed superior, we checked the results for both growth rates and filtered data to detrend the data. For the growth rate experiment, the slope is left in levels, 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, 2017). It is an alternative to the Hodrick-and-Prescott filter (Hodrick and Prescott, 1997). We set the parameters $h = 2$, and regress on the four most recent values. The filter is applied to consumption and to the following variables after scaling them by GDP: domestic credit, global credit, money, public debt, debt servicing, investment, and current account. Using changes scaled by GDP (our baseline) leads to consistently more accurate predictions than using growth rates or the Hamilton filter.

Additional variables. Next, we investigate how the performance changes if we use additional variables, e.g. adding the real long and short-term interest rates, or replacing the total credit growth variable by household and business loan growth separately. We also add house prices to the list of predictors. Adding the interest rates does not improve the performance of the models. This is in line with our previous observation that the slope of the yield curve captures the predictive power of both interest rates.[25]

Replacing total loans by household and business loans does not improve the performance of any model. Adding house prices does increase the performance of extreme trees and random forest by one and two percentage points, respectively. This is in line with the observation that credit growth after WW2 is strongly driven by the increase in mortgage debt and that house price bubbles are predictive of financial crises (Jordà

[24] We trained the robustness check and the baseline on those observations that are complete in both cases. If the pool of observations does not change in a robustness check (e.g. adding nominal rates) with respect to the baseline the results can be directly compared with the baseline in the first row.

[25] While we found a statistically significant interaction between the short-term interest and the slope when fitting the logistic regression to the whole dataset (Model 8 in Table IV), incorporating the interaction did not improve the out-of-sample performance.

et al., 2015). 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.[26] The inclusion of house prices also reduces the crisis sample considerable. Together, these findings led to the decision to exclude them from the baseline model.

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 consistently produced the most accurate results for all prediction models.

Importantly, extreme trees and random forest performed best in each of these additional experiments which confirms their superiority in this prediction problem and justifies our focus on the extreme trees in the analyses. All in all, the chosen baseline in terms of variables, transformations and horizon, delivers the best model performance in the great majority of cases, captures major economic signalling channels and provides a reasonable reaction time for eventual policy actions.

Cross-validation procedure There are several cross-validation procedures we can use for our problem. For example, in our main experiment, we make the constraint that the two observations of the same crisis (1–2 years before the actual crisis observation) are assigned to the same fold. In Appendix C.1 we compare four different approaches and assess the robustness of the results both in terms of performance and variable importance. Our results are qualitatively stable across the different types of cross-validation. Only the performance levels across all algorithms change in line with the information provided to a model during training.

4.3 Importance of individual variables

To assess the importance of the predictors across all observations, we compute the mean absolute Shapley values (1) for each predictor. We refer to this measure as the *predictive share* of a variable and show it in Figure IV for the different prediction models. The variables are ordered by decreasing predictive share of extreme trees. The two variables with the largest predictive shares are the global slope of the yield curve and global credit

[26] This also holds when we calibrate the model only on observations after WW2

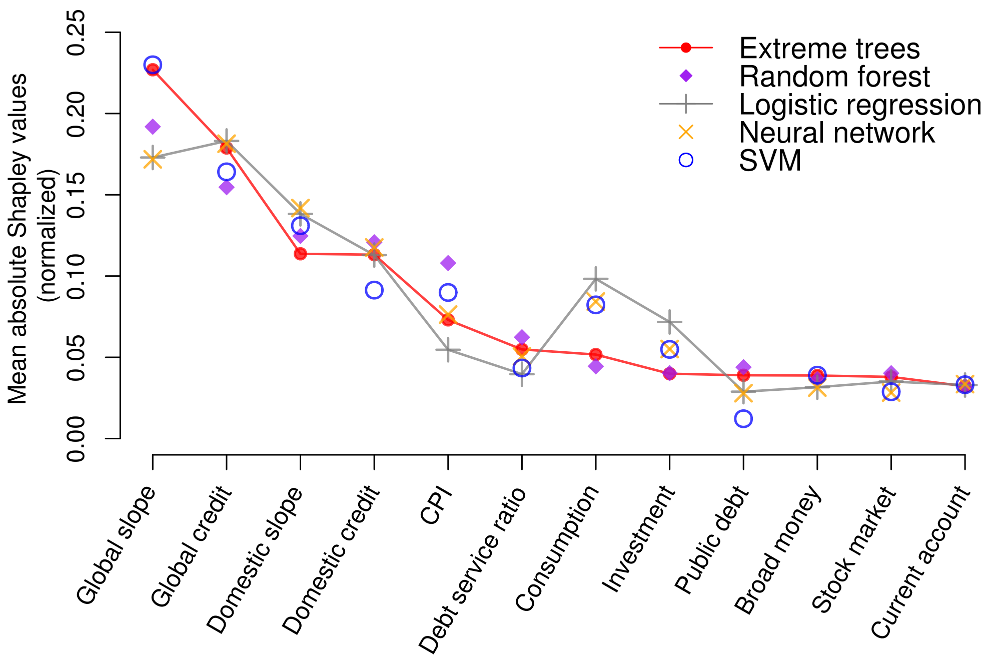

Figure IV: Mean absolute Shapley values.

growth. Both are consistently ranked as the top two across the five models. Domestic slope and credit follow after that. CPI, the debt servicing ratio, consumption and investment come next with predictive shares between five and ten percent. The remaining variables are less important to our models with average shares of less than five percent. Despite the large variety and different nature of the models, the importance ranking of the variables is quite consistent across models. Also, this ranking of the variables closely matches the strength of the predictors in the in-sample logistic regression (Table II) and strengthens the view that the key variables credit and slope are robust indicators for predicting financial crises.

4.3.1 The predictive power of the slope of the yield curve

Our results corroborate earlier findings that stressed the importance of domestic (Schularick and Taylor, 2012) and global credit to GDP growth (Alessi and Detken, 2018) as predictors for financial crises. Furthermore, the results indicate that asset price growth, money, and the current account are less predictive. The important role of the slope, both domestic, and global, is interesting and adds to the list of indicators prominently discussed in the literature. There could be several interlinked explanations for the slope's predictive power: (1) on a domestic level, a flat slope could indicate compressed net interest rate margins for banks (Borio et al., 2017). In this already fragile environment, any small shock to the financial system can seriously impact this already reduced bank profitability. This can lead to a contraction in credit supply and thus affect real economic activity and, in the worst case, even result in a financial crisis (Adrian et al., 2010). (2) A flat or inverted yield curve is usually accompanied by low risk premia, both domestically and globally. With low nominal rates and low risk premia, investors might have to search for riskier investment to achieve profitable returns without being properly compensated for their increased risk exposure. This system-wide build-up of under-priced risk makes diversification more difficult and could trigger a market correction with severe disruptions in financial markets, which again, in the worst case, could lead to a financial crisis.

(3) The spread of optimism/pessimism may lead to a collective, exaggerated belief in a low/high risk environment (Gennaioli et al., 2015). In the calm period before a crisis, a globally flat yield curve may be associated with collectively underestimated risk premia. This is in line with the view of shared narratives in financial markets (Gennaioli and Shleifer, 2018). (4) On a global level, a flattening of the yield curve could point towards a global economic slowdown, which could be a likely trigger for already present financial vulnerabilities. Overall, our findings suggest a combination of low risk perception, strong financial expansion and increasing financial fragility in the years preceding a crisis.

We will focus the subsequent on our best performing machine learning model, extreme trees, and state results for other models when necessary.

4.3.2 Importance of variables across time and across countries

The financial and economic system has changed substantially over the period of time covered in our dataset. We therefore expect that the prediction of crises is also subject to changes over time. Predictors relevant before World War II might not be relevant today. As we are interested in how well the predictors differentiate between a crisis and non-crisis observation, we compute the *Shapley difference*, the mean Shapley value of crisis observations subtracted by the mean Shapely values of non-crisis observations.

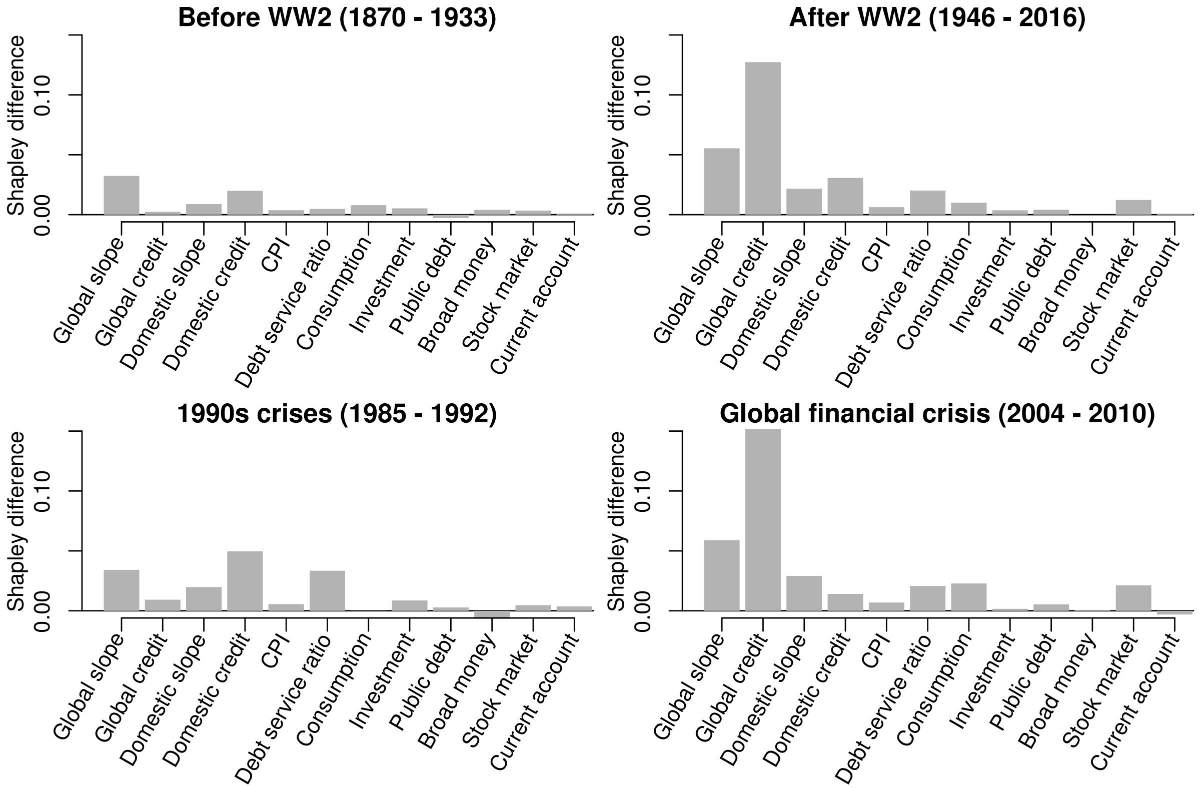

Figure V: Mean difference of Shapley values (crisis - non-crisis observations) for different periods

Figure V 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 2008).

Before World War II, primarily the global slope and domestic credit differentiated markedly between crisis and non-crisis observations. During the series of financial crises that occurred in the 1990s, domestic credit, global slope, and the debt service ratio were key predictors. During the global financial crisis, by far the most important predictor was global credit. While domestic and global slope and domestic credit were crucial predictors throughout our historic sample, the predictive power of global credit seems to be a very recent phenomenon.

Cleary, global financial conditions have an effect on domestic financial stability. As Cesa-Bianchi et al. (2018), we observe that the financial globalisation has magnified the importance of international credit growth. For example, Germany and Switzerland experienced negative domestic credit to GDP growth before the global financial crisis in

2008. Nevertheless, both countries experienced a financial crisis because their banking sectors have invested into the US subprime mortgages as well.

The global slope of the yield curve seems to be a crucial predictor across the whole time period covered by our data set. This might be explained by two regimes. First, pegged exchange rates established a close connection of macroeconomic policies between countries (Obstfeld et al., 2005). Later, the globalisation of the world economy, especially an enhanced global bond market integration (Diebold et al., 2008) cements the importance of a global yield curve.

Figure VI shows the Shapley decomposition at the most granular level for the United Kingdom and Spain, again for extreme trees. To retain legibility, only the Shapley values of the slope and credit (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 (dotted horizontal line) add up to the predicted value, shown by the black line. Crisis observations are highlighted with vertical bars in light red.

In the United Kingdom, the crisis in the 1990s is correctly predicted by a joint contribution of all four key indicators, while the global financial crisis in 2008 is mostly predicted by global credit growth. In Spain, the latter crisis is flagged by domestic and global credit growth. The marked false alarms in Spain's early 1990s can be explained by the effects of a severe recession affecting the Western world globally and Spain, locally.

In both the United Kingdom and Spain, the four key indicators correctly point to a low probability of crisis during the *golden age of capitalism* (Marglin and Schor, 1990) between the end of WW2 and the early 1970s.[27]

A similar pattern holds for the period after the global financial crisis in 2008. As noted before, the predictions before the second world war are less accurate and stable.

4.3.3 Shapley regressions

The above discussion of the drivers of machine learning models had been qualitative. Here we provide a rigorous statistical analysis of machine learning models. To our best

[27] Figure III shows that our model nearly makes no false positive predictions for any country in this period.

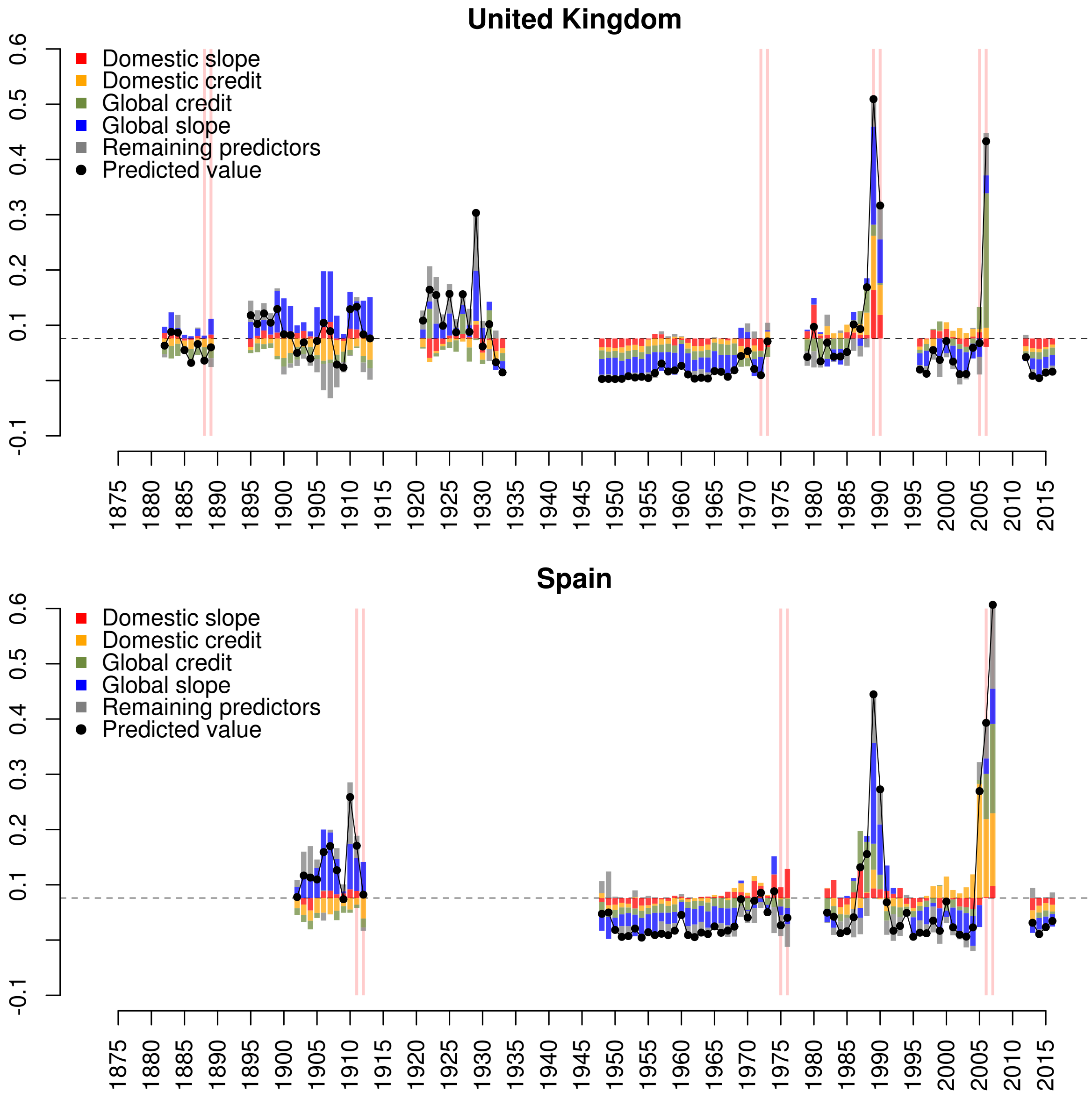

Figure VI: Shapley values as a function of time for the United States (top panel) and Spain (bottom panel).

knowledge this is the first time this type of analysis has been provided.

The left half of Table VIII shows the output of Shapley regressions (Joseph, 2019) where the crisis dummy is regressed on the model Shapley value decomposition (1). The coefficients represent the effects of the Shapley values for a one standard deviation change on the predicted log-odds. It is important to note that the sign of the coefficients does not indicate the sign of the association between the predictors and the crisis class label. Rather, the coefficients are expected to be positive because high Shapley values reflect an increase of the predictive probability of the positive class (crisis). A negative coefficient indicates that the model has not well learned the information in that variable and is considered non-significant. We obtained the direction of the association from our baseline logistic regression in Table II as a measure of linear alignment between a variable and the crisis label.

Consistent with our previous results, global and domestic credit and slope obtain the highest coefficients and lowest p-values (measured against the null hypothesis), accordingly. Consumption, investment and changes in stock market indices can also be judged 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 providing a useful signal. Variables like the debt servicing ratio, public debt and the current account balance have some predictive weight but their signals can not be differentiated from the null, i.e. there is no clear alignment with crises. The signs are in line with economic intuition and previous results, namely that credit growth and investment are positively associated with financial crises while the slope is negatively associated. Note, however, that the coefficients for consumption and CPI are significant with a negative sign. This points to real fragilities in the build-up of financial crises, which we will investigate in more detail below.

It is informative to compare the Shapley coefficients to the coefficients of our logistic regression using the raw inputs, not the Shapley values. The right hand side of Table VIII reproduces our baseline regression (Model 4 in Table II). This model also gives highest weight to the slope and credit but does produce some results which are qualitatively different for the remaining predictors. For instance, stocks and CPI are significant predictors in the Shapley regression but not significant in the logistic regression. These differences indicate that stocks and CPI are more predictive when they are not constrained to a linear relation.

	Shapley regression				Logit regression	
	Direction	Coefficient	p	Share	Coefficients	p
Global slope	-	0.55	0.000	0.23	-0.61	0.000
Global credit	+	0.33	0.000	0.18	0.67	0.000
Domestic slope	-	0.37	0.000	0.11	-0.58	0.000
Domestic credit	+	0.34	0.000	0.11	0.43	0.002
CPI	-	0.28	0.002	0.07	-0.24	0.160
Debt service ratio	+	0.06	0.244	0.05	0.16	0.347
Consumption	-	0.17	0.027	0.05	-0.42	0.003
Investment	+	0.18	0.005	0.04	0.32	0.016
Public debt	-	-0.05	0.295	0.04	-0.03	0.845
Broad money	+	-0.12	0.081	0.04	0.04	0.817
Stock market	-	0.16	0.020	0.04	-0.13	0.451
Current account	-	-0.05	0.296	0.03	-0.08	0.525

TABLE VIII: Left side: Shapley regression. Direction of alignment between predictor and target (same as sign of logistics regression), coefficients, p-values against the null hypothesis and predictive share of variable. Right side: Coefficients and p-values of a logistic regression.

4.4 Nonlinearities

Using Shapley values, we can depict the nonlinearities learned by the machine learning models. Figure VII shows the Shapley values of the key predictors as a function of the actual predictor values for extreme trees. A Shapley value greater than zero indicates an increase in the predicted probability of crisis. The crisis observations are marked in red, while the non-crisis observations are shown in grey. The black and blue line show a linear and cubic function fitted to the data.

To test the importance of nonlinearities, we fit linear (black line) and cubic polynomial (blue line) regressions to the input-Shapley value relations. As expected the goodness-of-fit is substantially better for nonlinear descriptions in terms of R^2. The nonlinear relationships are neither unintuitive nor complex. Rather, the relationships are generally monotonic and match our expectation. A severe flattening or inversion of the yield curve is associated with a higher probability of crisis, as are higher global and domestic credit growth.The importance of nonlinearity is particularly pronounced for global credit with little dispersion about the best-fit cubic regression.

Does modelling nonlinearity improve the predictive power of all predictors? To an-

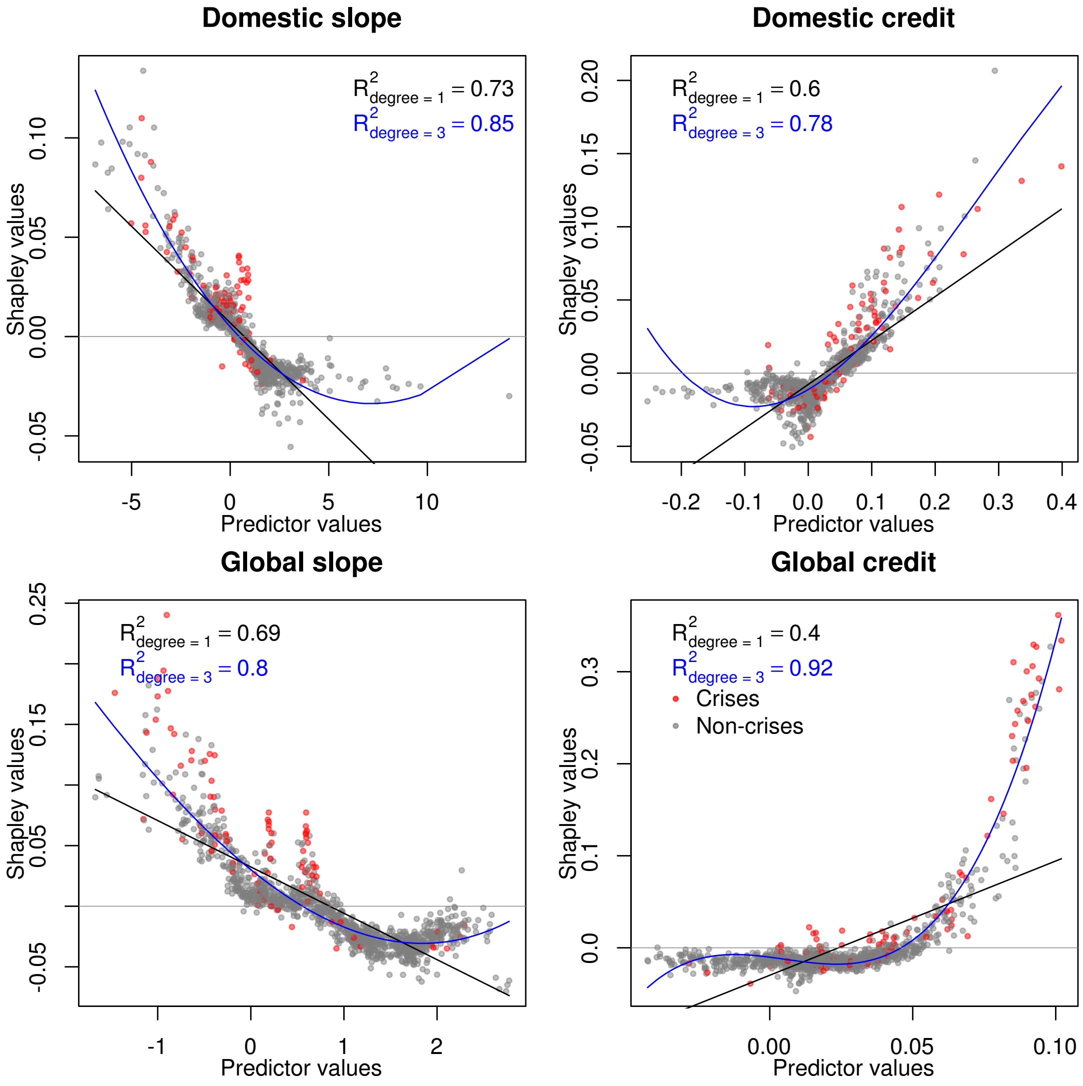

FIGURE VII: Indictor values plotted against Shapley values for the three most predictive indicators. Crisis observations are highlighted in red.

swer this question, we conduct a simple test and regressed the crisis outcome on each indictor independently, once on its actual values and once on the Shapley values.[28] Table IX compares both the AUC and log-likelihoods. We used the DeLong test (DeLong et al., 1988) and the Vuong's closeness test (Vuong, 1989) to evaluate whether the difference in AUC and log-likelihood are significant. For our key indicators, the goodness-of-fit is significantly better when we regress on Shapley values. The AUC score is also significantly different for most variables but not for domestic credit which suggests that the

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

nonlinearity in this variable is monotonic as the AUC score is not affected by a strictly monotonic transformation.

	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 IX: AUC and log likelihood of univariate logistic regressions. The crisis outcome is once regressed on the values of the variables (Linear) and once on the Shapley values extracted from extreme trees. Bold cells indicate a significant performance of one regression over the other ($p < 0.05$).

4.5 Interaction of predictors

The Shapley values shown in Figure IV show the overall effect of a variable on the prediction. It does not tell us how much of the effect can be attributed to that variable alone and how much to the interaction with other variables. Within the Shapley value framework, we can explicitly measure how much a particular interaction drives the prediction (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:

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

Interaction	Direction	Coefficient	p-values
Domestic slope x Domestic credit	-	0.08	0.154
Domestic slope x Investment	-	0.11*	0.070
Domestic slope x Consumption	+	0.17**	0.043
Domestic slope x CPI	+	0.04	0.365
Domestic slope x Stock market	+	0.09	0.109
Domestic credit x Investment	+	0.21***	0.005
Domestic credit x Consumption	-	-0.20	0.005
Domestic credit x CPI	+	0.17**	0.012
Domestic credit x Stock market	+	-0.17	0.009
Global slope x Global credit	-	0.32***	0.002
Global slope x Domestic slope	+	0.09	0.169
Global slope x Domestic credit	-	0.24***	0.004
Global slope x Investment	-	0.41***	0.000
Global slope x Consumption	+	0.13*	0.058
Global slope x CPI	+	0.26***	0.003
Global slope x Stock market	-	0.09	0.185
Global credit x Domestic credit	+	0.11*	0.083
Global credit x Domestic slope	-	0.18**	0.027
Global credit x Investment	+	0.14**	0.036
Global credit x CPI	-	0.39***	0.001
Global credit x Consumption	-	0.23***	0.002
Global credit x Stock market	+	0.23**	0.014

Table X: Shapley regressions including interactions. Each row is based on a different regression including the Shapley values of the respective interaction and the main effects of the 12 predictors. Significance levels: *p<0.1; **p<0.05; ***p<0.01.

We investigate Shapley interactions in the extreme trees model. We do not consider all $\frac{12\times 11}{2} = 66$ pairwise interactions but examine only interactions of global and domestic credit growth and the slope with variables that have a major predictive share and a significant main effect (see Table VIII).

To test the significance of the interactions we estimate Shapley regression models. Each regression includes the Shapley values of a single interaction as well as the general Shapley values of all other predictors.[29] We do not include all interactions at once because of collinearity issues which also reflect the dependence of different economic channels, e.g. total credit and investment.

[29] We control for the interactions when computing the Shapley values of the involved indicators. For example, when testing the interaction of domestic credit and domestic slope we adjust the main effect of the two variables: $\hat{\phi}_{\text{Dom. credit}} = \phi_{\text{Dom. credit}} - \phi^{\mathbf{i}}_{\text{Dom. credit}\times\text{Dom. slope}}$, $\hat{\phi}_{\text{Dom. slope}} = \phi_{\text{Dom. slope}} - \phi^{\mathbf{i}}_{\text{Dom. credit}\times\text{Dom. slope}}$.

Table X shows the summary statistics for the interaction terms in the Shapley regression models.[30] Three conditions emerge for the increased likelihood of a crisis: (i) Including global factors for both credit and the slope provide stronger interaction signals than domestic variables. This again suggests the importance of internationally shared narratives around risk perception in line with (Gennaioli and Shleifer, 2018).

(ii) The positively significant contributions of domestic and global credit and investment growth point to overconfidence about future expectations in financial markets and the supply side of the economy. (iii) This overconfidence seems to be matched by restraint on the demand side, as shown by the negatively significant interactions with consumption. Again, this is particularly true for the interaction between global and domestic factors pointing to the importance of both when assessing the likelihood of future financial crises.

The above findings are in line with historical accounts about the build-up of crises from exuberant expectations in credit markets (Aliber and Kindleberger, 2015) and the financial instability hypothesis (Minsky, 1977). However, not all credit booms result in a financial crisis (Dell'Ariccia et al., 2012). A credit boom might be more detrimental when recession expectations (negative slope of the yield curve) are high and consumption is declining than during normal global economic conditions (positive slope of the yield curve and consumption growth). Similarly, growth in investment is a higher risk factor for a crisis when the slope is low while a decline in investment might be problematic when the slope is high.

We show three significant interactions of the global slope of the yield curve and various domestic factors in Figure VIII. 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 red indicating a higher probability of crisis.

In the left and middle panel, we observe an increase in the predicted probability of crisis when the global slope is low and growth in credit and investment are high, respectively. Indeed most crises fall into the lower right quadrant of both figures highlighting the importance of conditions (i) and (ii). In the right panel we see that the combination

[30] As noted before, the sign of the coefficients are expected to be positive. We therefore do not interpret the interactions with a negative coefficients which point to a bad model fit for this variable. We estimate the sign by correlating the interaction of the input variables with the crisis outcome.

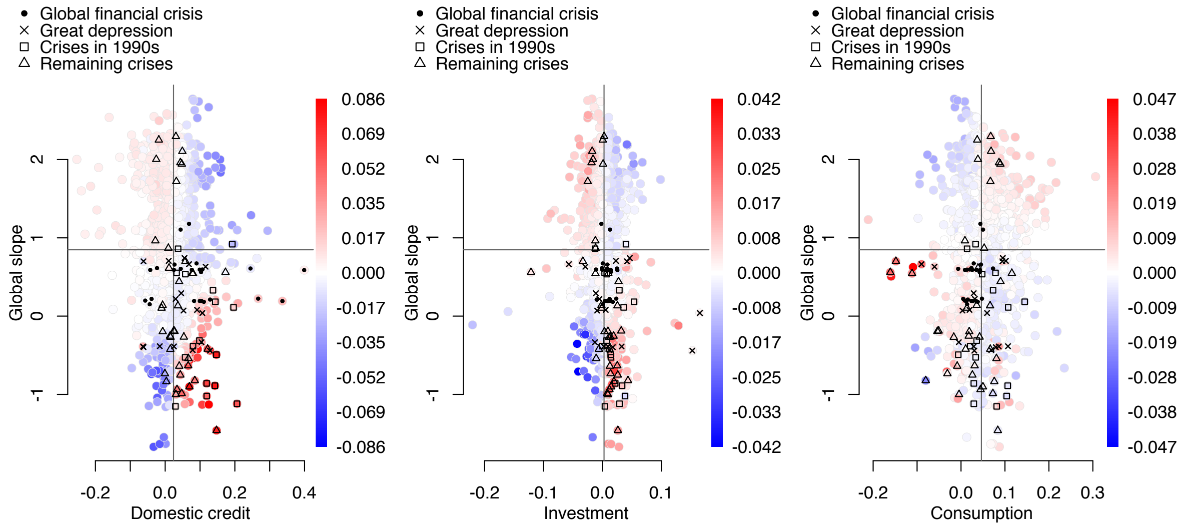

FIGURE VIII: Shapley interactions. Each scatter plot shows all observations (filled circles) as a function of their values on two predictors. The colour of the observations denotes the value of the Shapley interaction, with red indicating a higher predicted probability of crisis. Crisis observations are additionally highlighted with black symbols.

of a low global slope and sluggish consumption growth also indicates the build-up of financial crises (lower left quadrant). This highlights the importance of condition (iii).The weakening of consumption may be associated with an increased leverage of the household sector as evidenced by significantly elevated household credit and mortgage growth in the years before a crisis (see Table I).

It is important to stress again that the analyses in Table X and Figure VIII only shows interaction effects additional to main effects and higher order interactions. We therefore cannot expect that the interactions account well for all crises.

5 Forecasting experiment

All results shown so far are based on cross-validation. 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 time period, we predict crises in the past with data from the future. Therefore, our performance estimates do not reflect the performance of an early warning model at actual deployment that aims at predicting crises in the future. Notwithstanding, there are good reasons for using cross-validation instead of forecasting crises with past data only. For example, our analysis has shown that the global financial crisis is qualitatively different from previous crisis in that global credit plays a very 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. In general, due to the limited number of crises in the data, a comparison of the forecasting performance across models will suffer from low statistical power. Nevertheless, we conducted a recursive forecasting experiment to compare the different prediction methods.

We use all observations up to year $t - 2$ to train the models and test them on observations of year t[31], where $1946 \leq t \leq 2016$.

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. As the predicted probability, and therefore the AUC estimate, is highly sensitive to the proportion of classes in the training set, we resampled all training sets such that they contain the same number of crisis and non-crisis observations.[32]

Table XI compares the forecasting performance of the models. It shows the AUC on all observations between 1945–2016, and for the period before and after 2004.

Across the entire forecasting period, the best model is the neural network, followed by extreme tress and the SVM. Generally, the machine learning models—except the decision trees—perform better than logistic regression in forecasting, also when predicting the global financial crisis (2004–2016). However, it is important to note that the test set is small and all performance differences in all three periods are not significant ($p < 0.05$) according to a DeLong test (DeLong et al., 1988).

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

[32] 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. Using upsampling, we increase the number of crisis observations by drawing n_- observations with replacement. Using upsampling, we use a subsample (without replacement) of 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.

Figure IX shows the forecasting performance at a hit rate of 80% of the neural network.

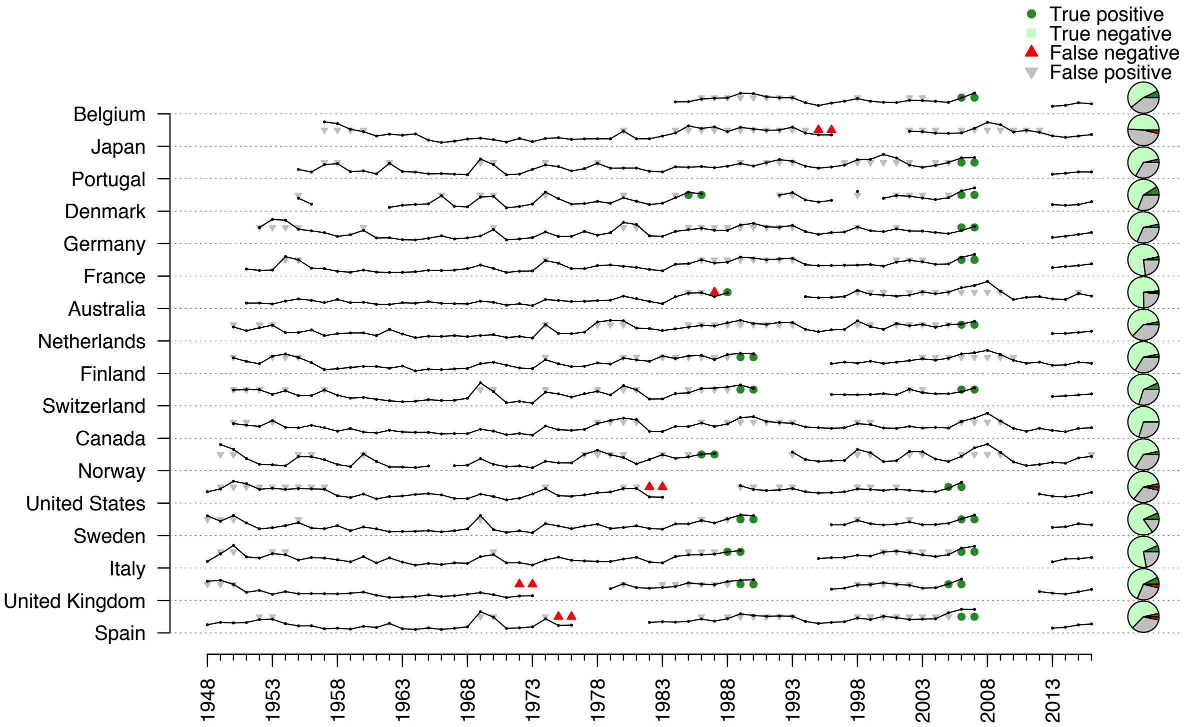

Figure IX: Forecasting performance of the neural network according to a prediction threshold of 0.40.

	Forecasting period		
	1945–2016	1945–2003	2004–2016
Extreme trees	0.813	0.748	0.870
Random forest	0.792	0.735	0.846
Logistic regression	0.789	0.704	0.867
SVM	0.808	0.700	0.911
Neural network	0.833	0.770	0.872
Decision tree	0.788	0.727	0.867

Table XI: Forecasting performance (AUC) on all observations after 1944 and those before and after the global financial crisis.

6 Conclusion

In this paper, we built early warning models based on machine learning approaches. We show that machine learning models outperform logistic regression in predicting financial crises in a macroeconomic dataset covering 17 countries between 1870–2016. The most accurate models are decision tree based ensembles, such as random forest, followed by support-vector machines and neural networks. A simple decision tree is the only machine learning model that falls behind the benchmark logistic regression. The gains in predictive accuracy justify the use of initially complex black box machine learning models. To understand their predictions, we use the Shapley value framework.

We observe that all models, including the logistic regression, consistently identify the same predictors for financial crises, namely domestic and global credit, as well as the domestic and global slope of the yield curve as key predictors. While the crucial role of credit is an established result, we want to highlight the predictive power of the slope, which has only obtained little attention as an early warning indicator for financial crises in the literature. With the Shapley regression approach we confirm that the contribution of our key predictors is statistically significant, even after controlling for the other covariates. We also inspect nonlinearities and interactions learned by the machine learning models. Nonlinearities are ubiquitous in the models but they are mostly relatively simple, e.g. monotonic kink-type form.

We observe several significant interactions point to a coherent narrative that might be useful for policy makers. Three common themes emerge around the build-up of financial crises. A shared global narrative as indicated by the importance of global variables, particularly the slope of the yield curve, throughout time. Prolonged high growth in credit relative to GDP and a flat/inverted yield curve are often go hand-in-hand before a financial crisis. A third observation is what one may call a disconnect between financial and real factors in the built-up phase of a crisis. Namely, consumption is generally slowing while credit continues to expand within a perceived low risk environment. The slowing of consumption and a potential weakening of demand may be associated with increased leverage in the household sector. Both loans to household and mortgage lending are significantly increased in the period preceding a crisis (see Table I).

Given the large time horizon of the data, it is not surprising that the importance of the variables change across time, reflecting historical change in the global monetary financial architecture but also more recent developments. The most salient time-varying effect is that of global credit growth, which is the most crucial indicator for the global crisis in 2008 but only marginally predictive before.

There are two appealing advantages of classical parametric modelling over machine learning modelling, namely the interpretability of the models and the availability of statistical inference tests. However, with the the application of the Shapley values and Shapley regressions, these shortcomings can be overcome. In fact, our study shows an important advantage of machine learning on top of the gains in predictive accuracy: The flexibility of the machine learning models helps us to uncover important patterns such as nonlinearities and interactions that would have been concealed using only a linear model. These patterns may now, after the fact, also be explicitly included in a standard model, such as logistic regression.

From an applied perspective, our results help policy makers to anticipate a financial crises in advance. The ability to do so is crucial given the enormous economic, political, and social consequences that financial crises entail. By having models and reliable indicators, policy makers can anticipate these events in advance and can put measures in place to avoid or at least lessen the consequences of a financial crisis.

References

Abbritti, Mirko, Mr Salvatore Dell'Erba, Mr Antonio Moreno, and Mr Sergio Sola (2013) *Global factors in the term structure of interest rates*, No. 13-223: International Monetary Fund.

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: John Wiley Hoboken, NJ.

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

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," SSRN Scholarly Paper ID 3205208, Social Science Research Network, Rochester, NY.

Bernanke, Ben S. (2009) "Asia and the Global Financial Crisis," October.

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) "Chapter 21 The financial accelerator in a quantitative business cycle framework," in *Handbook of Macroeconomics*, Vol. 1: Elsevier, pp. 1341–1393.

Betrán, Concha and María A Pons (2013) "Understanding Spanish Financial crises, 1850-2000: What determined their severity?" in *Paper presented at European Historical Society Conference*, Vol. 6, p. 7.

Beutel, Johannes, Sophia List, and Gregor von Schweinitz (2018) "An evaluation of early warning models for systemic banking crises: Does machine learning improve predictions?".

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, 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.

Cawley, Gavin C and Nicola LC Talbot (2010) “On over-fitting in model selection and subsequent selection bias in performance evaluation,” *Journal of Machine Learning Research*, Vol. 11, No. Jul, pp. 2079–2107.

Cecchetti, Stephen G, Marion Kohler, and Christian Upper (2009) “Financial crises and economic activity,”Technical report, National Bureau of Economic Research.

Cesa-Bianchi, Ambrogio, Fernando Eguren Martin, and Gregory Thwaites (2018) “Foreign booms, domestic busts: The global dimension of banking crises,” *Journal of Financial Intermediation*.

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.

Cortes, Corinna and Vladimir Vapnik (1995) “Support-vector networks,” *Machine Learning*, Vol. 20, No. 3, pp. 273–297.

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.

Dell’Ariccia, Giovanni, Deniz Igan, and Luc UC Laeven (2012) “Credit booms and lending standards: Evidence from the subprime mortgage market,” *Journal of Money, Credit and Banking*, Vol. 44, No. 2-3, pp. 367–384.

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*, 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.

Drehmann, Mathias and John Juselius (2012) “Do debt service costs affect macroeconomic and financial stability?” *BIS Quarterly Review*.

Drehmann, Mathias, Claudio Borio, and Kostas Tsatsaronis (2011) “Anchoring Countercyclical Capital Buffers: The Role of Credit Aggregates,” *International Journal of Central Banking.*

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.

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, Milton (1970) “Controls on interest rates paid by banks,” *Journal of Money, Credit and Banking*, Vol. 2, No. 1, pp. 15–32.

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 University Press.

Gennaioli, Nicola, Andrei Shleifer, and Robert Vishny (2015) “Neglected risks: The psychology of financial crises,” *American Economic Review*, Vol. 105, No. 5, pp. 310–14.

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

Hamilton, James D (2017) “Why You Should Never Use the Hodrick-Prescott Filter,”Technical report, National Bureau of Economic Research.

Hodrick, Robert J and Edward C Prescott (1997) “Postwar US business cycles: an empirical investigation,” *Journal of Money, credit, and Banking*, pp. 1–16.

Jordà, Òscar, Moritz Schularick, and Alan M Taylor (2015) “Betting the house,” *Journal of International Economics*, Vol. 96, pp. S2–S18.

——— (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," *Bank of England Staff Working Paper Series*, No. 784.

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, Saul Lizondo, and Carmen M Reinhart (1998) "Leading indicators of currency crises," *Staff Papers*, Vol. 45, No. 1, pp. 1–48.

Kindleberger, Charles P. (1978) *Manias, Panics and Crashes - A History of Financial Crises*: New York: Basic Books.

King, Mervyn (2010) "Speech at the University of Exeter," January.

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.

Marglin, Stephen A and Juliet B Schor (1990) *The golden age of capitalism: reinterpreting the postwar experience*: Oxford University Press.

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.

Niculescu-Mizil, Alexandru and Rich Caruana (2005) "Predicting good probabilities with supervised learning," in *Proceedings of the 22nd international conference on Machine learning*, pp. 625–632, ACM.

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.

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.

Platt, John et al. (1999) "Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods," *Advances in large margin classifiers*, Vol. 10, No. 3, pp. 61–74.

Quinlan, J Ross (1993) *C4. 5: programs for machine learning*: Elsevier.

Reinhart, Carmen M and Kenneth S Rogoff (2008) "Is the 2007 U.S. Sub-Prime Financial Crisis So Different? An International Historical Comparison," Working Paper 13761, National Bureau of Economic Research.

Reinhart, Carmen M. and Kenneth S. Rogoff (2009) *This Time Is Different: Eight Centuries of Financial Folly*: Princeton University Press. Google-Books-ID: ak5fLB24ircC.

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, ACM.

Rogers, William (1993) "Regression standard errors in clustered samples," *Stata Technical Bulletin*, Vol. 13, pp. 19–23.

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. 1029–61.

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

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.

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.

Vuong, Quang H (1989) "Likelihood ratio tests for model selection and non-nested hypotheses," *Econometrica: Journal of the Econometric Society*, 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.

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

Zadrozny, Bianca and Charles Elkan (2002) "Transforming classifier scores into accurate multiclass probability estimates," in *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 694–699, ACM.

Zaloom, Caitlin (2009) "How to read the future: the yield curve, affect, and financial prediction," *Public Culture*, Vol. 21, No. 2, pp. 245–268.

A Data Appendix

Figure A.I depicts which observations are discarded from the analysis. Note the order of plotting: the drop out due to missing feature values is plotted first, the post-crisis bias second, and the world wars and the aftermath of the great depression last.

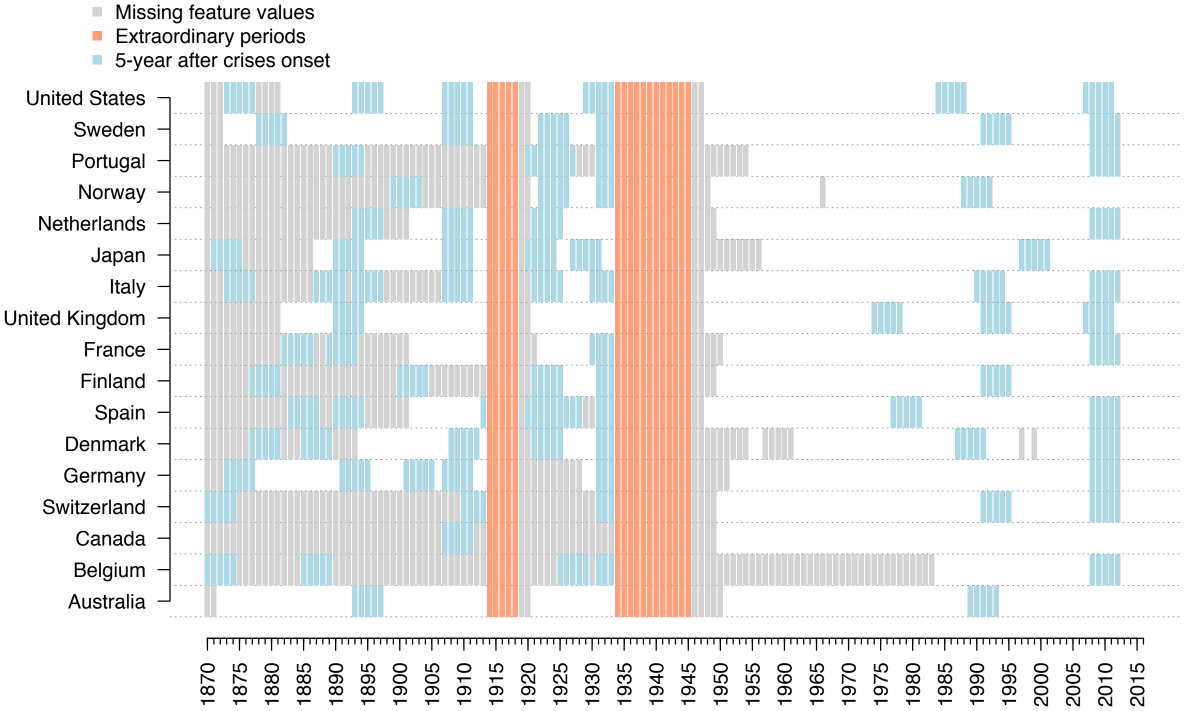

Figure A.I: Excluded data points.

B Models Appendix

B.1 Implementational details

We list 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. We tested all models described below in three configurations: (1) training the models on the training set as given, (2) randomly upsampling the minority class (financial crisis events) such that both classes have the same share in the training set, and (3) weighting the objects such that both classes contribute equally to the training set. Concretely, the objects in the positive class were weighted by $0.5/\frac{N_+}{N_+ + N_-}$ and the objects in the negative class by $0.5/(1 - \frac{N_+}{N_+ + N_-})$.

For each algorithm we only report the results for the configuration that performs best in the cross-validation experiment. If the models are not trained on the objects as given, we explicitly state it in the list below.

Logistic regression. We used the `SGDCClassifier` implementation from the Python package `sklearn` with *penalty* = None and *loss* = log. We also tried regularized 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. Random forests are known to be not very sensitive to the choice of hyperparameters. Nevertheless, we also tested a version of random forest 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. It did not improve the performance.

Extremely randomised trees. We used the `ExtraTreesClassifier` implementation from the Python package `sklearn` with *n_estimators=1000.* 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. However, 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* $\in \{2^{-10+13\times\frac{0}{9}}, 2^{-10+13\times\frac{1}{9}}, ..., 2^{-10+13\times\frac{9}{9}}\}$. We weight the objects such that both classes contribute equally to the training set. We trained 25 SVMS in each training sample. For each model we randomly upsampled the training data. 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* $\in \{n/3, n/2, n, (n, n/2), (n, n), (10, 5), (20, 10)\}$, where n is the number of features tested with numbers all round to the nearest integers. We trained 25 networks in the training sample. 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 objects such that both contribute equally to the training set.

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 results about CART in the study because it performed inferior to the other decision tree algorithm C5.0.

Gradient boosting We used the `SGBCClassifer` 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 training set when training this algorithm. We do not report results about CART in the study because it performed inferior to the other tree ensembles random forest and extremely randomised trees.

C Results Appendix

C.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. There are different constraints that can be made when assigning the observations to the folds. Here, we investigate whether these constraints qualitatively 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 make 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 was used in our baseline experiment. 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.

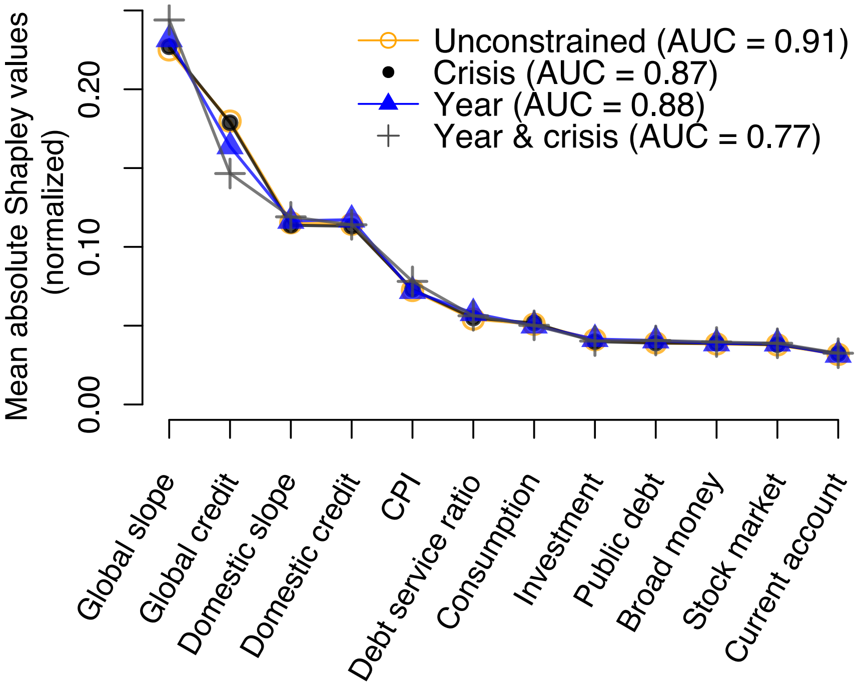

Figure C.I: Shapley difference for extreme trees for the four different cross-validation experiments.

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. We obtained the highest accuracy using unconstrained cross-validation (AUC = 0.91). This high performance compared to our baseline experiment (AUC = 0.87) confirms that the two years before a crisis are highly similar to each other and should be assigned to the same fold to avoid overly optimistic performance estimates. Assigning all observation of the same year to the same fold gives an AUC of 0.88. 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 2007–2008: Both years before the global financial (2005–2006) are in the same fold such that the models cannot learn the predictive importance of global credit growth for that crisis. The Shapley analysis in Figure C.I confirms this explanation. It shows the mean absolute Shapley values for the four types of cross-validation. Generally, the four types of cross-validation show highly similar patterns. However, the global credit variable is a less important predictor for the constrained cross-validation with the year & crisis constraint.

C.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 set.

For example, assume that half of the observations of year 2007 are in the training and set and the other half in the test set. As most countries in 2007 experienced a crisis, 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 form 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.

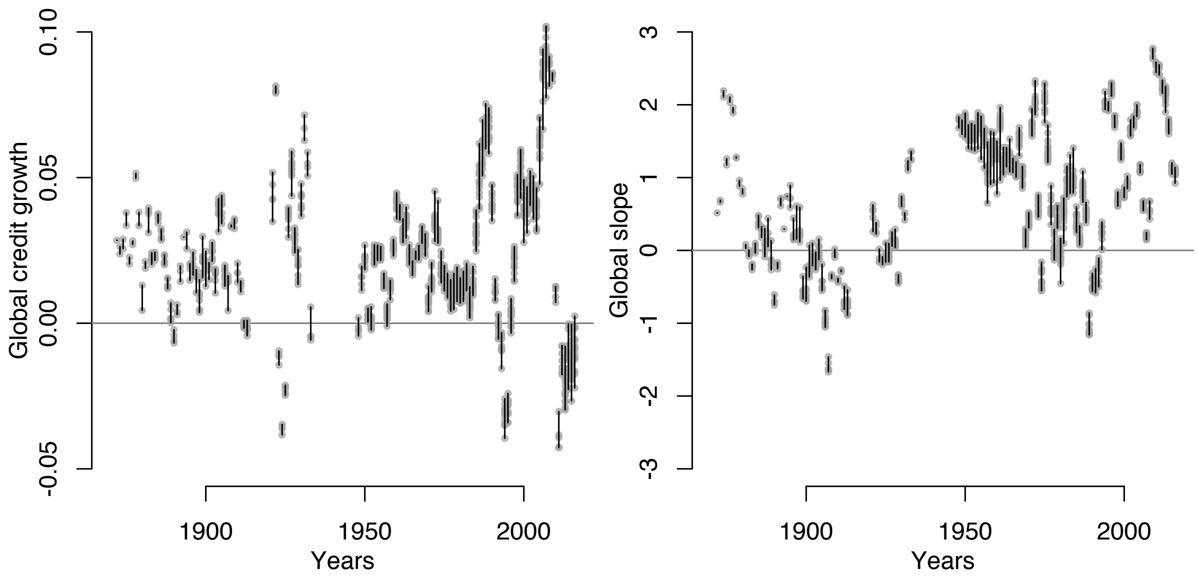

Figure C.II: Depiction of the global variables. The grey circles show the actual values, the vertical lines show the range of values in each year.

To avoid this effect, we defined the global variable for country c in year y as the value of the domestic variable 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 C.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 learns a smooth monotonic association between the actual value of the global variables and the probability of financial crises and not a direct mapping of values to probability of crisis.

Fourth, the constrained cross-validation (Figure C.I) and the forecasting experiment 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.

C.3 Logistic regression with real rates

Table C.I: Results

	(5)	(6)	(7)	(8)
Real short rate	0.552***		1.835***	0.307
	(0.151)		(0.303)	(0.189)
Real long rate		0.097	−1.607***	
		(0.156)	(0.330)	
Domestic slope				−0.780***
				(0.154)
Domestic slope × real short rate				0.156 (0.110)
+ Covariates	Domestic credit, global credit, CPI, debt service ratio, broad money, stock market, consumption, public debt, investment, current account,			
Observations	1,249	1,249	1,249	1,249
Log Likelihood	-269.507	-276.084	-257.062	-256.039
Akaike Inf. Crit.	563.014	576.168	540.123	540.077
Note:				*p<0.1; **p<0.05; ***p<0.01

Table C.II: Logistic regression model of the real short- and long-term interest rates and the other variables, except domestic and global slope of the yield curve.