*Towards better banking crisis prediction: could an automatic variable selection process improve the performance?**

ECONOMIC RECORD

ECONOMIC RECORD, VOL. 99, NO. 325, JUNE, 2023, 288–312

Towards Better Banking Crisis Prediction: Could an Automatic Variable Selection Process Improve the Performance?*

XIANGLONG LIU

Centre of Policy Studies, Victoria University, Melbourne, Victoria, Australia

This study proposes using the Least Absolute Shrinkage and Selection Operator (LASSO) method with cross-validation to automate the variable selection process of the conventional multivariate logit early warning system (EWS), the purpose being to improve the prediction of systemic banking crises. Using a dataset covering 23 OECD countries with quarterly data from 1970Q1 to 2018Q3, model performance is evaluated in a recursive out-of-sample *forecasting exercise, taking policy-makers' preference of missed crises and false alarms into account. The results suggest that the automatic variable selection process can enhance the predictive performance of the EWS. It also highlights the importance of extracting information from variable interactions and lags that may not be easily identified and accessed by typical subjective variable pre-selection. This simple approach is easy to interpret and is transparent, which are important aspects for effective policy communication. Five variables, namely credit growth, domestic and global credit gaps, real house price growth and the real effective exchange rate, are identified as the most important key indicators of systemic banking crises.*

I Introduction

As a result of the devastating economic and social consequences of the 2007–2008 global financial crisis (GFC), interest has grown in the prediction of systemic banking crises. Better early warning systems (EWSs) are called for to guide the activation of regulatory policies and guard against potential systemic events. These EWSs may draw on, for example, the remarkable similarities found in the data preceding financial crises (Reinhart & Rogoff, 2009).

The multivariate logit framework has emerged as one of the most popular models in the context of global EWS, and it can be used for both explanatory and forecasting purposes. Intuitively, multivariate EWS should include indicators that can capture the build-up

*I thank Tim Robinson, Guay Lim and Adrian Pagan for helping improving drafts of the paper. Comments from participants at the Macro brown bag seminar of the Faculty of Business and Economics (University of Melbourne), and the Australian Conference of Economists 2021 are greatly appreciated. I acknowledge financial support from the University of Melbourne's FBE Doctoral Program Scholarship for this research. This work was supported by ARC Grant DP160102654. Open access publishing facilitated by Victoria University, as part of the Wiley - Victoria University agreement via the Council of Australian University Librarians.

JEL classifications: C52, C54, G01, G17, G21

Correspondence: Xianglong Liu, Centre of Policy Studies, Victoria University, 300 Queen Street, Melbourne Victoria 3000, Australia. Email: locky.liu@vu.edu.au

doi: 10.1111/1475-4932.12721

of systemic risk and imbalances in the financial system. While a wide range of risk indicators and their different forms exists, EWS studies often select different sets of indicators based on economic intuition. However, the traditional intuition-based variable pre-selection has difficulty in determining the best set of variables to be included in the model from a large pool of risk indicators, which can be in many different forms and variants. This is of particular concern in the context of systemic event prediction for two reasons.

First, the choice of risk indicators, together with the transformations used, such as the growth rate, deviation from the long-run trend, choices of global variables, the number of lags and interaction terms, can all have a significant impact on the performance of the model (Davis & Karim, 2008; Duca & Peltonen, 2013). However, as systemic banking crises are rare events in history, the binary classes in the data structure are heavily imbalanced. Only a small number of covariates can be included in the model. If one attempts to expand the information set by incorporating more risk indicators to improve the model performance, it is likely to have an over-fitting problem. It means the model may fit the limited *in-sample* crises too well but lose its *out-of-sample* predictive power. For example, adding a couple of more lags and interaction terms to a multivariate logit EWS could easily lead to a perfect *in-sample* prediction, but this over-fitted model will predict poorly with a different sample.

Second, the optimal set of explanatory variables should depend on the objective of an EWS. If the model aims to serve an explanatory purpose, then the specification should be one that gives the best *in-sample* interpretation. If the model aims to serve a forecasting purpose, then the selection of predictors should optimise the *out-of-sample* performance. Even with the same modelling technique, a set of variables that provides good *in-sample* properties may not lead to the best *out-of-sample* prediction performance.

Furthermore, even if the objective of an EWS is to predict future crises, the model must maintain good interpretability and transparency, as communication is of key importance to policymakers. An EWS should not only have good predictive capability, but also be such that policymakers and the public could interpret it without difficulties. Models viewed as 'black-boxes' by policy-makers may only have a limited impact practically. This is particularly relevant for the recently proposed machine learning methods such as the 'Random Forest', which might sacrifice too much interpretability and transparency.

In summary, on the one hand, it is very important, but yet difficult in practice, to select the right set of predictors from a large pool of available variables. On the other hand, it is essential to keep the EWS transparent and interpretable if it is to have impact on policies. To the best of my knowledge, there has not been any attempt to systematically formalise the variable selection process in multivariate logit EWS for global systemic banking crisis prediction.

Motivated by the need to jointly address these two issues, this study proposes applying the Least Absolute Shrinkage and Selection Operator (LASSO) method with cross-validation to automate the variable selection process of the conventional multivariate logit model to better predict systemic banking crises. LASSO, originally proposed by Tibshirani (1996), is a model selection mechanism that drops variables with small coefficients and only keep the important predictors that contribute the most to models' performance. It is specifically designed to address the over-fitting issue and enhance the prediction accuracy. Cross-validation is a popular resampling technique to fine-tune the LASSO method and ensure the selected predictors have the best out-of-sample forecasting property.

Specifically, one can construct a variable pool that contains a large number of risk indicators and their various transformations such as interactions and lags. The LASSO method with cross-validation automatically selects a set of predictors which contains information that jointly has the best *out-of-sample* forecasting property from this variable pool to pin down the specification of the model, which I refer to as the 'LASSO logit model'.

The empirical analysis is conducted with the sample covering historical systemic banking crises of 23 OECD countries over 1970Q1 to 2018Q3. Through a classic pseudo-real-time recursive out-of-sample forecasting exercise, the predictive performance of the LASSO logit models is evaluated and compared with a benchmark multivariate logit model.

This study adopts the measure of usefulness proposed by Sarlin (2013) as the evaluation criteria for model performance comparison. The measure takes both (i) the preference of the policy-makers between Type I errors (missing the crises) and Type II errors (False alarms); and (ii) the imbalanced frequency of tranquil times and crises into consideration.

The results of the out-of-sample prediction exercise show that the predictive performance of multivariate logit models can be significantly improved by the LASSO method with cross-validation. The enhanced variable selection process can take advantage of information from lags and interaction terms that are not typically used by conventional approaches and improve the forecasting performance. It helps the EWS to produce more useful and accurate signals. The predictive performance also becomes more stable under varying policy-makers' preferences. The LASSO logit model identifies the most important predictors for the systemic banking crises to be domestic credit growth, the domestic and global credit-to-GDP gaps, real house price growth and the real effective exchange rate.

The LASSO method with cross-validation is closely related to systemic event prediction with machine learning approaches that are recently proposed in the literature. Manasse and Roubini (2009) pioneer the use of Classification and Regression Tree to predict sovereign debt crisis.[1] Alessi and Detken (2018) and Tanaka *et al*. (2016) propose the use of the 'Random Forest' technique to construct a completely automatic EWS for systemic banking crisis prediction. Introduced by Breiman (2001), the Random Forest method is a more sophisticated machine learning method based on classification trees.

As a popular modern classification tree ensemble technique, the Random Forest method often achieves great *in-sample* prediction performance and overcomes the issue of variable pre-selection. Tanaka *et al*. (2016) propose the use of it to predict failures at the level of individual banks, while Alessi and Detken (2018) focus on predicting systemic banking crises at country level. Both studies argue that the Random Forest method achieves better predictive performance than the conventional approaches. However, it may not be superior to the conventional multivariate logit model for three major reasons.

First, by comparing a benchmark logit approach to several recently proposed machine learning approaches, Beutel *et al*. (2019) find that machine learning methods are outperformed by the conventional logit approach in recursive out-of-sample evaluations, even though they often attain a very high *in-sample* fit. Their results suggest that the *out-of-sample* prediction performance of machine learning methods should not be taken as granted. They highlight the importance to establish a robust and valid *out-of-sample* prediction exercise.

Second, while interpretability and transparency are essential for an EWS to inform policy, it could be challenging to interpret machine learning models to the extent necessary for effective policy communication. For example, Alessi *et al*. (2015) acknowledged that their Random Forest model is inherently a black box model, which makes it difficult to defend its predictions, particularly if one wants to use it to support the activation of possibly unpopular policies. There have been some studies in the discipline of computer science attempting to interpret the results of Random Forests. Applications on rare crisis prediction, especially when the time dimension must be considered, however, are still at a developmental stage.

Furthermore, there could be potential transparency issues associated with machine learning approaches. Fine-tuning the hyperparameters through cross-validation is one of the most important steps when applying machine learning methods. As Hastie *et al*. (2009) point out, the cross-validation procedure must be correctly applied to the entire sequence of modelling steps. If a single cross-validation procedure is used for model tuning, estimation and evaluation, it would lead to misreporting of performance measures which would seriously overstate the performance of machine learning models. Neunhoeffer and Sternberg (2019) find that the performance of machine learning methods has been overestimated as a result of problematic cross-validation procedures in published studies of political science literature. They stress that such problem could be hard to identify.

In contrast, the LASSO method with cross-validation has a simple and transparent mechanism, which is also statistically closer to the conventional modelling approaches. The users have full control over the extent of penalisation on extreme parameter values with a single penalty parameter. The LASSO method only contains a single hyperparameter that needs to be fine-tuned by cross-validation, while some other machine learning methods may require fine-tuning multiple hyperparameters. Cross-validation is only applied for hyperparameter selection. Different data are used for model estimation and evaluation. Therefore, it is not subject to the concerns of

[1] The classification tree is an algorithm of binary recursive partitioning, which is an iterative process of splitting the data into binary partitions and then further splitting it up at each branch.

Hastie *et al.* (2009) and Neunhoeffer and Sternberg (2019). Nevertheless, a limitation is that the LASSO approach may not identify some of the non-linearities found to be important by more sophisticated machine learning methods.

This paper contributes to the strand of early warning literature of systemic banking crises by highlighting the importance of a systematic variable selection process. It proposes using the LASSO with cross-validation approach as a feasible, transparent and interpretable method to automate this process, and hence improve the out-of-sample predictive capability without compromising the crucial clarity of EWSs for policy communication.

The remainder of the paper is organised as follows. Section II reviews the relevant literature. Section III describes the data. Section IV introduces the empirical framework and the evaluation criteria. Section V presents and discusses the out-of-sample prediction results. Section VI conducts the sensitivity analysis. Section VII concludes the analysis presented here.Typesetter: Please check the section citations throughout this article and update per journal style (i.e. numbered or unnumbered section headings).

II Literature Review

There have been mainly two approaches to construct EWS for predicting systemic events since the 1990s. The first one is the signal extraction approach. Developed by Kaminsky and Reinhart (1999), the idea of the signal extraction approach is to warn of a potential systemic event if some leading indicators exceed their previously defined threshold level. Building on the work of Kaminsky and Reinhart (1999), authors such as Borio and Lowe (2002), Borio and Drehmann (2009) and Alessi and Detken (2011) further develop models based along the line of the signal extraction approach.

The second one is the multivariate probability approach, which assumes the probability of crisis occurrence to be a function of explanatory variables. Initially developed by Demirgüç-Kunt and Detragiache (1998) and (2000), multivariate logit models can be used to fit the data and transform the estimated crisis probability into a binary indicator of systemic banking crises. This econometric framework can be used to study the impact of various risk drivers on historical systemic events as well as produce predictions of future crises.

Demirgüç-Kunt and Detragiache (2005) emphasise the importance of interpretability of EWS and argue that multivariate logit models are more useful for policy-makers to identify factors associated with occurrences of crises. While multivariate logit models can be used to identify the common drivers of systemic banking crises, it cannot identify the heterogeneous drivers for each country. Davis and Karim (2008) find the signal extraction approach is better for developing country-specific EWS, while multivariate logit models are the most appropriate approach for a global EWS. Recent papers such as Duca and Peltonen (2013) and Behn *et al.* (2013) further develop the multivariate logit framework and use it to identify important early warning indicators and assess predictive performance. They find global variables and the interaction terms with domestic variables are significant predictors of systemic banking crises.

In the literature of multivariate logit EWS, it is common to make the dependent variable binary and omit the ‘in-crisis’ observations once a crisis occurs. This treatment aims to avoid the ‘post-crisis bias’ discussed by Bussiere and Fratzscher (2006). The behaviour of economic variables during a crisis would be significantly different from the tranquil times. Endogeneity problems could emerge from both crises and policies implemented to mitigate the crises.

To address the problematic ‘post-crisis bias’, one common approach many empirical studies adopted is to simply drop these ‘in-crisis’ observations that are considered to be uninformative about crisis prediction (Davis & Karim, 2008; Behn *et al.,* 2013; Duca & Peltonen, 2013). This approach eliminates the possibility of endogeneity at the cost of losing some observations.

Another strand of studies uses multinomial logit models to predict both the occurrence of crisis events and their duration. Studying systemic banking crises in low-income countries, Caggiano *et al.* (2014) find the multinomial logit model improves the predictive power compared to the binomial logit model. Taking a step further, Caggiano *et al.* (2016) compare the performance of binomial and multinomial logit models in the context of building EWS for systemic banking crises. They find the multinomial logit outperforms the binomial logit models, and the improvement increases with longer average duration of crises in the sample. Dawood *et al.* (2017) use a multinomial logit model to predict the occurrence of a sovereign crisis as well as its duration. However, they find the multinomial logit model does not predict better than binary logit models.

The LASSO method has been applied in the context of EWSs in many different disciplines. It is worth noting that Lang *et al*. (2018) apply this method to study bank stress from 1999Q1 to 2014Q4. This study differs from Lang *et al*. (2018) in three aspects. First, Lang *et al*. (2018) focus on the stress events of individual banks such as state aid cases, distressed mergers, defaults and bankruptcies, while this study focuses on predicting the systemic banking crises at a country level.[2] Second, while macro-financial indicators are included, their paper aims to use LASSO with cross-validation to select relevant predictors from a large set of bank-specific risk drivers related to financial statements. In this study, the LASSO with cross-validation approach is also used to capture the non-linear interactive relationship between variables and the information hidden in the lags. Last, their analysis attempts to predict the 'pre-distress events', while my research aims to predict the occurrence of systemic banking crises.

III *Data*

(i) Systemic Banking Crisis Database

The systemic banking crisis database in this study covers 23 OECD countries during the period 1970Q1 to 2018Q3. Table 1 summarises the identified systemic banking crises and their starting and ending date for each country.

The systemic banking crisis dataset used in this study is mainly based on Laeven and Valencia (2018), which updates a widely used global systemic banking crisis database developed by Laeven and Valencia (2013).

I also supplement the database with the work of Detken *et al*. (2014). Focusing on systemic banking crises associated with domestic credit and financial cycles in European countries, they update and amend the database of Babeckỳ *et al*. (2014) for European countries under the framework of the European Systemic Risk Board based on country experts' judgement.[3] The systemic banking dataset is therefore enriched by including past crises in European countries that are not recorded by Laeven and Valencia (2018).[4] This is in line with the purpose of early warning models, which is to identify the extremely vulnerable states of the financial system and to inform policy actions.

(ii) The Explanatory Variables

To predict systemic banking crises, we need to select a range of indicators that can capture the sources contributing to the build-up of systemic risk and macro-financial vulnerabilities. The selection of key variables in this study broadly follows the discussion of Behn *et al*. (2013) and Duca and Peltonen (2013). Details of the explanatory variables are presented in Appendix I: Table A1.

In the benchmark model, I include the same variables as the best performed model (Model 5) of Behn *et al*. (2013). Credit growth, the credit gap and their global counterparts are included to capture the domestic and global credit market development. Credit growth is the year-to-year rate, while the credit gap is calculated as the deviation of the credit-to-GDP ratio from its long-run backward-looking trend.

The credit gap is calculated using the modified Beveridge–Nelson (BN) filter proposed by Kamber *et al*. (2018). Many studies in the literature, such as Behn *et al*. (2013) and Alessi and Detken (2018), use an one-side Hodrick–Prescott (HP) filter with a very high smoothing parameter, $\lambda = 400{,}000$, to capture the long-run trend of the credit-to-GDP ratio. They follow the suggestion of Drehmann *et al*. (2011) that the financial cycle should be considered four times longer than the business cycle. This approach is recommended by the European Systemic Risk Broad[5] and the Basel Committee (BIS, 2010). However, it is subject to the critiques of Orphanides and Norden (2002) and Edge and Meisenzahl (2011), namely that the real-time output gap estimates by HP filter are unreliable.

The modified BN filter by Kamber *et al*. (2018) particularly addresses the key critique of Orphanides and Norden (2002). The output gap

[2] Lang *et al*. (2018) define the 'state aid cases' as direct capital injections, asset protection measures and loans/guarantees other than guaranteed bank bonds.

[3] This database also includes 'would-be' crises, which are the cases where systemic banking crises were likely to occur had it not been for policy action or external events that dampened the financial cycle. These 'would-be' crises are excluded from my dataset.

[4] This includes the following crises: Italy 1994–1995, Germany 2000–2003, the UK 1973–1975 and 1991–1995, Spain 1982–1985, France 1994–1995 and Denmark 1987–1993.

[5] Recommendation of the European Systemic Risk Broad of 18 June 2014 on guidance for setting countercyclical buffer rates (ESRB/2014/1).

Table 1
Systemic Banking Crisis: Summary

Country	Crisis	Start	End	Start	End	Start	End	Tranquil period	Crisis periods	Crisis share (%)
Australia	0							195	0	0.00
Austria	1	2008Q4	2012Q4					178	17	8.72
Belgium	1	2008Q4	2012Q4					178	17	8.72
Canada	0							195	0	0.00
Denmark	2	1987Q1	1993Q4	2008Q3	2012Q4			149	46	23.59
Finland	1	1991Q3	1995Q4					177	18	9.23
France	2	1994Q1	1995Q4	2008Q1	2009Q4			179	16	8.21
Germany	2	2000Q1	2003Q4	2008Q1	2009Q4			171	24	12.31
Greece	1	2008Q1	2012Q4					175	20	10.26
Ireland	1	2008Q1	2012Q4					175	20	10.26
Israel	1	1977Q1	1984Q2					195	0	0.00
Italy	2	1994Q1	1995Q4	2008Q3	2009Q4			181	14	7.18
Japan	1	1997Q4	2001Q4					178	17	8.72
South Korea	1	1997Q4	1998Q4					190	5	2.56
Netherlands	1	2008Q1	2009Q4					187	8	4.10
New Zealand	0							195	0	0.00
Norway	1	1991Q4	1993Q4					186	9	4.62
Portugal	1	2008Q3	2012Q4					177	18	9.23
Spain	2	1982Q2	1985Q3	2009Q2	2013Q2			164	31	15.90
Sweden	2	1990Q3	1993Q4	2008Q3	2009Q4			175	20	10.26
Switzerland	1	2008Q3	2009Q4					189	6	3.08
UK	3	1973Q4	1975Q4	1991Q1	1995Q2	2007Q1	2011Q4	148	47	24.10
USA	2	1988Q1	1988Q3	2007Q4	2011Q4			175	20	10.26

estimated by the modified BN filter is reliable, in the sense of more accurate in out-of-sample forecasting and subject to smaller revisions. It is also intuitive in the sense of 'being persistent, large in amplitude, and generally moving procyclically in terms of the NBER reference cycle' (Kamber *et al.*, 2018).

Macroeconomic-related variables such as the real GDP growth rate and inflation are included in the benchmark model to capture broad trends in the real sector, which may be closely related to the credit and financial cycle. Furthermore, the short-term rate and term spread are included in the pool of candidate variables to capture the risk appetite, as low cost of funding and large term spread may encourage risk-seeking activities for investors.

Asset prices usually show strong growth preceding systemic banking crises. Equity price growth and house price growth are therefore included in the benchmark model to capture the booms and busts in asset prices, which are tightly connected to systemic events as documented by Reinhart and Rogoff (2009). The house-price-to-income ratio, house-price-to-rental ratio and the deviation of these ratios from their long-run trends are also included as variables that can capture housing market imbalances. They are suggested to be good predictors for systemic events by many studies (Duca & Peltonen, 2013; Aldasoro *et al.*, 2018; Alessi & Detken, 2018).

As discussed by Kaminsky and Reinhart (1999), the 'twin crises', namely currency crises and banking crises, became closely linked following the financial liberalisation in many advanced countries in the 1980s. The current account relative to GDP and the real effective exchange rate are included to capture imbalances in the external sectors that may contribute to financial vulnerability. Davis *et al.* (2016) find the marginal effect of credit growth on the probability of banking crises greatly depends on the size of the current account deficit to GDP ratio. Their findings suggest the inclusion of both the current account relative to GDP and its interaction terms with credit growth and the credit gap. Reinhart and Rogoff (2009) also find that a widening current account imbalance preceded banking crises in OECD countries.

Furthermore, I include global variables to allow global imbalances to affect domestic financial stability through channels such as financial linkages and trade (Kaminsky & Reinhart, 2000). The global variables are calculated as the average of the US, UK, Japan and the Euro area. Following Behn *et al.* (2013), I include global variables for credit growth, the credit gap, real GDP growth, inflation, real house price growth and equity price growth.

Interaction terms are generally found to be very important in the literature on early warning models of systemic banking crises as they can improve the model's performance (e.g. Davis & Karim, 2008; Behn *et al.*, 2013; Duca & Peltonen, 2013). The best performing model of Behn *et al.* (2013) includes four interaction terms, which are the mutual interactions between the domestic and global credit growth and credit gap. However, as there is a large pool of variables and transformations, many interaction terms are possible. For example, Duca and Peltonen (2013) even included second-order interaction terms in their benchmark model.

The lags of possible explanatory variables may contain valuable information that can help with predicting systemic events. However, considering including more lags further widens the list of possible covariates to be included in the model. In practice, it can be troublesome to include more lags in the conventional multivariate logit model. The heavily imbalanced classes of the dependent variable naturally restrict the number of covariates that can be included in the model. Incorporating more explanatory variables may lead to a serious over-fitting problem or even perfect prediction in the extreme. This issue can be addressed by the LASSO with cross-validation approach which helps extract valuable information for prediction from a large variable pool.

IV Empirical Framework and Automatic Variable Selection

(i) Multivariate Logit Model

Following the literature, models are estimated in the multivariate logit framework. In the benchmark model, the probability of a systemic banking crisis for country i at time t follows the logistic distribution function.

$X_{i,t-h}$ is the vector of h-period lagged explanatory variables, which are indicators for financial vulnerability. The specific lag may vary across variables, so not all variables need to be at time $t-h$. The multivariate logit model is written below:

$$\ln\frac{Pr(Y_{i,t}=1)}{1-Pr(Y_{i,t}=1)} = domestic_{i,t-h}\beta_1 + global_{i,t-h}\beta_2 + interaction_{i,t-h}\beta_3 + c + \varepsilon_{i,t}. \quad (1)$$

The right-hand side of the logit equation includes domestic variables, global variables, interaction variables and a constant. β_1, β_2 and β_3 are the vectors of coefficients for corresponding variables. The benchmark model includes a range of pre-selected explanatory variables based on the best-performing model of Behn *et al.* (2013) (their model 5).[6] Specifically, the explanatory variables in the benchmark model are domestic and global credit growth and credit gaps, the mutual interactions of the four credit variables, inflation, domestic and global GDP growth, equity price growth and house price growth.

Set-up of the dependent variable

The construction of the dependent variable is carried out in two steps. First, I set the dependent variable to '1' in the beginning of a systemic event and to '0' otherwise. This differs from many EWS studies that define their dependent variables as the 'pre-crisis' periods, which is to set an early warning horizon and make the dependent variable be '1' in several quarters or years preceding the systemic events and be '0' otherwise (Behn *et al.*, 2013; Duca & Peltonen, 2013).

In discrete choice models, the 'pre-crisis' approach implicitly assumes that all observations are independent conditional on covariates, but the assumption of conditional independence is violated by the construction of the dependent variable, which would be inherently serial

[6] The benchmark model is not exactly the same model as the model 5 of Behn *et al.* (2013), as the dependent variable set-up and the treatment of fixed-effects are different. Modelling with alternative dependent variable set-ups and the inclusion of fixed-effects are discussed in the sensitivity analysis.

correlated for each country as most of it would be '0's. Considering the purpose of EWS, warning signals should be issued to reflect the transition from tranquil periods to the deteriorated crisis-incoming periods. If the dependent variable includes a list of sequential '1's in the 'pre-crisis' periods, it will mechanically produce autocorrelation in these '1's. It implies the warning signals are not issued independently, but depends on whether warning signals are issued in previous periods. This approach would lead to first-order autocorrelation in the residual and model misspecification. Intuitively this could be handled by incorporating the lagged dependent variable to capture the dynamic response in discrete choice models. However, this treatment is not feasible in the context of predicting systemic events, because crises are not really 'discrete choices' as the timing of crises cannot be predetermined.

To overcome this issue, my approach is to make the model directly predict the starting period of systemic banking crises rather than the 'pre-crisis' periods. Specifically, the model aims to catch the point of transition from tranquil periods to systemic banking crises, while the extent of 'early warning' is reflected in the interpretation and evaluation of the generated signals. Using this approach, the evidence of the model being correctly specified is that there is no serial correlation in the residuals.[7]

Second, following the common practice in the literature, I drop all the subsequent periods that have an ongoing crisis from the sample for each country. Once a crisis occurs, it would be followed by crisis deepening periods and economic recovery periods. Economic variables would be significantly affected by the crisis itself, the policies made to moderate the crisis, and people's sentiment and expectations. This creates an endogeneity problem in the model. The behaviour of economic variables in a crisis would be significantly different from the tranquil times during the adjustment process of the economy before reaching a more sustainable level or growth path. This is the so-called 'post-crisis bias' problem discussed by Bussiere and Fratzscher (2006).

To avoid the endogeneity problem and the post-crisis bias, one common approach for studies using binary logit models is to drop the in-crisis periods from the sample, such as Duca and Peltonen (2013), Behn *et al.* (2013) and Davis and Karim (2008). Some recent studies in systemic event analysis adopt multinomial logit models to address the 'post-crisis bias' problem and find that not discarding the 'in-crisis' information could improve the performance (Caggiano *et al.,* 2014, 2016). Notably, Caggiano *et al.* (2016) find that the longer the average duration of the crisis in the sample, the larger the improvement a multinomial logit model could make compared with a binomial logit model.

The multinomial logit models are valid under the implicit assumption of the Independence of Irrelevant Alternatives (IIA), which states that characteristics of one particular choice alternative do not affect the relative probabilities of choosing other alternatives. In the context of systemic event prediction, it means that the relationship between tranquil periods and the occurrences of crises should not be affected by the inclusion of in-crisis periods. The two most common IIA assumption tests, the Hausman–McFadden (HM) test by Hausman and McFadden (1984) and the Small–Hsiao (SH) test by Small and Hsiao (1985), are performed to test the IIA assumption. However, they provide conflicting information on whether the IIA assumption has been violated.[8] This is not surprising as simulation studies conducted by Fry

[7] Cumby–Huizinga panel serial correlation tests are conducted for the linear probability models under the two approaches of dependent variable construction over the full sample. Both models have the same specification as the benchmark model. The null hypothesis of no first-order serial correlation in the error is not rejected at the 5 per cent level of significance with the starting periods of crises serving as the dependent variable. In contrast, if the dependent variable is defined as the pre-crisis periods, there is serial correlation in the error, suggesting serious model misspecification. The results are robust to varying sample size and heteroskedasticity.

[8] The HM test gives evidence in favour of the null hypothesis that IIA holds. This is similar to the HM test results of Bussiere and Fratzscher (2006) and Caggiano *et al.* (2014, 2016). However, the SH test results strongly suggest that the IIA assumption does not hold. Such conflicted results exist for the full sample as well as the base training sample that is used to perform the out-of-sample forecasting in Section V.

and Harris (1996, 1998) Cheng and Long (2007) find that the IIA tests often produce inconsistent results based on different variations in test and data structure.[9]

With conflicted IIA test results, one further alternative could be the nested logit, which models the outcomes sequentially. However, this study focuses only on predicting the occurrence of crises, and the advantage of the proposed method is better demonstrated based on established EWS literature. Due to such consideration, I decide to use the conventional binary logit framework with in-crisis periods omitted from the sample to address the post-crisis bias problem in the main analysis. To reinforce the message delivered, a multinomial logit modelling framework is employed to examine the robustness of the results in the sensitivity analysis section.

By omitting all the in-crisis periods that are not informative on the transition from tranquil times to systemic events, this study aims to predict the occurrence of a systemic banking crisis, while leaving crisis duration outside the scope. Eventually, the dependent variable is binary as all tranquil periods have the value of 0. Periods that are the beginning of systemic banking crises have the value of 1.

Fixed effects or not?

There are two different estimation strategies among studies using multivariate logit models to predict systemic events, regarding whether the country fixed effects are considered or not.

One strand of studies estimates the pooled logit model without country fixed effects (Demirgüç-Kunt & Detragiache, 1998; Davis & Karim, 2008; Duca & Peltonen, 2013). The motivation for excluding the fixed effects is to avoid the selection bias. By including country fixed effects, all the countries that never experienced a systemic banking crisis in the sample have to be excluded from the estimation, because the country-specific dummy would be perfectly correlated with the banking crisis dummy for these countries. The estimation may therefore be biased due to the omission of these countries. In contrast, Behn *et al*. (2013) incorporate the fixed effect in their model nevertheless with the belief that the importance of addressing the unobserved time-invariant heterogeneity across countries outweighs the selection bias.

Van den Berg *et al*. (2008) argue that it is unlikely that systemic events are homogeneously caused by identical factors in their *in-sample* analysis. In contrast, with an extensive *out-of-sample* forecasting exercise, Fuertes and Kalotychou (2006) find a very weak association between *in-sample* fit and *out-of-sample* forecast performance in the context of EWS for sovereign debt crisis. They find the fixed-effect model describes the data better but perform poorly in out-of-sample prediction. Similarly, Dawood *et al*. (2017) also find that fixed-effect models fit the data better but pooled logit models perform significantly better in *out-of-sample* forecasting. As the objective of this study is to predict systemic events, I decided it is the best to use the pooled logit model without fixed effects in the main analysis. In the sensitivity analysis, I find the main results are still robust even if country fixed effects are incorporated.

(ii) LASSO with Cross-Validation

This study proposes to use the LASSO method with cross-validation to select variables to be included in the logit model specification for forecasting.

Among many machine learning techniques that could be used to enhance the variable selection process, LASSO is a relatively simple and straightforward one. In contrast with other more complicated techniques, LASSO is recommended and applied for two reasons. First, it is designed specifically for variable selection by addressing the over-fitting issue and enhancing the prediction accuracy. Second, LASSO is transparent and statistically closer to the conventional modelling approaches in comparison with a range of more complicated machine learning techniques.

Designed for the purpose of variable selection, the idea of LASSO regression is to drop variables with small coefficients and only keep variables that make a significant contribution to the model's performance.

[9] Fry and Harris (1998) find that the acceptance or rejection of the IIA assumption depends on which IIA test as well as variants of a given test that is used. Cheng and Long (2007) undertake Monte Carlo simulations to examine the properties of different IIA tests in a multinomial logit framework. They find the HM test has poor size properties even with a large sample size, while the SH test could have reasonable or poor size properties depending on the data structure. Based on such results, Cheng and Long (2007) and Long and Freese (2014) even go so far as to argue that 'tests of the IIA assumption that are based on the estimation of a restricted choice set are unsatisfactory for applied work' and provide no useful information.

In a generic form, the ordinary logit regression with binary response can be written as follows:

$$Pr(y_i = 1) = \pi_i = \frac{e^{y_i\beta}}{1 + e^{y_i\beta}}. \qquad (2)$$

The generic log-likelihood function can be written as in the following form accordingly:

$$\begin{aligned} L(\beta) &= \sum_{i=1}^{n}[y_i \log(\pi_i) + (1-y_i)\log(1-\pi_i)] \\ &= \sum_{i=1}^{n}[y_i x_i \beta - \log(1 + e^{x_i}\beta)]. \quad (3) \end{aligned}$$

In the standard logit regression, the parameter β is estimated by maximising the log-likelihood function $L(\beta)$. In the LASSO regression, the log-likelihood function is penalised with an additional term.

$$L_{\lambda}^{lasso}(\beta) = L(\beta) - \lambda \sum_{j=1}^{p} \|\beta\|_1. \qquad (4)$$

The choice of the penalty parameter λ is critical when LASSO is applied for variable selection. A higher penalty parameter indicates that more coefficients of variables would shrink to zero and therefore fewer variables are selected to be included in the model. As the purpose of this study focuses on prediction rather than model interpretation, it is favourable to select only those predictors with substantial coefficients to improve the prediction performance. The penalty parameter λ should be fine-tuned to fulfil this purpose. I use the K-fold cross-validation method to select the optimal penalty parameter λ that can maximise the out-of-sample prediction accuracy.

K-fold cross-validation is one of the most popular resampling techniques to evaluate the effectiveness of models in the machine learning literature. It is conducted by randomly splitting the sample into K folds of approximately equal size. One fold is taken out as the *out-of-sample* validation fold, while the other K-1 folds are used as training folds to estimate the model. The estimated parameters from the K-1 folds are used to predict the dependent variable in the one validation fold. This procedure is repeated for K times until every fold has served as the evaluation fold. The optimal λ is selected to be the one that the model is estimated with that minimises the estimated deviance, which measures the goodness of fit and represents the accuracy of the prediction.

It is important to note that the entire sample is *not* used in the cross-validation. A portion of the data is withheld so that the model's predictive performance can be evaluated with data not used in selecting its specification.

As the dependent variable is binary and the number of '1' is relatively rare, I conduct five-fold stratified cross-validation to select the optimal λ. Typical K-fold cross-validation is conducted with $K = 10$ or $K = 15$ to obtain a good bias-variance trade-off as suggested by James *et al.* (2013).[10] The cross-validation is stratified so that every fold has approximately the same proportion of observations of values of different classes; that is, the number of '1's and '0's is approximately the same across folds.

While the K-fold cross-validation technique is frequently applied in the recent machine learning studies to build EWS (e.g. Holopainen & Sarlin, 2017; Alessi & Detken, 2018), the validity of the cross-validation technique should not be considered as guaranteed. Cross-validation usually requires the data to be independent and identically distributed. However, if the data have time series structure with inherent possible serial correlation, the process of randomly reshuffling the observations will break the feature of time dependence and could lead to questionable results.

It is critical to adopt the correct cross-validation technique when constructing EWS for crisis prediction. The application of the standard K-fold cross-validation with time-series data is justified by Bergmeir *et al.* (2018). They find that a normal K-fold cross-validation procedure is valid if the residuals of the model are uncorrelated. In other words, if the model under-fits the data, it would lead to serially correlated errors. Therefore, the residuals of LASSO logit models should always be checked to ensure it is valid to perform K-fold cross-validation.

(iii) *Model Evaluation Criteria*

The criteria of model evaluation are of key importance in the prediction of systemic banking crises. This study follows Sarlin (2013) and adopts the measure of usefulness as the criteria

[10] Since systemic banking crises are very rare events, the heavily imbalanced class size of the binary dependent variable makes 10-fold cross-validation infeasible. Otherwise, there will not be enough observations of crises in each fold. Hence, the cross-validation is performed with five folds.

to evaluate the predictive performance of the model. This usefulness measure extends the one developed by Alessi and Detken (2011) and takes both (i) the preference of the policy-makers between Type I errors (missing the crises) and Type II errors (false alarms); and (ii) the imbalanced frequency of tranquil times and crises into consideration.

This approach is built upon the 'signalling approach' that was originally developed by Kaminsky *et al.* (1998) and Kaminsky and Reinhart (1999) and widely used in crisis prediction studies. The first step is to construct the contingency matrix (see Table 2).

The EWS issues a signal whenever the indicator passes a certain threshold. If a warning signal is issued by the model, then that period is predicted to be the starting period of a systemic banking crisis. If there is no signal given, the period is predicted to be a tranquil period. The share of Type I error is the number of missed crisis events relative to the total number of crisis events, which is represented as $T_1 = \frac{FN}{TP+FN}$. The share of Type II error is the number of false alarms relative to the total number of tranquil periods, which is represented as $T_2 = \frac{FP}{FP+TN}$. The hitting rate is therefore represented as $1-\frac{FN}{TP+FN}$, which is the share of correctly predicted crises relative to the total number of crises.

There are two issues related to predicting the starting periods of systemic banking crises. First, some degree of ambiguity is inevitable in the documentation of starting periods for systemic banking crises. There may be no consensus on the exact quarter of a banking crisis becoming systemic. Second, the prediction of systemic events should serve the purpose of early warning eventually, leaving time to inform and implement relevant policies. To accommodate these issues, a warning signal is taken to be correct if it is issued within six quarters before the occurrence of an actual crisis.

TABLE 2
The Contingency Matrix

	Crisis event (within six quarters)	Tranquil period
Signal	True positive *Correct signal*	False positive *False alarm* *(Type II error)*
No signal	False negative *Missed crisis* *(Type I error)*	True negative *Correct silence*

When evaluating the forecasting performance of the model, policy-makers are likely to put different weight on the Type I and Type II errors in the contingency matrix, because the cost of missing an incoming crisis is far larger than getting a false alarm. To take the policy-maker's relative preference for Type I and Type II errors into account, a loss function is defined following Sarlin (2013).

$$L(\mu) = \mu T_1 P_1 + (1-\mu) T_2 P_2. \qquad (5)$$

P_1 and P_2 are the unconditional probabilities of the vulnerable pre-crisis periods and tranquil periods, respectively. μ is the policy-makers' preference on the trade-off between issuing false alarm (Type II errors) and missing crises (Type I errors). It can be seen that the loss of policy-makers is a weighted average of Type I and Type II errors based on their preference for misclassification captured by the parameter μ, adjusting for the imbalance of classes in the panel.

The usefulness of model is defined as

$$U = min[\mu P_1, (1-\mu)P_2] - L(\mu). \qquad (6)$$

μP_1 is the policy-makers' loss if a model never issues a crisis signal. $(1-\mu)P_2$ is the policy-makers' loss if a model always issues crisis signals. $min[\mu P_1, (1-\mu)P_2]$ is therefore the loss if the model is ignored. The usefulness U represents the absolute gain for policy-makers using the model compared to the case if they completely ignore the model. Hence, U is expected to be positive for any useful model.

I further define the relative usefulness U_r as the ratio of the absolute usefulness relative to a perfectly performing model that achieves maximum possible usefulness. U_r is the normalisation of the usefulness U so that its magnitude is bounded between 0 and 1.

$$U_r = \frac{U_a}{\min[\mu P_1, (1-\mu)P_2]}. \qquad (7)$$

The performance of different models is evaluated by calculating and comparing their relative usefulness U_r. A model with higher U_r performs closer to the perfect model than other models and therefore is more useful to policy-makers.

V *Out-of-Sample Forecasting: Practice and Evaluation*

To evaluate and compare the performance of the benchmark model and the LASSO logit models, this section performs an *out-of-sample* forecasting evaluation on predicting the occurrence of systemic banking crises.

The full sample must be split into the training dataset and the evaluation dataset. The training set is the sample used to select the model specification with LASSO. The evaluation set is the sample used to evaluate the *out-of-sample* performance of models. The base training set covers the periods 1970Q1–2004Q4, while the evaluation set covers 2005Q1–2009Q4. As there is no systemic banking crisis occurring after the 2010s, I restrict the evaluation period to test the models' ability to predict the 2007–2008 GFC.

There are two distinct parts in the practice. Firstly, I adopt the LASSO with cross-validation approach to select the most important predictors from the pool of candidate variables over the base training sample. For this purpose, five-fold cross-validation is performed to find the best tuning parameter λ. Secondly, I produce recursive *out-of-sample* forecasts with the selected variables over the evaluation dataset. The forecasts are then evaluated and compared to the performance of the benchmark model. The *out-of-sample* forecasting exercises are performed in a pseudo-real-time manner. In each time period the information set only contains knowledge that is available up to that particular time period. To reiterate, fresh data are used to evaluate the model.

It is worth noting that the specification of the LASSO logit model is pinned down with the base training periods 1970Q1–2004Q4. It is then recursively estimated over the evaluation period. Although LASSO can choose a new model specification at each period with more up-to-date information, it would be hard for policy-makers to communicate to the public with a model that keeps changing specifications every quarter. Therefore, it is decided not to frequently re-specify the model specification for consistency in communication.

(i) *Variable Selection with the Training Set*

The first part of the practice is to select the most important predictors using LASSO with cross-validation using the training dataset. To evaluate whether the automatic variable selection would enhance the forecasting performance, the LASSO logit model is compared with the benchmark model. As mentioned in Section IV.(i), the benchmark model includes pre-selected explanatory variables based on the best-performing model of Behn *et al.* (2013), with all variables being lagged by one period.

The information in longer lags is usually not used in EWS studies for banking crisis. The heavily imbalanced class size of the dependent variable restricts the capability of conventional multivariate logit models to incorporate longer lags because of the potential over-fitting issue. For example, adding lags up to four quarters to the explanatory variables would already make the benchmark model produce a perfect prediction. The proposed LASSO method with cross-validation is capable of identifying and using such information by selecting only the important predictors in the longer lags. To make this point, I repeat the automatic variable selection using the same pool of candidate variables but with different lag orders.

Model 1 has the selected predictors chosen from a pool of candidate variables that include only one period lag, just like the benchmark model. Model 2 contains selected predictors from candidate variables with their lags up to four quarters. Table 3 summarises the selected predictors by Model 1 and Model 2.

It must be made clear that the handful of predictors selected by the automatic variable selection process with cross-validation are those that can achieve the best *out-of-sample* forecasting outcomes. This is appropriate given the goal of developing an EWS for prediction purposes. However, it is inappropriate to interpret and analyse these as a causal relationship when using predictive models. It is also infeasible to make any economic interpretation on why one particular variable is chosen over another among multiple highly correlated informative variables.[11]

[11] As with many variable selection techniques, the particular variables with non-zero coefficients selected by LASSO depend on the vagaries of sampling from the underlying population. While this method selects variables that jointly have the best out-of-sample forecasting property, there is no meaningful explanation on why it chooses a particular variable and omits the others when there are multiple highly correlated informative predictors. There is no clear economic interpretation on why an interaction term is selected but not the individual terms, or why lag 1 of a variable is selected but not lag 2.

TABLE 3
LASSO with Cross-Validation Selected Variables

Model 1		Model 2
lag 1	lag 1	
credit growth × credit gap		real house price growth
real effective real house × exchange rate price growth		global credit growth × global credit gap
credit growth × global credit gap		real effective exchange rate × real house price growth
credit gap × global credit gap	lag 3	credit gap × global credit gap
		credit growth × credit gap
		credit growth × current account to GDP
	lag 4	
		current account to GDP × global equity price growth
		global credit gap × global inflation
$\lambda = 6.867$		$\lambda = 6.021$

TABLE 4
Cumby–Huizinga Test for First-Order Serial Correlation

	P-value
Model 1	
Training set (1970Q1–2004Q4)	0.4310
Training set + Evaluation set(1970Q1–2009Q4)	0.0185**
Model 2	
Training set (1970Q1–2004Q4)	0.4290
Training set + Evaluation set(1970Q1–2009Q4)	0.8990

Nevertheless, the automatic variable selection process can deliver valuable information on what the most important early warning predictors are among a large number of variables and indices. With different samples and candidate variables, LASSO with cross-validation may select different predictors. If some predictors are frequently selected, even in different forms such as different lag order or interaction terms, they should be considered as key indicators that need particular attention and monitoring by policy-makers.

Model 1 and Model 2 share five common key indicators, namely credit growth, credit gap, global credit gap, real effective exchange rate and real house price growth. Section V.(iv) discusses how the development of these indicators may affect financial stability.

Before progressing to the next step, it is essential to check if the selected models are correctly specified. It should not be taken as guaranteed that no serial correlation will exist in the residuals. A serial correlation test should be performed not only on the training sample but also on the full sample because the model will be recursively estimated over the evaluation periods. The serial correlation test results are listed in Table 4.

Model 1 has no autocorrelated error in its residuals over the base training sample 1970Q1–2004Q4. However, in the recursive prediction exercise, as we keep expanding the size of the training sample, Model 1 starts to have autocorrelation in its residuals when the training sample rolls over the 2007–2008 financial crisis.

By restricting the lag order of the candidate variables to one, the LASSO method with cross-validation selects only four variables based on the base training sample 1970Q1–2004Q4. This specification is suitable for the training sample but could be overly parsimonious as the sample size extends. Therefore, Model 1 would be misspecified if we use it after the 2007–2008 financial crisis. The predictive result of Model 1 is reported in the next section for comparison purposes.

This problem can be solved using an extended training dataset to re-select the variables, but that would be unnecessary as we can simply expand the pool of candidate variables and select better models. Selecting predictors from a larger pool of candidate variables, Model 2 does not have this problem. Nevertheless, the forecasting practice is conducted for both Model 1 and Model 2.

(ii) Forecasting Results and Performance Evaluation

To evaluate and compare the predictive performance of models, the *out-of-sample* exercise is carried out recursively in the following way:

1. Starting with the base training sample 1970Q1–2004Q4, estimate the model on data that contains information only available up to that period. Predict the probability of a systemic banking crisis that is about to happen in the next period, which is out-of-sample.
2. Compute the *in-sample* relative usefulness U_r given the policy preference μ for all thresholds from 0 to 1. Apply the threshold that yields the highest relative usefulness to the predicted probability of a crisis in the next period and generate a 'real-time' warning signal of 0 or 1.
3. Expand the training set by one period forward and repeat steps 1–3 recursively.
4. Collect the warning signals produced over the entire evaluation period and calculate the relative usefulness based on the number of false signals and missed crises.

Relevant studies such as Behn *et al.* (2013) and Lang *et al.* (2018) set the value of policy preference parameter μ as 0.9. This is due to the consideration that the cost of missing a systemic event is far larger than issuing a false alarm, which would at least call for more attention to the current distress of banks. As discussed by Sarlin (2013), setting the policy parameter μ as 0.9 and adjusting for the unconditional probability of systemic events are equivalent to the approach of setting $\mu = 0.5$ without adjusting for the unconditional probability of systemic events.[12] In addition to setting the policy preference parameter μ as 0.9, the performance of models will be evaluated with the policy preference parameter μ being 0.8 and 0.7, which reflect cases that policy-makers may have relatively stronger preference to avoid false alarm rather than missing crises. This is to ensure the forecasting results are robust to varying policy preferences.

In addition to the measure of relative usefulness, the noise-to-signal ratio is also calculated for each model to supplement the forecasting evaluation from an objective perspective. The noise-to-signal ratio is calculated as the ratio of Type II error to the hitting rate. A smaller noise-to-signal ratio indicates a more precise prediction, regardless of policy-makers' preference.

Table 5 summarises the results of the recursive *out-of-sample* forecasting practice. Both Model 1 and Model 2 are LASSO logit models, whose predictors are selected by the LASSO method with cross-validation from pools of candidate variables. Model 1 only selects predictors from the pool of one-period lagged candidate variables. Model 2 selects predictors from a pool of candidate variables with their lags up to four quarters. The comparison of these models' results under different policy preferences can demonstrate how the LASSO method with cross-validation can use information in longer lags to produce more useful and more stable predictions. There are several notable results.

Regardless of different policy preferences, both Model 1 (LASSO lag 1) and Model 2 (LASSO lags 1–4) achieve higher relative usefulness, indicating they outperform the benchmark model in all cases. They also generate notably fewer false alarms than the benchmark model. The superior performance of Models 1 and 2 is primarily driven by the capability to produce significantly lower Type II errors consistently. The LASSO logit models all have very low noise-to-signal ratios, indicating a huge improvement in prediction precision compared to the benchmark model.

For each model, as the policy preference parameter μ becomes smaller, there would be more weight in the loss function on Type I errors but less on Type II errors, and result in lower relative usefulness. This suggests policy-makers would benefit from more warning signals so as to avoid the huge cost associated with systemic events.

At $\mu = 0.7$, the relative usefulness of the benchmark model is negative, suggesting this model is not useful for policy-makers to predict the 2007–2008 financial crisis. In comparison, Model 1 and Model 2 have positive usefulness in all cases.

The relative usefulness of Model 2 is only slightly lower than Model 1 at $\mu = 0.9$, but it outperforms Model 1 at lower μ. It also achieves a lower noise-to-signal ratio. It is quite obvious that the results of the benchmark model and Model 1 are relatively sensitive to varying policy preferences. As μ becomes smaller, policy-makers prefer fewer false alarms relative to missing crises. More heavily penalised Type II errors

[12] This approach has been used by Alessi and Detken (2018) and Duca and Peltonen (2013).

TABLE 5
Main Results: Forecast Evaluation

	Benchmark			Model 1			Model 2			Model 3		
				LASSO lag 1			LASSO lags 1–4			LASSO lags 1–4 with the linear terms		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Policy preference (μ)	0.9	0.8	0.7	0.9	0.8	0.7	0.9	0.8	0.7	0.9	0.8	0.7
Relative usefulness (U_r)	0.48	0.18	−0.09	0.67	0.47	0.27	0.64	0.55	0.51	0.50	0.33	0.22
Noise-to-signal ratio	0.16	0.16	0.14	0.04	0.03	0.02	0.04	0.02	0.01	0.08	0.05	0.04
Hitting rate	0.72	0.72	0.65	0.73	0.56	0.32	0.69	0.61	0.54	0.60	0.43	0.32
Type I error	0.28	0.28	0.35	0.27	0.44	0.68	0.31	0.39	0.46	0.40	0.57	0.68
Type II error	0.11	0.11	0.09	0.03	0.02	0.01	0.03	0.01	0.00	0.05	0.02	0.01

reduce the relative usefulness of the benchmark model. Model 1, while obtaining lower Type II errors, performs poorly in Type I errors. In contrast, Model 2 achieves a much more stable forecast under varying policy preferences.

Compared with the benchmark model and Model 1 that only include one period lagged variables, the results of Model 2 suggests that longer lags of the covariates contain valuable information that helps reinforce the stable predictive performance. This finding further justifies the use of LASSO with cross-validation to automate the variable selection process to extract information from a large variable pool for better prediction. The robustness of this finding is investigated in the sensitivity analysis section.

In Model 1 and Model 2, selected interaction terms are included in the model solely without the linear terms. As the ultimate objective is to achieve better forecasting performance rather than model interpretation, the implication of including the linear terms of the interaction terms is unclear. Columns (10)–(12) of Table 5 report the results of Model 3, which extends Model 2 by including the linear terms for the selected interaction terms. With the additional linear terms, Model 3 performs strictly worse than Model 2 in all cases.[13]

Overall, the results suggest that the automatic variable selection process using LASSO with cross-validation is useful to extract information that may not be accessible by the conventional multivariate logit models. Therefore, more useful, more stable and more precise predictive results are generated.

(iii) Forecast Horizon

The focus of my modelling approach is to capture the dynamics of the key transition from the tranquil periods to the systemic events and hence better predict the occurrence of crises. The feature of such a key transition is most apparent in the very period preceding the outbreak of a systemic banking crisis. Even if a warning signal should be interpreted as a likely incoming crisis within the next six quarters (as discussed in Section IV.(iii)), the out-of-sample forecasting exercise in Section V is essentially a one-period ahead prediction.

In practice, policy-makers may need a longer forecast horizon to allow more time for policy selection and implementation to reduce the probability of a crisis occurring. To make sure the LASSO with cross-validation approach could contribute to the purpose of early warning under different scenarios, I extend the forecast horizon and conduct h-step ahead direct forecasting exercise to evaluate the performance of the benchmark model and Model 2 (LASSO lags 1–4).[14] Table 6

[13] The same practice is conducted for Model 1, which is found to perform better without including the linear terms.

[14] As discussed in Section V.(i) and shown in Table 4, Model 1 has serial correlation problem when rolling over the evaluation set. This is because the restricted pool of candidate variables leads to a specification that is suitable for the training set, but is too parsimonious for larger sample.

Table 6
Varying Forecast Horizon

	Model 2			Benchmark		
	LASSO lags 1–4					
Forecast horizon	$t + 4$	$t + 8$	$t + 12$	$t + 4$	$t + 8$	$t + 12$
Relative usefulness	0.59	0.56	0.35	0.49	0.19	0.15
Noise-to-signal ratio	0.06	0.07	0.11	0.16	0.35	0.21
Hitting rate	0.69	0.67	0.46	0.76	0.81	0.28
Type I error	0.31	0.33	0.54	0.24	0.19	0.72
Type II error	0.04	0.05	0.05	0.12	0.28	0.06

presents the prediction results with forecast horizon being 4, 8 and 12 quarters ahead. The policy preference parameter μ is set to 0.9.[15]

As expected, with a longer forecast horizon, the prediction of both models becomes less useful and noisier. It reflects the difficulty of predicting the breakout of systemic events for longer periods ahead. Nevertheless, the LASSO logit model still strictly outperforms the benchmark model in all cases.

In the above exercise, Model 2 contains variables selected to optimise the one-step ahead forecasting results, yet it maintains good predictive performance in the h-step ahead forecast. One can select variables that optimise h-step ahead forecast performance and then conduct the exercise. However, this approach would omit too much valuable information considering that the variable dynamics just preceding the banking crises would be particularly important. Hence, the former approach is used to conduct the h-step ahead forecast exercise.

(iv) Implications for Monitoring to Preserve Financial Stability

As discussed in Section V.(i), the LASSO with cross-validation approach selects predictors that jointly achieve the best out-of-sample predictive performance. Regardless of the lag order and the interaction terms, variables that are frequently selected should be considered as the key early warning indicators that reveal existing and future financial vulnerability. By closely monitoring these early warning indicators, policy-makers would have a better understanding for adopting appropriate policies and for timing the implementation to mitigate the build-up of systemic risk, subsequently reducing the probability of potential systemic banking crises.

Table 7 presents the selected early warning indicators when the candidate variables contain different orders of lags across varying sample sizes. Columns (1) and (2) present the individual indicators selected by Model 1 and Model 2 using the pre-2005 sample. In various forms and order of lags, these indicators are used to conduct the forecasting performance evaluation in Section V.(ii). Columns (3) and (4) present the indicators selected by the LASSO with cross-validation method using the same set of candidate variables as in Columns (1) and (2) but with the full sample (1970Q1–2008Q3). The common selected key indicators are marked in bold.

Allowing for different lag orders of the candidate variables to be selected across different samples, five indicators are commonly selected by the LASSO with cross-validation approach. They are domestic credit growth, domestic credit gap, global credit gap, real effective exchange rate and real house price growth. It is important to understand how the development of these indicators may affect financial stability.

Credit-related variables, unsurprisingly, are selected as the most important indicators of systemic banking crises. Domestic credit growth and credit gap capture the systemic risk associated with domestic credit market development. Prolonged excessive credit growth is often observed preceding episodes of financial instability, especially the 2007–2008 GFC. In the expansionary phase of the financial cycle, rapid

[15] I also conduct the same exercise with policy preference μ being 0.8 and 0.7. The robustness of the results still holds.

TABLE 7
LASSO+CV Selected Indicators

	Pre-2005 sample		Full sample	
	(1) lag 1	(2) lags 1–4	(3) lag 1	(4) lags 1–4
Variable category				
Credit	credit gap credit growth global credit gap	credit gap credit growth global credit gap global credit growth	credit gap credit growth global credit gap global credit growth	credit gap credit growth global credit gap global credit growth
Housing market	real house price growth	real house price growth house-price-to-rental gap house-price-to-income gap	real house price growth house-price-to-rental gap house-price-to-rental ratio global house price growth	real house price growth house-price-to-rental gap house-price-to-rental ratio global house price growth
External sector	real effective exchange rate	real effective exchange rate current account to GDP	real effective exchange rate current account to GDP	real effective exchange rate current account to GDP
Macroeconomics		global inflation short rate	global inflation equity price growth inflation term spread	global inflation equity price growth

credit expansion is accompanied by credit risk building up in banks and non-bank financial institutions. This reflects rising market confidence and financial institutions becoming optimistic; consequently, they are willing to accept higher risks and increase leverage.

Active credit intermediation would also increase the degree of interconnectedness of the whole financial system. Banks and shadow banks, domestic financial institutions and foreign investors would be much more integrated. The impact of an adverse shock to a single bank is likely to spread through other financial intermediaries and become amplified. Studying the experience of 14 developed countries over 140 years, Jordà *et al*. (2011) conclude that excessive credit growth poses key risks to the financial system. They view credit growth as the best single indicator for financial instability. Monitoring credit growth would provide policy-makers with better insights into when to adopt policies to prevent credit market overheating.

The credit gap, which is defined as the gap between the credit-to-GDP ratio and its long-run trend, is widely acknowledged as one of the most useful early warning indicators for banking crises in a range of studies (Borio and Lowe 2002; Behn *et al.*, 2013; Drehmann & Tsatsaronis, 2014). The credit gap identifies the excessive portion of credit from the level justified by the development of fundamental economic factors. Basel III recommends policy-makers use the credit gap to guide the setting of macroprudential policies, such as adjusting countercyclical capital buffers for banks.

The global credit gap captures the systemic risk associated with imbalances in the development of worldwide credit. The degree of global financial integration has deepened considerably in recent years. Adverse shock to non-domestic financial institutions could spread across many countries, especially those with active cross-border credit flow, and threaten the stability of the domestic banking system.

The importance of monitoring real house price growth is reflected by the prominent role of property market development in previous systemic banking crises, especially the 2007–2008 financial crisis. During an asset price boom, a large amount of bank and non-bank credit flows into the property market and drives up house prices and household leverage. Banks usually perceive mortgages as safe assets as they are backed by tangible collateral. However, the potential credit risk, liquidity risk and concentration risk associated with the concentrated mortgage portfolio of banks should not be underestimated. Excessive house price growth may potentially cause an asset price bubble and lead to a boom-bust cycle in the property market, thereby threatening the resilience of the financial system. Macroprudential policies that target the demand side of housing credit such as putting caps on the loan-to-value ratio and debt-to-income ratio may be considered by policymakers to dampen the cycle.

The real effective exchange rate reflects market sentiment and investors' expectations regarding domestic economic development. It also captures the imbalances in the external sector and the pressure associated with a surge in capital inflows, which typically fuel a house price boom. Sudden adjustments could lead to an unexpected significant loss in the balance sheet of domestic and foreign financial institutions, while the loss is likely to spread across the financial system and exacerbate financial instability.

Frankel and Saravelos (2012) conduct a meta-analysis of 83 papers in the financial crisis literature and find that the real effective exchange rate stands out as one of the most useful leading indicators in explaining crisis incidence across different countries and episodes in the past. While advanced countries usually have much more resilient external sectors than developing countries, one should not ignore the build-up of vulnerabilities in these external sectors.

Some other frequently selected variables are the house-price-to-rental gap, the current account to GDP ratio, global credit growth and global inflation. They have proved to be very important in EWS. Briefly, global credit growth captures the impact of global credit expansion on the domestic banking system. The house-price-to-rental gap is an indicator of the housing market's strength. Excessive increases in the house-price-to-rental gap suggest properties are overvalued and the possibility of a sudden housing market correction. The inclusion of the current account to GDP ratio further highlights the importance of the external sector to financial stability. It is interesting to find that the global level of inflation is an important predictor for banking crises. Ciccarelli and Mojon (2010) study the global co-movement of inflation in detail and find that on average global inflation accounts for 70 per cent of the variability of inflation in industrialised countries from 1960 to 2008. It captures both the trend in the global price level and the fluctuations at global business cycle frequencies. As a proxy for the inflation expectations, it has good predictive ability in forecasting domestic inflation by capturing slow-moving trends in inflation rates (Ciccarelli & Mojon, 2010).

VI Sensitivity Analysis

(i) Longer Lags of Candidate Variables

One may wonder whether including even longer lags of the candidate variables will improve the model's predictive performance. In Model 4, the LASSO with cross-validation approach is used to select predictors from candidate variables with up to eight lags. The five common key variables are once again selected, indicating their importance to the EWS of systemic banking crises. There is no serial correlation in the residuals. The predictive performance is presented in Table 8.

The prediction of Model 4 is very close to Model 2 but not any better. It also strictly outperforms the benchmark model. It suggests that further expanding the lag order of the candidate variables may not be necessary. These results are still relatively insensitive to varying policy preference parameters.

(ii) Models with Fixed Effects

As discussed in Section 'Fixed effects or not?', to avoid the selection bias of dropping countries from the sample, I follow the common practice of early warning literature and estimate pooled logit models without fixed effects. Nevertheless, I conduct the same recursive *out-of-sample* forecasting practice to the fixed-effect model to ensure the superior predictive performance of my approach still holds regardless the treatment of fixed effects. To include the fixed effects, Australia, Canada and New Zealand are omitted from the sample because they never experienced any systemic banking crises. Table 9 summarises the results.

TABLE 8
Longer Lags of Candidate Variables

	Model 4			Benchmark		
	LASSO lags 1–8					
	(1)	(2)	(3)	(4)	(5)	(6)
μ	0.9	0.8	0.7	0.9	0.8	0.7
Relative usefulness	0.62	0.48	0.40	0.48	0.18	−0.09
Noise-to-signal ratio	0.04	0.04	0.01	0.16	0.16	0.14
Hitting rate	0.68	0.58	0.43	0.72	0.72	0.65
Type I error	0.32	0.42	0.57	0.28	0.28	0.35
Type II error	0.03	0.02	0.00	0.11	0.11	0.09

TABLE 9
Out-of-Sample *Performance of Logit Models: With Fixed Effects*

	Benchmark			Model 2			Model 4		
				LASSO lags 1–4			LASSO lags 1–8		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
μ	0.9	0.8	0.7	0.9	0.8	0.7	0.9	0.8	0.7
Relative usefulness	0.29	0.08	−0.19	0.48	0.37	0.26	0.68	0.50	0.42
Noise-to-signal ratio	0.27	0.20	0.18	0.05	0.05	0.05	0.04	0.02	0.01
Hitting rate	0.60	0.60	0.60	0.53	0.47	0.38	0.74	0.56	0.47
Type I error	0.40	0.40	0.40	0.47	0.53	0.62	0.26	0.44	0.53
Type II error	0.16	0.12	0.11	0.03	0.02	0.02	0.03	0.01	0.01

Selecting candidate variables up to four lags and eight lags respectively, both Model 2 and Model 4 achieve higher relative usefulness and lower noise-to-signal ratios than the benchmark model in all cases of policy preference. The LASSO logit models still strictly outperform the benchmark model even if we take the unobserved time-invariant heterogeneity of countries into account. With the inclusion of country fixed effects, Model 4 yields better and more stable predictive results compared to Model 2.

(iii) Multinomial Logit Model

Some studies in the EWS literature propose the use of multinomial logit models to address the 'post-crisis bias' problem (Bussiere & Fratzscher, 2006; Caggiano *et al.*, 2014, 2016). They find evidence suggesting that it could improve the model performance compared to omitting in-crisis periods in binary logit models. Even though the HM and SH test results (shown in Section 'Set-up of the dependent variable') provide conflicted information on the acceptance or rejection of the IIA assumption, I conduct the similar variable selection and out-of-sample forecasting exercises in a multinomial logit framework to test whether the main results still hold. Specifically, the dependent variable $Y_{i,t}$ can be in three states: the tranquil period $Y_{i,t} = 0$, the starting quarter of a crisis $Y_{i,t} = 1$ and the in-crisis period $Y_{i,t} = 2$.

Within a multinomial logit framework, the variable selection, out-of-sample forecasting exercise and evaluation process are conducted in a similar manner as they are done in the main analysis with binary logit framework in Section V. When applying the LASSO method with cross-validation, the best tuning parameter λ is selected in a grouped way so that it does not violate the nature of multinomial logit and select the same set of variables for different outcomes. Model 5 is the multinomial benchmark, which has the same set of variables as the binary benchmark. Models 6 and 7 contain selected predictors from the variable pool with lags up to four and eight quarters respectively. Table 10 summarises the results.

Comparing the relative usefulness of Models 5, 6 and 7 with varying policy preferences, it can be seen that the main results of Section V also hold in the multinomial logit framework. Using the LASSO with cross-validation method, both Models 6 and 7 produce better and more consistent prediction results than Model 5, the multinomial benchmark, under different policy preference scenarios. Such results further provide positive evidence for the advantage of using the proposed variable selection technique to enhance predictive performance of EWS.

It is found that the LASSO with cross-validation method tends to select more variables in the multinomial logit framework than in the binary logit model from the same variable pool. For example, Models 2 and 6 are binary logit and multinomial logit models that have variables selected from the same variable pool. Eight variables are selected for Model 2, while 22 are selected for Model 6. It may suggest an issue when applying the LASSO with cross-validation method to the multinomial logit EWS framework, as some lagged interaction terms could be selected because they are helpful in predicting the duration rather than the occurrence of a crisis. Nevertheless, the key EWS indicators selected, though in different forms, are almost identical in both binary and multinomial logit models. This finding further strengthens the method's capability to identify the most important key indicators for systemic events.

(iv) Alternative Dependent Variable Set-up

As discussed in Section 'Set-up of the dependent variable', many conventional early warning studies set up the dependent variable to include a list of sequential '1's in the 'pre-crisis' periods, which will lead to an autocorrelation problem. To avoid such issue, I use the start of systemic events as the dependent variable while allowing a window of six quarters before the actual crisis starting period to train and evaluate models in the spirit of early-warning. Nevertheless, to ensure the results are robust under different dependent variable set-ups, I evaluate the *out-of-sample* forecasting performance of the benchmark model, Models 1 and 2 with the dependent variable being set up as the 'pre-crisis' periods alternatively. Two exercises with different forecasting horizons are conducted, with the policy preference parameter set to 0.9. Specifically, the first exercise follows Duca and Peltonen (2013) and sets the

Table 10
Out-of-Sample *Performance of Multinomial Logit Models*

	Model 5			Model 6			Model 7		
	multinomial			multinomial			multinomial		
	benchmark			LASSO lags 1–4			LASSO lags 1–8		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
μ	0.9	0.8	0.7	0.9	0.8	0.7	0.9	0.8	0.7
Relative usefulness	0.48	0.05	−0.24	0.53	0.37	0.22	0.51	0.41	0.35
Noise-to-signal ratio	0.17	0.17	0.14	0.09	0.06	0.06	0.04	0.02	0.01
Hitting rate	0.82	0.78	0.72	0.68	0.55	0.52	0.63	0.48	0.42
Type I error	0.18	0.22	0.28	0.32	0.45	0.48	0.38	0.52	0.7
Type II error	0.14	0.13	0.10	0.06	0.03	0.03	0.05	0.01	0.01

Table 11
Alternative Dependent Variable Set-up with Different Forecast Horizon

	Pre-crisis: 1–6 quarters			Pre-crisis: 7–12 quarters		
	Benchmark	Model 1 LASSO lag 1	Model 2 LASSO lags 1–4	Benchmark	Model 1 LASSO lag 1	Model 2 LASSO lags 1–4
Relative usefulness	0.01	0.09	0.14	−0.30	0.05	0.08
NTSR	0.38	0.12	0.13	0.87	0.00	0.06
Hitting rate	0.44	0.13	0.21	0.26	0.05	0.09
Type I	0.56	0.87	0.79	0.74	0.95	0.91
Type II	0.17	0.02	0.03	0.23	0.00	0.01

dependent variable to ‘1’ in the six quarters preceding systemic events and ‘0’ to all the tranquil periods. The second exercise sets the ‘pre-crisis’ periods as 7–12 quarters preceding systemic events, following Behn *et al*. (2013).

As Table 11 presents, the results are generally robust to the alternative dependent variable set-up, even though autocorrelation and model misspecification problem compromises *out-of-sample* forecasting capacities of models. It also shows the difficulty to forecast too far ahead of crises.

(v) AUROC

Some early warning studies in banking crises employ the receiver operating characteristics (ROC) curves and the area under the ROC curve (AUROC) to evaluate model performance (see Caggiano *et al.*, 2014, 2016). Specifically, the ROC curve plots the true positive rate against the false positive rate for all value of thresholds. Hence, the AUROC is an aggregate measure of the signalling quality of EWS regardless of policy-makers’ preference over Type I and II errors. The value of AUROC approaching 1 indicates that the EWS is getting closer to the perfect classification. In contrast, the expected value of the AUROC for a random ranking is 0.5, which indicates the model being completely uninformative.

To complement the analysis with the use of AUROC as an alternative evaluation criteria, the in-sample prediction performance is evaluated and compared between the benchmark model and the LASSO logit models. The specifications of LASSO logit models are selected based on the sample used. The candidate variable pool includes variables with lag orders up to four.[16] Following the suggestion of Candelon *et al*. (2012), I use the AUROC comparison test proposed by DeLong *et al*. (1988) to compare the performance between EWS models. The null hypothesis is the equality of the estimated AUROC between models.

Table 12 reports the estimated AUROC of the benchmark model and the LASSO logit model over the full sample as well as the pre-GFC subsample. In both cases the LASSO logit model achieves a higher AUROC than the benchmark model. The rejection of the null hypothesis indicates that the difference between the estimated AUROC of two models is statistically significant.

VII Conclusion

This paper proposes using the LASSO method with cross-validation to automate the variable selection process of the conventional multivariate logit EWS. Through formalising and automating the variable selection process, the most important information, which may not be easily identified and accessed by subjective variable pre-selection, can be extracted to achieve better systemic banking crisis prediction.

The empirical analysis covers a set of 23 OECD countries with quarterly data. Through a classic pseudo-real-time recursive *out-of-sample* forecasting exercise, I evaluate the predictive

[16] I also test the AUROC for LASSO logit models selecting specification from candidate variables with up to eight lags. The expanded lag order of the variable pool improves the AUROC for LASSO logit models.

Table 12
AUROC and Comparison Test

	AUROC	Std. dev
Full sample: 1970Q1–2018Q3		
Benchmark model	0.91	0.02
LASSO logit model	0.96	0.02
AUROC comparison test	*P*	*P*-value 0.022
Subsample: 1970Q1–2004Q4		
Benchmark model	0.63	0.05
LASSO logit model	0.74	0.05
AUROC comparison test		*P*-value 0.0325

performance of a benchmark multivariate logit model with specification from Behn *et al*. (2013) against the LASSO logit models with automatic variable selection. The evaluation criteria take policy-makers' preferences into account. The results suggest that the proposed method can help produce more useful, more stable and more precise forecasting outcomes of systemic banking crises. Such results are robust to varying policy-makers' preferences, forecast horizon, lag length of candidate variables and the different treatment to the fixed effects.

With this straightforward add-on to the variable selection process, multivariate logit EWS can perform better in forecasting while retaining good interpretability and transparency for effective policy communication and activation. Five variables, namely credit growth, the domestic and global credit gaps, real house price growth and the real effective exchange rate, are identified as the most important early warning indicators of systemic banking crises. This study further highlights the importance of not only considering domestic early warning indicators but also variables that can capture global imbalances and the interactive relationship of these variables.

Acknowledgements

Open access publishing facilitated by Victoria University, as part of the Wiley - Victoria University agreement via the Council of Australian University Librarians.

Conflict of Interest

I declare that is no conflict of interest.

REFERENCES

Aldasoro, I., Borio, C. E., and Drehmann, M. (2018). 'Early Warning Indicators of Banking Crises: Expanding the Family', *BIS Quarterly Review*, March, 29–45.

Alessi, L., Antunes, A., Babecky, J., Baltussen, S., Behn, M., Bonfim, D., Bush, O., Detken, C., Frost, J., Guimaraes, R., Havranek, T., Joy, M., Kauko, K., Mateju, J., Monteiro, N., Neudorfer, B., Peltonen, T. A., Rusnak, M., Marques Rodrigues, P. M., Schudel, W., Sigmund, M., Stremmel, H., Smidkova, K., van Tilburg, R., Vasicek, B., and Zigraiova, D. (2015). 'C'omparing Different Early Warning Systems: Results from a Horse Race Competition among Members of the Macro-Prudential Research Network', MPRA Paper (No. 62194).

Alessi, L. and Detken, C. (2011), 'Quasi Real Time Early Warning Indicators for Costly Asset Price Boom/Bust Cycles: A Role for Global Liquidity', *European Journal of Political Economy*, **27**, 520–33.

Alessi, L. and Detken, C. (2018), 'Identifying Excessive Credit Growth and Leverage', *Journal of Financial Stability*, **35**, 215–25.

Babeckỳ, J., Havránek, T., Matěju, J., Rusnák, M., Šmídková, K. and Vašíček, B. (2014), 'Banking, Debt, and Currency Crises in Developed Countries: Stylized Facts and Early Warning Indicators', *Journal of Financial Stability*, **15**, 1–17.

Behn, M., Detken, C., Peltonen, T. A., and Schudel, W. (2013). 'Setting Countercyclical Capital Buffers Based on Early Warning Models: Would it Work?', ECB Working Paper (No. 1604).

Bergmeir, C., Hyndman, R.J. and Koo, B. (2018), 'A Note on the Validity of Cross-Validation for Evaluating Autoregressive Time Series Prediction', *Computational Statistics and Data Analysis*, **120**, 70–83.

Beutel, J., List, S. and von Schweinitz, G. (2019), 'Does Machine Learning Help us Predict Banking Crises?', *Journal of Financial Stability*, **45**, 100693.

BIS (2010), *Guidance for National Authorities Operating the Countercyclical Capital Buffer*. Basel Committee on Banking Supervision, BIS. https://www.bis.org/publ/bcbs187.pdf

Borio, C. and Lowe, P. (2002), 'Assessing the Risk of Banking Crises', *BIS Quarterly Review*, **7**, 43–54.

Borio, C. E. and Drehmann, M. (2009). 'Assessing the Risk of Banking Crises–Revisited', BIS Quarterly Review, March, 29–46.

Borio, C. E. and Lowe, P. W. (2002). 'Asset Prices, Financial and Monetary Stability: Exploring the Nexus', *BIS Working Paper*.

Breiman, L. (2001), 'Random Forests', *Machine Learning*, **45**, 5–32.

Bussiere, M. and Fratzscher, M. (2006), 'Towards a New Early Warning System of Financial Crises', *Journal of International Money and Finance*, **25**, 953–73.

Caggiano, G., Calice, P. and Leonida, L. (2014), 'Early Warning Systems and Systemic Banking Crises in Low Income Countries: A Multinomial Logit Approach', *Journal of Banking & Finance*, **47**, 258–69.

Caggiano, G., Calice, P., Leonida, L. and Kapetanios, G. (2016), 'Comparing Logit-Based Early Warning Systems: Does the Duration of Systemic Banking Crises Matter?', *Journal of Empirical Finance*, **37**, 104–16.

Candelon, B., Dumitrescu, E.-I. and Hurlin, C. (2012), 'How to Evaluate an Early-Warning System: Toward a Unified Statistical Framework for Assessing Financial Crises Forecasting Methods', *IMF Economic Review*, **60**, 75–113.

Cheng, S. and Long, J.S. (2007), 'Testing for Iia in the Multinomial Logit Model', *Sociological Methods & Research*, **35**, 583–600.

Ciccarelli, M. and Mojon, B. (2010), 'Global Inflation', *The Review of Economics and Statistics*, **92**, 524–35.

Davis, E.P. and Karim, D. (2008), 'Comparing Early Warning Systems for Banking Crises', *Journal of Financial Stability*, **4**, 89–120.

Davis, J.S., Mack, A., Phoa, W. and Vandenabeele, A. (2016), 'Credit Booms, Banking Crises, and the Current Account', *Journal of International Money and Finance*, **60**, 360–77.

Dawood, M., Horsewood, N. and Strobel, F. (2017), 'Predicting Sovereign Debt Crises: An Early Warning System Approach', *Journal of Financial Stability*, **28**, 16–28.

DeLong, E.R., DeLong, D.M. and Clarke-Pearson, D.L. (1988), 'Comparing the Areas under Two or more Correlated Receiver Operating Characteristic Curves: A Nonparametric Approach', *Biometrics*, **44**, 837–45.

Demirgüç-Kunt, A. and Detragiache, E. (1998), 'The Determinants of Banking Crises in Developing and Developed Countries', *IMF Staff Papers*, **45**, 81–109.

Demirgüç-Kunt, A. and Detragiache, E. (2000), 'Monitoring Banking Sector Fragility: A Multivariate Logit Approach', *The World Bank Economic Review*, **14**, 287–307.

Demirgüç-Kunt, A. and Detragiache, E. (2005), 'Cross-Country Empirical Studies of Systemic Bank Distress: A Survey', *National Institute Economic Review*, **192**, 68–83.

Detken, C., Weeken, O., Alessi, L., Bonfim, D., Boucinha, M. M., Castro, C., Frontczak, S., Giordana, G., Giese, J., Jahn, N., et al. (2014). 'Operationalising the Countercyclical Capital Buffer: Indicator Selection, Threshold Identification and Calibration Options', European Systemic Risk Board (ESRB) Occasional Paper Series (No. 5).

Drehmann, M., Borio, C. E., and Tsatsaronis, K. (2011). 'Anchoring Countercyclical Capital Buffers: The Role of Credit Aggregates', BIS Working Paper.

Drehmann, M. and Tsatsaronis, K. (2014). 'The Credit-to-Gdp Gap and Countercyclical Capital Buffers: Questions and Answers', *BIS Quarterly Review*, March.

Duca, M.L. and Peltonen, T.A. (2013), 'Assessing Systemic Risks and Predicting Systemic Events', *Journal of Banking and Finance*, **37**, 2183–95.

Edge, R.M. and Meisenzahl, R.R. (2011), 'The Unreliability of Credit-to-GDP Ratio Gaps in Real-Time: Implications for Countercyclical Capital Buffers', *International Journal of Central Banking*, **7**, 261–98.

Frankel, J. and Saravelos, G. (2012), 'Can Leading Indicators Assess Country Vulnerability? Evidence from the 2008–09 Global Financial Crisis', *Journal of International Economics*, **87**, 216–31.

Fry, T.R. and Harris, M.N. (1996), 'A Monte Carlo Study of Tests for the Independence of Irrelevant Alternatives Property', *Transportation Research Part B: Methodological*, **30**, 19–30.

Fry, T.R. and Harris, M.N. (1998), 'Testing for Independence of Irrelevant Alternatives: Some Empirical Results', *Sociological Methods & Research*, **26**, 401–23.

Fuertes, A.-M. and Kalotychou, E. (2006), 'Early Warning Systems for Sovereign Debt Crises: The Role of Heterogeneity', *Computational Statistics and Data Analysis*, **51**, 1420–41.

Hastie, T., Tibshirani, R. and Friedman, J. (2009), *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* 2nd edition. New York, NY: Springer.

Hausman, J. and McFadden, D. (1984), 'Specification Tests for the Multinomial Logit Model', *Econometrica: Journal of the Econometric Society*, **52**, 1219–40.

Holopainen, M. and Sarlin, P. (2017), 'Toward Robust Early-Warning Models: A Horse Race, Ensembles and Model Uncertainty', *Quantitative Finance*, **17**, 1933–63.

James, G., Witten, D., Hastie, T. and Tibshirani, R. (2013), *An Introduction to Statistical Learning*, vol. **112**. New York, NY: Springer.

Jordà, Ò., Schularick, M. and Taylor, A.M. (2011), 'Financial Crises, Credit Booms, and External Imbalances: 140 Years of Lessons', *IMF Economic Review*, **59**, 340–78.

Kamber, G., Morley, J. and Wong, B. (2018), 'Intuitive and Reliable Estimates of the Output Gap from a Beveridge-Nelson Filter', *Review of Economics and Statistics*, **100**, 550–66.

Kaminsky, G., Lizondo, S. and Reinhart, C.M. (1998), 'Leading Indicators of Currency Crises', *IMF Staff Papers*, **45**, 1–48.

Kaminsky, G.L. and Reinhart, C.M. (1999), 'The Twin Crises: The Causes of Banking and Balance-of-Payments Problems', *American Economic Review*, **89**, 473–500.

Kaminsky, G.L. and Reinhart, C.M. (2000), 'On Crises, Contagion, and Confusion', *Journal of International Economics*, **51**, 145–68.

Laeven, L. and Valencia, F. (2013), ‘Systemic Banking Crises Database’, *IMF Economic Review*, **61**, 225–70.

Laeven, M. L. and Valencia, M. F. (2018). ‘Systemic Banking Crises Revisited’, IMF Working Paper.

Lang, J. H., Peltonen, T. A., and Sarlin, P. (2018). ‘A Framework for Early-Warning Modeling with an Application to Banks’, ECB Working Paper

Long, J.S. and Freese, J. (2014), *Regression Models for Categorical Dependent Variables Using Stata*. College Station, TX: Stata Press.

Manasse, P. and Roubini, N. (2009), ‘"Rules of Thumb" for Sovereign Debt Crises’, *Journal of International Economics*, **78**, 192–205.

Neunhoeffer, M. and Sternberg, S. (2019), ‘How Cross-Validation Can Go Wrong and What to Do about it’, *Political Analysis*, **27**, 101–6.

Orphanides, A. and Norden, S.v. (2002), ‘The Unreliability of Output-Gap Estimates in Real Time’, *Review of Economics and Statistics*, **84**, 569–83.

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

Sarlin, P. (2013), ‘On Policymakers’ Loss Functions and the Evaluation of Early Warning Systems’, *Economics Letters*, **119**, 1–7.

Small, K.A. and Hsiao, C. (1985), ‘Multinomial Logit Specification Tests’, *International Economic Review*, **26**, 619–27.

Tanaka, K., Kinkyo, T. and Hamori, S. (2016), ‘Random Forests-Based Early Warning System for Bank Failures’, *Economics Letters*, **148**, 118–21.

Tibshirani, R. (1996), ‘Regression Shrinkage and Selection Via the LASSO’, *Journal of the Royal Statistical Society: Series B (Methodological)*, **58**, 267–88.

Van den Berg, J., Candelon, B. and Urbain, J.-P. (2008), ‘A Cautious Note on the Use of Panel Models to Predict Financial Crises’, *Economics Letters*, **101**, 80–3.

Appendix I

Table A1
Description of Explanatory Variables

Variable	Source	Description
Credit growth	BIS	The year-to-year growth rate of credit to private non-financial sector from all sectors at market value – US dollar – Adjusted for breaks
Credit-to-GDP gap	BIS	Calculated by applying the modified BN filter to the credit-to-GDP ratio, which is: The credit to private non-financial sector from all sectors at market value – Percentage of GDP – Adjusted for breaks
Real GDP growth	OECD	Year-to-year growth
Inflation	OECD	Calculated from OECD CPI index
Short rate	OECD	Three-month Interbank rate
Long rate	OECD	10-year government bond rate
Term spread	OECD	Calculated as the gap between the long rate and the short rate
Current account in percentage of GDP	OECD	
Real effective exchange rate	OECD	Index 2015 = 100
Equity price growth	OECD	Year-to-year growth, calculated from OECD share price index, 2010 = 100
Real house price growth	OECD	Year-to-year growth, calculated from OECD real house price index
House-price-to-income ratio	OECD	
House-price-to-rental ratio	OECD	
House-price-to-income gap	OECD	Calculated by applying the modified BN filter to the house-price-to-income ratio
House-price-to-rental gap	OECD	Calculated by applying the modified BN filter to the house-price-to-rental ratio