Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

FAKULTÄT
FÜR WIRTSCHAFTS- UND
SOZIALWISSENSCHAFTEN

New Forecasting Methods for an Old Problem: Predicting 147 Years of Systemic Financial Crises

Emile du Plessis
Ulrich Fritsche

WiSo-HH Working Paper Series
Working Paper No. 67
September 2022

WiSo-HH Working Paper Series
Working Paper No. 67
September 2022

New Forecasting Methods for an Old Problem: Predicting 147 Years of Systemic Financial Crises

Emile du Plessis, University of Hamburg
Ulrich Fritsche, University of Hamburg

ISSN 2196-8128

Font used: „TheSans UHH“ / LucasFonts

Kontakt:
WiSo-Forschungslabor
Von-Melle-Park 5
20146 Hamburg
E-Mail: experiments@wiso.uni-hamburg.de
Web: http://www.wiso.uni-hamburg.de/forschung/forschungslabor/home/

New Forecasting Methods for an Old Problem: Predicting 147 Years of Systemic Financial Crises

Emile du Plessis[a] and Ulrich Fritsche[b]

July 2022

[a]*University of Hamburg, Faculty of Business, Economics and Social Sciences, Hamburg, Germany, grensduplessis@gmail.com*

[b]*University of Hamburg, Faculty of Business, Economics and Social Sciences, Hamburg, Germany, ulrich.fritsche@uni-hamburg.de*

Abstract

A reflection on the lackluster growth over the decade since the Global Financial Crisis has renewed interest in preventative measures for a long-standing problem. Advances in machine learning algorithms during this period present promising forecasting solutions. In this context, the paper develops new forecasting methods for an old problem by employing 13 machine learning algorithms to study 147 year of systemic financial crises across 17 countries. It entails 12 leading indicators comprising real, banking and external sectors. Four modelling dimensions encompassing a contemporaneous pooled format through an expanding window, transformations with a lag structure and 20-year rolling window as well as individual format are implemented to assess performance through recursive out-of-sample forecasts. Findings suggest fixed capital formation is the most important variable. GDP per capita and consumer inflation have increased in prominence whereas debt-to-GDP, stock market and consumption were dominant at the turn of the 20th century. Through a lag structure, banking sector predictors on average describe 28 percent of the variation in crisis prevalence, real sector 64 percent and external sector 8 percent. A lag structure and rolling window both improve on optimised contemporaneous and individual country formats. Nearly half of all algorithms reach peak performance through a lag structure. As measured through AUC, F_1 and Brier scores, top performing machine learning methods consistently produce high accuracy rates, with both random forests and gradient boosting in front with 77 percent correct forecasts. Top models contribute added value above 20 percentage points in most instances and deals with a high degree of complexity across several countries.

JEL Classification: C14, C15, C32, C35, C53, E37, E44, G21

Keywords: machine learning, systemic financial crises, leading indicators, forecasting, early warning signal

1. Introduction

A decade after the Global Financial Crisis, its remnants are vividly illustrated by the lackluster pace of economic activity hampering progress in several advanced and developing countries. While financial crises have occurred periodically over centuries (Reinhart and Rogoff, 2009), the consequential high social, economic, and political costs (Chen et al., 2019; Funke et al., 2016; Laeven and Valencia, 2010; and Laeven and Valencia, 2018) necessitate an improved preventative framework to mitigate the next financial catastrophe. Recent advances in artificial intelligence in general and machine learning in particular present innovate approaches to revisit forecasting performance of financial crises and assess its contribution to the literature on preventative frameworks. A salient benefit of machine learning comprises its ability to accommodate non-linear interactions between crisis variables, which is useful as crises can have different precursors and during a volatile environment, crisis indicators generally fail to exhibit a linear trajectory. Another advantage is that machine learning methods are able to surface leading indicators. For policymakers, it is both a practical and straightforward approach given its proliferation across statistical programmes. In comparison to traditional macroeconomic tools such as probit or logit models, machine learning approaches have improved on forecasting performance (Alessi et al., 2015; Casabianca et al., 2019; Davis et al., 2011; Döpke et al., 2017; Fouliard et al., 2019). Yet, the applications of machine learning algorithms to study financial crises remain limited.

As new methods for an old problem, 13 machine learning models are developed to scrutinize systemic financial crises. Serving as a conventional non-parametric method, a baseline model is implemented. Commonly used linear and binomial regression models are compared to non-linear models. K-nearest neighbours and support vector machine are instance-based algorithms where the former classifies new observations according to the closest located known value in a dataset, and the latter applies kernels to enlarge the feature space to allow for non-linear relationships. As regularization algorithms, ridge and lasso reduce the contribution of less significant coefficients down to zero. Decision tree algorithms full tree and pruned tree are implemented given their ability to analyse large datasets, operate with missing values, and absent a predefined functional form, allow non-linear relationships, support interactions between variables and the identification of leading indicators. As a specific dimension reduction method, partial least squares approximate new features to the original features in the dataset, and given its relation to the outcome variable, explicate the outcome and predictors. Ensemble algorithms incorporate a set of weak learners to collectively construct a strong

learner, with the aim of enhancing the performance of a single forecast. This process involves training multiple models with the same algorithm. Two standard methods encompass bagging, featuring random forests and boosting, comprising adaptive boosting and gradient boosting. While both algorithmic approaches generate new data through sampling by replacement, procedurally there are two key distinctions. First, trees are formed in parallel in the bagging process, but sequentially for boosting. Second, bagging estimates strong learners using simple average across all the prediction trees, while boosting applies weighted average or learning rates.

Machine learning and forecasting algorithms frequently encounter a bias-variance trade-off. Accordingly, while reducing bias through a close fit to an existing dataset, it comes at the cost of higher variation when applied to a new dataset.

This paper studies 147 years of systemic financial crises, comprising a total of 17 present-day advanced economies which experienced a combined 90 crises between 1870 and 2016. Based on literature findings, this study features a vector of 12 leading indicators, encompassing real, banking and external sectors. In scrutinizing antecedents to financial crises, the relationships between these sectors are recurrently underscored in economic literature including Kindleberger (1978), González-Hermosillo et al. (1997), Hardy and Pazarbasioglu (1998), Kaminsky and Reinhart (1999), Reinhart and Rogoff (2009), Claessens et al. (2011). Real sector variables encompass gross domestic product per capita, consumption expenditure, fixed capital formation and capital output ratio, while banking sector indicators include total loans, debt, short-term and long-term interest rates, inflation and stock market, whereas external sector factors comprise exchange rates and current account balance.

Across four modelling dimensions, predictive strength of machine learning methods is assessed. These dimensions entail a contemporaneous pooled format with an expanding window, transformations with lag structure and a rolling window as well as in individual format. Across recursive out-of-sample performance, assessed measures include AUC, F_1 and Brier scores. Findings suggest that an expanding window with lag structure and rolling window generally improve on both the optimized contemporaneous and individual country formats. Six of the 13 algorithms reach highest accuracy through the lag structure. Top performing machine learning methods consistently produce high accuracy rates, on average above 70 percent for all derivates of the contemporaneous pooled format and frequently feature random forests and gradient boosting. Compared to a non-parametric baseline, all top models add accuracy value, above 20 percentage points for several countries. A measure of complexity underscores that

the models encountered a majority of complicated forecasting environments, holding both in contemporaneous pooled and individual formats.

In an analysis of important variables, fixed capital formation has the largest influence. GDP per capita and consumer inflation have risen in prominence over the last century, while debt-to-GDP, stock market and consumption expenditure had the highest influence at the turn of the 20th century. According to the lag structure, banking sector variables on average constitute 28 percent of the variation in crisis prevalence, real sector 64 percent and external sector 8 percent over the full period.

The structure of the paper is as follows. Section 2 provides an overview of the empirical literature. Section 3 describes the machine learning methodology. Section 4 highlights the data and variable selection and section 5 evaluates the findings. Section 6 concludes.

2. Empirical Literature

In recent years, the adoption of machine learning methods has proliferated given its processing ability to analyse Big Data, and deal with non-linear interactions between variables, both vital to identify the most important indicators and account for different precursors to crises. Furthermore, the evolution and central aim in the development of machine learning algorithms are found predominantly in out-of-sample forecasting performance. Estimation of pivotal tipping points presents another key benefit. Widespread inclusion of numerous algorithms in statistical programmes further broadens the utilisation of innovative methods. Yet, one drawback involves the inability of algorithms to compute marginal contributions of each predictor or confidence intervals for threshold levels (Joy et al., 2017).

Advanced by Breiman et al. (1984), classification and regression trees (CART) represent a prevailing set of machine learning techniques to study financial crises. Using binary recursive trees for currency crises during the period 1987 and 1999, Ghosh and Ghosh (2003) identify macroeconomic imbalances, high debt-equity ratios of organisations and weak governing institutions as key contributory factors. Analysing balance of payment crises from 1994 to 2005, Chamon et al. (2007) underscore the significance of international reserves, current account balance, short-term external debt, reserve cover, external indebtedness, and gross domestic product. Examining sovereign debt crises of emerging markets between 1970 and 2002, Manasse and Roubini (2009) highlight liquidity, solvency and macroeconomic imbalances, subsequently corroborated in an analogous investigation by Savona and Vezzoli

(2012), that further reveals the effects of contagion as key indicators. A shortfall of the CART approach is an intrinsic insensitivity to cross-sectional and time series features (Joy et al. 2017).

Surveying banking crises across a large group of countries during the period 1979 to 2003, Davis and Karim (2008), find domestic credit growth as most important predictor. Dattagupta and Cashin (2011) study crises in emerging markets between 1990 and 2005 and reveal the relevance of elevated inflation, severe currency depreciation and lackluster bank profitability. Spanning 20 countries in Asia and Latin America, Davis et al. (2011) compare the CART approach to a logistic regression. While varying by region, early warning predictors for Asian countries include national budget deficit and low domestic growth, and for Latin American countries involve currency depreciation and banking credit. Expanding on the generalized CART methodology in studying episodes of systemic risk, Alessi and Detken (2018) highlight random forests to be reliable in its identification of leading signals. Employing CART and random forests to scrutinise 36 advanced countries during the period 1970 to 2010, Joy at al. (2017) find tight interest rate spreads and inverted yield curves are leading predictors in the short-term, with house prices significant over the long term. Across a horse race involving nine forecasting models on 27 EU countries, Alessi et al. (2015) underscore high predictive strength of CART and random forests in comparison to probit and logit models, a signals and Bayesian model averaging approach. CART reveal a narrow yield curve, elevated money market rates and low bank profitability as precursors, yet for random forests, house price valuation constitutes the most significant factor, across short and long prediction horisons. Bank credit, government debt, long term yield and frail macroeconomic variables also serve as early warning signals. In another extension of the CART methodology, Casabianca et al. (2019) find adaptive boosting to outperform a logistic regression in forecasting financial crises between 1970 and 2017. Du Plessis (2022) highlight gradient boosting outperforming multiple outcome models including several machine learning models in crisis predictions. Fouliard et al. (2019) show decision-trees to outperform a regression model between 1985 and 2018, while Beutel et al. (2018) observe an opposite result. Ward (2017) and Bluwstein et al. (2020) make use of machine learning models to predict financial crises using the Macrohistory Database, underscoring improved out-of-sample performance.

3. Empirical Methods

This paper develops 13 machine learning models, all classified under the domain of supervised learning as it involves scrutinising a function that is mapping inputs to outputs based on a

training dataset. According to this process, algorithms search for crisis signals, informed by threshold values and rules that increase the likelihood of an event. As a result, machine learning fits as a combination of a non-parametric and parametric approach. The models in this paper include a non-parametric technique, regression algorithms, instance-based, regularisation and dimensionality reduction procedures, as well as decision tree methods and ensemble algorithms. Forecasting efficacy of these machine learning models is assessed through its performance on a test dataset of various dimensions. Descriptions of the methodologies and hyperparameter implementations feature in Section A (Appendix).

3.1 Benchmark Algorithms

To allow testing across a broad spectrum of models, simple to more complex algorithms are developed and employed. Serving as benchmark, a non-parametric model includes a baseline approach in the form of a conditional mean estimation whereas regression algorithms constitute a linear probability model and probit model. These modelling techniques are frequently utilised in forecasting financial distress. Formulations of the methods are discussed in Section A (Appendix).

3.1 Instance-Based Algorithms

Instance-based algorithms comprise k-nearest neighbours and support vector machine. Advanced by Cover and Hart (1967), the k-nearest neighbours (k-NN) algorithm involves the estimation of the conditional distribution of Y given X in order to categorise an observation according to the outcome class with highest estimated probability. Developed by Boser, Guyon and Vapnik (1992), the support vector machine (SVM) improves on the constraint of linear classifiers by accommodating non-linear relationships including quadratic and cubic terms. Achieved by employing kernels to enlarge the feature space of the predictors, the technique further improves computational efficiency as it does not explicitly execute in the enlarged feature space, but implicitly through its internal products of observations.

3.2 Regularisation Algorithms

Through a regularising procedure, coefficients of less relevant predictors shrink towards zero. Two algorithms feature in this paper, namely ridge and lasso.

The ridge procedure was developed and extended by Tikhonov (1943, 1963), Foster (1961), Phillips (1962) and Hoerl (1962). In contrast to the ordinary least squares statistical

technique which computes $\beta_0,\ \beta_1, \ldots, \beta_p$ by employing values which minimizes the residual sum of squared equation, ridge applies tuning parameter $\lambda \geq 0$, where the shrinkage penalty is small when $\beta_1, \ldots, \beta_p$ are near zero, so it reduces the estimates of β_j towards zero.

Analogous to ridge, lasso reduces the estimated coefficients of explanatory variables towards zero, but the penalty component forces some of the coefficients exactly to zero when the tuning parameter λ is adequately large. Through this procedure, lasso operates a variable selection technique and enhances interpretability of the model output, eventually also ensuring sparse models, which is an advantage in addressing variable correlation in the model. Furthermore, cross-validation is likewise integrated to estimate the optimal level of λ (James at al., 2013).

3.3 Dimensionality Reduction Algorithms

Introduced by Wold (1985), partial least squares regression (PLS) serves as dimension reduction method by detecting a new set of features $Z_1, \ldots, Z_m$ which are linear combinations of the initial features, and subsequently fitting a linear model using least squares. As PLS identify new features which approximate the original features that are associated with the outcome variable Y, it explicates the outcome and explanatory indicators (James at al. 2013).

3.4 Decision Tree Algorithms

3.4.1 Full Tree

Based on the seminal work of Breiman et al. (1984), the implementation of classification and regression trees (CART) accentuate several advantages. By following a semi-parametric framework, CART is not constrained by a predetermined functional form and can process various dimensions of data. Moreover, the method is suited to handle large and heterogenous datasets such as Big Data and can accommodate numerous predictors and operate with missing values. CART allow non-linear relationships, implement threshold levels and support interactions between variables. Resultantly, relationships between predictors could fluctuate given cross-sectional and time dimensions. By analysing all data observations, specification errors are minimized. Relevant to crisis literature, CART rank predictors according to their level of importance, thereby rendering leading indicators. Indeed, classification and regression trees are straightforward to interpret and a practical instrument for policymakers.

However, classification and regression trees encounter some limitations. As classification trees are predisposed to overfitting it could impact the accuracy of out-of-sample forecasts. Yet, it can be addressed through pruning, a technique that reduces branches of trees. Accordingly, during each iteration the model condenses the amount of data analysed from the full sample, which results in a local rather than global optimal. In comparison to regression models, as probability distributions are not operationalised, confidence intervals cannot be computed. Given that an individual probability value is allocated to all observations within a categorised set, marginal contributions of the explanatory variables are not estimated, even though the variation in probability of surpassing threshold levels is computed at each node. Finally, the ranking of variables could result in essential predictors being excluded from the final tree (Joy et al., 2017).

CART implement a top-down approach to partition data recursively, involving several predictors. Originally, through a partition with one predictor, a parent node is formed. Subsequently, divided into two homogenous child nodes, which are based on the discrete outcomes of the dependent variable, in this instance a systemic financial crisis or no-crisis. For every division, the algorithm chooses an optimal threshold value of the predictor. Child nodes are continually divided through this procedure until reaching a terminal node, which signify the final partitioning of data. This process can graphically be plotted as a decision tree. A forecasting model is computed as based on the decision path of each terminal node. Resultantly, this method analyses several divisions of predictors and selects those splits which best classify crisis and no-crisis episodes.

3.4.2 Pruned Tree

A shortfall of the full tree approach is the manifestation of over-fitting as all observations are considered. To lessen misclassification, pruning is employed as a general enhancement to the algorithmic framework. Centrally, pruning shrinks the size of a decision tree by transforming unreliable branch nodes into leaf nodes, and consequently by eliminating leaf nodes. Contextually, and according to the bias-variance trade-off, classification trees could fit the training data satisfactorily, yet become less accurate with new testing data.

3.5 Ensemble Algorithms

Ensemble algorithms operationalise a cohort of weak learners to jointly construct a strong learner, with the goal of improving on the performance of an individual forecast. This is

accomplished through a multi-classifiers approach, involving the training of multiple models using an identical algorithm. To lessen variance and bias, two prominent modelling frameworks comprise bagging and boosting. While both modelling approaches produce new data in the training environment through sampling by replacement, bagging assigns the same probability of replacements while boosting apportions weights, which thereby modifies replacement probabilities. In contrast, trees are formed independently and in parallel within the bagging process, but sequentially for boosting, the latter in order to enhance error rates by penalising misclassified observations or through shrinking a loss function. Strong learners are determined using simple average across every prediction tree for bagging, while in comparison, in the case of boosting, the weighted average is slanted towards better learners or inclusion of learning rates (Brownlee, 2016; James et al., 2013).

In this paper, two boosting algorithms are employed, namely adaptive boosting and gradient boosting, while the bagging algorithm is random forests.

3.5.1 Adaptive Boosting

Developed by Freund and Schapire (1997), adaptive boosting or adaboost represent one of the initial boosting algorithms. Distinctively, while classification and regression trees construct full trees on all observations, adaboost only builds stumps or weak learners. The error value obtained from one stump affects thereafter how the following stump is assembled based on a bootstrap sampling with replacement procedure. Each stump is also assigned a weight given its computed prediction error, which further denotes its contribution to the strong learner.

3.5.2 Gradient Boosting

As an extension of the adaboost approach, Friedman (2001) devises a boosting variation by employing a gradient descent procedure for regression and classification trees through a stepwise technique which solves for a loss function. Through this process, pseudo residuals are estimated to optimise every weak or base learner in a consecutive manner. The quantity of weak learners can be stipulated in the context of the bias-variance trade-off, with the aim of identifying the optimal quantum. Increasing the number of weak learners would lessen the bias as the model tracks the training data narrowly, but variance surges in the context of a noise factor, leading to reduced forecasting accuracy when new data is presented. Selecting fewer weak learners could result in higher bias, but a reduced probability of overfitting. A shrinkage parameter governs the learning rate of a weak learner, where a smaller value necessitates more

iterations to optimise and develop the final model (James et al., 2013).

3.5.3 Random Forests

Advanced by Breiman, (1984, 2001), random forests (RF) employ a group of weak learners to jointly create a strong learner, a process centred on bootstrapping and aggregation to enhance stability and accuracy. Executed in conjunction with a bagging procedure, a large quantum of regression trees is created through bootstrapped samples with replacements, obtained from the initial training sample. Nodes of trees are created based on a random selection of explanatory variables as well as the most optimal split amongst the predictors. Given that each tree renders a prediction, these predictions are averaged to calculate the final prediction.

A benefit of employing a large quantity of trees created from independent bootstrapped samples, comprises diminishing variance without increasing bias (Nyman and Ormerod, 2016). RF address the overfitting phenomenon of classification and regression trees by not processing all explanatory variables simultaneously, but by opting for the most important variables through majority votes and further only integrating the selected variables into the algorithm (Breiman et al., 1984). In contrast to individual trees, variable importance classifications of RF are more robust (Joy et al., 2017). Analogous to classification and regression trees, RF can process sizeable datasets, are not sensitive to outliers, model interactions between explanatory variables and are not limited by distributional assumptions.

Random forests algorithm permits optimisation through stipulation of tree complexity or depth, quantity of variables featuring in each tree, bootstrap sample size and the quantum of trees (Mullainathan and Spiess, 2017). Drawbacks of the approach comprise an inability to backwardly deduce interaction effects between variables due to the simple average procedure employed across a large number of decision trees (Joy et al., 2017), and a somewhat opaque framework given an algorithmic process executing across a multiplicity of bootstrap samples. Robust in-sample performance is intermittently not repeated with the addition of unseen observations (Alessi et al., 2015).

4. Data and Variable Selection

4.1 Data Composition

The classification and dating of systemic financial crises are centered on interpretation and judgement. This paper utilises the definition from Laeven and Valencia (2012), which describe a systemic financial crisis as a situation in which there are significant signs of financial sector

distress and losses in wide parts of the financial system that result in widespread insolvencies or significant policy interventions. In contrast to isolated banking failures, such as Herstatt Bank in Germany in 1974 or the termination of Baring Brothers in the United Kingdom in 1995, to be included as part of the definition, financial distress needs to be system-wide for instance the crises of 1890s, 1930s, Japanese banking crises in the 1990s and during the Global Financial Crisis. Dates on systemic financial crises are based on Jordà, Schularick, and Taylor (2013, 2017), which feature historical series from Bordo et al. (2001) and Reinhart and Rogoff (2009) for the period 1870 to 1970, and post-1970 from Laeven and Valencia (2008, 2012). Table 1 chronicles the systemic financial crises experienced by the countries in this study.

Table 1: Systemic Financial Crisis Dates by Country

Country	Crisis Dates
Australia:	1893, 1989
Belgium:	1870, 1885, 1925, 1931, 1934, 1939, 2008
Canada:	1907
Denmark:	1877, 1885, 1908, 1921, 1931, 1987, 2008
Finland:	1877, 1900, 1921, 1931, 1991
France:	1882, 1889, 1930, 2008
Germany:	1873, 1891, 1901, 1907, 1931, 2008
Italy:	1873, 1887, 1893, 1907, 1921, 1930, 1935, 1990, 2008
Japan:	1871, 1890, 1907, 1920, 1927, 1997
Netherlands:	1893, 1907, 1921, 1939, 2008
Norway:	1899, 1922, 1931, 1988
Portugal:	1890, 1920, 1923, 1931, 2008
Spain:	1883, 1890, 1913, 1920, 1924, 1931, 1977, 2008
Sweden:	1878, 1907, 1922, 1931, 1991, 2008
Switzerland:	1870, 1910, 1931, 1991, 2008
United Kingdom:	1890, 1974, 1991, 2007
United States:	1873, 1893, 1907, 1929, 1984, 2007

Each instance of systemic financial crisis is represented by a categorical variable, expressed by $Y_i = 0$ for a no-crisis episode and $Y_i = 1$ as proxy for a crisis event. While countries are selected for this study based on a key requirement to have experience with at least one systemic financial crisis, the preponderance of crisis episodes remains limited, with only 3.6 percent of all observations classified as $Y_i = 1$. Given that machine learning models represent novel approaches to deal with financial crises, the low prevalence of crisis episodes can be expected to be a constraint for some models to optimally function. Whereas the commonly used models might perform different in a setting with a higher proportion of each outcome of the categorical response variable, through the horse race of algorithms, fit for purpose models are expected to stand out.

Literature studies on financial crises underscore a solid relationship between macroeconomic factors and financial sector distress (Abiad, 2003; Berg et al., 2005; Claessens et al., 2011; Hardy and Pazarbasioglu, 1998; Vlaar, 2000). Specifically, Gonzalez-Hermosillo et al. (1997) find that banking sector factors reveal the probability of a bank failure, while real sector indicators impact its timing. Accordingly, for this study three classes of predictors are assessed, encompassing real, banking and external sectors.

Real sector indicators underscore the degree of efficient credit utilization in the economy and emphasise the ability of borrowers to settle their debt obligations. Particularly, this study assesses real gross domestic product per capita, real consumption expenditure, real fixed capital formation and capital output ratio. Gross domestic product per capita serves as a valuation of collective economic activity, which in conjunction with consumption and investment elicit credit demand. Capital output ratio functions as proxy for efficient use of investments. A severe credit boom as a result of unsustainable over-investment and consumption expenditure could portend an ensuing real sector slowdown. In turn, subdued gross domestic product per capita impacting on employment, aggregate output and income growth further encumbers the ability of household and corporate borrowers to repay outstanding debt. In this context, consumer spending represents a measure of economic health. Hardy and Pazarbasioglu (1998) find that banking distress is associated with a concurrent reduction in real gross domestic product growth and a drop in the capital output ratio.

Banking sector indicators comprise banking performance and inherent confidence, and include knowledge on total loans, debt-to-GDP, inflation, short-term and long-term interest rates and stock market levels. According to Reinhart and Rogoff (2009), credit booms and asset bubbles have frequently resulted in financial sector distress. While accelerating banking credit growth portends an ensuing lending boom with unsustainable debt levels, sharp fluctuations in stock market asset values could consolidate a loss of confidence and lead to further asset price deterioration. Consumer inflation and interest rates feature as shock variables affecting debt repayment and liability growth. Demirguc-Kunt and Detragiache (1998) highlight that higher interest rates and consumer inflation increase the probability of a crisis. In the context of diminishing income growth, rising inflation and interest rates hinder the repayment ability of debtors.

External sector indicators gauge regional spillovers and global contagion through the US dollar exchange rate and current account balance. A steep currency depreciation following reversals in capital flows could result in a slump in asset values and surge in the cost of imported goods which restrains the ability of borrowers to meet their periodic debt obligations.

Kaminsky and Reinhart (1999) point out that declining terms of trade is an antecedent to banking crises. A weakening in the current account balance results in a comparatively higher outflow of working capital.

Explanatory indicators feature in Table 2 (Appendix). Data are obtained from Jordà-Schularick-Taylor Macrohistory Database (Jordà et al., 2017) and consist of annual time series. Another consideration includes experience of a previous systemic financial crisis. The final sample spans the period 1870-2016 and consists of 17 advanced economies, which collectively experienced 90 systemic financial crises over a combined 2,499 years and with 12 parameters constitute 29,988 observations. A representative sample of countries stem from North America, Australasia and Europe. According to Table 3 (Appendix), the mean and median quantum of crises experienced by the countries amount to five, with Canada on one and Italy on nine.

Figure 1 illustrates the share of countries in crisis over the past 147 years. A higher crisis frequency is observed from 1870 to the Second World War and, which resumed in 1974 following the great moderation. In particular, the crises of 1907/8, 1929-31 and 2007/2008 were more ubiquitous and global in nature, impacting more than 50 percent of the sampled countries. The Global Financial Crisis had the largest scale, comprising 70 percent of all the countries.

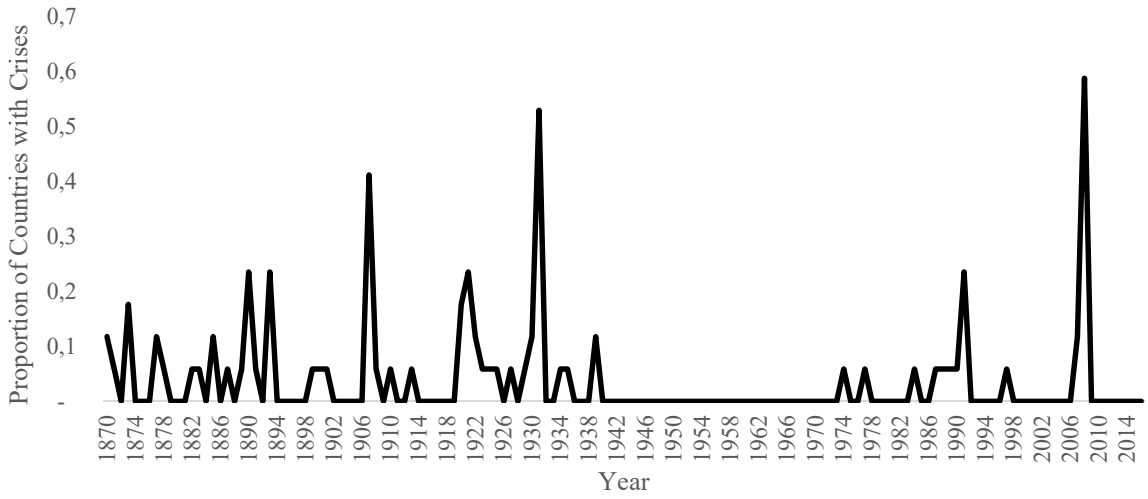

Figure 1: Proportion of Countries with Crises

To counter stationarity, ratios, first difference and log forms are employed, with lag based on statical significance, while real transformations confine the influence of inflation. Unit root tests produce satisfactory results as described in Table 4 (Appendix).

4.2 Significance of Individual Variables

The sample means for the three sets of indicators, encompassing real, banking and external

sectors are described in Table 5 (Appendix), and include a two-tailed t-test with significance levels.

Real sector indicators highlight a differentiated economic environment during a systemic financial crisis. Real gross domestic product per capita drops during a crisis. Similarly, real consumption expenditure and real investment are higher absent a crisis, with the latter turning negative during bouts of financial instability. Capital output ratio, as proxy for efficient use of investment capital, could be construed as reflecting diminishing returns in the build-up to a crisis due to an overinvestment boom, while the lower asset valuations during a catastrophe present higher forward-looking return rates for long-term investment projects.

Banking sector indicators accentuate banking performance. Debt, as ratio to gross domestic product drops sharply during a crisis as liquidity constraints, more stringent credit appetite and lower demand weigh on credit extensions, exemplified by a shrinkage in total loans. The lower revaluation of assets is reverberated by the decline in the stock market. Consumer inflation lowers during a crisis period due to a reduction in aggregate demand for goods and services. While real short-term interest rates increase during a crisis, partly due to lower inflation and also as a result of the higher cost to obtain and access credit, long-term rates also inch up, due to a risk-on environment, albeit more stable given its forward-looking characteristics. The more recent and post-crisis applications of quantitative easing would be picked up by non-crisis periods in the subsequent years.

External sector indicators underscore the spillover between trade partners. Real exchange rates depreciate in the wake of systemic events as capital flows follow safer havens. Current account weakens in response to more expensive imports and the impact of lower aggregate demand.

Results from the two-tail t-test show that all but two variables are significant, which accentuate a discernable environment between crisis and no-crisis periods. The null hypothesis of similarity between crisis and tranquil observations can be rejected for all individual real sector variables. For the banking sector, short-term rates are significantly different at a 99 percent confident level, consumer inflation, long-term rates and debt-to-GDP at 95 precent and total loans at 90 percent confidence levels. In the case of the external sector, current account is dissimilar at 95 percent confident levels.

While the yield curve features as harbinger of recessions (Benzoni et al., 2018) and recently also modelled in financial crisis literature (Alessi et al., 2015; Joy at al., 2017; Bluwstein et al., 2020), the inclusion of this factor has not resulted in improved forecasting performance, likely given its covariance with other variables such as short- and long-term rates,

as well as the smaller impact of interest rates compared to real and other banking sector variables and therefore do not appear in this study.

5. Empirical Results

Serving as new methods to study an old problem, a total of 13 machine learning algorithms are developed to model 147 years of systemic financial crises. Model fit and forecasts are assessed across four dataset dimensions. As main modelling dimension, and aimed at providing an immediate early warning signal, a standard one period lag structure is employed for all variables in the pooled format. Given the low prevalence of $Y = 1$, the machine learning algorithms are modeled across the pooled dataset which encompasses all the countries in this study. As all algorithms are classified as supervised methods, benefits include a larger sample size, more variance in the predictors, higher degrees of freedom with more crisis episodes, collective and faster algorithmic learning, and a practical approach to assess financial catastrophes given global interlinkages. Formally, and with variables captured in Table 2 (Appendix), this can be stated as

$$Y_{i,t} = c_i + \sum_{i=1}^{N}\sum_{t=1}^{T}\beta_1(GDP)_{i,t-1} + \beta_2(CE)_{i,t-1} + \beta_3(FCF)_{i,t-1} + \beta_4(COR)_{i,t-1} + \beta_5(DEBT)_{i,t-1} + \beta_6(LOANS)_{i,t-1} + \beta_7(STOCK)_{i,t-1} + \beta_8(CPI)_{i,t-1} + \beta_9(SR)_{i,t-1} + \beta_{10}(LR)_{i,t-1} + \beta_{11}(ER)_{i,t-1} + \beta_{12}(CA)_{i,t-1} + \varepsilon_{i,t},$$

where $Y_{i,t}$ is the crisis index, N the number of countries, T the full time period and $\varepsilon_{i,t}$ stochastic error term. Benefits would include faster response times as the release of annual data frequently follows after the commencement of a crisis in the same or previous year.

To verify whether a lag structure delivers the highest predictive strength, the second modelling dimension applies a contemporary structure and with optimised statistical properties as described in Table 2 (Appendix). Mathematically denoted as $Y_{i,t} = c_i + \sum_{i=1}^{N}\sum_{t=1}^{T}\sum_{j=1}^{K}\beta_j x_{j,i,t-l} + \varepsilon_{i,t}$, with x_j the jth explanatory variable given j = 1,…,K, and l the number of lags.

Thirdly, all the machine learning algorithms are modeled independently for each individual country, by using the optimised contemporary structure employed by the second modelling dimension. Technically described as $Y_{i,t} = c_i + \sum_{t=1}^{T}\sum_{j=1}^{K}\beta_j x_{j,i,t-l} + \varepsilon_{i,t}$, where i comprises the specific country. This allows a direct comparison between country-level forecasts based on individual crisis experience and communal experience from the second modeling framework. While it comes at a trade-off of a smaller sample size, advantages include

a study on heterogeneous method responses where individual country models are aimed at detecting idiosyncratic characteristics and nuances.

Fourthly, given the long-term nature of the data series, where structural breaks could occur or level of economic development are not comparable after several decades, a rolling window of 20 years is employed to assess forecasting performance. Also based on the optimised contemporary framework, this dimension can be formulated as $Y_{i,t}(w) = c_i(w) + \sum_{i=1}^{N}\sum_{t=w}^{T}\sum_{j=1}^{K}\beta_j(w)x_{j,i,t-l}(w) + \varepsilon_{i,t}(w)$, where w is a fixed window with 20 observations and $t = w, w+1,.., T$ with $T - w + 1$ the number of subsamples. Essentially, the aim of these four approaches is to verify whether long-term pooled, disaggregated, lag or period-bound datasets are more conducive to model accuracy, in the context of an inherent bias-variance machine learning trade-off and as measured by the error function and confusion matrix.

The performance of these novel methods is evaluated across recursive out-of-sample predictions, by adding one datapoint to the training set for each new iteration and forecasting one year ahead, until the end of the sample. Formally, $Y_{i,t+h} = c_i + \beta_j x_{j,i,t+h} + \varepsilon_{i,t+h}$, where h denotes the h-step ahead forecasting horizon, with $1 \leq h \leq T$, and $x_{j,i,t+h}$ a vector of regressors with time-varying parameters. An expanding window is used in contemporary and lag format as well as for individual countries. While the rolling window retains a consistent 20-year range, it likewise updates iteratively by adding one new year while simultaneously dropping the year furthest back. Performance of individual countries is modelled separately and reported in both individual and aggregated format for comparability. The starting date for all model forecasts is based on available degrees of freedom.

Serving as regression algorithm, a probit model, which is widely employed by policymakers to assess the likelihood of an adverse event occurring, also constitutes a valuable alternative to evaluate forecasting performance compared to more recently developed algorithmic frameworks. Coefficients and statistical significance for both the lag structure and optimised contemporaneous pooled probit models are described in Table 6 (Appendix). In the case of the former, half of all variables are significant whereas with the optimal model, with the exception of the US dollar exchange rate, all variables are significant.

Performance assessment criteria includes area under receiver operating characteristics curves (AUROC), F_1 measures and Brier scores. While receiver operator characteristics (ROC) constitute a visual representation of the true positive rate by false positive rate or $1 -$ specificity, area under curve (AUC) summarises the outcome into a single value. True positive rate (TPR) is also referred to as sensitivity or recall and comprise the ratio of correct predictions

(TP) to the summation of correct predictions and false negatives (FN) or type II errors, where $TPR = \frac{TP}{TP+FN}$. False negative is the incorrect acceptance of a false hypothesis. In an environment where the subject under study has a low prevalence as is the case with financial crises, AUC is shown to exhibit higher stability given its insensitivity to outcome imbalances. AUC scores range from 0 to 1, where the latter signifies a correct set of forecasts. (DeLong et al., 1988; Fawcett, 2006).

In comparison, F_1 score represents another measure of a model's accuracy for a given forecast. F_1 scores are a weighted average of recall and precision, the latter the ratio of true positives to the combined true and false positives rates. False positive rates (FPR) consist of false alarms (FP) as a ratio to the collective false alarms and true negatives (TN), denoted mathematically as $FPR = \frac{FP}{FP+TN}$. F_1 as a measure thereby takes into account both false positives and false negatives or type I and type II errors (Chinchor, 1992; Van Rijsbergen, 1979). F_1 score can formally be denoted as $F_1 = \frac{TP}{TP+\frac{1}{2}(FP+FN)}$, where a higher F_1 score highlights a more accurate forecast, with F_1=1 showing a perfect forecast. Predictions without true positive values would revert to F_1=0.

The Brier score in contrast is akin to a cost function, which measures the mean squared difference between the predicted probability and the actual outcome (Brier, 1950). Formally it can be stated as $Brier\ score = \frac{1}{N}\sum_{i=1}^{N}(f_t - a_t)^2$, where f is the forecasted value, a the actual outcome and N the number of forecasts. Brier scores also range from 0 to unity, with the inversion applicable in that a lower score is indicative of a lower error and thereby a higher accuracy.

5.1 Recursive Out-of-Sample Crisis Forecasts with Lag Structure

The main modelling dimension encapsulates all the data in a pooled format, with a key configuration in the lag structure of the predictors. Given the annual time series, and with the purpose of predicting an ensuing crisis at shortest lead time, all predictors are transformed using one lag. Through variable importance techniques, leading indicators are uncovered across nearly a century and a half, simultaneously providing insights into the workings of the machine learning models and serving as input into the policy making process to prevent and mitigate ensuing financial crises. Prediction strength for all the algorithms is assessed thorough recursive out-of-sample forecasts.

5.1.1 Variable Importance

A benefit of machine learning methods entails the identification of the most important explanatory indicators. This is achieved by analysing the prevalence of each variable used by the algorithm to make key decisions. When the selection of a variable at a split node results in better performance of the error function, the higher its relative importance becomes. A Gini index is employed to measure performance, based on a reduction in the sum of squared errors each time a variable is selected to split a tree or node (Brownlee, 2016). Based on the gradient boosting algorithm, which frequently outperform amongst machine learning methods in horse-race events (Nevasalmi, 2020), and selected for its classification and regression abilities, Figure 2 denotes all the predictors across the full period, while Figure 3 (Appendix) shows predictors individually on a scale of 50, and Figure 4 (Appendix) by sector. As robustness test, random forests variable importance, as denoted in Figure 5 (Appendix) employs two further measures. The first is the permutation measure, formally denoted as $VI(X_j) = \frac{\sum_{t \epsilon B} VI^{(t)}(X_j)}{ntree}$, where the importance measure for indicator X_j is estimated as the summation of the importance scores across all trees. Expressed as a percentage increase in mean squared error, it entails applying permutations to each individual variable to assess the resultant impact on the overall accuracy of predictions. Where a variable consists of random noise, permutations should not affect accuracy. Second is the increase in node purity. Mathematically describe as $VI(X_m) = \frac{1}{ntree}\sum_T \sum_{t \in T: v(s_t)=X_m} p(t)\Delta i(s_t, t)$, variable importance is based on the mean value determined across all trees T and all nodes t, where $p(t)$ shows the number of samples reaching node t and $v(s_t)$ signifies the variable utilised to split node t. This measure is analogous to the Gini index employed by gradient boosting, where a reduction in the sum of the squared error from the utilisation of a variable to split a node, results in a higher importance allocated to the associated variable (Breiman, 2001; Hjerpe, 2016). From an interpretation perspective, the scale is less relevant whereas relative values are indicative of inter-variable importance. A drawback of the random forests variable importance approach revolves around a higher influence of continuous and multiple outcome variables on importance measures (Strobl et al., 2007).

According to the findings for the panel with lag structure, fixed capital formation exerts the single most influence, from around 20 percent at the turn of the 20th century, spiking to 50 percent the year before the 1907 banking crisis, followed by a gradual increase over the subsequent decades, reaching above 40 percent in the 2010s. The second most influential

variable is gross domestic product per capita, which provides an overall gauge of economic activity adjusted for the size of the population, and that grew from low single digits in the 1880s to 15 percent at the end of the sample, hovering around 10 percent for most of the period under review. While these variables stood out, significant fluctuations in the levels of other variables are observed during specific developmental epochs. For instance, debt-to-GDP spiked above a 35 percent level of influence around the banking crisis in the 1890s while the stock market remained above 20 percent in the years leading up to the 1907 crisis. Consumption expenditure peaked at 20 percent around the first two decades of the 1900s. These findings are consistent with research on the role of fixed capital formation booms, vigorous consumption spending and escalating debt growth on the formation of financial crisis (Kindleberger, 1978, Reinhart and Rogoff, 2009), with the stock market instrumental as an indicator of existing vulnerabilities. Serving as a major leading indicator for most of the 1900s, inflation has been a pivotal indicator since the years before the Great Depression as cost-push pressures exert more influence on the repayment ability of debtors, while exchange rates peaked around both world wars.

Findings from the random forests robustness test broadly confirms the leading indicators. Fixed capital formation takes poll position in reducing the mean squared error and sum of squared and contributing to higher overall accuracy. Capital output ratio features in the top three most influential variables across both measures, while inflation is highlighted as having the second highest Gini index. Furthermore, total loans and short-term rates are also classified as important variables, while the inclusion of total loans and the current account increases overall accuracy.

On average, banking sector variables constitute 28 percent of the variation in crisis prevalence, real sector 64 percent and external sector 8 percent. After peaking around a 65 percent level of importance in the 1880s, banking sector predictors declined in prominence until the 1910s and drifted upwards above 30 percent in the lead up to the Great Depression, after which it fluctuated within a 20-30 percent band until the start of the 21st century. The real sector demonstrates an inverse trajectory, gradually increasing from around a 30 percent level of importance in the 1880s to over 60 percent in the years before the start of the Great Depression in 1929. During the subsequent eight decades, real sector variables remained on a high level of contribution to the underlying causes of financial crises, spiking to 70 percent at the start of the Global Financial Crisis. The lag structure of the panel model partly detects the real estate investment boom that contributed to the sub-prime crisis and eventually culminated in a fully-fledged financial crisis. Albeit more volatile, external sector influence increased

during the end of the 19th century and first two decades of the 20th century in tandem with the progression of globalisation, remaining in a narrow band during the subsequent decades, spiking again in the 1970s with the dissolution of the gold standard.

In a comparison applying a contemporaneous structure in Figure 4 (Appendix), banking sector influence increases to 50 percent with real sector dropping to 42 percent, and with external sector unchanged, emphasising the dynamic adjustments of leading indicators one year preceding a crisis compared to the year of a crisis.

The random forests variable importance measure for the lag structure further underscores a similar outcome as with the gradient boosting measure, with banking sector influence observed around 40 percent, real sector on 53 percent and external sector at 7 percent according to their contributions to overall model accuracy. The broadly comparable results between gradient boosting and random forests support a targeted mitigation approach from a policy making perspective.

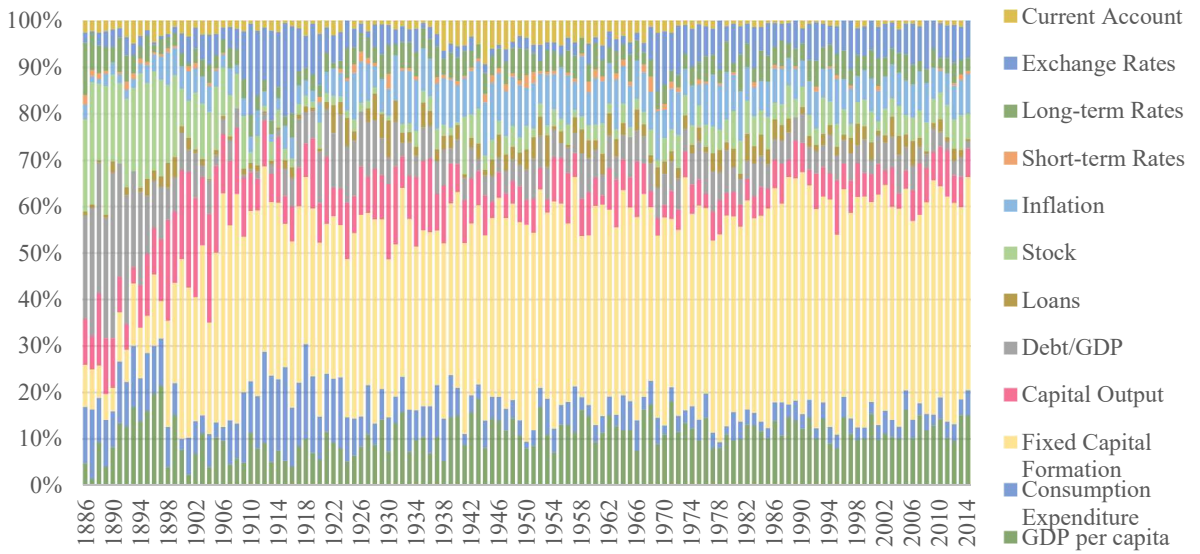

Figure 2: Variable Importance Over Time

Table 7 exhibits the recursive out-of-sample forecasts using AUC mean values, F_1 and Brier scores. The top performing methods using AUC are random forests, gradient boosting, probit regression, ridge, linear regression and adaptive boosting, all around the 70 percent level of accuracy. F_1 shows a comparable result, with pruned tree followed by gradient boosting and random forests. Brier scores are lowest for support vector machine, followed by full tree in reducing the mean squared error between actual and predicted values. Results are clustered within a 0.06 to 0.08 band for most algorithms. An overall ranking is estimated as a function of AUC + F1 Score – Brier Score, where the highest values are indicative of topmost predictive

strength, with random forests first, followed by gradient boosting and support vector machines. Overall AUC predictive accuracy across all algorithms reaches 64 percent for the lag structure.

Table 7: Recursive Out-of-Sample Forecasts with Lag Structure

Method	AUC	F_1 Score	Brier Score	Rank
Baseline	0.534 [0.474, 0.593] *0.030*	0.073	0.068	8
Linear Prediction	0.707 [0.655, 0.759] *0.026*	0.130	0.074	5
Probit Regression	0.732 [0.678, 0.787] *0.027*	0.073	0.840	11
K-Nearest Neighbours	0.498 [0.497, 0.499] *0.000*	0.000	1.004	13
Support Vector Machine	0.696 [0.640, 0.752] *0.028*	0.116	0.015	3
Ridge	0.707 [0.653, 0.761] *0.027*	0.137	0.071	4
Lasso	0.533 [0.473, 0.592] *0.030*	0.073	0.069	9
Partial Least Squares	0.500 [0.500, 0.500] *0.000*	0.000	1.000	12
Full Tree	0.617 [0.545, 0.690] *0.027*	0.136	0.063	7
Pruned Tree	0.605 [0.528, 0.682] *0.039*	0.181	0.064	6
Adaptive Boosting	0.697 [0.639, 0.756] *0.029*	0.071	0.500	10
Gradient Boosting	0.754 [0.702, 0.807] *0.026*	0.142	0.069	2
Random Forests	0.765 [0.719, 0.811] *0.023*	0.139	0.072	1

Variance of AUC is defined by DeLong et al. (1988) and estimated with algorithm specified by Sun and Xu (2014). AUC upper and lower bounds in squared brackets are based on 95 percent confidence intervals. Standard errors in italics.

5.2 Recursive Out-of-Sample Crisis Forecasts for All in Contemporaneous Format

As robustness tests, three more forecasting frameworks are developed. The second modelling dimension also entails combining all 17 countries in an optimised contemporaneous pooled format across the period under review. In contrast to the lag model, the aim is to verify if predictive accuracy improves without the benefit of enhanced lead time and instead through the use of contemporaneous indicators. Similar to the panel model with lag structure, this approach allows the machine learning methods to observe and learn from the experience of all countries and utilises a comprehensive dataset in the context of a low frequency event to build and calibrate each model in recursive manner to operationalise out-of-sample forecasting.

5.2.1 Contemporaneous pooled recursive out-of-sample results

Recursive out-of-sample results for all countries in optimised contemporaneous pooled format are summarised in Table 8. In terms of AUC, gradient boosting is the best performing model followed by random forests. Linear, ridge and probit models also perform above average. Assessing the F_1 scores, full tree is in first position, followed by gradient boosting and random forests. An analysis of Brier scores shows support vector machine with lowest error, followed by ridge. A comparison of the three measures shows that AUC correlates 67 percent of the time

with F_1 scores, with the latter showing a negative correlation with Brier score of 75 percent. Based on the combined ranking across all three measures, gradient boosting and random forests constitute the top two followed by ridge. In contrast to the lag structure dimension (overall 64 percent), the contemporaneous format a shown to register lower average results of 61 percent, emphasizing the forecasting benefits of detecting vulnerabilities with lead time.

Table 8: Recursive Out-of-Sample Forecasts in Contemporaneous Pooled Format

Method	AUC	F_1 Score	Brier Score	Rank
Baseline	0.564 [0.500, 0.627] *0.032*	0.068	0.076	7
Linear Prediction	0.687 [0.632, 0.742] *0.027*	0.108	0.084	5
Probit Regression	0.681 [0.615, 0.747] *0.033*	0.070	0.844	11
K-Nearest Neighbours	0.499 [0.498, 0.500] *0.000*	0.000	1.002	13
Support Vector Machine	0.637 [0.588, 0.687] *0.025*	0.098	0.020	4
Ridge	0.683 [0.629, 0.738] *0.027*	0.114	0.073	3
Lasso	0.528 [0.457, 0.598] *0.035*	0.082	0.076	8
Partial Least Squares	0.500 [0.500, 0.500] *0.000*	0.000	1.000	12
Full Tree	0.570 [0.496, 0.643] *0.037*	0.163	0.074	6
Pruned Tree	0.544 [0.486, 0.602] *0.029*	0.075	0.076	9
Adaptive Boosting	0.638 [0.577, 0.700] *0.031*	0.122	0.500	10
Gradient Boosting	0.750 [0.692, 0.808] *0.029*	0.137	0.081	1
Random Forests	0.696 [0.637, 0.755] *0.030*	0.126	0.084	2

Variance of AUC is defined by DeLong et al. (1988) and estimated with algorithm specified by Sun and Xu (2014). AUC upper and lower bounds in squared brackets are based on 95 percent confidence intervals. Standard errors in italics.

5.2.2 Individual country out-of-sample results

Individual country forecasts are implemented by taking the experience of other countries into account. The purpose is to authenticate if knowledge of rare events from other countries could improve forecasting performance of an individual country. While experience could be nuanced with unique predictors, findings from variable importance signify commonality across the full cohort of countries, which could underscore key learnings with broad-based applications. It also allows more variability in the predictors and increases the degrees of freedom. Mean AUC values are summarised for each country in Table 9 (Appendix).

Table 10 (Appendix) highlights the top performing model per country and the deviation to both baseline and across all models. Accuracy rates range from 60.3 percent in the case of Germany to 94.3 percent for Australia. Full tree, linear and probit regression each registers the highest accuracy rates across three countries, gradient boosting and random forests have the most correct predictions amongst two countries each and SVM, ridge and adaptive boosting each outperforms in one country.

The deviation between the top performing model and baseline confirms the value added by the best algorithm against a non-parametric benchmark, where a higher variance denotes a larger enhancement. Top models add value for all countries, and contribute above 20 percentage points for Finland, Italy, Australia, UK, Sweden and Switzerland.

Overall deviation serves to mark the variability of the models. A higher deviation would accentuate the complexity of modelling the underlying series for the specific country. Employing 10 percentage points as an arbitrary threshold, and assessing all models, a large degree of complexity was encountered for the majority of countries, with Australia and Netherlands at the top end of the spectrum.

5.3 Recursive Out-of-Sample Crisis Forecasts for Individual Countries

The third modelling dimension revolves around the individual experience of each country. In contrast to the pooled format with contemporaneous structure, models only take into account the knowledge of what transpired in a particular country, which ensures that idiosyncratic factors are ringfenced for the development of country specific models and in turn used for recursive forecasting. For comparability, results are reported in both individual country and aggregated format, the latter a combination of the former.

Table 11: Recursive Out-of-Sample Forecasts for Individual Countries in Aggregated Format

Method	AUC	F_1 Score	Brier Score	Rank
Baseline	0.501 [0.434, 0.567] *0.034*	0.073	0.08	7
Linear Prediction	0.602 [0.536, 0.668] *0.033*	0.092	0.22	8
Probit Regression	0.543 [0.474, 0.612] *0.035*	0.074	0.97	11
K-Nearest Neighbours	0.503 [0.490, 0.516] *0.006*	0.107	1.01	12
Support Vector Machine	0.559 [0.497, 0.621] *0.031*	0.082	0.04	3
Ridge	0.530 [0.481, 0.580] *0.025*	0.106	0.40	9
Lasso	0.582 [0.521, 0.643] *0.031*	0.075	0.09	4
Partial Least Squares	0.499 [0.497, 0.500] *0.000*	0.000	1.00	13
Full Tree	0.593 [0.535, 0.650] *0.029*	0.107	0.07	2
Pruned Tree	0.501 [0.432, 0.570] *0.035*	0.076	0.07	6
Adaptive Boosting	0.658 [0.597, 0.720] *0.031*	0.067	0.50	10
Gradient Boosting	0.647 [0.587, 0.708] *0.030*	0.094	0.10	1
Random Forests	0.537 [0.477, 0.597] *0.030*	0.076	0.10	5

Variance of AUC is defined by DeLong et al. (1988) and estimated with algorithm specified by Sun and Xu (2014). AUC upper and lower bounds in squared brackets are based on 95 percent confidence intervals. Standard errors in italics.

Shown in Table 11 and in aggregated format, the four best performing methods are adaptive and gradient boosting, linear regression, full tree and lasso, on average slightly under

or above 60 percent. In terms of F_1 scores, full tree, k-nearest neighbours, ridge and gradient boosting also reach high accuracy, whereas support vector machine, full and pruned tree reflect low Brier scores. Overall, the top three models are gradient boosting, full tree and support vector machine. In contrast to the contemporaneous pooled format, knowledge from other countries slightly improves the average aggregated outcomes to 61 percent from 56 percent. In comparison, the non-aggregated format displayed on an individual country-level by mean AUC in Table 12 (Appendix), highlights a narrowing in the deviation between the two approaches, at 61 percent to the 62 percent for the pooled format. However, when comparing the best performing models between the two approaches as shown in Table 13 (Appendix), the inverse transpires, yet at marginal levels with individual format on 81 percent to the pooled format on 80 percent. Across both formats, the top models therefore correctly predict at a high (80 percent) accuracy rate across the 147 years under investigation. In terms of the top performing models, adaptive boosting records the most accurate prediction across four countries, pruned tree across three countries, lasso, ridge, support vector machine and linear each for two countries with full tree and probit on one country each.

While the slightly lower deviation to baseline could be ascribed to less variability in the predictors, lasso in the case of Sweden and full tree for Germany added the most value. Although the complexity encountered is slightly less for the contemporaneous pooled format, with a difference of only one percentage point, the variability in predictors might result in models better equipped to handle more complex datasets. Similar to the contemporaneous pooled format, Australia is at the top of the list for complexity, but then followed by Canada. The lower prevalence of crises experienced by these two countries can be expected to contribute to the degree of complexity faced by the models.

5.4 Rolling Window Out-of-Sample Crisis Forecasts

As fourth modelling dimension, a new configuration is applied to the panel format. Instead of increasing the cumulative volume of the training set during each iterative procedure, a 20-year rolling window is employed. As the economic landscape evolves over time, and in the context of the extended historical series, comparability between contemporary events and occurrences that took place over a century ago might be limited, which could affect the forecasting performance when applied to a different epoch. Informed by the mid-point of the Kuznets infrastructural investment cycle, spanning 15-25 years (Black et al. 2012), and given the importance of fixed capital formation as leading indicator over the 147-year period, a

standardised 20-year period window is employed, executed on a rolling basis, through which the time-bound focus allows events to be modelled and forecasted around a comparable period.

Shown in Table 14, average mean values of 64 percent are comparable to the pooled format with a lag structure. Top performing methods as based on forecasted accuracy consist of random forests and gradient boosting, followed further down by probit and linear regressions and ridge. Random forests and gradient boosting generate high F_1scores with mid-tier Brier scores. Combined top models comprise gradient boosting, random forests and linear regression.

Table 14: Rolling Window Out-of-Sample Forecasts

Method	AUC	F_1 Score	Brier Score	Rank
Baseline	0.527 [0.467, 0.587] *0.032*	0.082	0.065	8
Linear Prediction	0.717 [0.652, 0.782] *0.033*	0.136	0.072	3
Probit Regression	0.731 [0.662, 0.800] *0.035*	0.066	0.869	11
K-Nearest Neighbours	0.499 [0.497, 0.500] *0.000*	0.000	1.003	13
Support Vector Machine	0.648 [0.590, 0.706] *0.029*	0.111	0.018	5
Ridge	0.696 [0.638, 0.755] *0.029*	0.116	0.084	6
Lasso	0.528 [0.469, 0.587] *0.030*	0.082	0.077	9
Partial Least Squares	0.500 [0.500, 0.500] *0.000*	0.000	1.000	12
Full Tree	0.660 [0.581, 0.739] *0.040*	0.160	0.068	4
Pruned Tree	0.536 [0.480, 0.592] *0.028*	0.087	0.069	7
Adaptive Boosting	0.665 [0.601, 0.729] *0.032*	0.066	0.500	10
Gradient Boosting	0.776 [0.717, 0.835] *0.030*	0.158	0.076	1
Random Forests	0.778 [0.726, 0.829] *0.026*	0.142	0.078	2

Variance of AUC is defined by DeLong et al. (1988) and estimated with algorithm specified by Sun and Xu (2014). AUC upper and lower bounds in squared brackets are based on 95 percent confidence intervals. Standard errors in italics.

5.5 Ranked Methods across Forecasting Models

Across the four modelling dimensions, from optimised contemporaneous pooled format to transformations with lag structure and a rolling window to the aggregation of individual countries, select machine learning methods performed at consistently high accuracy levels. These are inclusive of ensemble and decision tree algorithms as well as traditional regressions. The strength of the probit and linear regressions to perform above average is supported by studies on its comparative effectiveness such as Beutel et al. (2019). In several instances, further transformations improved on model performance as it becomes better equipped to model the underlying dataset and predict an ensuing crisis.

Summarised by highest AUC mean value for each method specific to the associated top dimension, Table 15 underscores the variability and improvements across the four dimensions. Accordingly, six of the 13 models reached highest predictive strength through the lag structure,

four through the standardised rolling window, two within the contemporaneous pooled format and one when employing the aggregated individual format. When applying this combination, average mean AUC values increase to 65 percent, with the top two algorithms featuring random forests and gradient boosting, both on 77 percent overall accuracy rates across 17 countries and 147 years. Notwithstanding, average AUC mean values increase above 80 percent for top individual country models, both in panel and aggregated format. Across all three measures, the top models are gradient boosting, random forests, and support vector machine. While both the lag structure and rolling window deliver 64 percent overall accuracy rates, the former encompass highest prediction strength for nearly half of the machine learning models. However, the two best performing models feature within a rolling window framework, underscoring the value of employing a diverse set of modelling tools for leaning against the wind to prevent cleaning up after the bust.

Table 15: Top Recursive Out-of-Sample Forecasts Across All Formats

Methods	Top Dimension	AUC	F_1 Score	Brier Score	Rank
Baseline	Pool	0.564	0.068	0.076	8
Linear Prediction	Window	0.717	0.136	0.072	4
Probit Regression	Lag	0.732	0.073	0.840	11
K-Nearest Neighbours	Individual	0.503	0.107	1.010	12
Support Vector Machine	Lag	0.696	0.116	0.015	3
Ridge	Lag	0.707	0.137	0.071	5
Lasso	Lag	0.533	0.073	0.069	9
Partial Least Squares	Pool	0.500	0.000	1.000	13
Full Tree	Window	0.660	0.160	0.068	6
Pruned Tree	Lag	0.605	0.181	0.064	7
Adaptive Boosting	Lag	0.697	0.071	0.500	10
Gradient Boosting	Window	0.776	0.158	0.076	1
Random Forests	Window	0.778	0.142	0.078	2

6. Conclusion

In developing new forecasting methods for an old problem, 13 machine learning algorithms are employed to study 147 years of systemic financial crises across 17 countries. The range of methods include a baseline model as non-parametric approach as well as linear and probit regressions to serve as common comparison. Instance-based algorithms comprise k-nearest neighbours, which categorises new observations according to their closest points in an existing dataset, and support vector machine that apply kernels to enlarge the feature space to allow for non-linear relationships. Regularisation algorithm ridge reduces less significant coefficients towards zero, while in the case of lasso, coefficient estimates equate to zero. Classification and

regression trees include full tree and pruned tree and accommodate non-linear relationships and allow interactions between variables. Partial least squares constitute a dimension reduction method that find new features which approximate the initial features and are related to the outcome variable. Ensemble algorithms operationalise a set of weak learners to communally build a strong learner, with the aim of improving the performance of an individual forecast. The algorithms span random forests which revolve around bagging as well as gradient boosting and adaptive boosting which make use of a boosting process.

This paper implements a set of 12 leading indicators, inclusive of real sector predictors such as gross domestic product per capita, consumption expenditure, fixed capital formation and capital output ratio, as well as banking sector predictors comprising debt, credit, stock market, inflation and interest rates, together with external sector predictors which consist of exchange rates and current account balance. A representative sample of countries across several regions are used.

Four modelling dimensions which encompass a contemporaneous pooled format, transformations with lag structure and a 20-year rolling window as well as in individual format are implemented to assess forecasting strength of machine learning methods. Recursive out-of-sample forecasting performance is assessed by means of AUC, F_1 and Brier scores. Findings highlight that an expanding window lag structure as well as rolling window increase overall accuracy rates in comparison to the contemporaneous pooled and individual format. Notwithstanding, some individual country forecasts improved on the pooled experience utilised for individual country level predictions. Random forests and gradient boosting are consistently top performing machine learning methods, both classifying 77 percent of forecasts correctly across 17 countries and 147 years. Traditional regression models probit and linear also perform above average at respectively 73 and 71 percent accuracy rates. All top models add accuracy value, reaching above 20 percentage points for several countries in comparison to a non-parametric baseline. A level of complexity is detected across the time series for most countries, the majority breaching an arbitrary 10 percentage points threshold level in pooled and individual formats.

In an analysis of leading indicators, fixed capital formation exhibits the largest influence, followed by GDP per capita according to gradient boosting and inflation by means of random forests variable importance measures. Debt-to GDP, stock market and consumption were highly influential at the turn of the 20th century, whereas inflation has increased in importance over the last several decades. On an average basis over the full period and using a

lag structure, banking sector variables constitute 28 percent of the variation in crisis prevalence, real sector 64 percent and external sector 8 percent.

The practicality of implementing machine learning algorithms, its ability to handle large datasets and deal with non-linear relationships allow policymakers a straightforward set of tools to study financial vulnerabilities with improved forecasting accuracy. Across a long history of systemic financial crises, machine learning methods represent novel methods that make a valued contribution to the literature on early warning crisis signals and emerging forecasting frameworks.

References

Abiad, A. 2003. "Early-Warning Systems: A Survey and a Regime-Switching Approach". *IMF Working Paper*, WP/03/32.

Alessi, L. and Detken, C. 2018. "Identifying Excessive Credit Growth and Leverage". *Journal of Financial Stability*, Vol 35, pp 215–225.

Alessi, L., Antunes, A., Babecký, J., Baltussen, S., Behn, M., Bonfim, D., Bush, O., Detken, C., Frost, J., Guimarães, R., Havránek, T., Joy, M., Kauko, K., Matějů, J., Monteiro, N., Neudorfer, B., Peltonen, T., Rodrigues, P.M.M., Rusnák, M., Schudel, W., Sigmund, M., Stremmel, H., Šmídková, K., Van Tilburg, R., Vašíček, B. and Žigraiová, D. 2015. "Comparing Different Early Warning Systems: Results from a Horse Race Competition among Members of the Macro-Prudential Research Network". *MPRA Paper*, No. 62194.

Benzoni, L., Chyruk, O. and Kelly, D. 2018. "Why does the yield curve slope predict recessions?", *Chicago Fed Letters*, Number 404, The Federal Reserve Bank of Chicago.

Berg, A., Borensztein, E. and Pattillo, Z. 2005. "Assessing Early Warning Systems: How Have They Worked in Practice?", *IMF Staff Papers*, Volume 52, Number 3.

Berge, T. J. and Jordà, O. 2011. "Evaluating the Classification of Economic Activity into Recessions and Expansions", *American Economic Journal: Macroeconomics* 3(2), 246–277.

Beutel, J., List, S. and Von Schweinitz, G. 2018. "An evaluation of early warning models for systemic banking crises: Does machine learning improve predictions?". *Deutsche Bundesbank Discussion Paper Series,* No. 48.

Black, J., Hashimzade, M. and Myles, G. 2012. *A Dictionary of Economics*, 4th edition. Oxford University Press: Oxford.

Bliss, C.I. 1934. "The Method of Probits". *Science*, Volume 79.

Bliss, C.I. 1935. "The Calculation of the Dosage-Mortality Curve (With an Appendix by Fisher, R.A.)", *Annals of Applied Biology*, Volume 22.

Bluwstein, K., Buckmann, M., Joseph, A., Kang, M., Kapadia, S. and Şimşek, Ö. 2020. "Credit growth, the yield curve and financial crisis prediction: evidence from a machine learning approach". Bank of England, *Staff Working Paper* No. 848.

Bordo, M.D., Eichengreen, B., Klingebiel, D. and Martinez-Peria, M.S. 2001. "Is the crisis problem growing more severe?" *Economic Policy,* Vol 16 (32), pp 53–83.

Boser, B.E., Guyon, I.M. and Vapnik, V.N. 1992. "A Training Algorithm for Optimal Margin Classifiers", *Proceedings of the Fifth Annual Workshop on Computational Learning Theory,* COLT.

Breiman, L., Friedman, J. H., Olshen, R.A. and Ston, C.J. 1984. *Classification and Regression Trees,* Wadsworth and Brooks/Cole, Monterey.

Breiman, L. 2001. “Random Forests”, *Machine Learning*, Volume 45(1).
Brier, G.W. 1950. “Verification of forecasts expressed in terms of probability”. *Monthly Weather Review*, Volume 78 (1).
Brownlee, J. 2016. *Master Machine Learning Algorithms: Discover How They Work and Implement Them from Scratch*. Australia: Machine Learning Mastery.
Casabianca, E.J., Catalano, M., Forni, L., Giarda, E. and Passeri, S. 2019. “An early warning system for banking crises: From regression-based analysis to machine learning techniques”, *Marco Fanno Working Papers* 235, Dipartimento di Scienze Economiche "Marco Fanno".
Chamon, M., Manase, P. and Prati, A. 2007. “Can We Predict the Next Capital Account Crisis?” *IMF Staff Papers,* Volume 54(2).
Chen, W., Mrkaic, M. and Nabar, M. 2019. “The Global Economic Recovery 10 Years After the 2008 Financial Crisis”. *IMF Working Paper*, WP/19/83. Washington. IMF.
Chinchor, N. 1992. *MUC-4 Evaluation Metrics, in Proc. of the Fourth Message Understanding Conference*, pp. 22–29.
Choi, I. 2001. “Unit Root Tests for Panel Data”. *Journal of International Money and Finance,* Volume 20, pp 249–272.
Claessens, S., Kose, M.A. and Terrones, M. 2011. “The Global Financial Crisis: How Similar? How Different? How Costly?” *Journal of Asian Economics*, Volume 21(3), pp 247-64.
Cover, T.M. and Hart, P.E. 1967. “Nearest Neighbor Pattern Classification”, *IEEE Transactions on Information Theory*, Volume 13(1).
Cramer, J. S. 2002. “The Origins of Logistic Regression”. *Technical Report,* No. 119, Tinbergen Institute.
Dattagupta, R. and Cashin, P. 2011. “Anatomy of Banking Crises in Developing and Emerging Market Countries.” *Journal of International Money and Finance*, No 30(2), pp 354–376.
Davis, E.P. and Karim, D. 2008. “Could Early Warning Systems Have Helped to Predict the Sub-Prime Crisis?” *National Institute Economic Review*, No 206.
Davis, E.P. and Karim, D. and Liadze, I. 2011. “Should Multivariate Early Warning Systems for Banking Crises Pool Across Regions?” *Review of World Economics,* No 147.
DeLong, E.R., DeLong, D.M. and Clarke-Pearson, D.L. 1988. “Comparing the Areas Under Two or More Correlated Receiver Operating Characteristic Curves: a Nonparametric Approach”. *Biometrics*, Volume 44, pp 837–845.
Demirguc-Kunt, A. and Detregiache, E. 1998. “The Determinants of Banking Crises in Developing and Developed Countries” *IMF Staff Papers,* Volume 45(1)*.* Washington: International Monetary Fund.
Döpke, J., Fritsche, U. and Pierdzioch, C. 2017. “Predicting Recessions with Boosted Regression Trees”, *International Journal of Forecasting*, Vol 33, pp 745-759.
Dueker, M. J. 2001. “Forecasting Qualitative Variables with Vector Autoregressions: A Qual VAR Model of U.S. Recessions”, Federal Reserve Bank of St. Louis*, Working Paper,* Number 012A.
Du Plessis, E. 2022. “Multinomial Modeling Methods: Predicting Four Decades of International Banking Crises”, Economic Systems, *forthcoming*. Available at https://doi.org/10.1016/j.ecosys.2022.100979.
Fawcett, T. 2006. “An introduction to ROC analysis”. *Pattern Recognition Letters*, Volume 27, Issue 8, pages 861-874.
Fechner, G.T. 1860. *Elemente der Psychophysik*. Breitkopft und Härtel: Leipzig.
Foster, M. 1961. "An Application of the Wiener-Kolmogorov Smoothing Theory to Matrix Inversion". *Journal of the Society for Industrial and Applied Mathematics*, Volume 9(3).

Fouliard, J., Howell, M. and Rey, H. 2019. “Answering the Queen: Machine learning and financial crises”, BIS, *Working paper*.

Freund, Y. and Schapire, R. E. 1997. “A Decision-theoretic Generalization of On-line Learning and an Application to Boosting”, *Journal of Computer and System Sciences,* Volume 55(1).

Friedman, J. H. 2001. “Greedy Function Approximation: A Gradient Boosting Machine”, *The Annals of Statistics*, Volume 29(5).

Funke, M., Schularick, M. and Trebesch, C. 2016. "Going to Extremes: Politics after Financial Crises, 1870–2014". *European Economic Review,* Volume 88, pp 227-260.

Gaddum, J.H. 1933. “Report on Biological Standards III: Methods of Biological Assay Depending on Quantal Response”. *Special Report Series of the Medical Research Council*, No 183. Medical Research Council: London.

Ghosh, S. R. and Ghosh, A.R. 2003. “Structural Vulnerabilities and Currency Crises.” *IMF Staff Papers*, Volume 50(3).

González-Hermosillo, B., Pazarbasioglu, C. and Billings, R. 1997. “Determinants of Banking Sector Fragility: A Case Study of Mexico”, *IMF Staff Papers,* Volume 44(3). Washington: International Monetary Fund.

Greene, W. E. 2008. *Econometric Analysis.* 6th Edition. New Jersey: Prentice Hall.

Hamilton, J. D. 1994. *Time Series Analysis.* Princeton: Princeton University Press.

Hardy, D.C. and Pazarbasioglu, C. 1998. “Leading Indicators of Banking Crises: Was Asia Different?” *IMF Working Paper*, WP/98/91.

Hastie, T., Tibshirani, R. and Friedman, J. 2009. *The Elements of Statistical Learning: Data Mining, Inference and Prediction.* Springer: New York.

Hjerpe, A. 2016. "Computing Random Forest Variable Importance Measures (VIM) on Mixed Continuous and Categorical Data", Master’s Thesis, *KTH Royal Institute of Technology*, *School of Computer Science and Communication*.

Hoerl, A.E. 1962. "Application of Ridge Analysis to Regression Problems". *Chemical Engineering Progress*, Volume 58(3).

James, G., Witten, D., Hastie, T. and Tibshirani, R. 2013. *An Introduction to Statistical Learning - with Applications in R*, Springer, Berlin.

Jordà, Ò., Schularick, M. and Taylor, A.M. 2013. “When credit bites back”. *Journal of Money, Credit and Banking*, Volume 45 (s2), 3–28.

Jordà, Ò., Schularick, M. and Taylor, A.M. 2016. “Sovereigns versus banks: credit, crises, and consequences”. *Journal of European Economic Association*, Volume 14 (1), 45–79.

Jordà, Ò., Schularick, M. and Taylor, A.M. 2017. “Macrofinancial History and the New Business Cycle Facts”. *NBER Macroeconomics Annual*, Volume 31, edited by Eichenbaum, M. and Parker, J.A. University of Chicago Press.

Joy, M., Rusnák, M., Šmídková, K. and Vašíče, B. 2017. “Banking and Currency Crises: Differential Diagnostics for Developed Countries”. *International Journal of Finance & Economics*, Vol 22 (1), pp 44–67.

Kaminsky, G.L. and Reinhart, C.M. 1999. “The Twin Crises: The Causes of Banking and Balance-of-Payments Problems”. *American Economic Review*, Volume 89.

Kindleberger, C. 1978. *Manias, Panics, and Crashes: A History of Financial Crises*, New York: Basic Books.

Laeven, L. and Valencia, F. 2008. “Systemic Banking Crises: A New Database”. *IMF Working Paper,* 08/224.

Laeven L. and Valencia F. 2010. “Resolution of Banking Crises: the Good, the Bad and the Ugly”. *IMF Working Paper,* 10/146.

Laeven, L. and Valencia, F. 2012. “Systemic Banking Crises Database: An Update”. *IMF Working Paper*, 12/163

Laeven, L. and Valencia, F. 2018. Systemic banking crises revisited. IMF Working Paper, WP/18/206.

Manasse, P. and Roubini, N. 2009. "'Rules of Thumb' for Sovereign Debt Crises". *Journal of International Economics,* No 78(2), pp 192–205.

Mullainathan, S. and Spiess, J. 2017. "Machine Learning: An Applied Econometric Approach", *Journal of Economic Perspectives,* Volume 31(2).

Nevasalmi, L. 2020. "Forecasting multinomial stock returns using machine learning methods". *The Journal of Finance and Data Science*, Vol 6, pp 86-106.

Nyman, R. and Ormerod, P. 2016. "Predicting Economic Recessions Using Machine Learning Algorithms", *University College London*.

Phillips, C.B. and Peron, P. 1988. "Testing for a Unit Root in Time Series Regression." *Biometrika*, Volume 75(2), pp 335–346.

Phillips, D. L. 1962. "A Technique for the Numerical Solution of Certain Integral Equations of the First Kind". *Journal of the ACM*, Volume 9.

Reinhart, C.M. and Rogoff, K.S. 2009. *This Time is Different: Eight Centuries of Financial Folly*, Princeton Press: New Jersey.

Santosa, F. and Symes, W.W. 1986. "Linear Inversion of Band-Limited Reflection Seismograms". *SIAM Journal on Scientific and Statistical Computing*, Volume 7(4).

Savona, R. and Vezzoli, M. 2012. "Multidimensional Distance-to-Collapse Point and Sovereign Default Prediction". *Intelligent Systems in Accounting, Finance and Management*, No 19(4).

Strobl, C., Boulesteix, A.L., Zeileis, A. and Hothorn, T. 2007. "Bias in random forest variable importance measures: Illustrations, sources and a solution". *BMC Bioinformatics,* Vol 8(25). https://doi.org/10.1186/1471-2105-8-25

Sun, X. and Weichao, X. 2014. "Fast Implementation of DeLongs Algorithm for Comparing the Areas Under Correlated Receiver Operating Characteristic Curves". *IEEE Signal Processing Letters*, Volume 21, pp 1389–1393.

Tibshirani, R. 1996. "Regression Shrinkage and Selection via the Lasso". *Journal of the Royal Statistical Society*, Volume 58(1).

Tikhonov, A.N. 1943. "On the Stability of Inverse Problems". *Doklady Akademii Nauk SSSR*, Volume 39(5).

Tikhonov, A.N. 1963. "Solution of Incorrectly Formulated Problems and The Regularization Method". *Soviet Mathematics*, Volume 4.

Van Rijsbergen, C.J. 1979. *Information Retrieval*. London: Butterworths.

Vlaar, P.J.G. 2000. "Currency Crises Models for Emerging Markets." *De Nederlandsche Bank Staff Report*, Number 45.

Wold, H. 1985. "Partial Least Squares". *Encyclopaedia of Statistical Sciences*. Wiley: New York.

Ward, F. 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.

Youden, W.J. 1950. "Index for rating diagnostic tests". *Cancer*, 3, 32–35. DOI: 10.1002/1097- 0142(1950)3:1<32::AID-CNCR2820030106>3.0.CO;2-3.

Appendix

Section A

A1 Non-Parametric

A1.1 Baseline Approach

As a non-parametric model, the baseline model functions as benchmark for the performance of all the algorithms. Based on a conventional modelling framework, the modelling approach studies mean values across the training dataset. Formally stated as $\hat{x}_i = \frac{1}{N}\sum_{i=1}^{N} x_i$, the model renders a straightforward non-parametric solution which is employed for predictions of the test dataset.

A2 Regression Algorithms

A2.1 Linear Probability Model

The linear probability model is an extension of the linear regression equation and operationalised as a generalised case of the binomial distribution. Thereby, underscoring a linear relationship between the predictors and discrete outcome variable *Y*. The probability of observing a systemic financial crisis ($Y = 1$) or non-crisis ($Y = 0$) is determined through vector x, mathematically stated as $Prob(Y = 1|x) = F(x, \beta)$ and $Prob(Y = 0|x) = 1 - F(x, \beta)$. Given that the β parameters express the response of fluctuations in x on the likelihood of a crisis episode, the marginal effects of predictors on the probability of the independent variable can be estimated. Following Greene (2008), by inserting the linear regression equation, $F(x, \beta) = x'\beta$, the linear probability regression can be denoted as $Y = E[y|x] + y - E[y|x] = x'\beta + \epsilon$. A shortfall of the linear probability modelling framework is that $x'\beta$ is not constrained to the 0 to 1 interval, and out of range results could inhibit clear interpretation (Greene, 2008).

A2.2 Probit Regression

As one of the oldest (see Fechner (1860), Gaddum (1933) and Bliss (1934, 1935)) and most popular statistical methods (Cramer, 2002) the probit regression, comparable to the linear probability method, models a binary outcome variable. Operationalised, by modelling an inverse standard normal distribution of the outcome variable as a linear relationship to the

explanatory variables. Based on Greene (2008), this can formally be denoted as $Y_i^* = x_i'\beta + \varepsilon_i$, where ε_i $N[0,1]$ and $y_i = 1$ if $y_i^* > 0$, else $y_i = 0$. Given that y_i follows a Bernoulli distribution, which consist of a single draw from a two-outcome binomial procedure, probability values can be described by $Prob(y_i = 1|x_i) = \phi(x'_i, \beta)$ and $Prob(y_i = 0|x_i) = 1 - \phi(x'_i, \beta)$. As nuance, the binary choice model in comparison to the linear probability model is estimated through maximum likelihood, which in combination with success probability $F(x'_i\beta)$ and independent and random observations can be defined through a joint probability as $L(y|X, \beta) = \prod_{i-1}^{n}[\phi(x'_i\beta)]^{y_i}[1 - \phi(x'_i\beta)]^{1-y_i}$.

A3 Instance-Based Algorithms

A3.1 K-Nearest Neighbours (k-NN)

Procedurally, and through a positive integer *k* and observation x_0, the k-NN classifier detects the *k* points in the dataset which are adjacent to x_0, characterized by N_0. Consequently, the conditional probability for class *j* is estimated as the proportion of datapoints in N_0 where the response values are identical to *j*, described formally as $pr(y = j|x = x_0) = \frac{1}{K}\sum_{i\in N_0} I(y_i = j)$.

Operationally, k-NN integrates Bayes' theorem to label the test observation x_0 as the outcome class with the highest probability. Subsequent to the classifier technique, the k-NN regression method is estimated, where $\hat{f}(x_0)$ is determined as the average of all the training responses in N_0, stated as $\hat{f}(x_0) = \frac{1}{K}\sum_{x_i\in N_0} y_i$. In setting *k*, the allowable error rate impacts on the bias-variance trade-off. Where *k* = 1 the error rate in the training dataset converges to zero, but the variance encountered in the test set would be large. By increasing the value of *k*, a higher quantity of errors would lead to higher bias, while the error count in the test dataset could shrink (James et al., 2013). In this paper, cross-validation consists of tenfold resampling, repeated ten times, with maximum number of *k* set to 9, and with distance set at 2.

A3.2 Support Vector Machine (SVM)

Kernels determine the level of relationship, which in turn finds support vector lines to classify the observations. Based on James et al. (2013), SVM is constructed using support vector classifiers, where a linear support vector classifier can be denoted as

$$f(x) = \beta_0 + \sum_{i=1}^{N} \alpha_i(x, x_i), \quad (1)$$

with N number of parameters α_i. To estimate the kernel, inner products of observations instead of actual observations are employed, represented by

$$(x_i, x_{i\,'}) = \sum_{j=1}^{p}(x_i, x_{i'\,j}), \quad (2)$$

for observations $(x_i, x_{i\,'})$. Consequently, parameters $\alpha_i, \dots, \alpha_n$ are computed using inner products $(x_i, x_{i\,'})$ of observations. Given that α_i only takes positive values for support vectors, α_i turn zero for all non-support vector observations. Where S constitutes the set of support points, equation (2) can be restated as

$$f(x) = \beta_0 + \sum_{i \in S} \alpha_i (x, x_i), \quad (3)$$

resulting in significantly fewer terms to consider. The inner product of the observations can be replaced with a generalized version $k(x_i, x_{i'})$, where k is a kernel, a function which measures the resemblance across a set of observations. Enhanced with a polynomial kernel of degree *d* so that

$$k(x_i, x_{i\,'}) = (1 + \sum_{j=1}^{p}(x_{ij}, x_{i'\,j})^d, \quad (4)$$

where $d > 1$, to support more flexible decision boundaries. Compared to the original feature space, through the polynomial, the kernel permits a higher-dimensional space. A support vector classifier in conjunction with a non-linear kernel result in a support vector machine and can mathematically be denoted as

$$f(x) = \beta_0 + \sum_{i \in S} \alpha_i k(x, x_i). \quad (5)$$

Where $d = 1$, the SVM and support vector classifiers are considered identical.

For the SVM algorithm, the radial kernel is used with gamma as 0.083, cost constraints (regularisation constant) set at 1 and insensitive loss-function (epsilon) to 0.1.

A4 Regularisation Algorithms

Ridge and lasso introduce some bias by adding a penalty to the regression, with the aim of dealing with the bias-variance trade-off encountered by machine learning.

A4.1 Ridge

In contrast to the ordinary least squares statistical technique which computes $\beta_0,\ \beta_1, \dots, \beta_p$ by employing values which minimizes the equation

$$RSS = \sum_{i=1}^{N}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2, \quad (1)$$

ridge coefficients are determined by minimising the following equation,

$$\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 + \lambda\sum_{j=1}^{p}\beta_j^2 = RSS + \lambda\sum_{j=1}^{p}\beta_j^2, \quad (2)$$

where $\lambda \geq 0$ represents a tuning parameter. According to component $\lambda\sum_j \beta_j^2$, the shrinkage penalty is small when, $\beta_1, \ldots, \beta_p$ are near zero, so it reduces the estimates of β_j towards zero.

Indeed, when $\lambda = 0$, the ridge regression will be comparable to least squares. However, in comparison to least squares, ridge produces a dissimilar group of coefficient estimates for distinctive values of λ. Choosing the optimal value of λ can be achieved through cross-validation. The shrinkage penalty is applied to $\beta_1, \ldots, \beta_p$, but not to the intercept. If the data matrix *X* has a zero mean, then the intercept becomes $\beta_0 = y_i = \sum_{i=1}^{n}\frac{y_i}{n}$.

The cross-validation process involves allotting all observations into λ folds, performed randomly and based on similar sizes. The first fold is considered the validation set, with the estimated model fitted on the remaining λ − 1 folds. Thereafter, the error value is calculated based on the model performance on the λ − 1 folds. Repeated λ times, the procedure treats a different fold as validation set every time. Consequently, the tuning parameter is chosen as based on the cross-validation rendering the smallest error. The final model applies the selected value of the tuning parameter in conjunction with the full set of observations.

Compared to ordinary least squares, ridge regression improves through the bias-variance trade-off, where a higher λ increases bias, but reduces variance. Given that the shrinkage penalty $\lambda\sum_j \beta_j^2$ reduces all coefficients towards zero, yet none set exactly to zero, a shortcoming of the ridge approach involves a final model comprising all explanatory variables, even if their impact is trivial, which in the context of a high number of variables, could impact interpretability of results (James at al., 2013).

In this study, hyperparameter settings employed include $\beta_j^2 = 0$ and $\lambda = 100$.

A4.2 Lasso

Overcoming the drawback of the ridge approach, Santosa and Symes (1986) and Tibshirani (1996), devised the Least Absolute Shrinkage and Selection Operator or Lasso algorithm. Operationalized by minimising the equation

$$\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 + \lambda\sum_{j=1}^{p}|\beta_j| = RSS + \lambda\sum_{j=1}^{p}|\beta_j|, \quad (1)$$

where the ridge penalty β_j^2 is replaced by the lasso penalty $|\beta_j|$.

For the lasso algorithm in the study, hyperparameter settings applied entail $|\beta_j| = 1$ and $\lambda = 100$.

A5 Dimensionality Reduction Algorithm

A5.1 Partial Least Squares (PLS)

The procedure involves estimating PLS directions. The first PLS direction is computed by normalising the predictors p and equating each $\emptyset_{jm}$ in equation $Z_m = \sum_{j=1}^{p} \emptyset_{jm} X_j$ to the coefficients from the linear regression of Y onto X_j. As a consequence, the coefficients are proportional to the correlation between Y onto X_j. In computing the equation $Z_1 = \sum_{j=1}^{p} \emptyset_{j=1} X_j$, PLS put larger weights on the explanatory variables which are best related to the outcome.

The second PLS direction is estimated by adjusting each predictor for Z_1, achieved by regressing each predictor on Z_1 and computing their residuals. These residuals signify the unexplained information from the first PLS direction. Following, Z_2 can be estimated with the same approach as Z_1, iterating M number of times to detect multiple new features, $Z_1, \dots, Z_m$. Once this process is complete, ordinary least squares are employed to fit a model predicting Y using $Z_1, \dots, Z_m$. The number M of partial least squares directions represents a tuning parameter which can be chosen using a cross-validation approach. If the predictors are highly correlated with each other, or if a smaller number of components accurately model the response, then the number of components in the PLS model would be less than the number of predictors. Dimension reduction procedure of PLS serves to reduce bias in existing datasets but faces lower accuracy when modelling new data.

In this study, the PLS algorithm incorporates 10-fold cross-validation, repeated 10-times, with the optimised number of principle components (ncomp) set to 3.

A6 Decision Tree Algorithms

A6.1 Full Tree

The aim of the splitting procedures is to minimize a loss function, which is computed and directed by the divergence from an exact partitioning of respective crisis and no-crisis observations into their identifiable nodes. As based on Joy et al. (2017), the quantity of observations of class c at node n is represented by $p(c|n)$. With binary outcomes of financial crises, class distribution can be denoted by $(p0, p1)$, in which case $p0$ signifies the probability of all no-crisis occurrences delineated into node n, while $p1$ demonstrates the probability of a crisis in node n. Divisions are estimated by the deviances within the child nodes. Skewer

distributions such as (0,1) comprise smaller deviances, with full divergence at (0.5,0.5). The Gini principle supports the dividing approach, with the aim to minimize the loss function $c(n)$: $c_{gini}(n) = \sum p0(n)p1(n)$. The latter is consequently minimized when terminal nodes include either of two classes of incidents, systemic financial crisis or no-crisis.

Tolerance levels of misclassification can be integrated through stipulation of weights, for instance not recognising a crisis, which could result in the identification of different predictors and their threshold levels. The partitioning process of forming tree branches ceases when the fall in the misclassification ratio is lower than the penalisation imposed on additionally produced terminal nodes. Analogously, this criterion is also employed to choose the best tree, with goodness of fit categorised as the optimal point between minimising the classification rate while bigger trees are penalised. Yet, terminal nodes are not always entirely uniform.

In this study, the full tree algorithm only attempts splits which reduces the overall lack of fit by at least 0.001 (complexity parameter). Lower values are expected to be pruned away in the subsequent procedure.

A6.2 Pruned Tree

Following James at al. (2013), the decision tree equation can formally be denoted as

$$\sum_{i=1}^{T} \sum_{x_i \in R_m} (y_i - \hat{y}_{Rm})^2 + \alpha|T|. \quad (1)$$

While creating a full tree, cost complexity pruning is applied to the large tree in order to obtain a series of solid subtrees, as a function of α. *K*-fold cross-validation is performed to select the value of α using the training data. By means of a forecast utilising the test dataset, the root mean squared error is obtained and assessed. Following, the average results across every value of α are estimated, and subsequently a value of α is selected that would minimize the average error. Lastly, the subtree associated with the chosen value of α can be identified. For this algorithm, the optimal size of tree nodes is set to 3 and used as part of the procedure to minimise the cross-validation error (xerror), and which determines the nodes to prune.

A7 Ensemble Algorithms

A7.1 Adaptive Boosting

Mathematically, the training dataset consists of $(x_1 + y_1), \dots, (x_N + y_N)$, with weight vector $w_i^1 = D(i)$ for $i = 1, \dots, N$ and for D the distribution over N. The quantity of iterations is

represented by $T = 1, 2, \ldots, T$. Initially, an equal set of weights w^t is applied across N, with distribution $p^t = \frac{w^t}{\sum_{i=1}^{N} w_i^t}$, estimated by standardising the weights. The weak learner applies the distribution p^t to produce a new prediction h_t. In a test on the efficacy of the forecast, an error of h_t is computed through $\epsilon_t = \sum_{i=1}^{N} p_i^t |h_t(x_i) - y_i|$. For every iteration, the weak learner with lowest error is elected. The error is applied to determine the new weights vector $w_i^{t+1} = w_i^t \beta_t^{1-|h_t(x_t)-y_i|}$, where $\beta_m = \frac{\epsilon^t}{1-\epsilon^t}$ is also incorporated to signify the contribution of the chosen weak learner to the last prediction of the strong learner. This process continues across T where predictions are determined by $h_f(x) = \begin{cases} 1 \;\; if \sum_{i=1}^{T} \left(log \frac{1}{\beta_t}\right) h_t(x) f(x) \geq \frac{1}{2} \sum_{i=1}^{T} \left(log \frac{1}{\beta_t}\right) \\ 0 \; otherwise \end{cases}$.

Parameter settings applied in this paper encompass number of trees on 100, with 10-fold cross-validation. The bootstrap sample of the training set is centred on the weights for every observation during each individual iteration.

A7.2 Gradient Boosting

Following Friedman (2001) and Döpke et al. (2017), mathematically, the algorithm bootstrap sample from training dataset $\{(x_i + y_i)\}_{i=1}^{N}$, with differentiable loss-function $L(y_i, F(x))$ to determine a negative gradient vector. The model is initialised with a constant, using $\mathrm{F}_0(\mathrm{x}) = argmin \sum_{i=i}^{N} \mathrm{L}(\mathrm{y}_i, \rho)$. For m = 1 to M, where the quantity of weak learners is capped, residuals are calculated for every sample $\widetilde{y_i} = -\left[\frac{\partial L(y_i, F(x_i))}{\partial F(x_i)}\right]_{F(x)=F_{m-1}(x)}$, given where $\left[\frac{\partial L(y_i, F(x_i))}{\partial F(x_i)}\right]$ denotes the gradient derivative and $\widetilde{y_i}$ pseudo residuals, computed using$\{(\widetilde{y_i} - g_m(x_i)\}_{i=1}^{N}$. The following step involves fitting the regression tree to the predicted residuals. Commencing with each leaf in every tree, output is determined that minimizes the function $\gamma_{jm} = argmin \sum_{x_i \in R_{ij}}^{N} \mathrm{L}(\mathrm{y}_i, F_{m-1}(x_i) + \gamma)$, achieved by adopting the previous prediction value and the selected sub-sample. For following trees, a learning rate described by ϑ, ranging from 0 to 1 is added to lessen the influence of a single tree on final output $\mathrm{F}_m(\mathrm{x}) = \mathrm{F}_{m-1}(\mathrm{x}) + \vartheta \sum_{j=1}^{J_m} \gamma_{jm} I(x \in R_{jm})$. Lastly, when m = M, the strong learner $\mathrm{F}_m(\mathrm{x})$ is computed as the sum of all weak learners, based on m = 0,…,M, which is adopted to make predictions using the out-of-bag sample.

To process the model, maximum tree depth is set to 1 which denotes an additive model. Minimum number of observations per final node equals 10 with a shrinkage parameter of 0.1. The procedure is simulated 100 times for purposes of statistical inference. Maximum quantity of base learners is set to 100. And 50 percent of the training data is randomly elected to create each new weak learner in the stepwise technique.

A7.3 Random Forests

Based on Hastie et al. (2009), a tree T_b using random forests is grown through the bootstrapped procedure until a minimum node size is reached. This process can be formulated as:

(1) Choosing *m* variables at random, from the *p* variables.
(2) Find the best split-point amongst the *m* variables.
(3) Subsequently, split the parent node into two child nodes.
(4) With the output of the trees encapsulated by $\{T_b\}_1^B$.

Predictions at each new point x can be executed through $F^B{}_{rf}(x) = \frac{1}{B}\sum_{b=1}^{B} T_b(x)$.

To operationalise the random forests algorithm, initially the quantity of trees is set to 1000, but the optimal number of trees necessary to calculate the minimum error estimate is consequently computed in the testing procedure and applied to predictions.

Table 2: Explanatory Indicators

Indicator	Definition	Category
GDP	GDP per capita, first difference in logs	Real
CE	Consumption expenditure, first difference in logs, two lags	Real
FCF	Fixed capital formation, first difference in logs, one lag	Real
COR	Fixed capital formation to GDP, first difference in logs	Real
DEBT	Debt relative to GDP, first difference in logs	Banking
LOANS	Total loans, in logs, one lag	Banking
STOCK	Stock market, first difference in logs	Banking
CPI	Consumer inflation, in logs	Banking
SR	Short-term interest rates	Banking
LR	Long-term interest rates	Banking
ER	Exchange rate, first difference	External
CA	Current account balance	External

Table 3: Number of Crises by Country

Country	Count of Crises
Australia	2
Belgium	7
Canada	1
Denmark	7
Finland	5
France	4
Germany	6
Italy	9
Japan	6
Netherlands	5
Norway	4
Portugal	5
Spain	8
Sweden	6
Switzerland	5
United Kingdom	4
United States	6

Table 4: Unit Root Tests

Indicator	Specification	Inverse chi-squared	Inverse normal	Inverse logit t	Modified inv. chi-squared
ADF – Fisher Test (Levels)					
GDP	c,4	461.772***	-19.166***	-31.049***	51.875***
CE	c,4	469.377***	-19.365***	-31.561***	52.797***
FCF	c,4	478.482***	-19.491***	-32.172***	53.901***
COR	c,4	623.583***	-22.787***	-41.930***	71.497***
DEBT	c,4	448.313***	-18.337***	-30.112***	50.242***
LOANS	c,4	4.535	7.547	8.158	-3.573
STOCK	c,4	532.674***	-20.703***	-35.817***	60.473***
CPI	c,4	2.424	7.429	7.794	-3.829
SR	c,4	31.307	-0.504	-0.486	-0.326
LR	c,4	31.743	-0.254	-0.359	-0.273
ER	c,4	460.365***	-19.160***	-31.919***	51.704***
CA	c,4	83.126***	-1.985**	-3.191***	5.957***
Phillips-Perron – Fischer Test (Levels)					
GDP	c,4	1165.212***	-32.556***	-78.350***	137.179***
CE	c,4	1168.767***	-32.613***	-78.590***	137.610***
FCF	c,4	1138.007***	-32.020***	-76.521***	133.880***
COR	c,4	1171.288***	-32.643***	-78.759***	137.916***
DEBT	c,4	1101.636***	-31.377***	-74.075***	129.469***
LOANS	c,4	3.095	8.190	9.194	-3.747
STOCK	c,4	1186.083***	-32.902***	-79.754***	139.710***
CPI	c,4	1.518	9.057	10.114	-3.938
SR	c,4	72.864***	-4.074***	-4.247***	4.713***
LR	c,4	30.969	-0.684	-0.633	-0.367
ER	c,4	1138.229***	-32.265***	-78.919***	133.907***
CA	c,4	50.684**	0.202	0.461	2.023**

Unit root tests are constructed using Augmented Dicky-Fuller (see Hamilton, 1994) and Phillips and Peron (see Phillips and Peron, 1988) procedures. Based on Choi (2001), four different methods are assessed to test the null hypothesis of a unit root across all panels, through an inverse χ^2, inverse-normal, inverse-logit transformation and a modification of the inverse χ^2 transformation of the p-values. The latter is appropriate for $N \rightarrow \infty$.
*** (**, *) denotes significance at 1%, (5%, 10%)

Table 5: Sample Means of Explanatory Indicators

Indicators	Y = 0	Y = 1	T-test
Real Sector			
Gross domestic product per capita	0.002	0.000	0.017**
Consumption expenditure	0.017	0.006	0.068*
Fixed capital formation	0.105	-0.029	0.001**
Capital output ratio	-0.001	0.019	0.003**
Banking Sector			
Debt to gross domestic product	0.007	-0.470	0.018**
Total loans	5.913	4.457	0.052*
Stock market	0.010	-0.076	0.215
Consumer inflation	1.447	0.047	0.005**
Short-term interest rates	4.806	6.053	0.000***
Long-term interest rates	5.591	5.655	0.003**
External Sector			
Exchange rates	0.052	0.069	0.682
Current account	-57591	-270219	0.012**

T-test p-values: ***/**/* denotes 10%, 5%, 1% rejection of null hypothesis.

Table 6: Probit Model Results

	Lag Structure	Optimal Contemporary Structure
No. of observations:	2,431	2,414
Constrained log-likelihood:	-375.147	-374.526
Max. log-likelihood:	-354.944	-338.919
LR-chi^2:	40.40***	71.21***
AIC:	0.303	0.292
BIC:	-18140.981	-18023.648
Variable	dy/dx	dy/dx
Gross domestic product per capita	0.556 (0.864)	-1.212 (0.539) **
Consumption expenditure	-0.353 (0.238)	0.404 (0.173) **
Fixed capital formation	0.007 (0.002) **	0.005 (0.002) **
Capital output ratio	-0.056 (0.061)	0.110 (0.049) **
Debt to gross domestic product	0.003 (0.001) **	-0.004 (0.001) ***
Total loans	0.001 (0.000)	0.001 (0.000) *
Stock market	-0.001 (0.005)	-0.007 (0.004) *
Consumer inflation	-0.003 (0.001) **	-0.003 (0.001) ***
Short-term interest rates	0.007 (0.002) ***	0.010 (0.002) ***
Long-term interest rates	-0.006 (0.002) ***	-0.009 (0.002) ***
Exchange rates	-0.015 (0.025)	-0.001 (0.005)
Current account	-0.000 (0.000) *	0.000 (0.000) *

Margins with standard errors in brackets; *** (**, *) denotes significance at 1%, (5%, 10%)

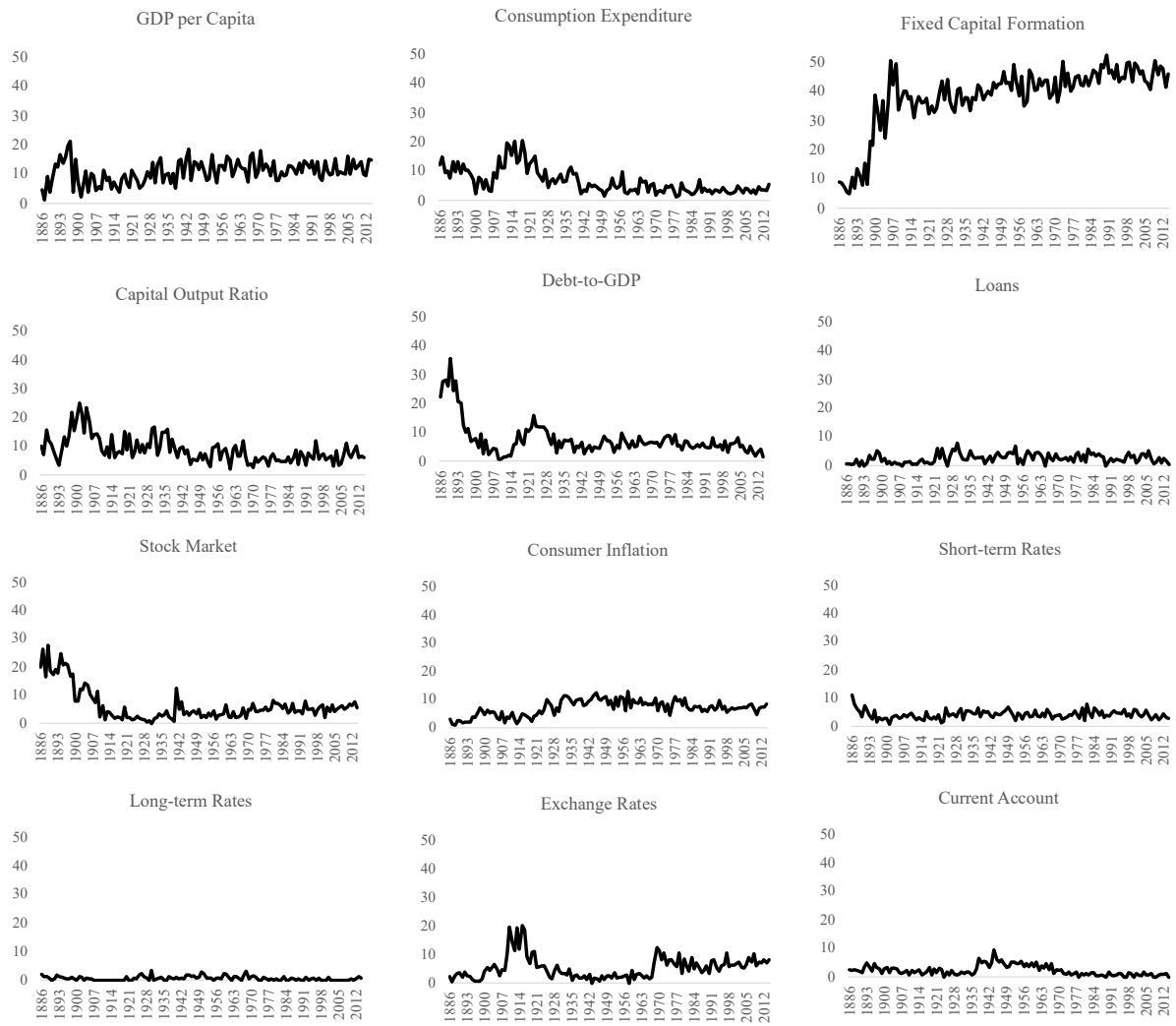

Figure 3: Gradient Boosting: Variable Importance by Indicator

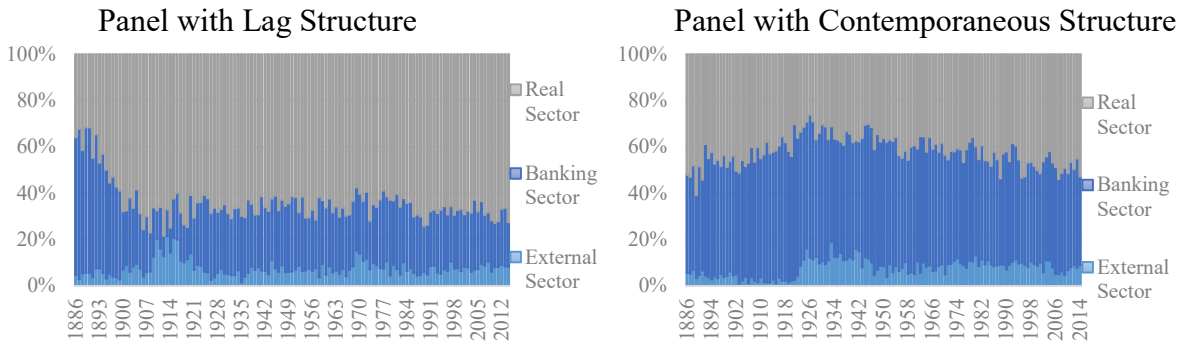

Figure 4: Gradient Boosting: Variable Importance by Sector

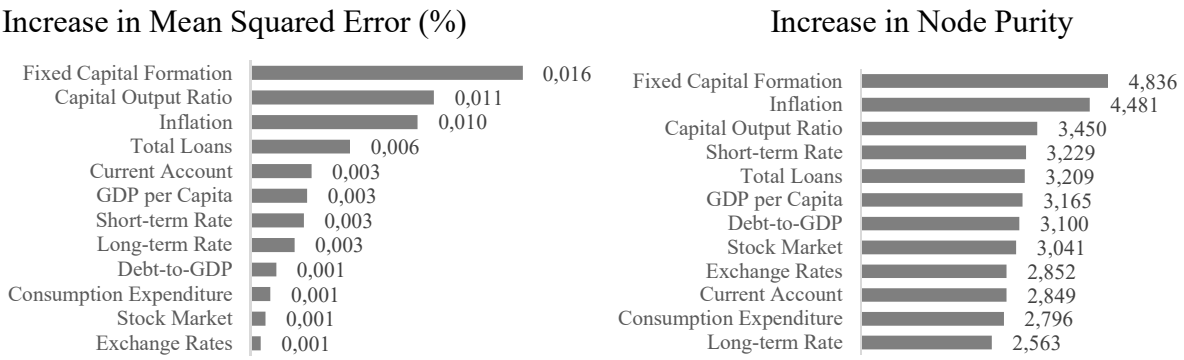

Figure 5: Random Forests: Variable Importance

Table 9: Individual Country Recursive Out-of-Sample Forecasts in Pooled Format

Country	Base	Linear	Probit	kNN	SVM	Ridge	Lasso	PLS	Tree	Prune	Ada	Boost	RF
Australia	0.610	0.918	0.943	0.500	0.752	0.674	0.539	0.500	0.498	0.557	0.621	0.436	0.656
Belgium	0.583	0.614	0.674	0.500	0.493	0.552	0.529	0.500	0.736	0.708	0.622	0.722	0.636
Canada	0.577	0.669	0.662	0.500	0.648	0.690	0.577	0.500	0.810	0.563	0.627	0.599	0.739
Denmark	0.572	0.536	0.612	0.500	0.678	0.599	0.570	0.500	0.824	0.514	0.558	0.817	0.676
Finland	0.445	0.835	0.752	0.500	0.715	0.843	0.537	0.500	0.560	0.491	0.561	0.729	0.811
France	0.581	0.636	0.711	0.500	0.735	0.420	0.483	0.500	0.593	0.543	0.713	0.609	0.825
Germany	0.482	0.603	0.543	0.500	0.458	0.589	0.529	0.500	0.502	0.505	0.503	0.562	0.535
Italy	0.491	0.877	0.797	0.496	0.728	0.842	0.561	0.500	0.539	0.520	0.636	0.738	0.703
Japan	0.648	0.563	0.514	0.500	0.633	0.564	0.600	0.500	0.628	0.600	0.701	0.673	0.565
Netherlands	0.814	0.814	0.881	0.500	0.652	0.601	0.853	0.500	0.520	0.875	0.483	0.675	0.599
Norway	0.574	0.517	0.586	0.500	0.689	0.656	0.511	0.500	0.532	0.541	0.660	0.550	0.596
Portugal	0.567	0.665	0.765	0.496	0.565	0.609	0.747	0.500	0.591	0.607	0.662	0.568	0.614
Spain	0.579	0.745	0.772	0.500	0.683	0.596	0.642	0.500	0.477	0.541	0.730	0.728	0.603
Sweden	0.543	0.734	0.748	0.500	0.658	0.748	0.553	0.500	0.627	0.572	0.731	0.844	0.836
Switzerland	0.509	0.657	0.478	0.493	0.619	0.488	0.508	0.500	0.539	0.553	0.700	0.810	0.719
UK	0.485	0.748	0.771	0.500	0.709	0.580	0.580	0.500	0.723	0.526	0.654	0.748	0.843
USA	0.689	0.837	0.697	0.500	0.701	0.633	0.686	0.500	0.587	0.645	0.812	0.627	0.821

Table 10: Top Country Models from Contemporaneous Panel Format

Country	Top Model	Top Model Accuracy	Deviation to Baseline	Overall deviation across all models
Australia	Probit	0.943	0.235	0.158
Belgium	Full Tree	0.736	0.108	0.087
Canada	Full Tree	0.810	0.165	0.089
Denmark	Full Tree	0.824	0.178	0.109
Finland	Ridge	0.843	0.281	0.146
France	Random Forests	0.825	0.173	0.117
Germany	Linear	0.603	0.086	0.042
Italy	Linear	0.877	0.273	0.140
Japan	Adaptive Boosting	0.701	0.037	0.065
Netherlands	Probit	0.881	0.047	0.155
Norway	Support Vector Machine	0.689	0.081	0.064
Portugal	Probit	0.765	0.140	0.081
Spain	Probit	0.772	0.136	0.102
Sweden	Gradient Boosting	0.844	0.213	0.121
Switzerland	Gradient Boosting	0.810	0.213	0.108
UK	Random Forests	0.843	0.253	0.121
USA	Linear	0.837	0.105	0.109

Table 12: Individual Country Recursive Out-of-Sample Forecasts in Independent Format

Country	Base	Linear	Probit	kNN	SVM	Ridge	Lasso	PLS	Tree	Prune	Ada	Boost	RF
Australia	0.888	0.576	0.647	0.500	0.546	0.961	0.508	0.500	0.422	0.888	0.945	0.470	0.686
Belgium	0.628	0.674	0.580	0.496	0.529	0.547	0.606	0.496	0.577	0.593	0.685	0.482	0.515
Canada	0.918	0.527	0.781	0.500	1.000	0.844	0.918	0.500	0.449	0.918	0.879	0.875	0.633
Denmark	0.696	0.689	0.573	0.500	0.682	0.771	0.696	0.500	0.621	0.629	0.931	0.747	0.732
Finland	0.749	0.694	0.504	0.500	0.652	0.747	0.749	0.500	0.663	0.777	0.641	0.646	0.632
France	0.578	0.399	0.550	0.500	0.452	0.810	0.579	0.496	0.550	0.567	0.647	0.746	0.678
Germany	0.469	0.721	0.508	0.596	0.634	0.661	0.561	0.500	0.735	0.423	0.589	0.668	0.634
Italy	0.635	0.515	0.635	0.500	0.468	0.419	0.635	0.500	0.620	0.407	0.489	0.565	0.603
Japan	0.585	0.477	0.752	0.500	0.510	0.644	0.585	0.500	0.473	0.585	0.512	0.466	0.648
Netherlands	0.751	0.583	0.670	0.500	0.351	0.698	0.686	0.500	0.371	0.751	0.562	0.740	0.412
Norway	0.745	0.569	0.556	0.492	0.418	0.746	0.661	0.500	0.494	0.719	0.844	0.588	0.521
Portugal	0.700	0.776	0.615	0.484	0.595	0.494	0.627	0.492	0.749	0.614	0.850	0.710	0.640
Spain	0.690	0.785	0.642	0.496	0.621	0.540	0.691	0.500	0.584	0.675	0.450	0.754	0.485
Sweden	0.428	0.653	0.550	0.496	0.626	0.710	0.762	0.500	0.599	0.685	0.748	0.706	0.597
Switzerland	0.748	0.642	0.519	0.500	0.814	0.547	0.690	0.500	0.440	0.736	0.578	0.606	0.671
UK	0.741	0.600	0.614	0.488	0.564	0.478	0.469	0.500	0.465	0.871	0.787	0.606	0.384
USA	0.677	0.821	0.618	0.496	0.555	0.729	0.579	0.492	0.626	0.614	0.718	0.606	0.568

Table 13: Top Country Models from Independent Format

Country	Top Model	Top Model Accuracy	Deviation to Baseline	Overall deviation across all models
Australia	Ridge	0.961	0.052	0.197
Belgium	Adaptive Boosting	0.685	0.040	0.067
Canada	Support Vector Machine	1.000	0.058	0.198
Denmark	Adaptive Boosting	0.931	0.166	0.116
Finland	Pruned Tree	0.777	0.020	0.097
France	Ridge	0.810	0.164	0.115
Germany	Full Tree	0.735	0.188	0.096
Italy	Lasso	0.635	0.000	0.083
Japan	Probit	0.752	0.118	0.086
Netherlands	Pruned Tree	0.751	0.000	0.146
Norway	Adaptive Boosting	0.844	0.070	0.127
Portugal	Adaptive Boosting	0.850	0.106	0.113
Spain	Linear	0.785	0.067	0.109
Sweden	Lasso	0.762	0.236	0.104
Switzerland	Support Vector Machine	0.814	0.047	0.113
UK	Pruned Tree	0.871	0.092	0.143
USA	Linear	0.821	0.102	0.094