du Plessis, Emile; Fritsche, Ulrich

New forecasting methods for an old problem: Predicting 147 years of systemic financial crises

Journal of Forecasting

Provided in Cooperation with:
John Wiley & Sons

Suggested Citation: du Plessis, Emile; Fritsche, Ulrich (2024) : New forecasting methods for an old problem: Predicting 147 years of systemic financial crises, Journal of Forecasting, ISSN 1099-131X, Wiley Periodicals, Inc., Hoboken, NJ, Vol. 44, Iss. 1, pp. 3-40, https://doi.org/10.1002/for.3184

This Version is available at:
https://hdl.handle.net/10419/313771

Received: 8 November 2023 | Revised: 26 May 2024 | Accepted: 21 July 2024

DOI: 10.1002/for.3184

RESEARCH ARTICLE

WILEY

New forecasting methods for an old problem: Predicting 147 years of systemic financial crises

Emile du Plessis | **Ulrich Fritsche**

Faculty of Business, Economics and Social Sciences, University of Hamburg, Hamburg, Germany

Correspondence
Emile du Plessis, Faculty of Business, Economics and Social Sciences, University of Hamburg, Hamburg, Germany.
Email: grensduplessis@gmail.com

Abstract

This paper develops new forecasting methods for an old and ongoing problem by employing 13 machine learning algorithms to study 147 years of systemic financial crises across 17 countries. Findings suggest that 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. A lag structure and rolling window both improve on optimized contemporaneous and individual country formats. Through a lag structure, banking sector predictors on average describe 28% of the variation in crisis prevalence, the real sector 64%, and the external sector 8%. 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% correct forecasts, and consistently outperform traditional regression algorithms. Learning from other countries improves predictive strength, and non-linear models generally deliver higher accuracy rates than linear models. Algorithms retaining all variables perform better than those minimizing the influence of variables.

KEYWORDS

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

JEL CLASSIFICATION

C14, C15, C32, C35, C53, E37, E44, G21

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. Financial crises have further increased in prominence, first, through the awarding of the Sveriges Riksbank Prize in Economic Sciences in Memory of Alfred Nobel 2022 (Nobelprize, 2022) for research on banking and financial crises, and secondly, due to a spate of recent banking failures and troubles in 2023. First Republic Bank, Silicon Valley Bank, and

Signature Bank are respectively the second, third, and fourth biggest US banks to fail (FDIC, 2023), while troubled bank Credit Suisse, the second largest in Switzerland was taken over by UBS, the largest bank in the country (FSO, 2023). While financial crises have occurred periodically over centuries (Reinhart & Rogoff, 2009), the consequential high social, economic, and political costs (Chen et al., 2019; Funke et al., 2016; Laeven & Valencia, 2010; Laeven & 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 innovative 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 policy makers, it is both a practical and straightforward approach given its proliferation across statistical programs. In comparison to traditional macroeconomic tools such as probit or logit models, machine learning approaches have improved forecasting performance (Alessi et al., 2015; Casabianca et al., 2019; Davis et al., 2011; Döpke et al., 2017; Fouliard et al., 2021a). With the ongoing banking distress, crisis prevention has become more urgent. Can these new forecasting methods better predict and thereby assist policy makers to prevent an old and ongoing problem that has caused havoc over centuries and continues to confront both large and small economies?

This paper contributes to the literature by studying 147 years of systemic financial crises, comprising a total of 17 present-day advanced economies that experienced a combined 90 crises between 1870 and 2016. Given that some countries resembled emerging markets during the period under review, results are not limited only to advanced economies and support generalizable implications. Based on economic theory, 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), and du Plessis (2022a, 2022b). 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 modeling dimensions, the predictive strength of machine learning methods is assessed. These dimensions entail an optimized contemporaneous panel format with an expanding window, transformations with lag structure, and a rolling window as well as in individual country format. Across recursive out-of-sample performance, assessed measures include AUC (area under the curve), F_1 and Brier scores. Findings suggest that an expanding window with lag structure and rolling window generally improve on both the optimized contemporaneous panel and individual country formats.

The algorithms in this paper include a non-parametric technique, regression algorithms, instance-based, regularization, and dimensionality reduction procedures, as well as decision tree methods and ensemble algorithms. Six of the thirteen algorithms reach the highest accuracy through the lag structure. Top-performing machine learning methods consistently produce high accuracy rates, on average above 70% for all derivates of the panel format, and frequently feature random forests and gradient boosting, the latter consistently outperforming commonly used regression algorithms. Compared to a non-parametric baseline, all top models add an 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 panel and individual country formats. Further contributions highlight how learning from other countries improves predictive strength, and non-linear models generally deliver higher accuracy rates than linear models. Algorithms keeping all variables perform better than those minimizing or excluding the influence of variables.

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% of the variation in crisis prevalence, the real sector 64%, and the external sector 8% over the full period.

The remaining 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 provides policy implications and Section 7 concludes.

2 | EMPIRICAL LITERATURE

In recent years, the adoption of machine learning methods has proliferated given its processing ability to analyze 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 programs further broadens the utilization of innovative methods. Yet, one drawback involves the inability of algorithms to compute the marginal contributions of each predictor or confidence interval for threshold levels (Joy et al., 2015). Furthermore, financial crises are regarded as rare events, but machine learning has demonstrated its ability to forecast rare events including pandemics (Coulombe et al., 2021).

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 organizations, and weak governing institutions as key contributory factors. Analyzing 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., 2015).

Surveying banking crises across a large group of countries during the period 1979 to 2003, Davis and Karim (2008), found domestic credit growth as the 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 bank credit. Expanding on the generalized CART methodology in studying episodes of systemic risk, Alessi and Detken (2018) highlight random forests to be reliable in their identification of leading signals. Employing CART and random forests to scrutinize 36 advanced countries during the period 1970 to 2010. Joy et al. (2017) found 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 the high predictive strength of CART and random forests in comparison to probit and logit models, signals and a Bayesian model averaging approach. CART reveals 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 horizons. 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, 2022) find adaptive boosting to outperform a logistic regression in forecasting financial crises between 1970 and 2017. du Plessis (2022a) highlights gradient boosting outperforming multiple outcome models including several machine learning models in crisis predictions. Fouliard et al. (2021a) show decision trees to outperform a regression model between 1985 and 2018, while Beutel et al. (2019) observe an opposite result. Ward (2017), Bluwstein et al. (2020), and Fouliard et al. (2021b) employ machine learning models to predict financial crises using the Macrohistory Database. While Ward (2017) applies classification tree ensembles against commonly used early warning models, Bluwstein et al. (2020) employ decision trees, random forests, support vector machines, and neural networks together with a logistic regression model. An enhancement by Ward (2017) entails creating a bigger number of derivations of the predictors to assess the value of modeling a larger number of variables compared to fewer variables. Bluwstein et al. (2020) reduced the sample size and focus on pre-crisis periods, with cross-validation as the main forecasting dimension, and using Shapley values to determine leading indicators. Fouliard et al. (2021b) likewise instate a narrower sample size with pre-crisis signals forecasted recursively out-of-sample. These three studies each use between five and six algorithms, with forecasting assessed using AUC. While Fouliard et al. (2021b) also consider RMSE to assess accuracy, Bluwstein et al. (2020) further identify leading indicators.

This paper differs from other studies and contributes to the literature in several ways. First, a horse race is instituted spanning 13 machine learning algorithms and thereby covering the modeling literature more broadly than studies only employing a few forecasting methods. Second, across four modeling dimensions, the predictive

strength of machine learning methods is assessed. These dimensions entail an optimized contemporaneous panel format with an expanding window, transformations with lag structure, and a rolling window as well as in individual format. This expands beyond one main dimension employed by other studies, thereby providing a richer understanding of the predictive value of the machine learning algorithms. Third, the paper investigates the leading indicators of the top-performing models, and provides an in-depth assessment of the economic drivers of financial crises, thereby verifying, and addressing some of the research lacunas, and highlighting prominent factors for ongoing monitoring. Fourth, the study builds more dynamic and comprehensive models and underscores the evolution of the economic drivers over a nearly 150-year period, with fixed capital formation found to be the most dominant factor. Fifth, an examination of economic sectors, by means of a panel format with a lag structure, highlights banking sector predictors describing on average 28% of the variation in crisis prevalence, the real sector 64%, and the external sector 8%. Sixth, several measures are employed to assess the forecasting performance of the machine learning algorithms, thereby improving on inquiries with a reliance on a single forecasting measure and including receiver operating characteristics with the area under curve estimates (AUROC), F_1, F_2, and $F_{0.5}$ scores, Brier scores, and a novel complexity measure. Seventh, algorithmic learning is assessed to determine if learning from other countries holds more benefits than learning from unique and own experiences. Eight, linear models are compared to non-linear models to determine optimal frameworks for rare events such as financial crises. Lastly, based on the workings of the algorithms, where some operate on all variables and others retain only the most important variables, the strength of these different modeling approaches is assessed.

3 | EMPIRICAL METHODS

This paper develops 13 machine learning models, all classified under the domain of supervised learning as it involves scrutinizing 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. Although the instance-based algorithm leans towards unsupervised learning through its clustering output and can be used as unsupervised learning, the optimizable distance estimator within the algorithm allows an option for supervisory input. Further, supervised learning is necessary given the objective of predicting financial crises, which requires the algorithm to learn from labeled observations.

Machine learning can also be used for classification and regression. Where machine learning models simplify functions to a known functional form, these can be classified as a parametric approach, whereas algorithms handling different functional forms would be categorized as a non-parametric approach. The models in this paper include a non-parametric technique, regression algorithms, instance-based, regularization, and dimensionality reduction procedures, as well as decision tree methods and ensemble algorithms. While neural networks were initially considered, reduced comparative accuracy to other machine learning methods in a recent study on banking crises by du Plessis (2022a), which could be attributed to the low frequency of crisis observations, resulted in its exclusion. Forecasting efficacy of these machine learning models is assessed through their performance on a test dataset of various dimensions. Mathematical descriptions of the methodologies and hyperparameter implementations feature in Appendix B. To optimize the machine learning algorithms, cross-validation is employed on the in-sample or training dataset. The process generally entails stratifying the training dataset into a number of folds, where during each iteration, one of the folds is used as the validation set while the remaining folds are used to train the algorithm. All folds are eventually used as a validation set. Based on the performance of the trained models on each of the validation sets, algorithms can further be enhanced through optimizing their hyperparameters in constructing the final model, and in turn used for recursive out-of-sample forecasting. While the cross-validation process is comparable between the different algorithms, there are nuances given the workings and nature of the optimization parameters, with some requiring initial parameters to be set to initiate the process by which the algorithms can tune and optimize hyperparameters. Hyperparameters are summarized in Table B1 (Appendix B).

3.1 | Benchmark algorithms

To allow testing across a broad spectrum of models, simple to more complex algorithms are developed and employed. Serving as a 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, where the dependent variable is, respectively, a linear and non-linear function of the regressors, also with dissimilar distributional assumptions. The baseline model

generates an approximation of future expectations based on past values and is less sensitive to low and high outliers. As the baseline model continuously recalibrates after each forecast based on new information, it serves as a dynamic benchmark that is continuously learning from incoming data. The output signifies a long-term trend, which is also useful to describe a fundamentals-based forward-looking path.

Although the linear probability model encounters statistical challenges for dichotomous dependent variables, practically, results can be comparable to logistic models especially if most of the modeled probabilities are between 0.20 and 0.80, which then overlap with the logit model, and further assist interpretability (Hellevik, 2009). While a logit model was also considered, notwithstanding little difference in the outcome of probit models (Greene, 2012) its theoretical limitations dealing with random variations of independent variables and ability to handle panel data when unobserved factors are correlated over time, which are all solved by probit, eventuated in the selection of the latter. One drawback of probit is the assumption that unobserved components are normally distributed (Train, 2002). Furthermore, the selected modeling techniques are frequently utilized in forecasting financial distress. Formulations of the methods are discussed in Appendix B.

3.2 | Instance-based algorithms

Instance-based algorithms comprise k-nearest neighbors and support vector machines. The k-nearest neighbors (k-NN) algorithm involves the estimation of the conditional distribution of Y given X in order to categorize an observation according to the outcome class with the highest estimated probability. 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.

3.3 | Regularization algorithms

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

In contrast to the ordinary least squares statistical technique which computes slope coefficients $\beta_0, \beta_1, ..., \beta_p$ by employing values that minimize the residual sum of the squared equation, ridge applies tuning parameter $\lambda \geq 0$, where the shrinkage penalty is small when $\beta_1, ..., \beta_p$ are near zero, so it reduces the estimates of β_j toward zero.

Analogous to the 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 the 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 et al., 2013).

3.4 | Dimensionality reduction algorithms

Partial least squares (PLS) regression serves as a dimension reduction method by detecting a new set of features $Z_1, ..., Z_m$, which are linear combinations of the initial features, and subsequently fitting a linear model using least squares. As the PLS approach identifies new features, which approximate the original features that are associated with the outcome variable Y, it explicates the outcome and explanatory indicators (James et al., 2013; Wold, 1985). Principal component analysis was also considered, a comparable yet more limited method to PLS, which estimates fewer linear combinations of the independent variables through a parsimonious summarization approach, but in contrast to PLS, does not take into account how each predictor is related to the outcome variable (Bair et al., 2006; Kleinbaum et al., 1998; Rosipal & Krämer, 2006). As it would entail a weaker understanding of how the crisis outcome is influenced by the economic information, resultantly only PLS was retained in the study.

3.5 | Decision tree algorithms

3.5.1 | Full tree

The implementation of the classification and regression trees (CART) algorithm accentuates 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 operates with missing values. The CART algorithm allows non-linear relationships, implements threshold levels, and supports interactions between variables. Resultantly, relationships between predictors could fluctuate given cross-sectional and time dimensions. By analyzing all data observations, specification errors are minimized. Relevant to crisis literature, the CART method ranks predictors according to their level of importance, thereby rendering leading indicators. Indeed, results from classification and regression trees are straightforward to interpret and a practical instrument for policy makers.

However, the classification and regression trees approach encounters 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 analyzed from the full sample, which results in a local rather than global optimum. In comparison to regression models, as probability distributions are not operationalized, confidence intervals cannot be computed. Given that an individual probability value is allocated to all observations within a categorized set, marginal contributions of the explanatory variables are not estimated, even though the variation in the 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., 2015).

CART algorithm implements a top-down approach to partition data recursively, involving several predictors. Originally, through a partition with one predictor, a parent node is formed. Subsequently, dividing 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 signifies 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 that best classify crisis and no-crisis episodes.

3.5.2 | 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.

3.6 | Ensemble Algorithms

Ensemble algorithms operationalize a cohort of weak learners to jointly construct a strong learner, with the goal of improving 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 modeling frameworks comprise bagging and boosting. While both modeling approaches produce new data in the training environment through sampling by replacement, bagging assigns the same probability of replacement 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 penalizing misclassified observations or through shrinking a loss function. Strong learners are determined using a 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.

3.6.1 | Adaptive boosting

Adaptive boosting or AdaBoost represents one of the initial boosting algorithms. Distinctively, while the classification and regression trees algorithm constructs 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 the replacement procedure. Each stump is also assigned a weight given its computed prediction error, which further denotes its contribution to the strong learner.

3.6.2 | Gradient boosting

As an extension of the AdaBoost approach, gradient boosting is a variation that employs 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 optimize 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 optimize and develop the final model (James et al., 2013).

3.6.3 | Random forests

Random forests (RF) algorithm employs a group of weak learners to jointly create a strong learner, a process centered 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 & Ormerod, 2016). RF method addresses 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., 2015). Analogous to classification and regression trees, the RF algorithm can process sizeable datasets, is not sensitive to outliers, models interactions between explanatory variables, and is not limited by distributional assumptions.

Random forests algorithm permits optimization through stipulation of tree complexity or depth, the quantity of variables featuring in each tree, bootstrap sample size, and the quantum of trees (Mullainathan & Spiess, 2017). Drawbacks of the RF 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., 2015), 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 judgment. This paper utilizes the definition from Laeven and Valencia (2012), which describes 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 et al. (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.

Each instance of systemic financial crisis is represented by a categorical variable, expressed by $Y_{it} = 0$ for a no-crisis episode and $Y_{it} = 1$ as a proxy for a crisis event, where $Y_{it} = 1$ is not limited to the onset of a crisis, but is based on duration, and references all time periods where a crisis is present and in progress. 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% of all observations classified as $Y_{it} = 1$. Given that machine learning models represent novel approaches to dealing with financial crises, the low prevalence of crisis episodes can be expected to constrain some models to function optimally. Whereas the commonly used models might perform differently 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.

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

While more than two outcomes were considered, such as the post-crisis period as employed by Bussière and Fratzscher (2006) and du Plessis (2022a), the main focus on crisis prevention, which is aimed at mitigating its severe impact and resultant costs, is concerned with optimizing correct signals for actual crisis events. By instituting a lag structure and optimized contemporaneous structure as two of the modeling dimensions in this paper, both the pre-crisis and crisis periods can be studied to ensure timely early warning signals.

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; du Plessis, 2022a, 2022b; Hardy & Pazarbasioglu, 1998; Vlaar, 2000). Specifically, González-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 emphasize 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, elicits credit demand. The capital output ratio functions as a proxy for the 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 employment, aggregate output, and income growth, further encumbers the ability of households 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 of 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 bank 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. Demirgüç-Kunt and Detregiache (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. Banking crises can be multinational in nature, with weaknesses spread across interlinkages between countries, as conveyed through the external sector variables. A steep currency depreciation, following reversals in capital flows, could result in a slump in asset values, and a 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 declining terms of trade are 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 A1 (Appendix A). Data are obtained from Jordà-Schularick-Taylor Macrohistory Database (Jordà et al., 2017) and consist of an annual time series. Another consideration includes the 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 variables constituting 29,988 observations. As all countries in the database experienced financial crises, no countries have been excluded from this study. While the annual time series nature of the database could be a limitation for modeling more immediate events, the advantage of the database is its long-term horizon spanning a century and a half. This allows the models to learn from more diverse and richer experiences. To provide lead time, a lag structure is used as the primary modeling dimension. The final dataset is available from the Harvard Dataverse (du Plessis & Fritsche, 2023). A representative sample of countries stems from Australasia, Europe, and North America. According to Table A2 (Appendix A), 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 was observed from 1870 to the Second World War, which resumed in 1974, following the great moderation. In particular, the crises of 1907–1908, 1929–1931, and 2007–2008 were more ubiquitous and global in nature, impacting more than 50% of the sampled countries. The Global Financial Crisis had the largest scale, comprising 70% of all the countries.

As the first step, all variables in the Jordà-Schularick-Taylor Macrohistory Database were considered, but only those variables that were congruent with economic theory and literature findings on precursors to financial crises were used. One of the critiques of machine learning is that it is more challenging to determine causal relationships compared to traditional regression models. Therefore, by relying on economic theory, the potential limitation of clearly attributing causality is addressed.

The second step thereafter retained variables in the model if their statistical power in a traditional probit regression model is found to be significant. The development of the benchmark for comparability was deemed important in this study in order to directly evaluate the performance of the machine learning models in contrast to commonly used regression models. As a key aim of the study is to determine the predictive accuracy of machine learning models and to contextualize its performance relative to traditional probabilistic models, where the latter are informed by diagnostic assessments such as stationarity and statistical significance, and variables are transformed based on the diagnostic results, and subsequently variables with poor levels of statistical significance are excluded. This serves to demonstrate whether machine learning models are better than a best-in-class probit model, and if affirmative, by how much the former improves on commonly used models.

To counter stationarity, ratios, first difference, and log forms are employed, with a number of lags included based on adequate statistical significance from *p*-values below 0.1 and so within the acceptance region, while real transformations confine the influence of inflation. Unit root tests for the probit model produce satisfactory results as described in Table A3 (Appendix A).

These steps ensure that findings are based on economic theory, causality is implicit, variables are selected based on statistical significance, and the methods can be compared to commonly used econometric regression models.

4.2 | Significance of individual variables

To verify if variables are significantly different between crisis and no-crisis periods, and test for equality of means, a two-tail *t*-test is used. The sample means for

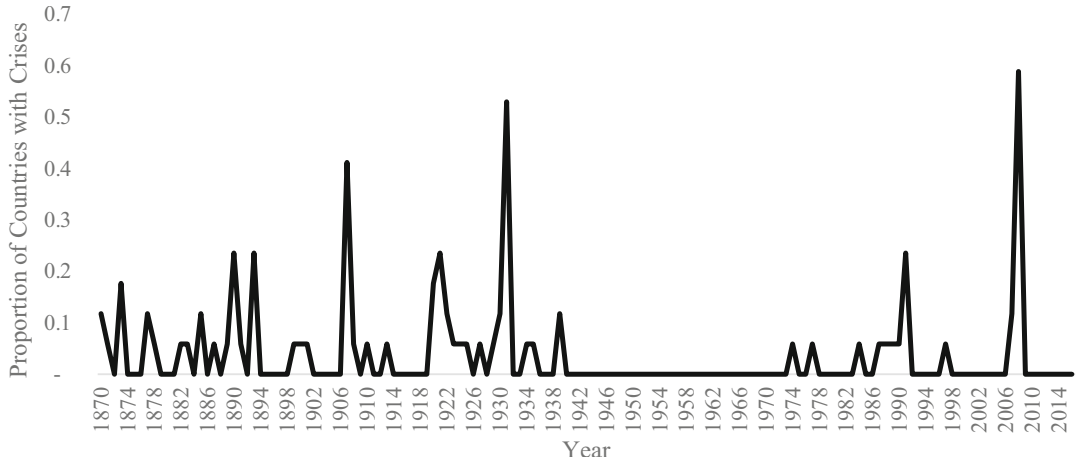

FIGURE 1 Proportion of countries with crises.

the three sets of indicators, encompassing real, banking, and external sectors are described in Table A4 (Appendix A), 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 a 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 a 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. The 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 accentuates 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% confidence level, consumer inflation, long-term rates, and debt-to-GDP at 95%, and total loans at 90% confidence levels. In the case of the external sector, the current account is dissimilar at 95% confidence levels. The robust statistical significance between the crisis and non-crisis observations further demonstrates the limited influence that a post-crisis period could have on no-crisis observations, which was identified by Bussière and Fratzscher (2006) as a useful period to consider in the modeling process. The low influence could also be explained by the small amount of 90 post-crisis observations in the 2,499 observations dataset.

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

5 | EMPIRICAL RESULTS

5.1 | Modeling dimensions

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. All these modeling dimensions are mathematically described in this section, as based on the benchmark probit regression. While all machine learning models employ the same transformed variables, relationships are not expected to be linear as in the traditional regression equations or to feature every variable where only some are retained as in the case of regularization and dimension reduction algorithms and pruned trees. The formal equations for each of the four modeling dimensions denote the commonly used probit benchmark and describe the input–output framework, while the inner workings and mapping vary by machine learning method.

As the main modeling dimension, and aimed at providing an immediate early warning signal, a standard one-period lag structure is employed for all variables in the panel format. Given the low prevalence of $Y = 1$, the machine learning algorithms are modeled across the panel dataset, which encompasses a time series of the same cross sections, the latter comprising all the countries in this study. Collective and faster algorithmic learning is enabled through a larger sample size, more variance in the predictors, higher degrees of freedom with more crisis episodes, and underscores a practical approach to assess financial catastrophes given global interlinkages. Formally, and with variables captured in Table A1 (Appendix A), the benchmark regression model can be stated as

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

where $Y_{i,t}$ is the crisis index, N the number of countries, T the entire time period and $\varepsilon_{i,t}$ stochastic error term. Benefits of the lag structure include faster response times as the release of annual data could follow after the commencement of a crisis in the same or previous year, and leveraging off pre-crisis signals to transmit more expeditious early warnings.

To verify whether the main panel format with lag structure delivers the highest predictive strength, the second modeling dimension applies a contemporaneous structure with optimized statistical properties as described in Table A1 (Appendix A). Transformation of variables in this model is done based on achieving optimal statistical significance of the commonly used probit model to ensure higher statistical power and robustness. Most variables enter the probit model absent any lags, thereby reflecting a contemporaneous structure. This ensures that an optimal version of a benchmark model is constructed to compare forecasting performance against the machine learning models, and in the case of outperformance by the latter, to contextualize the effectiveness of the novel methods given a frequently used alternative as standard. Mathematically, the traditional probit equation is denoted as $Y_{i,t} = c_i + \sum_{i=1}^{N}\sum_{t=1}^{T}\sum_{j=1}^{K}\sum_{l=1}^{L}\beta_j x_{j,i,t-l} + \varepsilon_{i,t}$, 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 applying the optimized contemporaneous structure employed by the second modeling dimension. Technically, the traditional probit regression is 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 the level of economic development is not comparable after several decades, a rolling window of 20 years is employed to assess forecasting performance. Also based on the optimized contemporaneous framework, this dimension can be formulated for the benchmark probit 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 the long-term panel, in optimized, disaggregated, lag, or period-bound dimensions is 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.

5.2 | Performance assessments

5.2.1 | Recursive out-of-sample forecasts

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, and as based on the commonly used probit model, $Y_{i,t+h} = c_{it} + \beta_j x_{j,i,t+h} + \varepsilon_{i,t+h}$, 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 optimized contemporaneous and lag formats, as well as for individual countries. The expanding window structure employed for the first three modeling dimensions simulates the policy-making process which is based on available information up to a point in time. Given the low frequency of crises, an updating dataset also provides richer information from which the models could learn. 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. A benefit of a rolling window is a focus on a specific crisis period for instance the Asian Financial Crisis of the 1990s or the Global Financial Crisis in the late 2000s. The performance of individual countries is modeled separately and reported in both individual and aggregate formats for comparability. The starting date for all model forecasts is based on available degrees of freedom and commences from 1886 onwards, with

recursive forecasts conducted on an annual basis until the end of the period under review.

5.2.2 | Benchmark probit model

Serving as a regression algorithm, a probit model, which is widely employed by policy makers 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 optimized contemporaneous probit models are described in Table A5 (Appendix A). 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. Gauged through diagnostic tests, statistically significant properties ensure that the standard probit model is adequately constructed and can serve as an appropriate benchmark for comparative performance.

Given the interrelatedness of macroeconomic variables, Granger causality is implemented to gauge the degree to which an explanatory variable, x, can predict the dependent variable, y. This is operationalized by adding lagged values of x, to an autoregression of y, so that $y_t = c + A_1 y_{t-1} + B_1 x_{t-1} + \ldots + A_p y_{t-p} + B_p x_{t-m} + e_t$. The number of lags is based on the lowest values of AIC and BIC and is limited to four lags. Granger causality is an F-test on joint significance, where $H_0 : B_1 = B_2 = \ldots = B_p = 0$. For the panel data, the Granger causality test is based on Dumitrescu and Hurlin (2012), which allows individual coefficients across cross-sections, so test statistics are estimated for each individual country to account for heterogenous factors, and then averaged across all countries. Rejecting the null hypothesis of no Granger causality signifies that a variable influences another variable (Granger, 1969). The procedure is repeated to determine the reverse causality between the dependent variable and the explanatory variable.

During a crisis, as measured by the dependent variable, the impact of a financial meltdown has been shown to impact in turn the level and direction of macroeconomic variables (Behringer et al., 2017). This is described in Table A6 (Appendix), where real sector variables consumption expenditure and fixed capital formation exert influence on the formation of a crisis across three lags, with reverse causality transpiring with the fourth lag. Debt shows a comparable trajectory. Capital output ratio features an inverse result, which can be explained by the inclusion of gross domestic product within the ratio, the latter observed to be affected over the short-term by a crisis. Loans underscore bidirectional influence across most lags, whereas exchange rates only with long lags. Crises are further shown to influence the stock market, inflation, long-term rates, and current account balance.

5.2.3 | Performance measures

Performance assessment criteria include area under the receiver operating characteristics curves (AUROC), F_1 measures and Brier scores. Employing a range of measures avoids an overreliance on one main measure. Notwithstanding, results exhibit a high correlation between the measures, which are constructed to individually emphasize distinctive parameters. While receiver operating characteristic (ROC) curves constitute a visual representation of the true positive rates by false positive rates, the area under curve (AUC) summarizes the outcome into a single value. True positive rate (TPR) is also referred to as sensitivity or recall and comprises 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. False positive rates (FPR) or 1-specificity consist of false alarms (FP) as a ratio to the collective false alarms and true negatives (TN), denoted mathematically as $FPR = \frac{FP}{FP+TN}$. 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 score signifies a correct set of forecasts. (DeLong et al., 1988; Fawcett, 2006). As AUC is an aggregate measure that averages over all possible thresholds, it is not dependent on classification thresholds. The result is based on a 95% confidence interval and computed with 100 bootstrap iterations. Yet it is possible for policy makers to derive a threshold value that maximizes sensitivity and specificity by employing measures such as the Youden Index (Youden, 1950). AUC serves as a general performance measure to indicate overall predictive strength, where the policy responses to a false crisis and a missed crisis are both costly and could result in recessions. When classes are imbalanced, true and false positive rates can be individually assessed, and other measures such as F_1 scores can also be considered.

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. 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. In contrast to AUC, F_1 focusses more on the performance of the positive class, which is a useful measure when the positive class is rare and false positives are not as costly as false negatives. To compute the threshold that results in the optimal balance between precision and recall, F_1 allocates equal weight to precision and recall, to express the harmonic mean of the two fractions as a single value. Maximizing precision, minimizes false positives, and maximizing recall minimizes false negatives. As robustness tests, two further F measures are developed. Formally, in $F_\beta = \frac{(1+\beta^2)*Precision*Recall}{(\beta^2)*Precision*Recall}$, where F_1 has a beta equal to 1, the two additional measures F_2 incorporates a beta of 2 and $F_{0.5}$ accordingly a beta of 0.5. While F_2 places more weight on recall, thereby on minimizing false negatives or the failure to signal a crisis, $F_{0.5}$ assigns more weight to precision, thereby minimizing false positives or false alarms.

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.

To synthesis the three sperate measures, an overall ranking is estimated as a function of AUC + F_1 Score - Brier Score, where the highest values are indicative of topmost predictive strength. While all three measures are commonly used to assess predictive strength, with measures falling between zero to one, each highlights a nuanced aspect of performance that is deemed more essential. For AUROC and F_1 score, the aim is to measure the share of crises signaled or missed. For Brier, the difference between actual and predicted values, and through a distance estimator, likewise underscores the difference between crises correctly or incorrectly called. While AUROC can be considered the main individual measure given its wide adoption in the machine learning literature, there is a high level of correlation, exceeding two-thirds of the time, between the three measures.

5.3 | Recursive out-of-sample crisis forecasts with lag structure

The main modeling dimension encapsulates all the data in a panel 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 the 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 an input into the policy-making process to prevent and mitigate ensuing financial crises. Predictive strength for all the algorithms is assessed through recursive out-of-sample forecasts.

5.3.1 | Variable importance measures

A benefit of machine learning methods entails the identification of the most important explanatory indicators. This is achieved by analyzing 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). An importance score is estimated for every individual decision tree, based on the value by which a variable employed at a split point enhances the performance measure, which is weighted by the number of observations where that variable is used. Technically, $I_{\ell k}$ indicates the significance of the variable X_ℓ in partitioning the class k observations from other classes. The overall expediency of X_ℓ is computed by averaging over all of the classes, where $I_\ell^2 = \frac{1}{K}\sum_{k=1}^{K}(I_{\ell k})^2$ (Hastie et al., 2009). The Gini index expresses the importance of variables relative to each other and is constructed on a 0–1 scale, where one has a higher relative importance. The Gini index of the gradient boosting algorithm is employed as the variable importance measure.

As a robustness test, random forest variable importance 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 described 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 utilized 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 utilization 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).

5.3.2 | Variable importance results

Based on the Gini index of the gradient boosting algorithm, which frequently outperforms among 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 sample by means of the panel format and which are recursively estimated by adding one new period at a time to forecast the next period as based on an expanding window, while Figure A1 (Appendix A) shows predictors individually on a scale of 50.

According to the findings from the panel with lag structure, fixed capital formation exerts the single most influence, from around 20% at the turn of the 20th century, spiking to 50% the year before the 1907 banking crisis, followed by stabilization and a gradual increase over the subsequent decades, reaching above 40% 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% at the end of the sample, hovering around 10% for most of the period under review. While these variables stand out, significant fluctuations in the levels of other variables are observed during specific developmental epochs. For instance, debt-to-GDP spiked above a 35% level of influence around the banking crises in the 1890s, while the stock market remained above 20% in the years leading up to the 1907 crisis. Consumption expenditure peaked at 20% 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 crises (Kindleberger, 1978; Reinhart & 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.

On average, as summarized in Figure A2 (Appendix A), and given the lag format, banking sector variables constitute 28% of the variation in crisis prevalence, the real sector 64%, and the external sector 8%, highlighting the impact of real sector developments in contributing to banking sector vulnerabilities, thereby underscoring its consequential role. After peaking around a 65% level of importance in the 1880s, banking sector predictors declined in prominence until the 1910s and drifted upwards above 30% in the lead-up to the Great Depression, after which it fluctuated within a 20–30% band until the start of the 21st century. The real sector demonstrates an inverse trajectory, gradually increasing from around a 30% level of importance in the 1880s to over 60% in the years before the start of the Great Depression in 1929. During the subsequent eight decades, real sector variables continually contributed on a large scale to the underlying causes of financial crises, spiking to 70% at the start of the Global Financial Crisis. The lag

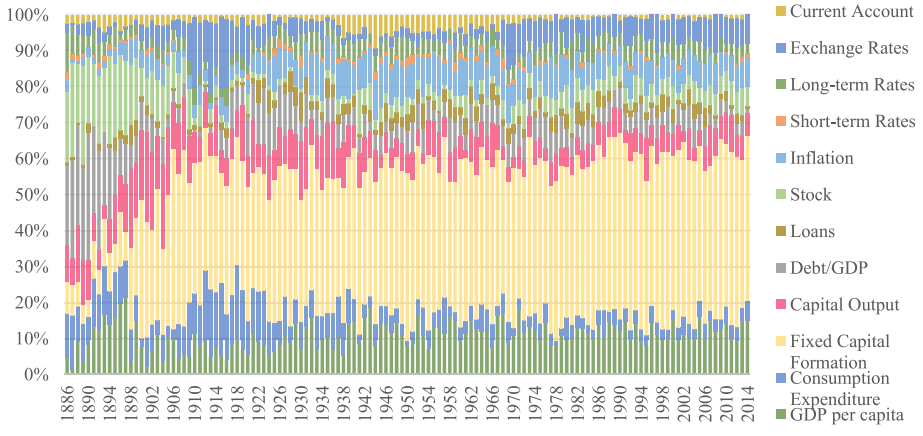

FIGURE 2 Variable importance over time.

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

Findings from the robustness test using random forests, as denoted in Figure A3 (Appendix A), broadly confirm the leading indicators. Fixed capital formation takes a poll position in reducing the mean squared error and sum of squared error 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.

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%, real sector on 53%, and the external sector at 7% 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.

5.3.3 | Recursive out-of-sample forecasts results

Table 2 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% level of accuracy, against a non-parametric yet dynamic baseline of 53%. F_1 shows a comparable result, with pruned trees followed by gradient boosting and random forests. Brier scores are lowest for the support vector machine, followed by a 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 + F_1 Score - Brier Score, where the highest values are indicative of topmost predictive strength, with random forests first, followed by gradient boosting and support vector machine. Although the individual assessment measures are generally comparable, in combination these measures provide a broader evaluation of overall performance as constructed through the underlying and nuanced criteria of each measure. Overall AUC predictive accuracy across all algorithms reaches 64% for the lag structure. In Table A7 (Appendix A), a robustness test employing the three different F measures, and in comparison to harmonic mean F_1, highlights a reduction in F scores when using $F_{0.5}$, weighted in order to minimize false alarms, and an increase in scores when applying F_2, geared towards minimizing failure to signal a crisis. All three measures move in tandem, with a high correlation of 95.2% between F_1 and F_2, and a near perfect correlation of 99.8% between F_1 and $F_{0.5}$. Overall ranking in combination with AUC and Brier scores when substituting the F measures, underscores an unchanged outcome. Resultantly, F_1 as harmonic mean features as the selected F measure in the remaining assessments.

5.3.4 | Robustness test: Recursive out-of-sample forecasts of economic downturns

Given that financial crises are rare events, a robustness test is instituted to model economic downturns using the same dataset and lag structure. The latter to provide signals in advance. While financial crises happen on average once every 28 years, economic downturns, categorized as periods with negative real gross domestic product growth, transpire every four and a half years according to the longitudinal dataset. Benefits of modeling downturns include more observations to train and test the models. It features a dependent variable series which is readily available as based on market performance, in contrast to the financial crisis series which is classified by in-depth studies and assessments. To ensure comparability to the existing modeling framework, the same variables used for financial crises are employed. Based on findings in Table A8 (Appendix), the same top three models, namely support vector machine, random forests, and gradient boosting features with the highest predictive accuracy. The lower overall accuracy of 61%, using AUC, compared to 64% for financial crises (in Table 2) using the same modeling approach, while comparable, could be due to variable selection, which is informed by literature findings on precursors to financial crises. Yet the consistency in modeling performance highlights both the models' accuracy and practical relevance.

5.4 | Recursive out-of-sample crisis forecasts in contemporaneous format

As robustness tests, three more forecasting frameworks are developed. The second modeling dimension also entails combining all 17 countries in an optimized panel format across the period under review. In contrast to the

TABLE 2 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 Neighbors	0.498 [0.497, 0.499] *0.000*	0.000	1.000	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

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

lag model, the aim is to verify if predictive accuracy improves without the benefit of enhanced lead time, and instead through the use of predominantly contemporaneous indicators, as informed by the optimal statistical significance of the variables in the probit model. Similar to the panel model with lag structure, this approach allows the machine learning methods to observe and learn from the experiences of all countries, and utilizes a comprehensive dataset in the context of a low-frequency event, to build and calibrate each model in a recursive manner in order to operationalize out-of-sample forecasting.

5.4.1 | Optimized contemporaneous panel: Variable importance results

Comparatively, in applying an optimized contemporaneous structure, which is geared towards identifying predictors at the time of an actual crisis, by means of the gradient boosting algorithm, banking sector influence increases to 50% while the real sector reduces to 42%, and the external sector remains unchanged, thereby emphasizing the dynamic adjustments of leading indicators one year preceding a crisis compared to the year of a crisis. In contrast to the lag format, where the real sector appears more dominant, assessing the results in conjunction with the contemporaneous structure, highlights an interplay between the real sector and banking sector over time, which exemplifies the sequential role of vulnerabilities in the real sector propagating to the banking sector in the lead up to the crisis.

5.4.2 | Optimized contemporaneous panel: Recursive out-of-sample results

Recursive out-of-sample results for all countries in optimized contemporaneous panel format are summarized in Table 3. In terms of AUC, gradient boosting is the best-performing model followed by random forests. Linear probability, 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 the lowest error, followed by ridge. A comparison of the three measures demonstrates that AUC correlates 67% of the time with F_1 scores, with the latter showing a negative correlation with Brier score of 75%. Based on the combined ranking across all three measures, gradient boosting and random forests constitute the top two algorithms followed by ridge. In contrast to the lag structure dimension (overall AUC of 64%), the optimized contemporaneous format is shown to register lower average results of 61% across all algorithms, emphasizing the forecasting benefits of detecting vulnerabilities with lead time.

5.4.3 | Individual countries: 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 the same type of rare events from other countries could improve forecasting performance for an individual country. While

TABLE 3 Recursive out-of-sample forecasts in an optimized panel 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 Neighbors	0.499 [0.498, 0.500] *0.000*	0.000	1.000	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

Notes: 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% confidence intervals. Standard errors in italics.

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 graphically summarized for each country in Figure A4 (Appendix A), amounting to an average of 62% across all the countries.

Table A9 (Appendix A) highlights the top-performing model per country and the deviation to both baseline and across all models. Accuracy rates range from 60.3% in the case of Germany to 94.3% 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 support vector machine, 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 Australia, Finland, Italy, Sweden, Switzerland, and the UK.

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

5.5 | Recursive out-of-sample crisis forecasts for individual countries

The third modeling dimension revolves around the individual experience of each country. In contrast to the optimized panel format in the second modeling dimension, models only take into account the knowledge of the experiences that 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 aggregate format, the latter a combination of the former.

5.5.1 | Individual countries: Variable importance results

Variable importance results from the gradient boosting algorithm are combined across all years for each individual country and summarized in Figure A5 (Appendix). Fixed capital formation emerges as a leading indicator across most countries, similar to the overall findings (in Figure 2). Other variables include the stock market, consumption expenditure, debt, capital output ratio, and short-term interest rates. While there exists similarity across most countries, there are notable exceptions such as the strength of gross domestic product and current account balance in the United Kingdom, loans and exchange rates for France, and short-term rates together with fixed capital formation in the United States. These nuances highlight the ability of the models to detect

influential variables causing banking crises on an individual country level, underscoring the models' regional applicable and practical value.

5.5.2 | Individual countries: Recursive out-of-sample results

As shown in Table 4 and in aggregate format, the five best-performing methods are adaptive and gradient boosting, linear regression, full tree, and lasso, on average slightly under or above 60%. In terms of F_1 scores, full tree, k-nearest neighbors, ridge, and gradient boosting 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. With average AUC results of 56% across all models, in contrast, the panel format in the second modeling dimension shows how knowledge from other countries somewhat improves the average aggregate outcomes to 61% (from Table 3). In comparison, the non-aggregate format displayed graphically on an individual country level by mean AUC in Figure A6 (Appendix A), highlights a narrowing in the deviation between the two approaches, at 61% to 62% (as shown in Figure A4) for the panel format. However, when comparing the best-performing models between the two approaches as shown in Table A10 (Appendix A), the inverse transpires, yet at marginal levels, with the individual country format on 81% to the aggregate country results from the panel format on 80%. Across both formats, the top models therefore correctly predict at a high accuracy rate (80%) 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. Although the experience of other countries improves accuracy rates overall and for most countries, Germany is an exception with better results observed from models tailored to the country's individual experiences.

In comparison to the aggregate results for individual countries in the panel format, 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 add the most value. Although the complexity encountered is slightly less for the panel format, with an average difference of only one percentage point to that of individual country format, the variability in predictors might result in models becoming somewhat better equipped to handle more complex datasets. Similar to the individual country results derived through the panel 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.

5.6 | Rolling window out-of-sample crisis forecasts

As the fourth modeling dimension, a new configuration is applied to the panel format. Instead of increasing the

TABLE 4 Recursive out-of-sample forecasts for individual countries in aggregate format.

Method	AUC	F_1 Score	Brier Score	Rank
Baseline	0.501 [0.434, 0.567] *0.034*	0.073	0.080	7
Linear Prediction	0.602 [0.536, 0.668] *0.033*	0.092	0.220	8
Probit Regression	0.543 [0.474, 0.612] *0.035*	0.074	0.970	11
K-Nearest Neighbors	0.503 [0.490, 0.516] *0.006*	0.107	1.000	12
Support Vector Machine	0.559 [0.497, 0.621] *0.031*	0.082	0.040	3
Ridge	0.530 [0.481, 0.580] *0.025*	0.106	0.400	9
Lasso	0.582 [0.521, 0.643] *0.031*	0.075	0.090	4
Partial Least Squares	0.499 [0.497, 0.500] *0.000*	0.000	1.000	13
Full Tree	0.593 [0.535, 0.650] *0.029*	0.107	0.070	2
Pruned Tree	0.501 [0.432, 0.570] *0.035*	0.076	0.070	6
Adaptive Boosting	0.658 [0.597, 0.720] *0.031*	0.067	0.500	10
Gradient Boosting	0.647 [0.587, 0.708] *0.030*	0.094	0.100	1
Random Forests	0.537 [0.477, 0.597] *0.030*	0.076	0.100	5

Notes: 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% confidence intervals. Standard errors in italics.

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 contemporaneous events and occurrences that took place over a century ago might be limited, and 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 the leading indicator over the 147-year period, a standardized 20-year window is employed, executed on a rolling basis, through which the time-bound focus allows events to be modeled and forecasted around a comparable period.

As shown in Table 5, average mean values of 64% across all algorithms are similar to the panel format with a lag structure. Top performing methods as based on forecasted accuracy consist of random forests and gradient boosting, followed by probit and linear probability regressions and ridge. Random forests and gradient boosting generate high F_1 scores with mid-tier Brier scores. Combined top models comprise gradient boosting, random forests, and linear probability regression. According to AUC, a probit regression achieves a 73% accuracy compared to the 77% of gradient boosting and random forests. The four-percentage point difference between the models over 17 countries and close to 2,499 annual forecasts translates to around 100 more forecasts correctly called by the top two ensemble algorithms compared to the traditional probit regression.

5.7 | Ranked methods across forecasting models

Across the four modeling dimensions, from optimized contemporaneous panel 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 probability 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 model performance as it became better equipped to model the underlying dataset and predict an ensuing crisis.

Summarized by the highest AUC mean value for each method specific to the associated top dimension, Table 6 underscores the variability and improvements across the four dimensions. Accordingly, six of the thirteen models reach the highest predictive strength through the lag structure, four through the standardized rolling window, two within the optimized contemporaneous panel format, and one when employing the aggregate format comprising individual countries. When applying this combination, average mean AUC values increase to 65%, with the top two algorithms featuring random forests and gradient boosting, both on 77% overall accuracy rates across 17 countries and 147 years. Notwithstanding, average AUC mean values increase above 80% for top individual country models, both in panel and aggregate

TABLE 5 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 Neighbors	0.499 [0.497, 0.500] *0.000*	0.000	1.000	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

Notes: 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% confidence intervals. Standard errors in italics.

Methods	Top Dimension	AUC	F_1 Score	Brier Score	Rank
Baseline	Panel	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 Neighbors	Individual	0.503	0.107	1.000	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	Panel	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

TABLE 6 Top recursive out-of-sample forecasts across all formats.

format. Across all three measures, the top models are gradient boosting, random forests, and support vector machines. While both the lag structure and rolling window deliver 64% overall accuracy rates, the former encompasses the 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 modeling tools for leaning against the wind to prevent cleaning up after the bust.

Overall, most models outperform the baseline model, which is a dynamic model but comparatively more stable than the other models. In an environment where the outcome is rare, relying on a baseline model which provides a more fundamental view of developments, is expected to deliver less false positives and more true positives of non-crisis events. Yet overall, the highest performance reaches 56%.

Non-linear models are shown to deliver higher accuracy rates compared to linear models, which can be expected given that relationships are non-linear in the formation of assets, debt, fixed capital formation, and consumption bubbles, and break down during episodes of distress. SVM performs better than its linear equivalent k-NN, also exacerbated by the low frequency of crisis events. In addition, models geared towards dimension reduction, higher weighted regularization, and tree pruning, which results in the exclusion of variables are also shown to exhibit comparatively lower performance strength. This explains why ridge outperforms lasso, PLS underperforms the baseline model and full tree generally delivers better results than a pruned tree. A finding corroborated in a study by Ward (2017), showing larger numbers of variables improve predictions.

Traditional regression methods, probit, and linear probability models, perform above average according to mean AUC values, respectively third and fourth best across the lag and rolling window structure, and third and fifth best across the optimized panel format. However, performance is more mixed for individual countries. Based on the overall results combining three assessment measures, the linear probability model reaches a top four position after gradient boosting, random forests, and support vector machine.

6 | POLICY IMPLICATIONS

Systemic financial crises are rare events, yet with debilitating ramifications. Its unique nature requires novel modeling approaches. The interplay between the real sector and the banking sector exemplifies the sequential role of vulnerabilities in the real sector propagating to the banking sector over time, with more stress in the banking sector contributing to failures. While capital formation constitutes an important variable in contributing to the formation of banking crises over the long run, other real, as well as banking and external sector variables continue to evolve in prominence over time and require ongoing monitoring.

New methods in the form of machine learning algorithms are shown to improve the prediction of an old problem. Across four modeling dimensions, a lag structure and rolling window are more conducive to optimize forecasting performance. A panel format, that is based on the communal experience of a large group of countries, generates higher accuracy levels of forecasts for individual countries, given more episodes to learn from the

experience of other countries, and exposure to diverse environments. Furthermore, non-linear models deliver higher accuracy rates compared to linear models, which underscores the deterioration of relationships during episodes of distress. In addition, models accommodating more variables, and not excluding the influence of variables generally deliver higher performance strength. Ensemble algorithms in general, and gradient boosting, and random forests in particular, are consistently top-performing models over a long-run horizon and provide policymakers with an enhanced modeling toolkit. Yet, commonly used regression methods, probit, and linear probability models perform above average. Variations in performance in individual countries underscore the value of employing a diverse set of modeling tools for leaning against the wind to prevent cleaning up after the bust. The study on 147 years of crises highlights the robust performance of novel modeling methods over the long run. Higher predictive strength could further reduce and contain resolution costs.

The old and ongoing problem of recurring financial crises, as quantified in the historical dataset spanning a century and a half, highlights that most countries continue to experience further financial crises and are not able to graduate from these crises. In 2023, countries including the USA and Switzerland are experiencing several bank failures. So, the aim of this study is to address this old and ongoing problem with novel modeling tools, in the form of machine learning methods, to better forecast financial crises over a long-run horizon.

Based on the forecasting accuracy of these novel methods, it would have been beneficial for a policymaker in the 1800s and 1900s to use machine learning rather than commonly used alternatives such as the baseline and probit models. Robust results over the long term, recursively forecasted as if policymakers lived during the past century and a half, underscores the potential of these novel models for the long run future.

7 | 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 includes a baseline model as a non-parametric approach as well as linear probability and probit regressions to serve as a common benchmark. Instance-based algorithms comprise k-nearest neighbors, which categorize 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. Regularization algorithm ridge reduces less significant coefficients towards zero, while in the case of lasso, some coefficient estimates equate to zero. Classification and regression trees include full tree and pruned trees and accommodate non-linear relationships and allow interactions between variables. Partial least squares constitute a dimension reduction method that finds new features that approximate the initial features and are related to the outcome variable. Ensemble algorithms operationalize 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 short-term and long-term interest rates, together with external sector predictors which consist of exchange rates and current account balance. A representative sample of countries across several regions is used.

Four modeling dimensions, which encompass an optimized contemporaneous panel format, transformations with lag structure, and a 20-year rolling window as well as an individual country format, are implemented to assess the 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 a rolling window increase overall accuracy rates in comparison to the optimized contemporaneous panel and individual country format. Notwithstanding, some individual country forecasts improve on the panel experience utilized for individual country-level predictions. Random forests and gradient boosting are consistently top-performing machine learning methods, both classifying through AUC, 77% of forecasts correctly across 17 countries and 147 years. Traditional regression models probit and linear also perform above average at respectively 73% and 71% 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, with the majority breaching an arbitrary 10 percentage points threshold level in panel and individual country 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% of the variation in crisis prevalence, real sector 64%, and the external sector 8%.

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

ACKNOWLEDGEMENTS

The authors thank conference participants at the 2023 Annual Congress of the Swiss Society for Economics and Statistics, in Neuchâtel, Switzerland; the Royal Economic Society and Scottish Economic Society 2023 Annual Conference in Glasgow, UK; and the 2022 Ecomod International Conference on Economic Modeling and Data Science, Ljubljana, Slovenia for valuable comments and suggestions. Open access funding enabled and organized by Projekt DEAL.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are openly available in Harvard Dataverse at https://doi.org/10.7910/DVN/ZJ7JBF.

ORCID

Emile du Plessis https://orcid.org/0000-0002-9519-5829

REFERENCES

Abiad, A. (2003). Early-warning systems: A survey and a regime-switching approach. *IMF Working Paper*, *3*(32), 1. https://doi.org/10.5089/9781451845136.001

Alessi, L., & Detken, C. (2018). Identifying excessive credit growth and leverage. *Journal of Financial Stability*, *35*, 215–225. https://doi.org/10.1016/j.jfs.2017.06.005

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., … Žigraiová, D. (2015). Comparing different early warning systems: Results from a horse race competition among members of the macro-prudential research network. *SSRN 2566165. SSRN*).

Bair, E., Hastie, T., Debashis, P., & Tibshirani, R. (2006). Prediction by supervised principal components. *Journal of the American Statistical Association*, *473*, 119–137. https://doi.org/10.1198/016214505000000628

Behringer, J., Stephan, S., & Theobald, T. (2017). Macroeconomic factors behind financial instability: Evidence from Granger causality tests. *IMK Working Paper 178*, 1–60. Hans Boeckler Foundation, Macroeconomic Policy Institute.August 2016

Benzoni, L., Chyruk, O., & Kelly, D. (2018). Why does the yield curve slope predict recessions? In *Chicago Fed Letters* (Vol. 404). The Federal Reserve Bank of Chicago.

Berg, A., Borensztein, E., & Pattillo, Z. (2005). Assessing early warning systems: How have they worked in practice? *IMF Staff Papers*, *52*(3), 462–502. https://doi.org/10.2307/30035972

Beutel, J., List, S., & Von Schweinitz, G. (2019). Does machine learning help us predict banking crises? *Journal of Financial Stability*, *45*, 100693. https://doi.org/10.1016/j.jfs.2019.100693

Black, J., Hashimzade, M., & Myles, G. (2012). *A dictionary of economics* (4th ed.). Oxford University Press. https://doi.org/10.1093/acref/9780199696321.001.0001

Bliss, C. I. (1934). The method of probits. *Science*, *79*, 409–410. https://doi.org/10.1126/science.79.2037.38

Bliss, C. I. (1935). The calculation of the dosage-mortality curve (with an appendix by Fisher, R.A.). *Annals of Applied Biology*, *22*, 134–167. https://doi.org/10.1111/j.1744-7348.1935.tb07713.x

Bluwstein, K., Buckmann, M., Joseph, A., Kang, M., Kapadia, S., & Şimşek, Ö. (2020). Credit growth, the yield curve and financial crisis prediction: Evidence from a machine learning approach. *Bank of England, Staff Working Paper 848*, 1–63. https://doi.org/10.2139/ssrn.3520659

Bordo, M. D., Eichengreen, B., Klingebiel, D., & Martinez-Peria, M. S. (2001). Is the crisis problem growing more severe? *Economic Policy*, *16*(32), 53–83. https://doi.org/10.1111/1468-0327.00070

Boser, B. E., Guyon, I. M., & Vapnik, V. N. (1992). A training algorithm for optimal margin classifier. In *Proceedings of the Fifth Annual Workshop on Computational Learning Theory (COLT 92)*, Pittsburgh, 27–29 July 1992, 144–152.

Breiman, L., Friedman, J. H., Olshen, R. A., & Ston, C. J. (1984). *Classification and regression trees*. Wadsworth and Brooks/Cole.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32. Kluwer Academic Publishers.

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, *78*(1), 1–3. https://doi.org/10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2

Brownlee, J. (2016). *Master machine learning algorithms: Discover how they work and implement them from scratch*. Machine Learning Mastery.

Bussière, M., & Fratzscher, M. (2006). Towards a new early warning system of financial crises. *Journal of International Money and Finance*, *25*, 953–973. https://doi.org/10.1016/j.jimonfin.2006.07.007

Casabianca, E. J., Catalano, M., Forni, L., Giarda, E., & Passeri, S. (2019). An early warning system for banking crises: From regression-based analysis to machine learning techniques. In *Marco Fanno Working Papers, 235*. Dipartimento di Scienze Economiche.

Casabianca, E. J., Catalano, M., Forni, L., Giarda, E., & Passeri, S. (2022). A machine learning approach to rank the determinants of banking crises over time and across countries. *Journal of*

International Money and Finance, *129*, 102739. https://doi.org/10.1016/j.jimonfin.2022.102739

Chamon, M., Manase, P., & Prati, A. (2007). Can we predict the next capital account crisis? *IMF Staff Papers*, *54*(2), 270–305. https://doi.org/10.1057/palgrave.imfsp.9450012

Chen, W., Mrkaic, M., & Nabar, M. (2019). The global economic recovery 10 years after the 2008 financial crisis. In *IMF Working Paper* (Vol. 19(83)). IMF.

Chinchor, N.A. (1992). MUC-4 evaluation metrics. In *Proc. of the Fourth Message Understanding Conference*, McLean, Virginia, 16–18 June 1992, 22–29.

Choi, I. (2001). Unit root tests for panel data. *Journal of International Money and Finance*, *20*, 249–272. https://doi.org/10.1016/S0261-5606(00)00048-6

Claessens, S., Kose, M. A., & Terrones, M. (2011). The global financial crisis: How similar? How different? How costly? *Journal of Asian Economics*, *21*(3), 247–264. https://doi.org/10.1016/j.asieco.2010.02.002

Coulombe, P. G., Marcellino, M., & Stevanovic, D. (2021). Can machine learning catch the COVID-19 recession? CIRANO Working Papers 2021s-09, 1–39. Center for Interuniversity Research and Analysis on Organisations.

Cover, T. M., & Hart, P. E. (1967). Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, *13*(1), 21–27. https://doi.org/10.1109/TIT.1967.1053964

Cramer, J. S. (2002). *The origins of logistic regression*. *Technical Report, 119*. Tinbergen Institute.

Dattagupta, R., & Cashin, P. (2011). Anatomy of banking crises in developing and emerging market countries. *Journal of International Money and Finance*, *30*, 354–376. https://doi.org/10.1016/j.jimonfin.2010.08.006

Davis, E. P., & Karim, D. (2008). Could early warning systems have helped to predict the sub-prime crisis? *National Institute Economic Review*, *206*, 35–47. https://doi.org/10.1177/0027950108099841

Davis, E. P., Karim, D., & Liadze, I. (2011). Should multivariate early warning systems for banking crises pool across regions? *Review of World Economics*, *147*, 693–716. https://doi.org/10.1007/s10290-011-0102-1

DeLong, E. R., DeLong, D. M., & Clarke-Pearson, D. L. (1988). Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach. *Biometrics*, *44*, 837–845. https://doi.org/10.2307/2531595

Demirgüç-Kunt, A., & Detregiache, E. (1998). The determinants of banking crises in developing and developed countries. In *IMF Staff Papers* (Vol. 45(1)) (p. 81). International Monetary Fund. https://doi.org/10.2307/3867330

Döpke, J., Fritsche, U., & Pierdzioch, C. (2017). Predicting recessions with boosted regression trees. *International Journal of Forecasting*, *33*, 745–759. https://doi.org/10.1016/j.ijforecast.2017.02.003

du Plessis, E. (2022a). Multinomial modeling methods: Predicting four decades of international banking crises. *Economic Systems*, *46*(2), 100979, 1–34. https://doi.org/10.1016/j.ecosys.2022.100979

Dumitrescu, E.-I., & Hurlin, C. (2012). Testing for Granger non-causality in heterogeneous panels. *Economic Modelling*, *29*(4), 1450–1460. https://doi.org/10.1016/j.econmod.2012.02.014

du Plessis, E. (2022b). Dynamic forecasting of banking crises with a Qual VAR. *Journal of Applied Economics*, *25*(1), 477–503. https://doi.org/10.1080/15140326.2020.1816132

du Plessis, E., & Fritsche, U. (2023). Replication data for: New forecasting methods for an old problem: Predicting 147 years of systemic financial crises. In *Harvard Dataverse*. Harvard University. https://doi.org/10.7910/DVN/ZJ7JBF

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, *27*(8), 861–874. https://doi.org/10.1016/j.patrec.2005.10.010

FDIC. (2023). *Bank data and statistics*. Federal Deposit Insurance Corporation. https://www.fdic.gov/bank/statistical

Fechner, G. T. (1860). *Elemente der psychophysik*. Breitkopft und Härtel.

Fix, E., & Hodges, J. L. (1951). Discriminatory analysis. Nonparametric discrimination: consistency properties. In *USAF School of Aviation Medicine, Project 21–49-004(4)*. Randolph Field.

Foster, M. (1961). An application of the Wiener-Kolmogorov smoothing theory to matrix inversion. *Journal of the Society for Industrial and Applied Mathematics*, *9*(3), 387–392. https://doi.org/10.1137/0109031

Fouliard, J., Howell, M., & Rey, H. (2021a). Answering the queen: Machine learning and financial crises. *BIS Working Paper 926*. Bank for International Settlements.

Fouliard, J., Rey, H., & Stavrakeva, V. (2021b). *Is this time different: Financial follies across the centuries*. Keynes Lecture.

Freund, Y., & Schapire, R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, *55*(1), 119–139. https://doi.org/10.1006/jcss.1997.1504

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, *29*(5), 1189–1232. https://doi.org/10.1214/aos/1013203451

FSO. (2023). *Federal statistical office*. Swiss National Bank. https://www.bfs.admin.ch/bfs/en/home/statistics

Funke, M., Schularick, M., & Trebesch, C. (2016). Going to extremes: Politics after financial crises, 1870–2014. *European Economic Review*, *88*, 227–260. https://doi.org/10.1016/j.euroecorev.2016.03.006

Gaddum, J. H. (1933). Report on biological standards III: Methods of biological assay depending on quantal response. In *Special Report Series of the Medical Research Council* (Vol. 183). Medical Research Council.

Ghosh, S. R., & Ghosh, A. R. (2003). Structural vulnerabilities and currency crises. *IMF Staff Papers*, *50*(3), 481–506. https://doi.org/10.2307/4149942

González-Hermosillo, B., Pazarbasioglu, C., & Billings, R. (1997). Determinants of banking sector fragility: A case study of Mexico. In *IMF Staff Papers* (Vol. 44(3)). International Monetary Fund.

Granger, C. W. J. (1969). Investigating causal relations by econometric models and cross-spectral methods. *Econometrica*, *37*(3), 424–438. https://doi.org/10.2307/1912791

Greene, W. H. (2012). *Econometric analysis* (7th ed.). Pearson.

Hamilton, J. D. (1994). *Time series analysis*. Princeton University Press. https://doi.org/10.1515/9780691218632

Hardy, D. C., & Pazarbasioglu, C. (1998). Leading indicators of banking crises: Was Asia different? *IMF Working Paper*, *98*(91), 1. https://doi.org/10.5089/9781451951745.001

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference and prediction*. Springer. https://doi.org/10.1007/978-0-387-84858-7

Hellevik, O. (2009). Linear versus logistic regression when the dependent variable is a dichotomy. *Quality and Quantity*, *43*, 59–74. https://doi.org/10.1007/s11135-007-9077-3

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*, *58*(3), 54–59.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning, with applications in R*. Springer.

Jordà, Ò., Schularick, M., & Taylor, A. M. (2013). When credit bites back. *Journal of Money, Credit and Banking*, *45*(2), 3–28. https://doi.org/10.1111/jmcb.12069

Jordà, Ò., Schularick, M., & Taylor, A. M. (2017). Macrofinancial history and the new business cycle facts. In M. Eichenbaum & J. A. Parker (Eds.), *NBER Macroeconomics Annual* (Vol. 31). University of Chicago Press.

Joy, M., Rusnák, M., Šmídková, K., & Vašíče, B. (2015). Banking and currency crises: differential diagnostics for developed countries. In *Working Paper Series* (Vol. 1810). European Central Bank.

Joy, M., Rusnák, M., Šmídková, K., & Vašíček, B. (2017). Banking and currency crises: Differential diagnostics for developed countries. *International Journal of Finance and Economics* *22*(1), 44–67.

Kaminsky, G. L., & Reinhart, C. M. (1999). The twin crises: The causes of banking and balance-of-payments problems. *American Economic Review*, *89*, 473–500. https://doi.org/10.1257/aer.89.3.473

Kindleberger, C. (1978). *Manias, panics, and crashes: A history of financial crises*. Basic Books. https://doi.org/10.1007/978-1-349-04338-5

Kleinbaum, D. G., Kupper, L. L., Muller, K. E., & Nizam, A. (1998). *Applied regression analysis and multivariable methods* (3rd ed.). Brooks/Cole Publishing Company.

Laeven, L., & Valencia, F. (2008). Systemic banking crises: A new database. *IMF Working Paper 8/224*, 1–78.

Laeven, L., & Valencia, F. (2010). Resolution of banking crises: The good, the bad and the ugly. *IMF Working Paper*, *10*/146, 1–35. https://doi.org/10.5089/9781455201297.001

Laeven, L., & Valencia, F. (2012). Systemic banking crises database: An update. *IMF Working Paper 12/163*, 1–32. https://doi.org/10.2139/ssrn.2096234

Laeven, L., & Valencia, F. (2018). Systemic banking crises revisited. *IMF Working Paper*, *18*(206), 1. https://doi.org/10.5089/9781484376379.001

Manasse, P., & Roubini, N. (2009). 'Rules of thumb' for sovereign debt crises. *Journal of International Economics*, *78*, 192–205. https://doi.org/10.1016/j.jinteco.2008.12.002

Mullainathan, S., & Spiess, J. (2017). Machine learning: An applied econometric approach. *Journal of Economic Perspectives*, *31*(2), 87–106. https://doi.org/10.1257/jep.31.2.87

Nevasalmi, L. (2020). Forecasting multinomial stock returns using machine learning methods. *The Journal of Finance and Data Science*, *6*, 86–106. https://doi.org/10.1016/j.jfds.2020.09.001

Nobelprize. (2022). The Sveriges Riksbank Prize in economic sciences in memory of Alfred Nobel 2022. *NobelPrize.org*. https://www.nobelprize.org/prizes/economic-sciences/2022/summary/

Nyman, R., & Ormerod, P. (2016). *Predicting economic recessions using machine learning algorithms*. University College London.

Phillips, C. B., & Peron, P. (1988). Testing for a unit root in time series regression. *Biometrika*, *75*(2), 335–346. https://doi.org/10.1093/biomet/75.2.335

Phillips, D. L. (1962). A technique for the numerical solution of certain integral equations of the first kind. *Journal of the ACM*, *9*, 84–97. https://doi.org/10.1145/321105.321114

Reinhart, C. M., & Rogoff, K. S. (2009). *This time is different: Eight centuries of financial folly*. Princeton Press. https://doi.org/10.1515/9781400831722

Rosipal, R., & Krämer, N. (2006). Overview and recent advances in partial least squares. In C. Saunders, M. Grobelnik, S. Gunn, & J. Shawe-Taylor (Eds.), *Subspace, latent structure and feature selection 3940*, 34–51. Springer-Verlag. https://doi.org/10.1007/11752790_2

Santosa, F., & Symes, W. W. (1986). Linear inversion of band-limited reflection seismograms. *SIAM Journal on Scientific and Statistical Computing*, *7*(4), 1307–1330. https://doi.org/10.1137/0907087

Savona, R., & Vezzoli, M. (2012). Multidimensional distance-to-collapse point and sovereign default prediction. *Intelligent Systems in Accounting, Finance and Management*, *19*(4), 205–228. https://doi.org/10.1002/isaf.1332

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

Sun, X., & Xu, W. (2014). Fast implementation of DeLongs algorithm for comparing the areas under correlated receiver operating characteristic curves. *IEEE Signal Processing Letters*, *21*, 1389–1393. https://doi.org/10.1109/LSP.2014.2337313

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society*, *58*(1), 267–288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x

Tikhonov, A. N. (1943). On the stability of inverse problems. *Doklady Akademii Nauk SSSR*, *39*(5), 176–179.

Tikhonov, A. N. (1963). Solution of incorrectly formulated problems and the regularization method. *Soviet Mathematics Doklady 4*, 1035–1038.

Train, K. (2002). *Discrete choice methods with simulation*. Cambridge University Press.

Van Rijsbergen, C. J. (1979). *Information retrieval*. Butterworths.

Vlaar, P. J. G. (2000). Currency crises models for emerging markets. In *De Nederlandsche Bank Staff Report 45*, 253–274. De Nederlandsche Bank N.v.

Ward, F. (2017). Spotting the danger zone: Forecasting financial crises with classification tree ensembles and many predictors. *Journal of Applied Econometrics*, *32*(2), 359–378. https://doi.org/10.1002/jae.2525

Wold, H. (1985). *Partial least squares. Encyclopaedia of statistical sciences*. Wiley.

Youden, W. J. (1950). Index for rating diagnostic tests. *Cancer 3*, 32–35. https://doi.org/10.1002/1097-0142

AUTHOR BIOGRAPHIES

Emile du Plessis is a banking executive at Standard Bank in South Africa. As senior economist, he advises business divisions and clients on economic strategies. His forecasts are published weekly by Bloomberg and Thomson Reuters, and he received the 2019 and 2022 LSEG/Refinitiv StarMine Award for Most Accurate Forecasting of a foreign exchange rate. He supports banking regulatory requirements with through-the-cycle modelling of credit portfolios for IFRS 9. Emile served on the Academic Management Board of the Bureau of Market Research at UNISA from 2016–2018. He established and leads an applied Behavioural Economics capability across 17 African countries since 2015. He holds a PhD from the University of Hamburg in Germany, for which he was awarded a science prize from the Deutsche Bundesbank in 2024. Currently, Emile is completing an Executive MBA at the University of Cape Town in South Africa. He previously received his Bachelor of Arts (Politics, Philosophy, Economics), Bachelor of Commerce with Honours and Master of Commerce in Economics degrees from the University of Stellenbosch in South Africa. His research interests include banking crises, systemic financial risks, early warning indicators, econometric modelling, forecasting, machine learning, behavioural and experimental economics.

Ulrich Fritsche holds a PhD from the Free University of Berlin and began his career in macroeconomic forecasting at the German Institute for Economic Research (DIW Berlin). Since October 2009, Ulrich Fritsche has been a university professor of economics, in particular applied economics, at the Department of Socioeconomics at the University of Hamburg. He has also been a research fellow at the Swiss Institute for Business Cycle Research (KOF) at the ETH Zurich since 2010 and a member of the Research Program on Forecasting at George Washington University, Washington, D.C., since 2013. His research interests are in the areas of applied macroeconomic research, in particular forecasting and expectations analysis, expectations formation in macroeconomic models, European integration, machine learning and computational linguistics.

How to cite this article: du Plessis, E., & Fritsche, U. (2025). New forecasting methods for an old problem: Predicting 147 years of systemic financial crises. *Journal of Forecasting*, *44*(1), 3–40. https://doi.org/10.1002/for.3184

APPENDIX A

TABLE A1 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 A2 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 A3 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**

Note: Unit root tests are constructed using Augmented Dicky-Fuller (see Hamilton, 1994) and Phillips and Peron (see Phillips & 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 $N \to \infty$.

*** (**, *) denotes significance at 1%, (5%, 10%).

TABLE A4 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	−57,591	−270,219	0.012**

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

TABLE A5 Probit model results.

	Lag Structure	Optimal Contemporaneous 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) *

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

TABLE A6 Probit model: Granger causality.

Y	X	Lag = 1 Z-bar tilde	Lag = 2 Z-bar tilde	Lag = 3 Z-bar tilde	Lag = 4 Z-bar tilde
Real Sector					
Crisis	GDP per capita	0.392	0.392	0.749	−0.179
GDP per capita	Crisis	2.611***	2.611	1.539	1.040
Crisis	Consumption expenditure	2.566	2.566*	3.247***	2.613***
Consumption expenditure	Crisis	−0.808	−0.808	−0.026	1.644*
Crisis	Fixed capital formation	5.913***	5.913***	3.674***	3.035***
Fixed capital formation	Crisis	0.153	0.153	0.895	2.354**
Crisis	Capital output ratio	0.402	0.402	1.082	1.859*
Capital output ratio	Crisis	3.708***	3.708**	2.998***	1.527
Banking Sector					
Crisis	Debt to GDP	−0.718	−0.718*	−2.311**	−2.980***
Debt to GDP	Crisis	−1.378	−1.378	1.700*	2.711*
Crisis	Total loans	−0.668	−0.668**	2.710*	4.254***
Total loans	Crisis	6.233***	6.233***	3.156***	3.397***
Crisis	Stock market	1.037	1.037	−0.348	−1.014
Stock market	Crisis	3.820***	3.820***	2.965***	2.036**
Crisis	Consumer inflation	−0.751	−0.751	−0.864	0.601
Consumer inflation	Crisis	3.941***	3.941**	1.392	−0.006
Crisis	Short-term rates	0.957	0.957	−0.172	−0.679
Short-term rates	Crisis	11.940***	11.940***	10.671***	7.924***
Crisis	Long-term rates	−0.335	−0.335	−0.381	−0.339
Long-term rates	Crisis	−0.514	−0.514**	0.710	0.882
External Sector					
Crisis	Exchange rates	−0.869	−0.869	0.819	4.153***
Exchange rates	Crisis	0.914	0.914	−0.535	−1.658*
Crisis	Current account	−0.517	−0.517	−0.891	−1.128
Current account	Crisis	3.187***	3.187***	4.465***	4.770***

Note: Granger causality for panel models using Dumitrescu and Hurlin (2012); *** (**, *) denotes p-value significance of Z-bar tilde at 1%, (5%, 10%).

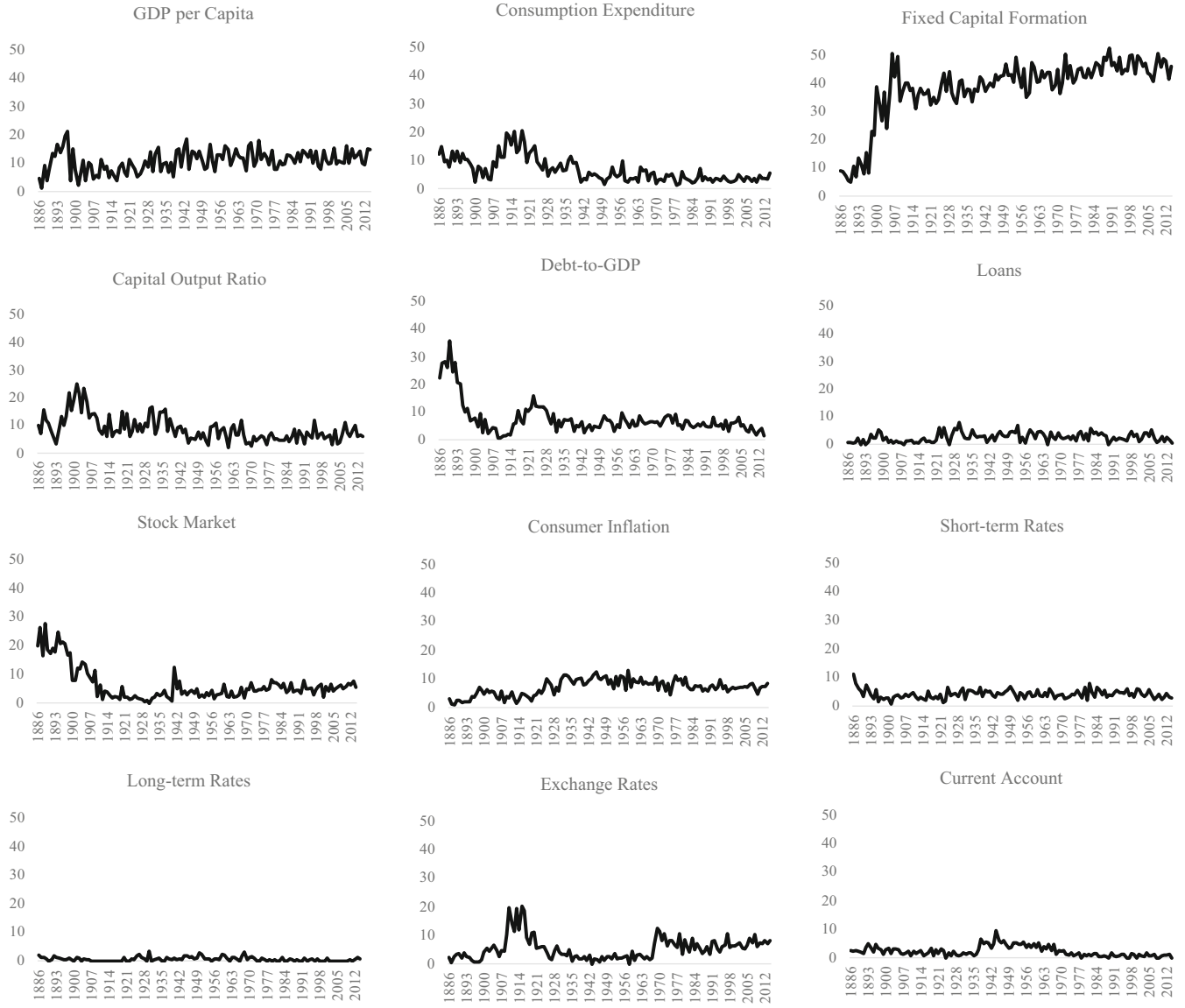

FIGURE A1 Gradient boosting: Variable importance by indicator.

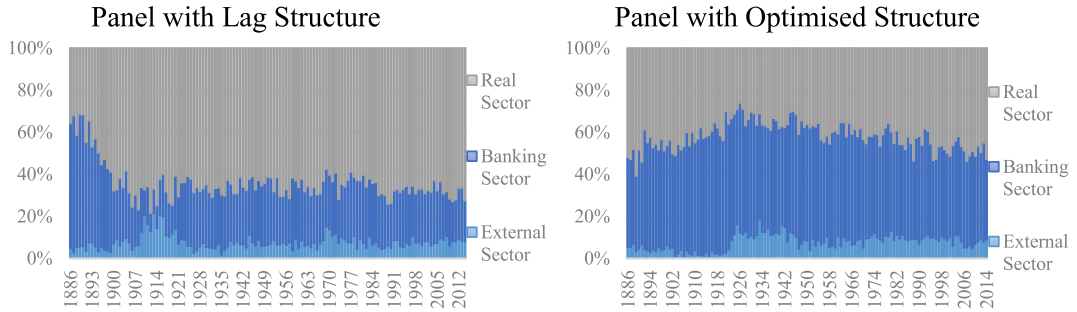

FIGURE A2 Gradient boosting: Variable importance by sector.

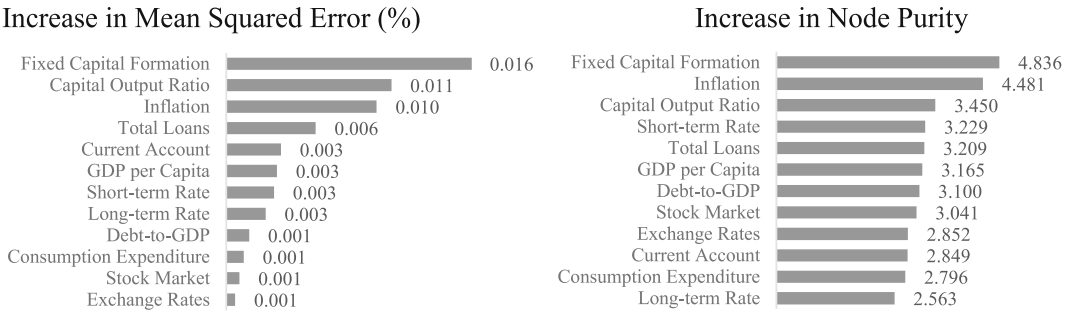

FIGURE A3 Random forests: Variable importance.

TABLE A7 Robustness test of *F*-measures using lag structure.

Method	F_1 score	F_2 score	$F_{0.5}$ score	Rank incl F_1	Rank incl F_2	Rank incl $F_{0.5}$
Baseline	0.073	0.162	0.051	8	8	8
Linear Prediction	0.130	0.259	0.089	5	5	5
Probit Regression	0.073	0.164	0.051	11	11	11
K-Nearest Neighbors	0.000	0.000	0.000	13	13	13
Support Vector Machine	0.116	0.232	0.079	3	3	3
Ridge	0.137	0.268	0.093	4	4	4
Lasso	0.073	0.162	0.051	9	9	9
Partial Least Squares	0.000	0.000	0.000	12	12	12
Full Tree	0.136	0.231	0.090	7	7	7
Pruned Tree	0.181	0.250	0.115	6	6	6
Adaptive Boosting	0.071	0.161	0.050	10	10	10
Gradient Boosting	0.142	0.277	0.096	2	2	2
Random Forests	0.139	0.278	0.095	1	1	1

TABLE A8 Recursive out-of-sample forecasts of downturn with lag structure.

Method	AUC	F_1 score	Brier score	Rank
Baseline	0.539 [0.511, 0.568] *0.014*	0.366	0.351	9
Linear Prediction	0.604 [0.575, 0.633] *0.014*	0.386	0.374	7
Probit Regression	0.593 [0.563, 0.622] *0.015*	0.082	0.687	11
K-Nearest Neighbors	0.557 [0.539, 0.576] *0.009*	0.000	1.000	13
Support Vector Machine	0.652 [0.624, 0.679] *0.013*	0.418	0.130	1
Ridge	0.604 [0.575, 0.633] *0.012*	0.384	0.353	6
Lasso	0.530 [0.531, 0.559] *0.014*	0.366	0.347	10
Partial Least Squares	0.497 [0.493, 0.501] *0.001*	0.000	1.000	12
Full Tree	0.635 [0.606, 0.664] *0.014*	0.432	0.309	4
Pruned Tree	0.625 [0.596, 0.654] *0.014*	0.423	0.324	5
Adaptive Boosting	0.673 [0.647, 0.700] *0.013*	0.440	0.500	8
Gradient Boosting	0.688 [0.661, 0.714] *0.013*	0.456	0.314	3
Random Forests	0.703 [0.678, 0.728] *0.012*	0.457	0.321	2

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

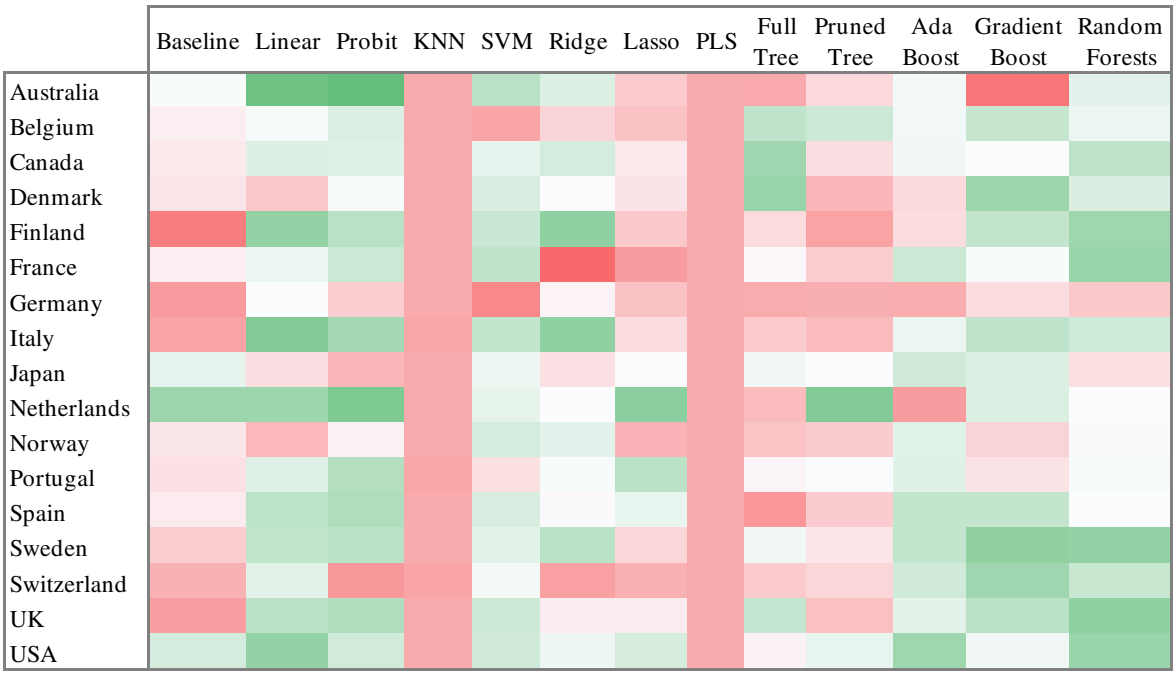

FIGURE A4 Individual countries: Recursive out-of-sample forecasts in panel format. *Note*: AUROC low (red) to high (green) gradient results.

TABLE A9 Top country models from optimized 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

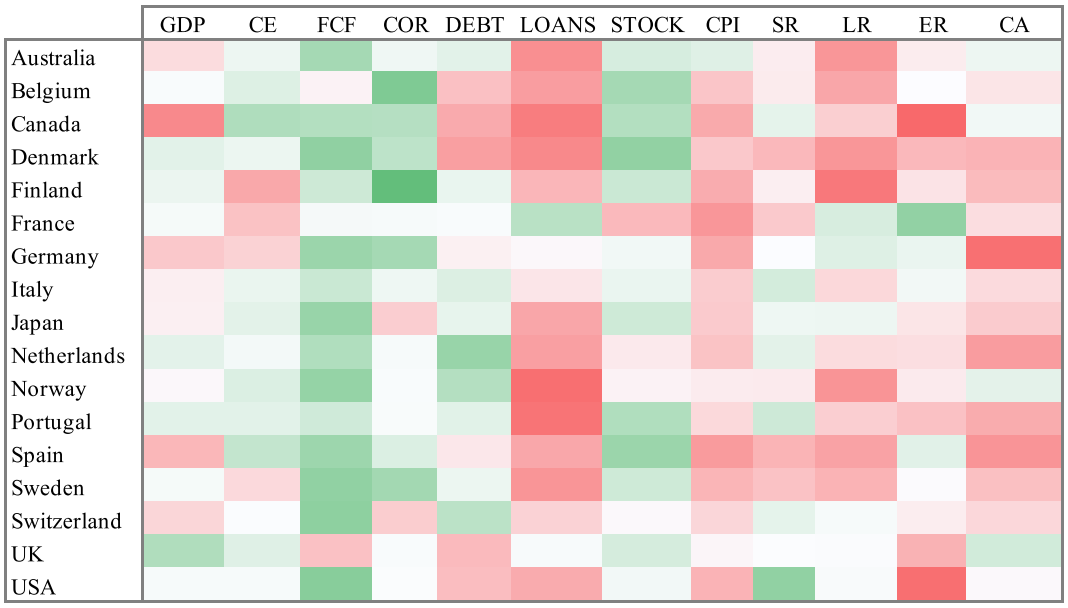

FIGURE A5 Individual countries: Variable importance. *Note*: low (red) to high (green) gradient results.

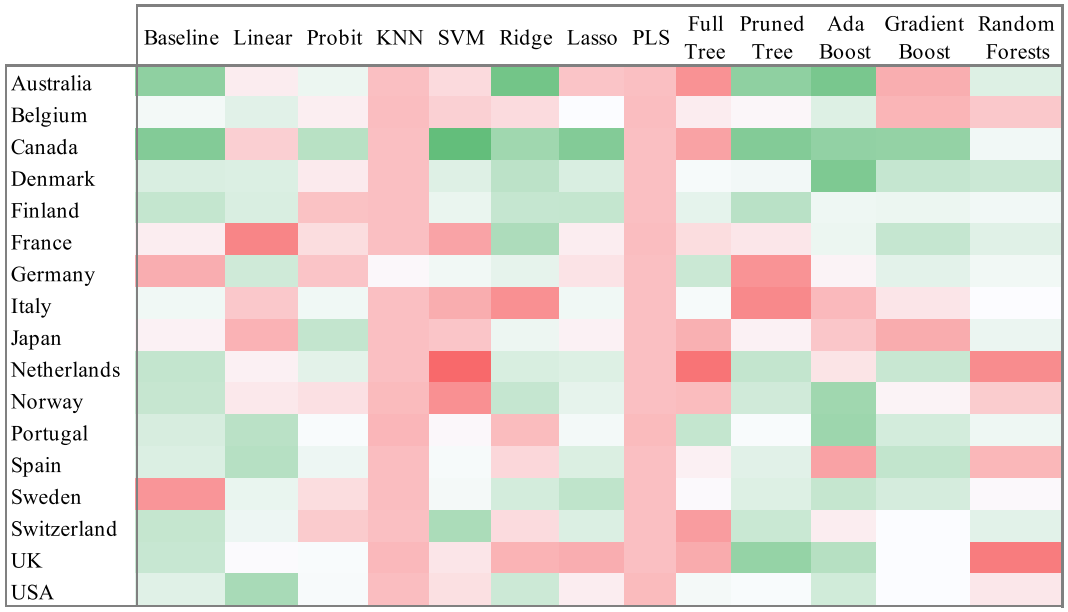

FIGURE A6 Individual countries: Recursive Out-of-sample forecasts in independent format. *Note*: AUROC low (red) to high (green) gradient results.

TABLE A10 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

APPENDIX B

B.1 | Non-parametric

B.1.1 | Baseline approach

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

B.2 | Regression algorithms

B.2.1 | Linear probability model

The linear probability model is an extension of the linear regression equation and operationalized as a generalized 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 (2012), 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 modeling framework is that $x'\beta$ is not constrained to the 0 to 1 interval, and out-of-range results could inhibit clear interpretation (Greene, 2012).

B.2.2 | Probit regression

As one of the oldest (see Bliss, 1934, 1935; Fechner, 1860; Gaddum, 1933) and most popular statistical methods (Cramer, 2002) the probit regression, comparable to the linear probability method, models a binary outcome variable. Operationalized, by modeling an inverse standard normal distribution of the outcome variable as a non-linear relationship to the explanatory variables. Based on Greene (2012), this can formally be denoted as $Y_{it}{}^{*} = x'_{it}\beta + \varepsilon_{it}$, where $\varepsilon_{it}\, N[0,1]$ and $y_{it} = 1$ if $y_{it}{}^{*} > 0$, else $y_{it} = 0$. Given that y_{it} follows a Bernoulli distribution, which consists of a single draw from a two-outcome binomial procedure, probability values can be described by $Prob(y_{it} = 1|x_{it}) = \phi(x'_{it},\beta)$ and $Prob(y_{it} = 0|x_{it}) = 1 - \phi(x'_{it},\beta)$. Distinctively, the binary choice model in comparison to the linear probability model is estimated through maximum likelihood, which in combination with success probability $F(x'_{it}\beta)$,

and independent and random observations, can be defined through a joint probability as $L(y|X,\beta)=\prod_{i-1}^{n}\left[\phi\left(x'_{it}\beta\right)\right]^{y_{it}}\left[1-\phi\left(x'_{it}\beta\right)\right]^{1-y_{it}}$.

B.3 | Instance-based algorithms

B.3.1 | K-nearest neighbors (k-NN)

Advanced by Fix and Hodges (1951) and expanded by Cover and Hart (1967), procedurally, and through a positive integer k and observation x_0, the k-NN classifier detects the k points in the dataset that 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 $pr(y=j|x=x_0)=\frac{1}{K}\sum_{i\in N_0}I(y_{it}=j)$.

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 $\widehat{f}(x_0)$ is determined as the average of all the training responses in N_0, stated as $\widehat{f}(x_0)=\frac{1}{K}\sum_{x_{it}\in N_0}y_{it}$. 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 10 times, with a maximum number of k estimated as 9, and with distance as 2.

B.3.2 | Support vector machine (SVM)

As developed by Boser et al. (1992), with a support vector machine, 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

$$f(x)=\beta_0+\sum_{i=1}^{N}\alpha_i(x,x_{it}), \tag{B1}$$

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

$$(x_{it},x_{it'})=\sum_{j=1}^{p}\left(x_{it},x_{it'j}\right), \tag{B2}$$

for observations $(x_i,x_{i'})$. Consequently, parameters $\alpha_i,\ldots,\alpha_n$ 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 (B2) can be restated as

$$f(x)=\beta_0+\sum_{i\in S}\alpha_{it}(x,x_{it}), \tag{B3}$$

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

$$k(x_{it},x_{it'})=\left(1+\sum_{j=1}^{p}\left(x_{itj},x_{it'j}\right)^{d}\right), \tag{B4}$$

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 results in a support vector machine and can mathematically be denoted as

$$f(x)=\beta_0+\sum_{i\in S}\alpha_{it}k(x,x_{it}). \tag{B5}$$

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

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

B.4 | Regularization 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.

B.4.1 | Ridge

The ridge procedure was developed and extended by Tikhonov (1943, 1963), Foster (1961), Hoerl (1962) and Phillips (1962). In contrast to the ordinary least squares statistical technique which computes slope coefficients $\beta_0,\beta_1,\ldots,\beta_p$ by employing values which minimises the equation

$$RSS = \sum_{i=1}^{N}\left(y_{it} - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ijt}\right)^2, \tag{B6}$$

ridge coefficients are determined by minimizing the following equation,

$$\sum_{i=1}^{n}\left(y_{it} - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ijt}\right)^2 + \lambda\sum_{j=1}^{p}\beta_j^2 = RSS + \lambda\sum_{j=1}^{p}\beta_j^2, \tag{B7}$$

where $\lambda \geq 0$ represents a tuning parameter. According to component $\lambda\sum_j \beta_j^2$, the shrinkage penalty is small when, $\beta_1,...,\beta_p$ are near zero, so it reduces the estimates of β_j toward 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 $\beta_1,...,\beta_p$, but not to the intercept. If the data matrix X has a zero mean, then the intercept becomes $\beta_0 = y_{it} = \sum_{i=1}^{n}\frac{y_{it}}{n}$.

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 $\lambda - 1$ folds. Thereafter, the error value is calculated based on the model performance on the $\lambda - 1$ folds. Repeated λ times, the procedure treats a different fold as validation set every time. Consequently, the tuning parameter is chosen 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 et al., 2013).

In this study, to estimate hyperparameter settings, initially $\beta_j^2 = 0$ and $\lambda = 100$.

B.4.2 | 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 minimizing the equation

$$\sum_{i=1}^{n}\left(y_{it} - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ijt}\right)^2 + \lambda\sum_{j=1}^{p}|\beta_j| = RSS + \lambda\sum_{j=1}^{p}|\beta_j|, \tag{B8}$$

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

For the lasso algorithm in the study, to estimate hyperparameters, initially $|\beta_j| = 1$ and $\lambda = 100$.

B.5 | Dimensionality reduction algorithm

B.5.1 | Partial least squares (PLS)

Introduced by Wold (1985), the partial least squares procedure involves estimating PLS directions. The first PLS direction is computed by normalizing the predictors p and equating each $\emptyset_{jm}$ in equation $Z_{mt} = \sum_{j=1}^{p}\emptyset_{jm}X_{jt}$ to the coefficients from the linear regression of Y onto X_{jt}. As a consequence, the coefficients are proportional to the correlation between Y onto X_{jt}. In computing the equation $Z_1 = \sum_{j=1}^{p}\emptyset_{j=1}X_{jt}$, PLS method puts larger weights on the explanatory variables that 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, $Z_1,...,Z_m$. Once this process is complete, ordinary least squares are employed to fit a model predicting Y using $Z_1,...,Z_m$. The number M of partial least squares directions represents a tuning parameter that 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. The dimension reduction procedure of PLS serves to reduce bias in existing

datasets but faces lower accuracy when modeling new data.

In this study, the PLS algorithm incorporates 10-fold cross-validation, repeated 10-times, with the optimized number of principle components determined through cross-validation.

B.6 | Decision tree algorithms

B.6.1 | Full tree

Decision tree algorithm is based on the seminal work of Breiman et al. (1984). 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. (2015), 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. Skewed 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 of minimizing 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 recognizing 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 penalization imposed on additionally produced terminal nodes. Analogously, this criterion is also employed to choose the best tree, with goodness of fit categorized as the optimal point between minimizing the classification rate while bigger trees are penalized. Yet, terminal nodes are not always entirely uniform.

The full tree algorithm only attempts splits which reduces the overall lack of fit by the numerical value of the complexity parameter, the latter determined by cross-validation. Values that are lower than the complexity parameter are expected to be pruned away in the subsequent procedure.

B.6.2 | Pruned tree

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

$$\sum_{i=1}^{T} \sum_{x_{it} \in R_m} (y_{it} - \widehat{y}_{Rm})^2 + \alpha \mid T \mid . \tag{B9}$$

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 utilizing a test or holdout 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. The optimal size of tree nodes is estimated by the procedure that minimizes the cross-validation error, which also determines the nodes to prune.

B.7 | Ensemble Algorithms

B.7.1 | Adaptive boosting

Adaptive boosting was developed by Freund and Schapire (1997). Mathematically, the training dataset consists of $(x_1 + y_1), ..., (x_N + y_N)$, with weight vector $w_i^1 = D(i)$ for $i = 1, ..., N$ and for D the distribution over N. The quantity of iterations is represented by $S = 1, 2, ..., S$. Initially, an equal set of weights w^s is applied across N, with distribution $p^s = \frac{w^s}{\sum_{i=1}^{N} w_{it}^s}$, estimated by standardizing the weights. The weak learner applies the distribution p^s to produce a new prediction h_s. In a test on the efficacy of the forecast, an error of h_s is computed through $\epsilon_t = \sum_{i=1}^{N} p_i^s \mid h_s(x_{it}) - y_{it} \mid$. For every iteration, the weak learner with lowest error is elected. The error is applied to determine the new weights vector $w_{it}^{s+1} = w_{it}^s \beta_t^{1-|h_s(x_s)-y_{it}|}$, where $\beta_m = \frac{\in^s}{1-\in^s}$ is also incorporated to signify the contribution of the chosen weak learner to the last prediction of the strong learner. This process continues across S where predictions are determined by

$$h_f(x) = \begin{cases} 1 \ if \sum_{i=1}^{S} \left(log \frac{1}{\beta_s} \right) h_s(x) f(x) \geq \frac{1}{2} \sum_{i=1}^{S} \left(log \frac{1}{\beta_s} \right) \\ 0 \ otherwise \end{cases}.$$

Initial 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 centered on the weights for every observation during each individual iteration.

B.7.2 | Gradient boosting

Gradient boosting was devised by Friedman (2001). Following Friedman (2001) and Döpke et al. (2017), mathematically, the algorithm bootstrap sample from the training dataset $\{(x_{it}+y_{it})\}_{i=1}^{N}$, with differentiable loss-function $L(y_{it},F(x))$ to determine a negative gradient vector. The model is initialized with a constant, using $\mathrm{F}_0(\mathrm{x}) = argmin\sum_{i=i}^{N}\mathrm{L}(\mathrm{y}_{it},\rho)$. For $\mathrm{m}=1$ to M, where the quantity of weak learners is capped, residuals are calculated for every sample $\tilde{y}_i = -\left[\frac{\partial L(y_{it},F(x_{it}))}{\partial F(x_{it})}\right]_{F(x)=F_{m-1}(x)}$, given where $\left[\frac{\partial L(y_{it},F(x_{it}))}{\partial F(x_{it})}\right]$ denotes the gradient derivative and $\tilde{y}_i$ pseudo residuals, and computed using $\{(\tilde{y}_i - g_m(x_{it})\}_{i=1}^{N}$. 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 $\gamma_{jm} = argmin\sum_{x_{it}\in R_{ijt}}^{N}\mathrm{L}(\mathrm{y}_{it},F_{m-1}(x_{it})+\gamma)$, achieved by adopting the previous prediction value and the selected sub-sample. For the 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 $\mathrm{F}_m(\mathrm{x}) = \mathrm{F}_{m-1}(\mathrm{x}) + \vartheta\sum_{j=1}^{J_m}\gamma_{jm}I\left(x\in R_{jm}\right)$. Lastly, when $\mathrm{m}=\mathrm{M}$, the strong learner $\mathrm{F}_m(\mathrm{x})$ is computed as the sum of all weak learners, based on $\mathrm{m}=0$, ...,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. Robustness tests done with different initial values produced comparable results. And 50% of the training data is randomly elected to create each new weak learner in the stepwise technique.

B.7.3 | Random forests

Advanced by Breiman (1984, 2001), and also 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 $\{\mathrm{T}_b\}_1^B$.

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

To operationalize the random forests algorithm, initially, the quantity of trees is set to 1,000, but the optimal number of trees necessary to calculate the minimum error estimate is consequently computed in the testing procedure and applied to predictions.

B.8 | Hyperparameters

TABLE B1 Hyperparameters.

Method	Optimized Hyperparameter	Value
Partial Least Squares	Cross-validation (fold)	10
K-Nearest Neighbors	Maximum K-number of neighbors	9
	Distance	2
Support Vector Machine	Gamma	0.5
	Regularization constant (cost constraints)	1
	Insensitive loss function (epsilon)	0.1
Ridge	Shrinkage penalty	0
	Cross-validation (folds)	10
Lasso	Shrinkage penalty	1
	Cross-validation (folds)	10
Adaptive Boosting	Cross-validation (fold)	10
	Initial number of trees	100
Gradient Boosting	Learning rate (shrinkage)	0.1
	Maximum tree depth (additive model)	1
	Base learners	100
	Simulation	100
Random Forests	Minimum number of observations in a terminal node (nodesize)	5
	Number of variables randomly sampled at each node (mtry)	4
	Initial quantity of trees	1,000