The Wisdom of the Crowd: Using ensemble machine learning techniques as an early warning indicator for systemic banking crises

Helena Patterson (NIS: 187961) Gabriela Lavagna (NIS: 187950)

Robizon Razmadze (NIS: 187894)

30 May 2021

Abstract

We develop early warning models for systemic crisis prediction using machine learning techniques on macrofinancial data for 36 countries for quarterly data spanning 1970-2013. Machine learning models outperform logistic regression in out-of-sample predictions under the recursive window forecasting mechanism. In particular, using the ensemble random forest algorithm for both feature selection and prediction substantially outperforms the logit models. We identify the key economic and financial drivers of our models using the random forest framework by extracting each feature's Gini impurity and corresponding information gain. Throughout the time period, the most important predictors are credit, foreign liabilities, asset prices and foreign currency reserves.

1. Introduction

Financial crises can negatively affect the social, economic and political environment of any country. The economic losses are usually higher when it comes to the systemic financial risks, as the disruption to the economy often leads to capital losses and the need for government intervention. The associated costs also tend to be high and persistent, as shown by several studies in the literature (Reinhart and Rogoff [2009] Cerra and Saxena [2008], JORDÀ et al. [2013]).

Due to these negative implications, it is essential for decision makers to foresee the potential risks stemming from these crises and implement the necessary policies in order to reduce the periodicity and extremity of these events. Thus, possessing viable prediction models is an essential prerequisite for implementing efficient macroprudential policies.

There is a wide amount of literature regarding this topic, due to its importance from an economic and welfare perspective. The two more traditional approaches are the signals approach and the use of binary

models (logit and probit). However, the first one is limited by the fact that it considers only one variable at a time and is constrained by the linearity assumption between the variables, and can suffer the "curse of dimensionality".

Recent studies have incorporated machine learning techniques to overcome these issues. In most of the cases, these types of models outperform the logit specifications (Ward [2017], Fouliard et al. [2021], Bluwstein et al. [2020]). One study however finds that they outperform logit models in sample but not when tested out-of-sample as a result of overfitting (Beutel et al. [2018]).

Our aim in this study is to construct a machine learning methodology to improve the predictive ability of systemic crises models. In order to achieve this, we consider an ensemble machine learning algorithm, the random forest, in addition to the K-nearest neighbours methodology. Aside from this, we also perform a set of preprocessing techniques widely used in this literature, and make use of the ensemble algorithm for feature selection based on the notion of Gini Impurity. Additionally, we use this method to determine which are the most relevant variables of the model. We find that throughout the time period, the most important predictors are credit, foreign liabilities, asset prices and foreign exchange reserves.

The remainder of the paper is structured as follows. Section 2 describes the literature review, putting emphasis on the recent work related to machine learning models. Section 3 presents the theoretical framework: we present the models applied and the way to assess their performance. Section 4 describes the data and its preprocessing (interpolation, scaling and encoding). Section 5 presents our proposed methodology for feature selection, our proposed models and the way to assess out of sample evaluation. Section 6 presents the results of the out of sample exercise and an analysis of the most relevant variables. Section 7 concludes.

2. Literature Review

There are two main classic methodologies which have been used predominantly in the past for predicting crises. The first is a signals approach introduced by Kaminsky and Reinhart [1999] and defines a signal as a departure of a leading indicator from its "normal" behaviour. If a crisis occurs after the indicator has produced a signal, then it is considered a good signal. Otherwise, it is considered a false alarm or noise. This approach has the advantage of being simple to construct and interpret, whilst also giving a clear idea of which variables relate to future potential crises. On the other hand, this method constrains the accuracy of the predictions, by only taking into consideration one variable at a time.

The second approach is the use of discrete choice models (logit and probit) used widely in the literature

(Schularick and Taylor [2012]). These models have been shown to outperform the signals approach to forecast banking crises (Davis and Karim [2008]), whilst also having the added advantage of possessing simple, interpretable coefficients. Nonetheless, they are constrained by the linearity assumption between the variables, and can suffer the "curse of dimensionality".

More recently, a few studies have introduced some modern machine learning approaches in order to overcome these problems. Below we present the review of the academic literature which examines the implications of using machine learning in forecasting. The papers presented use a variety of data sources (both developed and developing economies) and for differing timeframes. Most studies compare the predictable performance of the machine learning models against a more simplistic logit benchmark, with the majority revealing that the machine learning models outperform.

In the paper Spotting the Danger Zone - Forecasting Financial Crises with Classification Tree Ensembles and Many Predictors – Ward [2017] uses classification tree ensembles to produce a binary banking crisis forecast. The aim of this paper is to demonstrate whether the use of multiple predictors, coupled with machine learning algorithms, can improve financial crises forecasts, when compared to current early warning systems which are based on logit models. In order to initially evaluate the performance of the models, the author uses bagging and randomisation techniques. Whilst this is a common approach in machine learning algorithms, its use in time-series data is limited as it fails to take into account any serial dependencies in the data. In other words, future unseen observations could be used to predict the past.

As a supplement, the author performs a case-study test and evaluates the models' ability to accurately forecast the 2007-2008 global financial crisis out-of-sample. For this purpose the data was split in the year 1997, with pre-1997 data being used as the training sample and the remaining for testing. Again, bagging and the use of many predictors allowed classification tree ensembles to significantly increase the forecasting performance compared to current logit models. In addition to this, machine learning algorithms more clearly signaled that financial crises were expected in developed economies; however, the performance of the model was not favorable in the case of individual countries.

Other authors have adopted similar methodologies to the Ward [2017] paper. For example, the performance of logit and machine learning models were also compared by Beutel et al. [2018] - An evaluation of early warning models for systemic banking crises: Does machine learning improve predictions? In the first stage, the authors compared the logit model to four different machine learning algorithms: Random forest, k-nearest neighbors (KNN), Decision trees, support vector machine (SVM).

Whilst the machine learning algorithms were found to show better in-sample performance than the cor-

responding logit models, when tested out-of-sample the authors' found that these models were prone to overfitting and therefore performed more poorly out-of-sample when compared to logistic regressions. They concluded this overfitting could be a primary reason why machine learning algorithms cannot outperform the logit models. This is somewhat surprising given the results of the Ward [2017] paper where the only main difference is that Beutel et al. [2018] focus on more macroeconomic variables as opposed to the mix with financial as in Ward [2017]. In conclusion, the authors suggest that due to the instability of machine learning algorithms, they should be used with caution when evaluating real-world out-of-sample predictions.

We turn next to some more recent literature by Fouliard et al. [2021] - Answering the Queen: Machine Learning and Financial Crises. The authors aim to build an early warning system based on machine learning algorithms, and detect financial crises 3-years ahead in out-of-sample data, so as to be able to inform policymakers in advance of impending financial crises. As with previous literature, the authors analysed and compared both logit and machine learning algorithms in an attempt to understand which models and predictors provide most information regarding crises for individual countries. For the forecasting of the financial crisis they used a framework of sequential prediction (online machine learning) and, to allow the model flexibility, they did not use any assumption of the true model of the economy. Moreover, compared to the classical statistical approaches, in the online machine learning pre-crisis is not a specified mechanism (which could be either stochastic or deterministic) and this provides the authors additional flexibility as they do not need to make assumptions regarding how data is generated. As the driving factors of the true models are unknown and can vary over time, online machine learning models have an added advantage of being able to analyse sequential data in order to make the optimal prediction at any given point in time.

Similar to Ward [2017], the authors consider both economic and financial predictors to foresee the financial crisis. In the online machine learning framework, predictions are made based on previous t-1 observations, meaning models are adaptable over time. The optimal model is chosen based on different predictive model performance (set of experts). These models could be both statistical and machine learning algorithms, and the final outcome is a combination of all with differing weights. The authors concluded that the online machine learning framework results varied for each country. The algorithm successfully detected financial crises 3 years ahead out-of-sample for France, UK, Germany and Italy.

Finally, we examine the paper Bluwstein et al. [2020] in their article Credit growth, the yield curve and financial crisis prediction: evidence from a machine learning approach, which combines many of the ideas noted above. Again, the authors aim to compare machine learning algorithms against the traditional logit classification. In order to predict the crisis, the authors solved a classification problem to detect the best predictors of the financial crisis. After running different sets of machine learning classification algorithms

(decision trees, random forests, neural networks) the researchers used cross-validation to verify the out-of-sample performance of the models. By using the random selection, they split the data into the test (20% of observations) and training dataset (80% of observations). The authors' note that this approach of splitting the data comes with some caveats. Notably, due to such a random division of test and training datasets, it might appear that some future data points are used to train the model for the prediction of past years' crises. Despite this issue, the authors' claim that this approach is still the optimal due to data limitations.

Therefore, they choose to assess the robustness of their findings by performing a recursive forecasting approach. Additionally, the authors investigate the importance of the predictors using Shapley values. They concluded that, compared to logit models, machine learning algorithms can take into account both non-linear relationships between predictors, in addition to the interactions of variables and obtain a superior performance when compared to logit models in out-of-sample prediction.

3. Theoretical Underpinnings

3.1 Binary Classification Models

Binary classification, simply put, is a type of supervised learning where observations are split into two potential categories. Models are then used to categorise new observations into the predefined categories. For models of crisis prediction, the dependent variable will typically take the form of 1 or 0, where 1 indicates the presence of a financial crisis, and 0 otherwise. There are many types of models which could be used for such a problem. We define three of the most common below:

1. Logistic Regression (logit) As mentioned before, the logit model has been widely used in the literature for predicting crises. In these specifications there is a latent linear model $y_i^* = x_i'\beta + u_i$, where y_i^* is a continuous unobservable variable. In turn, we observe:

$$y_i = \begin{cases} 1 & \text{if } y_i^* > \theta, \\ 0 & \text{otherwise.} \end{cases}$$

The logit model assumes that $P(y_i = 1|x_i) = \frac{e^{x_i'\beta}}{1+e^{x_i'\beta}}$. and is estimated via maximum likelihood.

2. Random Forest The Random Forest is a typical machine learning technique, often used in classification models. A decision tree is the building block of a random forest model where each observation is separated into classes according to their composition. In simple terms, we can think of the decision tree as sorting an observation based on a specific criterion at each node. The node in the tree is the point where the path diverges and observations that meet the criterion in question will go down the 'yes' branch and those that do not will go down the 'no' branch.

Figure 1 below provides a visual representation of the more common CART model (Classification and Regression Tree) as summarised in Hastie et al.. Simply put, the tree considers a regression problem with continuous response variable Y and inputs X_1 and X_2, each taking values in the unit interval. We first split the data at the node $X_1 = t_1$, corresponding to the region $X_1 < t_1$. Observations satisfying the condition are assigned to the left branch and the others to the right branch. The data is then split at the $X_2 = t_2$ node and so on. The result of this is the split of the data into 5 R partitions. One of the main advantages of representing the model in this way is its interpretability as the feature space is fully defined by a single tree.

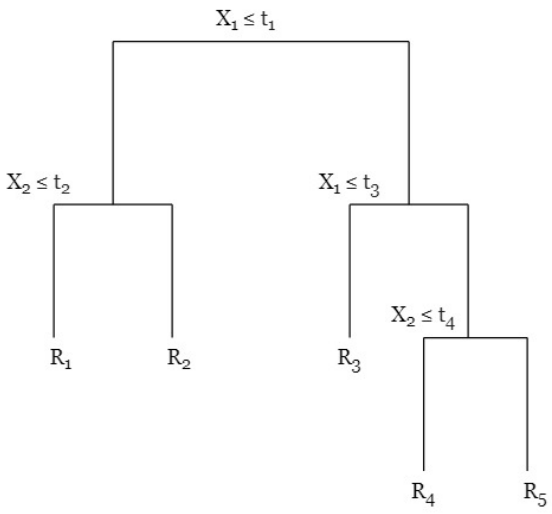


Figure 1 CART Tree

The random forest model consists of a large number of individual decision trees, operating as an ensemble. The reason why it is commonly used in predictor models is that each individual tree is assumed to have low correlation since not every tree will see every feature or observation; on average however, the trees will correctly classify. This is otherwise known as 'the wisdom of crowds', a powerful phenomenon in prediction and decision-making.

3. K-Nearest Neighbours (KNN) The k-nearest neighbours algorithm takes a slightly different approach to that of the random forest. This methodology is constructed in such a fashion that it attempts to predict the probability of an observation belonging to a certain class based on the "vote" of the k closest observations (neighbours). The closeness between two observations $x_1 = (x_{1,1}, x_{1,2}, .., x_{1,p})$ and $x_2 = (x_{2,1}, x_{2,2}, .., x_{2,p})$ is determined by their euclidean distance:

$$||x_1 - x_2|| = \sqrt{\sum_{i=1}^{p}(x_{1,i} - x_{2,i})^2}$$

Taking the k nearest observations to x_0, we estimate the probability of Y (class label) being 1 conditional on X being x_0 as follows:

$$P(Y = 1|X = x_o) = \frac{1}{k}\sum_{i \in N_0} I(y_i = 1)$$

Where N_0 is the set of k-nearest observations and $I(y_i = 1)$ is the indicator function, which is equal to 1 when an observation is of class 1.

3.2 The Training Test Split

As with any predictor models, maintaining an adequate split between the training and test data is crucial for the integrity of the predictors and reliability of the model. By training the model and evaluating in-sample a forecaster runs the risk of overfitting resulting in the model being unable to predict on any other sample other than the one it was trained on. In cross-sectional data these issues are overcome by undertaking randomised splits in the data where a certain fraction is assigned as training data and the remaining as the test sample.

However, the issue of the split becomes more complicated when a forecaster is dealing with time-series data. Methodologies whereby time series data is split in a random nature contains two major flaws. Firstly, any split of this type would fail to take into account any potential serial dependence in the data. Secondly, and more importantly, by removing random observations from the data the forecaster runs the risk of using future, supposedly unseen, observations to predict the past. This would draw into question both the reliability and accuracy of a model as any results obtained in this manner would not be viewed as a true out-of-sample evaluation.

Typically econometricians will alleviate these issues by opting for either a rolling window or an expansive window forecasting methodology. The rolling window estimation scheme assumes the forecaster has split

the data into a fixed sample size R. At the first forecast origin the observations $t = 1, 2....R$ are used to estimate the model's h-step ahead forecasts. The window is then shifted one period ahead and the sample $t = 2, 3....R + 1$ is used to generate the h-step ahead forecasts. This process continues until the last forecast origin is reached (T-h). The recursive window estimation scheme begins in the same way where the observations $t = 1, 2....R$ are used to estimate the initial h-step ahead forecast. The difference arises in the next period where the window is expanded by one period so the sample $t = 1, 2, ...R, R + 1$ is used to estimate the next forecast. This process continues until the last forecast origin is reached (T-h).

One of the main issues with using these methodologies in a binary classification model is that the forecaster may not have observed enough 1s in the initial period R to adequately train the model. This is especially true in the case of financial crises which are historically rare. Many authors therefore choose to ignore the time series dimension of the data adding a caveat to their findings (for example both Bluwstein et al. [2020] and Ward [2017] opt for this methodology).

3.3 Assessing performance

There are many methods which can be used to assess prediction performance of the models, but typically a binary classification of the outcomes is undertaken: when the predicted probability is over 50%, it is assigned a 1 (crisis period - positive class). Otherwise, it gets assigned a 0 (non crisis period - negative class). Using this binary classification the following possibilities are obtained for each of the observations:

	Positive	Negative
Positive	True Positive (TP)	False Negative (FN)
Negative	False Positive (FP)	True Negative (TN)

That is, the observation can be correctly classified (a crisis is classified as 1 (true positive) and non crisis as 0 (true negative)), or misclassified (non crisis is assigned a 1 (false positive) or a crisis is assigned a 0 (false negative)).

Based on this, we can calculate the following indicators to assess the model's performance: the accuracy, defined by the number of observations correctly classified with respect to the total amount of observations $((TP + TN)/(TP + FN + FP + TN))$, the sensitivity, defined as the proportion of the positive class that got correctly classified $(TP/(TP + FN))$, and the false positive and false negative rate $(FP/(TN + FP)$ and $FN/(TP + FN)$ respectively).

Another way to measure a model's performance is by constructing the Receiver Operator Characteristic (ROC) curve, which is a probability curve that plots the true positive rate against the false positive rate for various threshold values to determine the class of the observations. We can then calculate the Area Under the Curve (AUC - area under the ROC), which serves as a summary of the ROC. An AUC close to 1 implies that the false positive rate is near zero and the true positive rate is near 1. That is, most observations were correctly classified. An AUC close to 0.5 implies that the model cannot distinguish between classes. In summary, the higher the AUC, the more accurate the model.

4. Data

4.1 Sources and Transformations

We base our analysis on a quarterly panel data of 36 developed economies (as defined by the UN World Economic Situation and Prospects 2014 - see Annex for full country list) for the period spanning 1970 and 2013, constructed by Ward [2017]. The author uses the definitions of financial crises as produced by Laeven and Valencia [2013], along with a series of financial and macroeconomic indicators obtained from the IMF IFS-database (GDP, consumer prices, exchange rates, bank loans, stock prices, house prices, interest rates, foreign liabilities and foreign exchange reserves) and from the BIS database (total credit to the non-financial sector). These base indicators are then used to construct 113 predictors by applying the following transformations: deflating nominal series to convert into real terms, constructing growth rates, trend deviations, ratios in terms of GDP, global averages (non-weighted) and constructing real exchange rates and interest rates differentials (see Annex for full list of variables considered).

As the aim is forecasting, Ward [2017] constructs the trend deviations as the deviation from a slowly adjusting one-sided HP-trend ($\lambda = 1600$). Moreover, the ratios in terms of GDP are standardised when there is no cross country comparability between the series (as in the case of loan aggregates). The real exchange rate is defined as follows:

$$RER_{i,t} = NER_{i,t} * P_{i,t}/P_t^*$$

Where $NER_{i,t}$ is the nominal USD-exchange rate in price notation of country i at time t, $P_{i,t}$ is the domestic CPI and P_t^* is the GDP-weighted average CPI of all countries with non-missing values at time t. Finally, the interest rate differential is calculated as the difference between global average interest rates and country

specific interest rates.

For our dependent variable, the objective is to predict systemic crises twelve quarters ahead, under the understanding that this provides a sufficient time period so as to enable policymakers to act appropriately. Thus, following Ward [2017] and Fouliard et al. [2021] we define our variable of interest as a pre-crisis indicator $I_{n,t}$ (for country n at time t) in the following manner:

$$I_{n,t} = \begin{cases} 1 & \text{if } \exists h \in H = [1, 12] \text{ such that } \ C_{n,t+h} = 1 \\ 0 & \text{otherwise.} \end{cases}$$

where $C_{n,t}$ is the systemic crisis characteristic function based on the crisis definition by Laeven and Valencia [2013], defined as follows:

$$C_{n,t} = \begin{cases} 1 & \text{if there is a systemic crisis in country n at time t} \\ 0 & \text{otherwise.} \end{cases}$$

4.2 Pre-processing and cleansing

A clean dataset is crucial in order to ensure accuracy of any supervised learning techniques. We outline below in full the methodologies undertaken before passing data to the models.

4.2.1 Missing variables

As a first step in preparing our data for the models we look to fill any missing values. Generally with prediction models, the larger the number of observations the more likely it is to gain a more accurate result. Nonetheless, there is a trade-off associated. By removing entire rows which contain some missing values we could be losing valuable information. Conversely, by filling missing information with approximations could be adding unnecessary noise and thus limiting the model's ability to effectively learn.

In order to assess the optimal interpolation method, we perform a time series cross-validation exercise using different interpolation techniques on only half our sample, before the year 1990. By undertaking the evaluation of the optimal methodology only on the dataset before 1990, we maintain the integrity of the train/test split as outlined in the previous section. In a usual cross-sectional cross-validation approach, the algorithm will remove data points at random, creating a series of pseudo test sets, each consisting of a single observation and the remaining observations acting as the psuedo train set. The time series approach

is slightly more sophisticated as it takes into account any potential serial correlation in our data. The pseudo test sets will be created in the usual way, however the corresponding training sets will consist only of observations that occured prior to the single observation that forms the test set. In other words, the time series nature of the data is preserved.

By undertaking this approach instead of opting to run the different interpolation methods using the full sample, we avoid any potential overfitting and our forecast horizon (1990 onwards) remains fully unseen still. We visually inspect the gaps in our dataset and, based on the evaluation undertaken by Lepot et al. [2017] on different interpolation techniques, we identify two potential methods which could be used to interpolate our data - 1) Linear interpolation and 2) Nearest neighbours interpolation. We also considered other methods such as piecewise polynomial and spline but ruled these out due to the discontinuity associated with our missing values which would make these methods less effective. As a third approach, we also consider purely dropping any row which contains at least one missing value in any column.

Method 1 - Linear Interpolation This method is undertaken by searching for a straight line that passes through the two points x_A and x_B, where x_B is the last observed point before the gap and x_A is the last observed point after the gap. Each missing (X_i) is filled using the approximation below, where a and b are the indexes of x_A and x_B:

$$X_i = \frac{x_A - x_B}{a - b}(i - b) + x_B$$

We run this methodology for all our missing values and then undertake the cross-validation exercise to assess the accuracy of the predictor. Whilst we assess multiple forecasting models in this paper, for the purposes of evaluating interpolation we opt for the random forest as a benchmark. The performance is evaluated on the basis of the 'accuracy' measure, which is defined by the number of observations correctly classified with respect to the total amount of observations, as outlined in section 3.3. Performing the cross-validation using the linear method provides us with an accuracy score of 0.870.

Method 2 - Nearest neighbours The nearest neighbours method is often viewed as a fairly simplistic approach and according to Hyndman and Athanasopoulos can often yield the worst results. Nonetheless, this method could potentially work well with our sample given the discontinuity of our missing values. This method works on the basis of filling any missing value with its closest known (existing) neighbour. This approach is depicted mathematically in the equation below:

$$X_i = \begin{cases} x_B & \text{if } i < \frac{a+b}{2}, \\ x_A & \text{if } i >= \frac{a+b}{2} \end{cases}$$

Again, we undertake the cross-validation and gain an accuracy score of 0.863.

Method 3 - Removing all missing values As a final approach, we run the cross-validation algorithm to see whether the accuracy is compromised purely by dropping any row which contains at least one missing value in any column. This is the more conservative approach as it ensures all data points are observed and not based on a model's 'best guess' estimation. Nonetheless, there are many potential informational disadvantages that could arise from discarding valuable information from other columns. As in the cases before, we run the cross-validation model using data where all missing values have been fully removed, which results in an accuracy of 0.863. The linear method of interpolation results in a marginally higher accuracy score and, as such, we proceed with this pre-processing.

4.2.2 Standardisation

As a next step we look to standardise our continuous features by removing the mean and scaling to a unit variance. This is because many of the models will assume that features are distributed as a standard normal. If a certain feature had a variance of a magnitude larger than others, it could potentially dominate the objective function and provide unnecessary weight to the algorithm's ability to properly learn from the features. Each variable is therefore standardised according to $z = \frac{(x-u)}{s}$ where u is the mean of the training sample and s is the standard deviation of the training sample. The mean and standard deviation are then stored to be used later to transform each test sample.

4.2.3 Encoding

We next need to account for the fact that our data is of a panel nature and transform the categorical country variables in order to capture country specific characteristics. A common approach would be to implement fixed effects by including dummy variables for all 36 countries. However, one of the main issues with this approach is that it increases the dimensionality of the model, potentially providing unnecessary noise and hindering the model's ability to learn effectively.

As such, we opt for the mean target encoding methodology, which is a common approach used in machine learning in order to capture these specific characteristics. Simply put, it is the process of replacing any

categorical value with the mean of the target variable. Each category (in our case country) is assigned a different 'label' which is directly correlated with the target variable (the crisis dummy). For example, if a certain country had witnessed 1 year of a crisis, but 9 years without over the space of a 10-year period, its mean-encoded value would be 0.1. Again, we maintain the integrity of the train/test split by only calculating the mean-encoded value based on the training data and then the transformation is applied to the test set. This ensures the dependent variable for the test set remains unseen.

5. Our Methodology

5.1. Out-of-Sample Forecasting Approach

In order to preserve this structure and avoid 'looking to the future', it is imperative we train our model purely on the information we have available up to a certain point. Therefore, we opt for a pseudo out-of-sample evaluation based on the recursive window estimation mechanism, as this methodology is most reflective of what a policymaker would observe in real-time. We initially take the year 1990 as our forecast origin.

There is a trade-off associated when opting for this methodology, notably the fact that the sample is imbalanced since the number of crises observed between the years 1970-1989 is fairly limited. Nonetheless, we believe the size of our panel and choice of countries allows for a sufficient number of crisis observations so as to enable our models to accurately learn.

We therefore train our models based on the data available until 1989 and perform a three year ahead forecast based on the data available in 1990 (that is, using the data available in 1990 we forecast the probability of a crisis in 1993). We repeat this exercise for all the years until 2013. In this manner, no future observation will ever be used in constructing the forecasts. That is, the model solely uses information that the policymaker or forecaster in question will hold up until a certain point in time.

Figure 2 illustrates a series of training and test sets, where the blue observations depict the training sets, the orange the test sets and the pink is the period for which the prediction is being made.

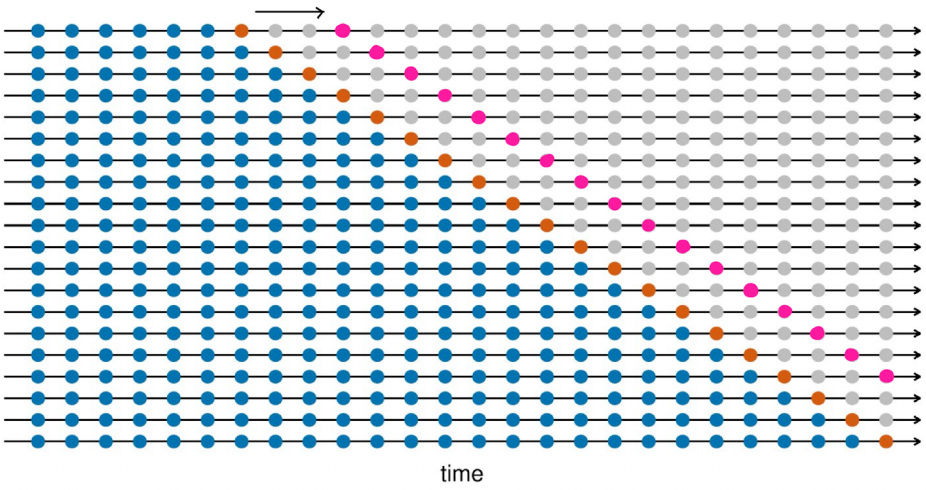


Figure 2 Recursive window forecasting

5.2 Feature Selection

Appropriate feature selection is important for any forecasting model, in particular machine learning algorithms. Not only does it reduce the complexity of the model and thus enables the machine to learn faster, but it also lowers overfitting risk and improves the accuracy of the model if the right subset is chosen.

Figure 3 provides an indication of the importance of appropriate feature selection. The dark pink sections indicate where variables are highly correlated with each other, leading to an issue of high dimensionality and thus potentially reducing the accuracy of our model. Therefore we note potential scope to trim variables in order to maintain a relevant subset whilst avoiding overfitting.

We opt for an ensemble machine learning algorithm, specifically the Random Forest methodology, as a means of selecting our features. The theory behind this model for prediction purposes was discussed in section 3.1, but the way in which the forest is formulated means it is also possible to use this methodology for feature selection.

The reasoning is that, at every node in the tree, there is a condition on how to split the values so that similar observations end up within the same classification. This condition is based on the notion of Gini impurity/entropy information gain (a measure of the likelihood of incorrect classification). Therefore, when training a tree, it is possible to calculate the degree to which each feature contributes in decreasing the relative impurity. In other words, how does adding or removing certain features contribute to the accuracy of our model. The exact formula for Gini impurity is shown below where $P(i)$ is the probability of a certain classification i, per the training data set.

$$GI = \Sigma_{i=1}^{J} P(i) * (1 - P(i))$$

The Gini gain is then calculated when building the decision tree to help determine which attribute provides the most information in order to determine which class a new data point belongs to. In order to calculate this, a weighted sum of the Gini impurity is taken at each branch of the decision tree.

$$\textit{Gini Gain}\ (\textit{Feature}) \ = \ \textit{Total Impurity} \ - \ \textit{Remaining Impurity}\ (\textit{Feature})$$

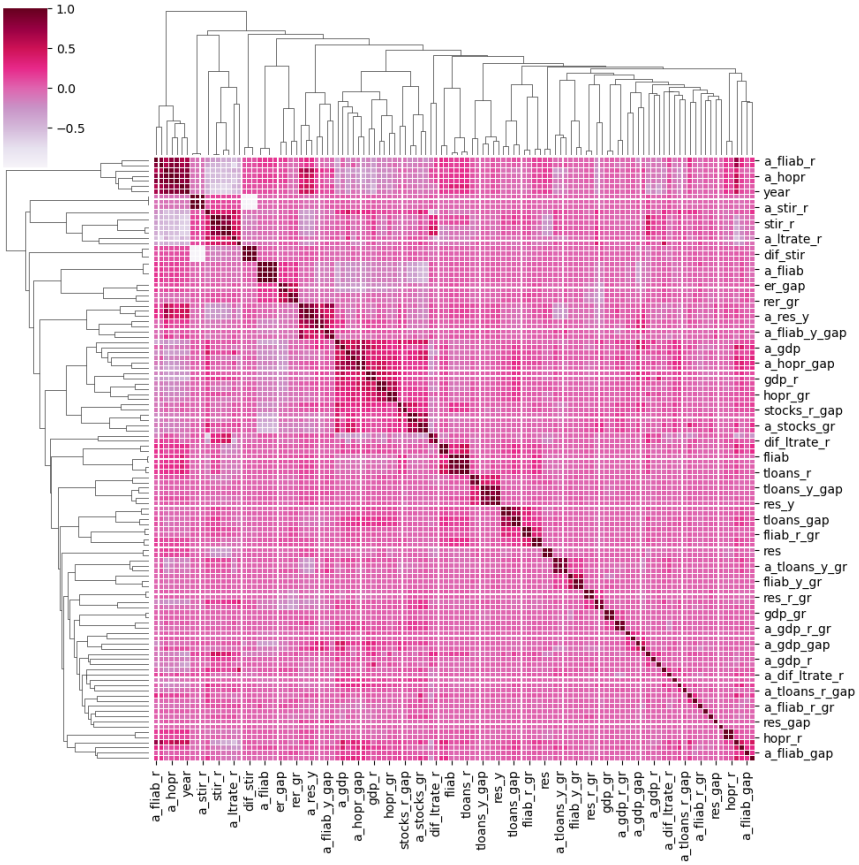


Figure 3 Correlation heatmap

As noted in section 5.1, we define our training period as running from 1970 to 1989, with our first forecast target being 1993 (based on the information available in 1990). Given the importance of maintaining the dependent variables in our test periods as unseen, we run the random forest variable selection on only the train dataset at each recursive window.

For each training sample the Gini gain for each variable is calculated, with the model then selecting only

those variables with contributions higher than the mean level. In this manner we are able to reduce the dimensionality of the training sample and are left with approximately 40 distinct variables each year, a significant reduction from our 113 original.

5.3 The Models

Our aim is to assess the differences in out-of-sample performance between a traditional logistic classification model and an ensemble machine learning algorithm, specifically the random forest classification model. We also run a K-nearest neighbours approach as a form of robustness to verify whether our results are specifically related to ensemble models, or whether other machine-learning techniques could also be applied.

We consider three different specifications for the logit model. The first is a more standard specification, based on Ward [2017] which has as explanatory variables: Loans (gap in real terms), Loans (growth, in real terms), globally weighted average of Foreign Liabilities (gap), globally weighted average of the Long Term Interest Rate (in real terms), globally weighted average of real GDP (gap), globally weighted average of real GDP (growth), Inflation, Exchange Rate (gap), Foreign Exchange Reserves (gap, in real terms); and the following interactions: Loans/GDP with GDP (growth) and Loans/GDP (gap) with Exchange Rate (gap), to capture effects from exchange rate devaluations on the banking system.

As it was mentioned in section 5.1, we perform an out of sample exercise running a recursive forecast for seven different models: three logit models (basic specification, considering all the 113 features, and only considering the variables chosen by our feature selection algorithm), and two random forest and knn models (considering all the variables and only considering the selected variables for both cases).

5.4 Accuracy Evaluation

In order to assess the model's performance we undertake the usual AUC and ROC calculations as described in section 3.3. As we are dealing with a very imbalanced classification problem, the accuracy indicator may be misleading. This is because a high accuracy ratio can be easily achieved by classifying all non crisis correctly, but this information is not useful to asses the model's ability to predict a crisis. Thus the sensitivity, false positive and false negative indicators are more useful in assessing the models' performance.

6. Results

6.1 Prediction Accuracy

We now apply all the preprocessing outlined thus far and evaluate the model's performance using the recursive window forecasting approach to undertake a full out-of-sample evaluation.

As we can see in the table below, the three logit specifications are the ones which obtain the lowest performance scores, with the AUC ranging between 0.68 and 0.694. It is worth noting the fact that increasing the number of features considered in the baseline specification generates a major improvement in the amount of crises predicted by the model. While the base specification was able to classify correctly only 22% of the crises, the models considering all the variables and the set of variables chosen by our feature selection algorithm were able to predict approximately 51% and 44% of the crises respectively. However, by including all the 113 features, we also increase the false positive rate. This explains why this is the model which obtains the worst AUC score, even though it has a better performance considering both the true positive and the false negative rates. In other words, this model misses less crises than the base model but in turn has more false alarms. Arguably, from a policy maker point of view this is not trivial: not being able to predict a crisis may be less desirable than falsely predicting one.

We can also see that all the machine learning models outperform all of the logit specifications, with the AUC score ranging from 0.77 to 0.81. Considering these models, the specifications that consider all 113 features obtain an out-of-sample AUC of 0.783 and 0.777 for the random forest and the knn model, whereas using approximately only 40 variables (only those above the mean Gini gain) we obtain an out-of-sample AUC score of 0.810 and 0.800 respectively. Therefore, not only have we managed to improve the performance of our model by undertaking concise feature selection but we have also managed to reduce unnecessary noise.

The best out of sample performance in terms of AUC is the random forest that considers only the variables obtained by our feature selection algorithm. Taking a closer look at the model's results, we can see that it correctly classifies approximately 66.67% of crises and has a false positive rate of 4.75%. This is an important improvement with respect to all of the logit specifications, as the maximum true positive rate obtained by these models was 51.69%.

Finally, as it has been mentioned before, we can see that in this case taking the accuracy as a measure of the model's performance can be misleading. For all of the models the accuracy is above 80%, as the magnitude of the class imbalance makes predicting correctly a large amount of the non crisis observations quite simple, but this is not informative of the ability of the model in predicting crises.

Model	AUC	True positive rate	False positive rate	False negative rate	Accuracy
Random Forest - selected features	0.8096	0.6667	0.0475	0.3333	0.9212
KNN - selected features	0.8007	0.6715	0.0701	0.3285	0.9016
Random Forest - all features	0.783	0.6087	0.0428	0.3913	0.919
KNN - all features	0.769	0.6135	0.0755	0.3865	0.8905
Logit - selected features	0.6937	0.4444	0.057	0.5556	0.8884
Logit - basic specification	0.6901	0.2222	0.0826	0.7778	0.8413
Logit - all features	0.6809	0.5169	0.1551	0.4831	0.809

Table 1 Results

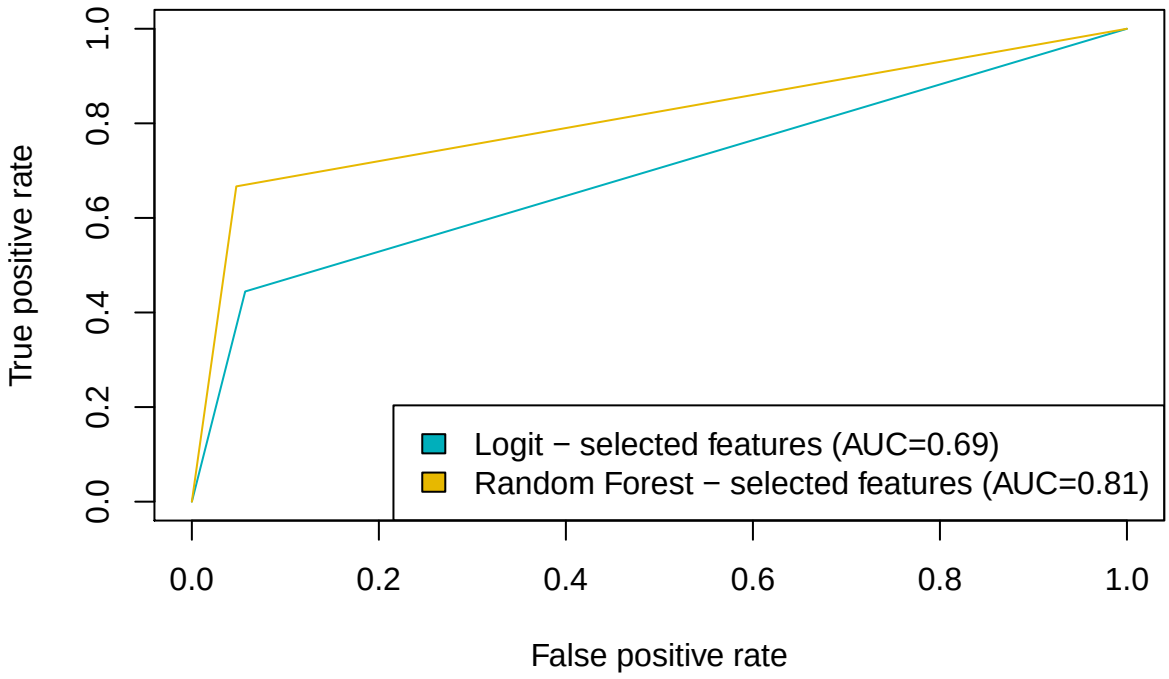


Figure 4 AUC scores for logit and random forest

6.2 Important Crisis Contributors

As we saw in the previous section, by using either of the machine learning models (KNN of random forest), the accuracy of our model increases significantly. Moreover, the variable selection obtained by applying the random forest algorithm for feature selection improved the results even further, which implies that it was able to recognise the features which were adding only noise and not new information to the model.

In spite of these benefits, some of the recent skepticism regarding machine learning techniques as predictor models has arisen as a result of the black box concern, particularly in the context of policy makers and their

public obligation to remain transparent and accountable with regards to decisions. From this point of view, apart from being able to improve the accuracy of the model we would also be interested in being able to interpret which are the most relevant variables.

To address this problem, we supplement our recursive window forecasting methodology by extracting the Gini gain of each feature at every forecast horizon, and averaging across each feature over two distinct time periods (the 1990s and post-2000s). Figure 5 below highlights the results of this analysis. The Gini gain from each variable transformation is summed and an average taken over the two time periods. The Annex provides a full breakdown of every variable and movements over time.

As we can see from figure 5, whilst there appear to be some changes, the variables that have more predictive power are relatively stable over time. Noticeably, the variables related to stocks, loans foreign liabilities and reserves are the ones that seem to be of most importance through the whole period. First, we can see that the variables that have a higher Gini are the ones related to credit (bank loans). This result is not surprising. The importance of loans as a predictor follows the idea of financial crises as "credit booms gone bust" widely found in the literature (Schularick and Taylor [2012]; JORDÀ et al. [2013];Mian et al. [2017];Greenwood et al. [2020]).

The second group of variables, in terms of importance, are those relating to foreign liabilities. The relevance of this variable links to the theory that suggests that foreign credit and financial integration is related to the occurrence of a banking crisis (Reinhart and Rogoff [2009], Cesa-Bianchi et al. [2019]). It is worth noting that most of the literature that points to this type of mechanism focuses on emerging markets. Our results suggest that this may be relevant also for advanced economies.

Similarly, we also see that variables relating to reserves also have a significant information contribution. As described by Dominguez et al. [2011], when a country faces sudden capital outflows, there tends to be pressure to depreciate the currency. In cases where the pressure is strong, it may encourage capital outflows, which in turn could lead to a systemic crisis. In order to avoid this, the authorities can resort to the use of foreign reserves to absorb capital outflow pressure and to moderate the speed of currency depreciation. The authors find that higher reserve accumulations prior to the crisis are associated with higher post-crisis GDP growth.

Our results also suggest that the variables related to stock prices have a high Gini gain, and that these are more relevant than those related to housing prices. This result may seem surprising, given that many authors have found that predictive power of asset prices is limited. However, Greenwood et al. [2020] show that when combined with credit, the forecasting accuracy improves. This could be the case in our model,

given that the random forest predictor model takes into account interactions between variables.

Finally, we note the importance of the long-term interest rate relative to the short-term rate post-2000s. There are a number of research papers that focus on the importance of the term spread as an early warning indicator for crises. For example, Bluwstein et al. [2020] note the importance of the slope of the yield curve as an early warning predictor of financial crises. Our results suggest that this variable became more prominent post-2000s.

Our main proposal therefore from the feature analysis would be to highlight the benefits of the use of the random forest methodology as a mechanism for identifying the key economic and financial drivers as early warning signals and predictors of systemic banking crises. In particular, for the case of policymakers, implementing these methodologies would allow machine learning techniques to be incorporated into the policy-makers' toolkit, whilst still enabling them to maintain full accountability over any decisions.

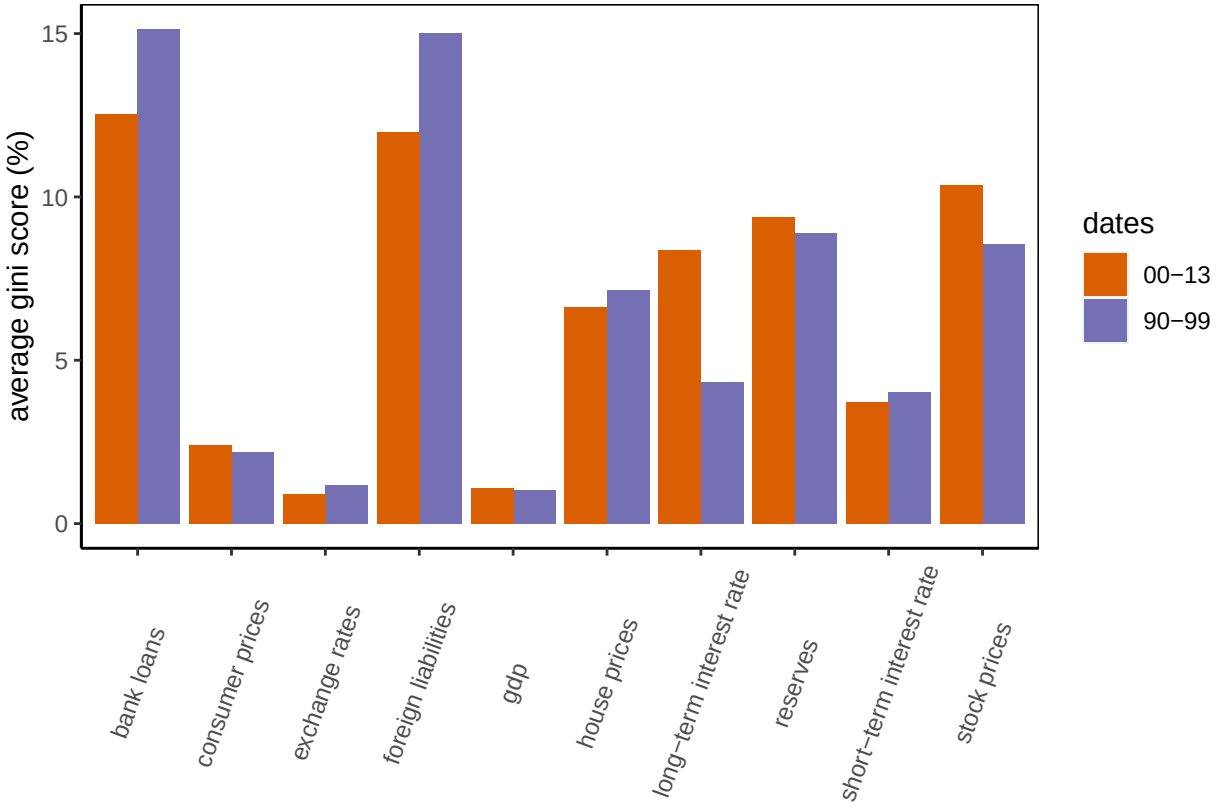


Figure 5 Gini gain for selected features

9. Conclusion

The aim of this study was to construct a machine learning methodology to improve the predictive ability of systemic crises models. The results of the paper show that prediction of the financial crisis is more accurately obtained via machine learning algorithms as opposed to logit models. To undertake the analysis we used macrofinancial quarterly data spanning the years 1970-2013. The study covered 36 developed countries. In the machine learning models we ran a recursive window forecasting mechanism which outperformed the logistic models in out-of-sample predictions (obtaining an AUC score in a range of 0.77-0.81). These gains in predictive accuracy of machine learning approaches confirm that opting for this approach is more beneficial in planning macroprudential policies.

During the analysis our goal was not only to be able to improve the predictive power of the models but also to be able to select the most relevant and concise predictor variables. While we considered the Random Forest and K Nearest Neighbours models for prediction purposes, we also applied the random forest ensemble algorithm undertake feature selection (of which the selection was based on a measure of Gini gain). As a result, out of available 113 variables we selected only 40 features, which had a Gini gain greater than the mean value. This approach allowed us to increase out-of-sample AUC scores of Random Forest and K Nearest Neighbours models to 0.810 and 0.800 respectively. Moreover, the analysis showed that over the years the variables such as credit, foreign liabilities, asset prices and currency reserves were the most important as, despite the changes during the time period, they remained relatively stable whilst maintaining strong predictive power.

The value-add derived from the models developed by us can be viewed in two main directions: Firstly, the use of time-series data in predicting crises has meant many authors in the past have been unable to avoid the 'looking to the future' issue. We managed to alleviate this risk by using a recursive window estimation mechanism. The main benefit of the methodology is that it allows policymakers to observe the predictors in real-time.

The second main advantage of our models is that by ranking variables based on Gini Gain we were able to reveal the key economic and financial drivers which should be used by the policy makers in evaluating any pressing risks of systemic crises (credit, foreign liabilities, asset prices and foreign currency reserves as mentioned above).

Overall, the paper proves the advantage of using machine learning models compared to the traditional logit classification as it gives higher predictive power to the policy makers. Moreover, approaches adopted in this paper have enabled us to identify the key economic variables essential for the prediction of financial

crisis thus making it possible to integrate machine learning models into the transparent nature of the policy decision-making process.

Annex 1 - Complete List of Countries

Developed countries

European Union	New EU member States	Other Europe	Other countries
Austria	Bulgaria	Iceland	Australia
Belgium	Croatia	Norway	Canada
Denmark	Cyprus	Switzerland	Japan
Finland	Czech Republic		New Zealand
France	Estonia		United States
Germany	Hungary		
Greece	Latvia		
Ireland	Lithuania		
Italy	Malta		
Luxembourg	Poland		
Netherlands	Romania		
Portugal	Slovakia		
Spain	Slovenia		
Sweden			
United Kingdom			

Annex 2 - Complete List of Variables and Transformations

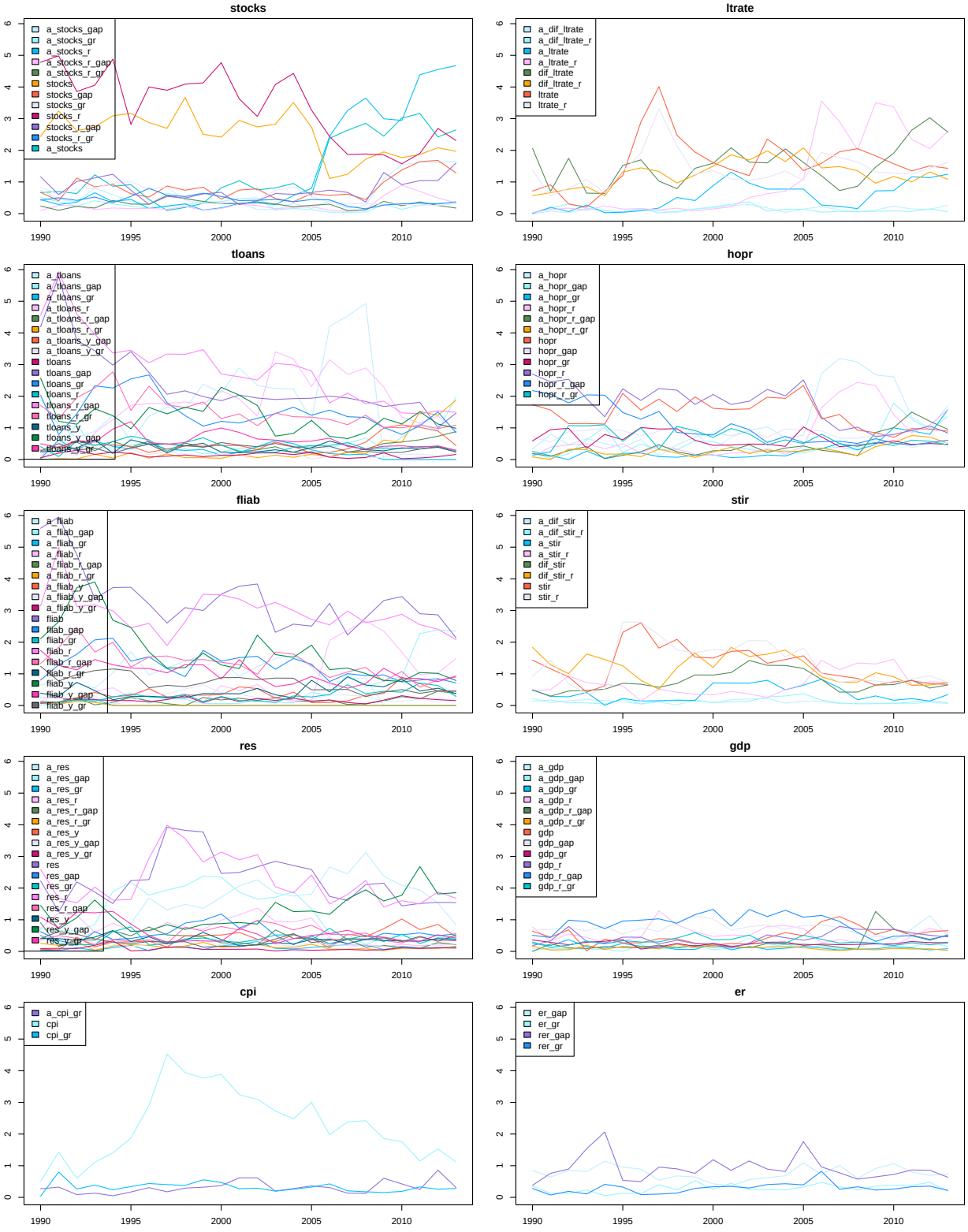
Exchange Rate (er)	Foreign Exchange Reserves (res)	Foreign Liabilities (fliab)
Exchange Rate (gr)	Reserves	Foreign Liabilities
Exchange Rate (gap)	Reserves (r)	Foreign Liabilities (r)
Exchange Rate (r)(gr)	Reserves (y)	Foreign Liabilities (y)
Exchange Rate (r)(gap)	Reserves (gr)	Foreign Liabilities (gr)
	Reserves (r)(gr)	Foreign Liabilities (gap)
	Reserves (y)(gr)	Foreign Liabilities (r)(gap)
	Reserves (gap)	Foreign Liabilities (a)(gap)
	Reserves (r)(gap)	Foreign Liabilities (a)
	Reserves (y)(gap)	Foreign Liabilities (gr)(a)
	Reserves (a)	Foreign Liabilities (r)(gr)
	Reserves (a)(r)	Foreign Liabilities (r)(gap)(a)
	Reserves (a)(y)	Foreign Liabilities (r)(a)
	Reserves (gr)(a)	Foreign Liabilities (r)(gr)(a)
	Reserves (a)(gap)	Foreign Liabilities (y)(gap)
	Reserves (a)(r)(gap)	Foreign Liabilities (y)(a)
	Reserves (y)(gap)(a)	Foreign Liabilities (y)(gr)
	Reserves (r)(gr)(a)	Foreign Liabilities (y)(gr)(a)
	Reserves (gr)(y)(a)	Foreign Liabilities (y)(gap)(a)

GDP	House Prices (hopr)	Long Term Interest Rate (ltrate)	CPI
GDP	House Prices	LT Interest Rate	CPI
GDP (r)	House Prices (r)	LT Interest Rate (a)	CPI (gr)
GDP (gr)	House Prices (gr)	LT Interest Rate (r)	CPI (gr)(a)
GDP (r)(gr)	House Prices (r)(gr)	LT Interest Rate (r)(a)	
GDP (gap)	House Prices (gap)	LT Interest Rate (dif)	
GDP (r)(gap)	House Prices (r)(gap)	LT Interest Rate (dif)(r)	
GDP (gap)(a)	House Prices (gap)(a)	LT Interest Rate (dif)(a)	
GDP (a)	House Prices (a)	LT Interest Rate (dif)(r)(a)	
GDP (gr)(a)	House Prices (gr)(a)		

GDP	House Prices (hopr)	Long Term Interest Rate (ltrate)	CPI
GDP (r)(gap)(a)	House Prices (r)(gap)(a)		
GDP (r)(a)	House Prices (r)(a)		
GDP (r)(gr)(a)	House Prices (r)(gr)(a)		

Short Term Interest Rate (stir)	Stock Prices (stocks)	Bank Loans (tloans)
ST Interest Rate	Stock Prices	Loans
ST Interest Rate (r)	Stock Prices (r)	Loans (gap)
ST Interest Rate (a)	Stock Prices (gr)	Loans (gap)(a)
ST Interest Rate (r)(a)	Stock Prices (r)(gr)	Loans (a)
ST Interest Rate (dif)	Stock Prices (a)	Loans (gr)
ST Interest Rate (dif)(a)	Stock Prices (gap)	Loans (gr)(a)
ST Interest Rate (dif)(r)	Stock Prices (r)(gap)	Loans (r)
ST Interest Rate (dif)(r)(a)	Stock Prices (gap)(a)	Loans (r)(gap)
	Stock Prices (r)(a)	Loans (r)(gap)(a)
	Stock Prices (gr)(a)	Loans (r)(a)
	Stock Prices (r)(gap)(a)	Loans (r)(gr)
	Stock Prices (r)(a)(gr)	Loans (r)(gr)(a)
		Loans (y)
		Loans (y)(gr)
		Loans (y)(a)
		Loans (y)(gap)
		Loans (y)(a)(gap)
		Loans (y)(gr)(a)

Notes: (r): real, (gap): percentage deviation from (one-sided) HP-trend ($\lambda = 1600$), (a): global GDP-weighted average, (gr): growth, (y): as a percentage of GDP, (dif): interest rate differential.


stocks
a_stocks_gap
a_stocks_gr
a_stocks_r
a_stocks_r_gap
a_stocks_r_gr
stocks
stocks_gap
stocks_gr
stocks_r
stocks_r_gap
stocks_r_gr
a_stocks
ltrate
a_dif_ltrate
a_dif_ltrate_r
a_ltrate
a_ltrate_r
dif_ltrate
dif_ltrate_r
ltrate
ltrate_r
tloans
a_tloans
a_tloans_gap
a_tloans_gr
a_tloans_r
a_tloans_r_gap
a_tloans_r_gr
a_tloans_y_gap
a_tloans_y_gr
tloans
tloans_gap
tloans_gr
tloans_r
tloans_r_gap
tloans_r_gr
tloans_y
tloans_y_gap
tloans_y_gr
hopr
a_hopr
a_hopr_gap
a_hopr_gr
a_hopr_r
a_hopr_r_gap
a_hopr_r_gr
hopr
hopr_gap
hopr_gr
hopr_r
hopr_r_gap
hopr_r_gr
fliab
a_fliab
a_fliab_gap
a_fliab_gr
a_fliab_r
a_fliab_r_gap
a_fliab_r_gr
a_fliab_y
a_fliab_y_gap
a_fliab_y_gr
fliab
fliab_gap
fliab_gr
fliab_r
fliab_r_gap
fliab_r_gr
fliab_y
fliab_y_gap
fliab_y_gr
stir
a_dif_stir
a_dif_stir_r
a_stir
a_stir_r
dif_stir
dif_stir_r
stir
stir_r
res
a_res
a_res_gap
a_res_gr
a_res_r
a_res_r_gap
a_res_r_gr
a_res_y
a_res_y_gap
a_res_y_gr
res
res_gap
res_gr
res_r
res_r_gap
res_y
res_y_gap
res_y_gr
gdp
a_gdp
a_gdp_gap
a_gdp_gr
a_gdp_r
a_gdp_r_gap
a_gdp_r_gr
gdp
gdp_gap
gdp_gr
gdp_r
gdp_r_gap
gdp_r_gr
cpi
a_cpi_gr
cpi
cpi_gr
er
er_gap
er_gr
rer_gap
rer_gr
0
1
2
3
4
5
6
1990
1995
2000
2005
2010

References

Johannes Beutel, Sophia List, and Gregor Von Schweinitz. An evaluation of early warning models for systemic banking crises: Does machine learning improve predictions? 12 2018.

Kristina Bluwstein, Marcus Buckmann, Andreas Joseph, Miao Kang, Sujit Kapadia, and Özgür Simsek. Credit growth, the yield curve and financial crisis prediction: evidence from a machine learning approach. Bank of England working papers 848, Bank of England, 2020. URL https://EconPapers.repec.org/RePEc:boe:boeewp:0848.

Valerie Cerra and Sweta Chaman Saxena. Growth dynamics: The myth of economic recovery. *American Economic Review*, 98(1):439–57, March 2008. doi: 10.1257/aer.98.1.439. URL https://www.aeaweb.org/articles?id=10.1257/aer.98.1.439.

Ambrogio Cesa-Bianchi, Fernando Eguren Martin, and Gregory Thwaites. Foreign booms, domestic busts: The global dimension of banking crises. *Journal of Financial Intermediation*, 37:58–74, 2019. ISSN 1042-9573. doi: https://doi.org/10.1016/j.jfi.2018.07.001. URL https://www.sciencedirect.com/science/article/pii/S1042957318300615.

E Davis and Dilruba Karim. Comparing early warning systems for banking crises. *Journal of Financial Stability*, 4(2):89–120, 2008. URL https://EconPapers.repec.org/RePEc:eee:finsta:v:4:y:2008:i:2:p:89-120.

Kathryn Dominguez, Yuko Hashimoto, and Takatoshi Ito. International reserves and the global financial crisis. *Journal of International Economics*, 88, 08 2011. doi: 10.1016/j.jinteco.2012.03.003.

Jérémy Fouliard, Michael Howell, and Hélène Rey. Answering the queen: Machine learning and financial crises. Technical report, National Bureau of Economic Research, 2021.

Robin Greenwood, Samuel G Hanson, Andrei Shleifer, and Jakob Ahm Sørensen. Predictable financial crises. Working Paper 27396, National Bureau of Economic Research, June 2020. URL http://www.nber.org/papers/w27396.

Trevor Hastie, Trevor Hastie, Robert Tibshirani, and J.H. Friedman. The elements of statistical learning: Data mining, inference, and prediction.

R.J. Hyndman and G. Athanasopoulos. Forecasting: principles and practice. OTexts.com/fpp2.

ÒSCAR JORDÀ, MORITZ SCHULARICK, and ALAN M. TAYLOR. When credit bites back. *Journal of Money, Credit and Banking*, 45(s2):3–28, 2013. doi: https://doi.org/10.1111/jmcb.12069. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/jmcb.12069.

Graciela Kaminsky and Carmen Reinhart. The twin crises: The causes of banking and balance-of-payments problems. *American Economic Review*, 89:473–500, 06 1999. doi: 10.1257/aer.89.3.473.

Luc Laeven and Fabián Valencia. Systemic banking crises database. *IMF Economic Review*, 61(2):225–270, 2013. doi: 10.1057/imfer.2013.12. URL https://doi.org/10.1057/imfer.2013.12.

Mathieu Lepot, Jean-Baptiste Aubin, and François HLR Clemens. Interpolation in time series: An introductive overview of existing methods, their performance criteria and uncertainty assessment. *Water*, 9(10): 796, 2017.

Atif Mian, Amir Sufi, and Emil Verner. Household debt and business cycles worldwide. *The Quarterly Journal of Economics*, 132(4):1755–1817, 2017. URL https://EconPapers.repec.org/RePEc:oup:qjecon:v:132:y:2017:i:4:p:1755-1817.

Carmen M. Reinhart and Kenneth S. Rogoff. The aftermath of financial crises. *American Economic Review*, 99(2):466–72, May 2009. doi: 10.1257/aer.99.2.466. URL https://www.aeaweb.org/articles?id=10.1257/aer.99.2.466.

Moritz Schularick and Alan M. Taylor. Credit booms gone bust: Monetary policy, leverage cycles, and financial crises, 1870——2008. *The American Economic Review*, 102(2):1029–1061, 2012. ISSN 00028282. URL http://www.jstor.org/stable/23245443.

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