Expert Systems With Applications 112 (2018) 353–371

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

Forecasting stock market crisis events using deep and statistical machine learning techniques

Sotirios P. Chatzis[a,*], Vassilis Siakoulis[b], Anastasios Petropoulos[b], Evangelos Stavroulakis[b], Nikos Vlachogiannakis[b]

[a] Department of Electrical Engineering, Computer Engineering and Informatics, Cyprus University of Technology, Limassol 3036, Cyprus
[b] Bank of Greece, Banking Supervision Division, 3 Amerikis Str., Athens 10250, Greece

ARTICLE INFO

Article history:
Received 10 April 2018
Revised 13 June 2018
Accepted 14 June 2018
Available online 28 June 2018

Keywords:
Stock market crashes
Forecasting
Random forests
Support vector machines
Deep learning
XGBoost

ABSTRACT

This work contributes to this ongoing debate on the nature and the characteristics of propagation channels of crash events in international stock markets. Specifically, we investigate transmission mechanisms across stock markets along with effects from bond and currency markets. Our approach comprises a solid forecasting mechanism of the probability of a stock market crash event in various time frames. The developed approach combines different machine learning algorithms which are presented with daily stock, bond and currency data from 39 countries that cover a large spectrum of economies. Specifically, we leverage the merits of a series of techniques including Classification Trees, Support Vector Machines, Random Forests, Neural Networks, Extreme Gradient Boosting, and Deep Neural Networks. To the best of our knowledge, this is the first time that Deep Learning and Boosting approaches are considered in the literature as a means of predicting stock market crisis episodes. The independent variables included in our data contain information regarding both the two fundamental linkage channels through which financial contagion can be initiated: returns and volatility. We apply a suite of machine learning algorithms for selecting the most relevant variables out of a large set of proposed ones. Finally, we employ bootstrap sampling for adjusting the imbalanced nature of the available fitting dataset. Our experimental results provide strong evidence that stock market crises tend to exhibit persistence. We also find significant evidence of interdependence and cross-contagion effects among stock, bond and currency markets. Finally, we show that the use of Deep Neural Networks significantly increases the classification accuracy, while offering a robust way to create a global systemic early warning tool that is more efficient and risk-sensitive than the currently established ones. Thus, central banks may use these tools to early adjust their monetary policy, so as to ensure financial stability.

© 2018 Elsevier Ltd. All rights reserved.

1. Introduction

A global financial crisis can emerge from a series of local or/and regional market shocks, which evolve into a worldwide economic crisis due to the interconnectedness of the financial markets. For example, the Asian crisis in 1997 initially originated in Thailand; subsequently, it propagated to other Asian countries, and eventually made it to the financial markets of the United States of America and Europe (Kaminsky, Lizondo, & Reinhart, 1998). In other cases, a crisis may start from a single economy whose size is large enough to generate turbulence in other countries. This is the case, for instance, with the subprime crisis that started in the United States and evolved into a sovereign debt crisis in several European countries.

The observation that an economic crisis is manifested by a subsequent recession (Barro & Ursua, 2009; Bluedorn, Decressin, & Terrones, 2013; Estrella & Mishkin, 1996; Farmer, 2012) renders reliable Early Warning Systems (EWSs) valuable tools for policymakers, in their effort to curtail contagion risk and, in extreme cases, even preempt a global economic crisis. An EWS must be capable of producing clear signals as to whether an economic crisis is imminent, complementing the expert judgment of policymakers. Hence, EWS systems facilitate policy makers in unveiling vulnerabilities of the economy and taking precautionary actions to diminish the risks that can trigger a crisis. Certainly, there is always a trade-off between developing EWSs that are capable of predicting a lot

* Corresponding author.
E-mail addresses: sotirios.chatzis@cut.ac.cy, soteri0s@me.com (S.P. Chatzis), VSiakoulis@bankofgreece.gr (V. Siakoulis), apetropoulos@bankofgreece.gr (A. Petropoulos), EStavroulakis@bankofgreece.gr (E. Stavroulakis), nvlachogiannakis@bankofgreece.gr (N. Vlachogiannakis).

https://doi.org/10.1016/j.eswa.2018.06.032
0957-4174/© 2018 Elsevier Ltd. All rights reserved.

of alarms for an imminent crisis, at the expense of some of them being wrong (false-alarms), and EWSs that predict rather too few signals of impending crises, at the expense of missing a major crisis event. Optimally, an EWS should let no crisis events go unnoticed, while minimizing the number of generated false-alarms. It goes without saying that the cost of not signaling a global crisis is significantly higher than that of an incorrect alarm.

At the same time, the incorporation of the probability of a worldwide crisis in decisions related to asset al.location (Kole, Koedijk, & Verbeek, 2006) can substantially benefit investors. Indeed, this is the case since a global crisis significantly curtails diversification benefits, as worldwide markets move in the same direction. In addition, any hedging strategies may become ineffective (Ibragimov & Johan, 2007) due to the structural changes in the observed correlations among asset classes. Indeed, during periods of high volatility in bear markets, correlations increase across assets (Longin & Solnik, 2001). Thus, as the markets cannot quickly correct any disruptions in their function, it becomes even more imperative for regulators to intervene so as to restore financial stability.

The extent to which a series of local crises can lead to a global economic crisis has not been extensively studied in the current literature. One representative study is performed by Markwat, Kole, and Van Dijk (2009), who examine whether a domino effect at the local level may evolve to regional and then to global stock market crashes. Our work contributes to the current literature by investigating the predictive performance of a group of state-of-the-art statistical machine learning techniques, including Deep Learning and Boosting algorithms, in the classification problem of future market crash detection on the global level, extending the investigated period up to 2017. In this context, our work offers a multilevel, multi-component modeling framework for global crisis events forecasting, driven by machine learning algorithms combined with application of appropriate ensemble generalization techniques. This is a novel solution to the problem of forecasting global stock market crashes, that has never been reported in the past. We perform an exhaustive experimental evaluation of our system, using a test dataset of 7-year stock market daily returns performance.

As already mentioned, a key novelty of our approach consists in the utilization of Deep learning techniques (LeCun, Bengio, & Hinton, 2015). Deep learning is a recent trend in machine learning that offers higher flexibility in learning nonlinear dynamics in large datasets. These methods have gained a significant momentum due to their state-of-the-art performance in various scientific fields, including computer vision, natural language processing, etc. Deep Learning techniques essentially comprise a new generation of artificial neural networks that employ statistical modeling arguments to overcome problems that plague traditional ones, such as the vanishing gradients problem, and their overfitting tendencies.

Finally, a main goal of this paper is to explain the causes of negative co-exceedances in stock markets. We effect this analysis by introducing financial economic covariates to account for interdependence in normal times and external shock variables; these allow to study contagion effects in crisis periods. On this basis, we employ a series of machine learning techniques to forecast tail events in the global stock markets. This is in contrast to existing early warning systems, which usually employ error-prone, biased, and oversimplistic heuristically defined macro-indicators. On the contrary, our suggested approach is considerably more robust in modeling nonlinear behavior in financial data, it can better model the interaction effects between leading financial turbulence indicators, and can capture long and subtle temporal patterns that are elusive to human analysts.

In a nutshell, the proposed innovation lies in five areas:

- We avoid the ad-hoc determination of crisis episodes by considering a volatility-linked mechanistic way of crisis definition. This way, a crisis event is connected to high volatility levels that may cause extreme loss to a long or short position.
- We use advanced machine learning techniques, including Deep Leaning and Extreme Gradient Boosting, while comparing with traditional methods of econometrics.
- We use daily financial market data, which tend to be more responsive than macro-data, since they are reported in higher frequencies.
- We combine multiple machine learning techniques by leveraging ensemble methods, as a way to minimize model risk and increase accuracy.
- We provide a thorough out-of-sample evaluation of the proposed novel approach. Specifically, we collect data over a long time-period which spans more than 20 years. The long period considered under our experimental setup allows exploring system performance under volatility-stress periods, and smooth trending periods alike. In addition, it also allows for examining whether financial crises exhibit persistence and clustering.

The remainder of this paper is organized as follows. In Section 2, we provide a brief overview of related empirical work, and explore the most commonly used modeling techniques for crisis forecasting and EWSs. In Section 3, we elaborate on the dataset used for developing and evaluating our system. In Section 4, we provide a thorough description and technical analysis of all the employed machine algorithms. In Section 5, we elaborate on our experimental setup, and present our empirical results. We assess our empirical findings by analyzing the obtained crisis prediction accuracy (correct classification rates), computed on a test sample that spans a long time-frame. Finally, in Section 6 we draw our conclusions, while also indicating directions for future research.

2. Literature review

Crisis and distress prediction is a very useful tool, as it may be effectively utilized in banking, finance, business, and other areas. This is the reason why financial contagion and crisis forecasting are very popular topics in academic research during the last years, especially after the significant impact of the 2008 global economic crisis. Forecasting of financial variables with the use of advanced modeling techniques has been one of the most extensively studied topics in the academic literature. In particular, Bagheria, Mohammadi, and Akbaric (2014) propose a hybrid neural based system to predict FX time series, which proves to be helpful and efficient in price forecasting. Adopting a different perspective, Cervelló-Royo, Guijarroa, and Michniukab (2015) identify trading patterns that provide evidence that the Efficient Market Hypothesis is not confirmed. Furthermore, Chiang, Enke, Wu, and Wang (2016) suggest that traders can generate higher returns when using their proposed decisions adaptive decision support system. Further, the results of Enke and Thawornwong (2005) imply that classification algorithms based on neural networks produce higher risk-adjusted profits, due to their capacity to adequately capture nonlinearities. Specifically, Ghazali, Hussain, and Liatsis (2011) provide evidence for the superior predictive performance of Dynamic Ridge Polynomial Neural Network (DRPNN) relative to other neural network specifications. In another strand of literature, Zhong and Enke (2017) attest that a combination of methods, that is, ANN and PCA, can yield improved outcomes. Similarly, the Kim and Chang (2018) propose a series of combined long short-term memory (LSTM) and GARCH models to forecast stock price volatility, and verify that these models lead to the lowest prediction errors. In short, it appears that the combination of advanced deep learn-

ing and statistical machine learning techniques is the optimal approach in terms of predictive performance in financial forecasting.

A large category of financial contagion studies focuses on predicting the reaction of markets to unexpected news by making use of factor models (Dungey & Martin, 2006). The model is a system of equations which uses a vector of country indexes as endogenous variable, and a world factor and a vector of country factors as exogenous variables. The effect of contagion in crisis periods is modeled by testing whether the parameter of the third country factor effect is significant. Other methodologies incorporate correlation coefficients as a measure of contagion. This way, after a crisis event, the already established macroeconomic links between 2 countries can either strengthen (correlation increase), break (correlation decrease) or remain the same (Forbes & Rigobon, 2002). The first attempt of using multinomial discrete choice methodologies for modeling financial contagion in the tails can be found in Bae, Karolyi, and Stulz (2003). Similar studies can be found in Christiansen and Ranaldo (2009); they used a similar econometric framework to analyze the financial integration of the new European Union (EU) member states' stock markets in comparison to old EU member countries. Finally, Markwat et al. (2009) used ordered logit regression in order to classify stock market crashes as local, regional or global, and examine their domino effects.

Turning to crisis forecasting, classical approaches develop macro-indicators that can be leveraged by an early warning system. šíč, Zingraoiva, Hoeberichts, Smidkova, Vermuelen, and De Haan (2017) applied Bayesian Model Averaging on data stemming from 25 OECD countries to identify indicators for predicting financial crisis. A significant result of this study is that linear models cannot capture financial dynamics and predict financial stress in an out-of-sample setup. Bussiére and Matthieu (2013) showed that the performance of early warning systems decreases substantially when seeking to predict the exact date of the event. Other studies employing classical econometric techniques in forecasting banking and economic crisis, e.g. multinomial logit, include: Babecký et al. (2014), who applied Bayesian model averaging for identifying early warning indicators of crises; and Faranda, Flavio, Giachino, Vaienti, and Dubrulle (2015), who analyzed US and Europe stock indices by means of autoregressive moving average models to identify early warning indicators of financial crises.

The use of Machine Learning Techniques in the development of early warning systems for financial crisis is rather limited in the existing literature. Cuneyt, Oztekin, Ozkan, Serkan, and Erkam (2014) developed three different early warning systems, based on artificial neural networks (ANN), decision trees, and logistic regression, and tested them on the Turkey economy; artificial neural networks yielded the best performance in their analyses. Atsalakis, Protopapadakis, and Valavanis (2016) focused on 1-day stock market forecasting during stress periods, and employed a neuro-fuzzy modeling methodology. Oztekin, Kizilaslan, Freund, and Iseri (2016) also focused on prediction of daily stock price. Their work deployed and integrated adaptive neuro-fuzzy inference systems, artificial neural networks, and support vector machines. Döpke, Fritsche, and Pierdzioch (2017) implemented boosted regression Trees for predicting recessions. Finally, Dabrowski, Beyers, and De Villiers (2016) investigated dynamic Bayesian network models and showed that they can provide significantly more precise early-warnings compared to logistic regression.

In view of this summary of the existing literature, it is easy to observe that the approach outlined in this paper offers significant novelty and advantages. Specifically, we leverage the merits of a very diverse variety of modeling approaches, while also including a wide range of explanatory variables. Specifically, the following models are employed in our analysis: Logistic Regression (LogR), Random Forests (RF), Support Vector Machines (SVMs), Neural Networks (NNs), CART, Extreme Gradient Boosting (XGBoost), and Deep Learning Techniques (MXNET). To the best of our knowledge, XGBoost and Deep Learning have never been explored in the past in the context of financial crisis forecasting. Furthermore, we consider more than 2700 explanatory variables on the basis of their potential predictive value in the context of the developed system. Such an in-depth analysis of the available types of explanatory variables that can be used for the purpose of crash prediction has never been performed in the past. Finally, note that it is the first time that a study on crisis prediction employs a dataset reaching up to Q4 2017. Thus, our work offers a timely set of empirical outcomes, missing from the existing literature.

3. Data collection and processing

3.1. Data collection

In order to perform modeling of stock market crisis events, we have collected information on various financial indicators from a number of different sources and databases. We focus on liquid markets, where the transmission of extreme events is better depicted in the pricing patterns. At the same time, we select a sample of 39 countries around the globe, based on two criteria:

- Provide adequate data sample for the examined period (22 years).
- Provide an adequate number of countries by region.

Under this framework, we exclude countries with relative short time-series, as well as some countries in regions which were overrepresented in our sample. The same logic regarding the length of the coveredtime period is applied in the additional financial indices employed in our analysis.

Specifically, the used data include the Stock Price Index, the yield of the 10-year government bond of 39 countries across the globe, the exchange rates of 18 currencies against the United States dollar, as well as additional financial indices such as Oil, Gold, and Vix (we elaborate on these in Section 3.2). These selected financial indices include the most sensitive market variables, covering the three most important financial markets, namely America, Asia and Europe. Therefore, we are able to identify a number of different crises in terms of nature and severity. The data was sourced from the FRED database and the SNL (S&P Global Market Intelligence) website, and covers the period 10/01/1996 – 15/12/2017, measured on a daily basis. That is, a 22-year period which includes a number of economic crises of different nature and severity. This way, we eventually obtain a dataset with more than 5000 records, which offers the capability of modeling contagion dynamics and financial markets interdependencies across countries and across time. In addition, we are able to classify local and global crises, based on the impact of the identified events.

Fig. 1 illustrates the number of countries exhibiting an extreme stock market fall during the selected 22 years period. These events have been identified based on the number of stock market negative co-exceedances (less than 1% percentile of the empirical distribution) across the 39 countries in the sample. We observe at least 10 events with global impact, with the most severe being the global economic crisis of 2008. During the more recent observations from 2011 to 2017, it seems that the global market is more stable, as we observe less variability, even though we still observe some events.

The main financial crises recorded over the last 20 years (covered by the data sample used in this study) which had severe contagion effects across the global stock markets comprise:

- The devaluation of the Mexican Peso (20/12/1994), which triggered the Latin America crisis known as the "Tequila Crisis."
- The devaluation of the Thai Baht (2/7/1997), which triggered the crisis in Asia known as the "Asian Flu."

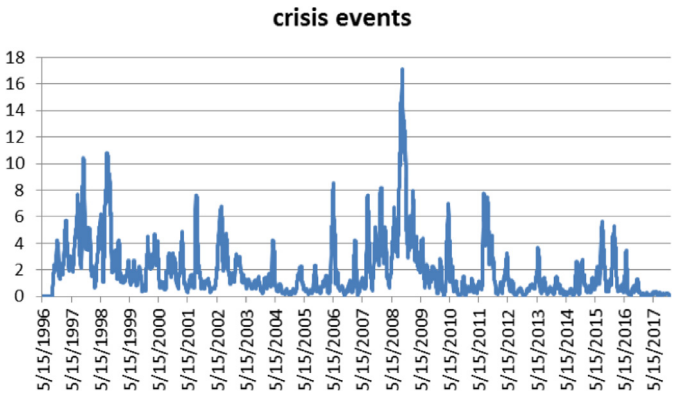

Fig. 1. Number of countries exhibiting an extreme stock market fall (exceedance less than 1% percentile of the empirical distribution).

- The Russian default (17/8/1998), also known as the "Russian Virus," which caused severe liquidity problems to global markets (Baig & Goldfajn, 2000).
- The beginning of Long Term Capital Management (LTCM) recapitalization (23/9/1998).
- The Hong-Kong stock market crash (28/10/1998).
- The currency devaluation in Brazil (13/1/1999).
- The collapse of Argentine currency board (20/12/2001).
- The US and EU dot.com collapse (April 2000).
- The Brazilian elections (October 2002).
- The Brazilian run-up to presidential elections (2003).
- The global financial crisis, which started in stock markets in 2007, and led to a recession in the real sector of the economy (2007).
- Greek Sovereign crisis (2010).
- European Sovereign crisis (2011, 2012).

Besides the above-mentioned events, there were also crises which led to more modest increases in the volatility of international stock markets:

- US fiscal cliff (2013).
- FED taper tantrum (2014).
- US Elections (2016).
- Brexit Vote (2016).

In view of these facts, we split our dataset into two large parts. An in-sample dataset (full in-sample), comprising the data spanning until the end of 2010, and an out-of-sample dataset that spans over the years 2011–2017. The former is used for model fitting, while the latter for performance evaluation. Data splitting is a common approach in model development; it allows to assess performance in a different sample which ensures robust predictions. In our case, the selection of the out-of-sample period enables us to assess whether the developed model can predict events that had both severe (i.e. sovereign crisis) and modest (e.g. 2016 US elections, Brexit) contagion effects in the international stock markets.

In all cases, we consider *two predicted/dependent variables*. These constitute binary indicators that take the value of *one in case there is a global stock market crash event in the next day*, or in the *next 20 days*, respectively, and take the value of zero otherwise. To perform model selection, we employed five-fold cross-validation, using predictive accuracy as our model selection criterion, as we shall describe in a subsequent section. Performance evaluation results were assessed on the available test sample, to allow for evaluating the generalization capacity of the developed models.

3.2. Data processing

As already discussed, we created two *binary* dependent variables, one pertaining to an one-day predictive horizon, and one pertaining to a 20-day predictive horizon. We calculated these dependent variables on the basis of the following hierarchical process, which we graphically illustrate in Fig. 2.

Initially, we identified a "crisis event" for each country, at each working day, if the return of the Stock Index was below the first percentile of the associated *empirical distribution of returns*. By employing a percentile-based mechanistic way of crisis detection, we avoid the ad-hoc selection of crisis episodes whereas we connect them mostly to high volatility levels. This defines a negative volatility exceedance event which could cause extreme loss to a long or short position. In addition, multiple exceedance events (co-exceedances) are linked to extreme losses in more than one market.

We narrow the traditional negative exceedance definition of 5% quantile to 1% so as to focus on the extremities of the return distribution where the successful detection of an extreme event becomes a challenging task. From supervisory point of view, the 5% quantile captures crash events which can be adhered to in the daily crisis management process, whereas the 1% quantile recognizes a limited number of extreme crashes (on average, daily fall in the stock exchange more than 4%). The latter could have important repercussions; hence, their prediction may function as a solid ground for the imposition of proactive measures.

We calculated the initial empirical distribution of returns (pertaining to the first day in our analysis) based on the stock index returns of the first 200 observations (covering the period 10/01/1996–15/10/1996). For each subsequent record (day in the examined period), we recalculated the empirical distribution of returns in order to incorporate the new observation. Then, an event was identified if the return was below the first percentile of the new empirical distribution. Thus, for the last observation in the sample (i.e. 15/12/2017), the empirical distribution of returns was based on the period: 10/01/1996–14/12/2017.

Having used the daily movement of the corresponding stock indices to identify the " crisis events" occurring in each of the considered 39 countries, we proceeded to event aggregation on re-

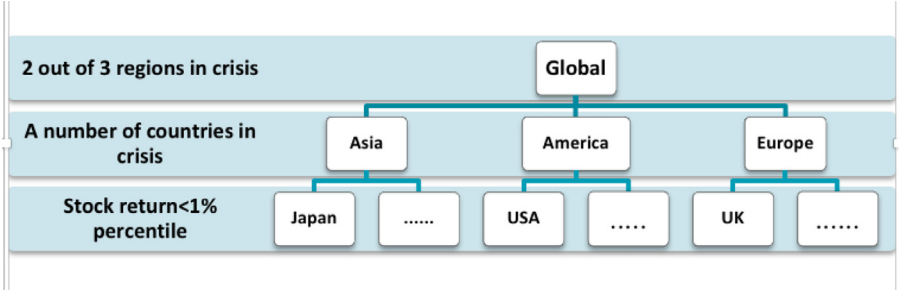

Fig. 2. Outline of the dependent variable construction process.

Table 1
Obtained categories of raw measurements.

Stock Variables	Curency and Bond Markets	Other Variables
•Sample countries stock market returns and volatilities. •Binary variables of extreme movement events in country level. •Discrete variables counting number of extreme events on a day globaly and by region (e.g. Asia). •Binary variables representing regional or past global events on different horizons (5, 20, 40 and 60 days).	•10 year sovereign bond yield changes. •Changes in basic exchange rates. •Volatilities of sovereign bond yields and basic currencies. •Binary variables of extreme movement events in country level for sovereign bonds and currencies. •Discrete variables counting number of extreme movements (>99% percentile) on a day globaly.	•Libor rate •VIX index •Gold Price •TED spread •Oil price •Effective Federal Funds Rate •High yield bond returns.

gional level, i.e. America, Asia, Europe, as well as Globally. To this end, we calculated the aggregate number of events per day in the America, Asia, and Europe regions, as well as on the Global scale. On this basis, we derived a set of intermediate binary predictor variables, identifying whether the number of events per day exceeded a threshold, at each separate region or globally. Our definition of regional and global "crisis events" essentially reflects the negative co-exceedance, i.e. the simultaneous abrupt falls in the stock markets on a given day across countries (for the regional variables) or regions (for the global variable). Therefore, in the remainder of this paper the terms "crisis event" and "negative co-exceedances" will be used interchangeably since they refer to the same think. As such, the postulated binary variables essentially try to capture the "significant events" within a region, i.e. events that have had a collective impact on many stock markets in the region or globally.

The selected thresholds were determined in relation to size (i.e. number of stock markets available in each region) in a way that about half of the countries in each region exhibit negative co-exceedance. Specifically, we postulated the following thresholds:

- America: At least 3 events per day (out of a total of 7 countries).
- Asia: At least 6 events per day (out of a total of 13 countries).
- Europe: At least 8 events per day (out of a total of 19 countries).
- Global: At least 2 regions are in negative co-exceedance mode on a daily basis.

Based on the above outcomes, we created two classes of target variables to forecast. The first measures whether there is a significant global crisis event in the next working day, and the second measures whether there is a significant global crisis event during the next 20 working days (approximately 1 calendar month).

On the other hand, the used categories of raw measurements are summarized in Table 1. To create the independent variables presented to the developed models, we examined an extended set of exploratory variables that can be derived from these raw measurements. These are outlined in Table 2. Finally, besides these country-specific variables, we also explored whether the identification of extreme movements in a specific country or region has any statistical correlation with observed crises in the past. To this end, we created additional exploratory variables that constitute lagged values of the extracted predictive indicators (binary variables). These variables can be considered systemic shock indicators, inclusion of which may potentially facilitate capturing contagion effects among different regions, as well as of crisis clustering effects (i.e., a crisis in one day causes a second crisis the next day).

Eventually, on the basis of the outlined exploratory variables, we proceeded to construct the independent variables presented

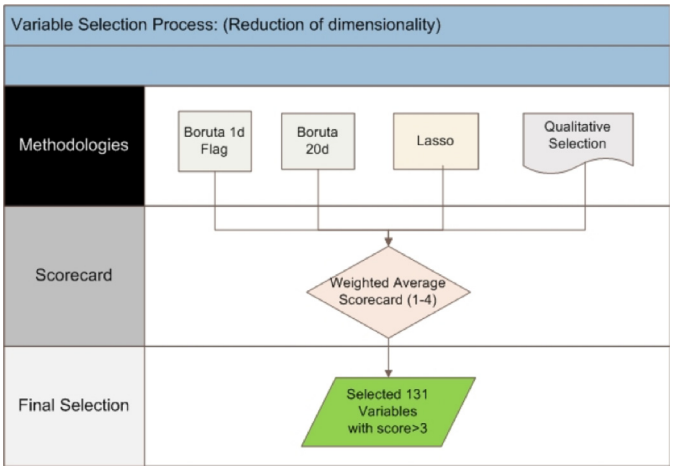

Fig. 3. Variable scoring and selection process.

to the develop machine learning models. Specifically, in an effort to capture subtler dynamics and dependencies, we consecutively computed the following transforms (next, we present in brackets the naming convention pertaining to each type of transformation):

(i) We derived the continuously compounded rate of daily returns, by calculating the log returns [ln]. This transformation was performed for all variables under the categories of Stock Indices, Bond Yields and Currency Exchange Rates.

(ii) We calculated the daily volatility of return [ln2], by squaring the log-returns.

(iii) We calculated lagged variables on a daily basis for each crisis indicator (regional/global), starting from 1 up to 5 days, 20 days, 40 days and 60 days [lag1, lag2,..., lag5, lag20, lag40, lag60].

(iv) We computed the average number of significant events during the last 5 working days and the last 20 days [L5D, L20D], based on previous values of the *binary predictor variable* (regional/global) which identifies whether the*number of events exceeded a threshold* at some day (as described previously).

(v) We computed the average number of events during the last 5 working days and the last 20 days [L5D, L20D], based on the *total number of events* on a daily basis.

This independent variable generation process led to a set of almost 2700 predictors that may be used for presenting the developed machine learning models with.

3.3. Variable selection

The constructed dataset comprises an excessive number of independent variables, which is clearly disproportional to the size of the dataset, i.e. the number of days available in the sample; specifically, we are dealing with around 2700 variables over around 5400 days (data points). Fitting a machine learning model to such a huge number of independent variables (relative to the size of the dataset) is doomed to suffer from the so-called curse of dimensionality problem. That is, the fitted classifier may seem to yield very good performance in the training dataset, but it turns out to generalize very poorly, yielding a catastrophically low performance outcome in the test data. Thus, to ensure a good performance outcome, we need to implement a robust independent variable (feature) selection stage, so as to limit the number of used features to the absolutely necessary. Besides, apart from increasing the generalization capabilities of the fitted models, such a reduction is also important for increasing the computational efficiency of the explored machine learning algorithms.

Fig. 3 provides an overview of the adopted feature selection procedure. It comprises three phases: In the first phase, we employ

Table 2
Examined exploratory variables.

Stock markets	Bond markets	Exchange rates	Additional variables
America	USA 10-year bond Yield	CNYUSD: Chinese Yuan to U.S. Dollar	Libor rate
USA Stock Index	Japan 10-year bond Yield	DKKUSD: Danish Kroner to U.S. Dollar	VIX index
Canada Stock Index	Germany 10-year bond Yield	HKDUSD: Hong Kong Dollars to U.S. Dollar	Gold price
Argentina Stock Index	UK 10-year bond Yield	INRUSD: Indian Rupees to U.S. Dollar	TED spread
Peru Stock Index	France 10-year bond Yield	JPYUSD: Japan Yen to U.S. Dollar	Oil price
Brazil Stock Index	India 10-year bond Yield	KRWUSD: South Korean Won to U.S. Dollar	Effective Federal Funds Rate
Chile Stock Index	South Korea 10-year bond Yield	MYRUSD: Malaysian Ringgit to U.S. Dollar	High yield bond returns
Mexico Stock Index	Russia 10-year bond Yield	MXNUSD: Mexican Peso to U.S. Dollar	
Europe	Spain 10-year bond Yield	NOKUSD: Norwegian Kroner to U.S. Dollar	
Germany Stock Index	Mexico 10-year bond Yield	SEKUSD: Swedish Kronor to U.S. Dollar	
UK Stock Index	Indonesia 10-year bond Yield	CHFUSD: Swiss Francs to U.S. Dollar	
France Stock Index	Netherlands 10-year bond Yield	TWDUSD: New Taiwan Dollars to U.S. Dollar	
Russia Stock Index	Switzerland 10-year bond Yield	THBUSD: Thai Baht to U.S. Dollar	
Spain Stock Index	Taiwan 10-year bond Yield	EURUSD: U.S. Dollars to Euro	
Netherlands Stock Index	Sweden 10-year bond Yield	GBPUSD: U.S. Dollars to British Pound	
Switzerland Stock Index	Poland 10-year bond Yield	CADUSD: Canadian dollar to U.S. Dollar	
Sweden Stock Index	Belgium 10-year bond Yield	AUDUSD: U.S. Dollars to Australian Dollar	
Poland Stock Index	Thailand 10-year bond Yield	NZDUSD: U.S. Dollars to New Zealand Dollar	
Belgium Stock Index	Austria 10-year bond Yield		
Austria Stock Index	Norway 10-year bond Yield		
Norway Stock Index	Honk Kong 10-year bond Yield		
Israel Stock Index	Israel 10-year bond Yield		
Denmark Stock Index	Denmark 10-year bond Yield		
Ireland Stock Index	Philippines 10-year bond Yield		
Greece Stock Index	Malaysia 10-year bond Yield		
Czech Republic Stock Index	Ireland 10-year bond Yield		
Hungary Stock Index	Greece10-year bond Yield		
Slovakia Stock Index	Czech Republic 10-year bond Yield		
Asia	Hungary 10-year bond Yield		
China Stock Index			
Japan Stock Index			
India Stock Index			
South Korea Stock Index			
Indonesia Stock Index			
Taiwan Stock Index			
Thailand Stock Index			
Honk Kong Stock Index			
Philippines Stock Index			
Malaysia Stock Index			
Pakistan Stock Index			
Australia Stock Index			
New Zealand Stock Index			

three popular methodologies that independently assign importance to the available features: Boruta (Kursa & Rudnicki, 2010), LASSO (Tibshirani, 1996), and a qualitative criteria-driven filter method. In the second phase, a balanced score is produced for each variable. In the third phase, we impose a heuristically determined cut-off score, and discard all features that do not reach this score. This way, a total of 131 explanatory variables are eventually selected to be retained.

The Boruta algorithm is based on a Random Forest model. Based on the inferences of this Random Forest, features are removed from the training set, and model training is performed anew. Boruta infers the importance of each independent variable (feature) in the obtained predictive outcomes by creating shadow features. This helps in the identification of all attributes which are in some circumstances relevant for classification, the so-called all-relevant problem. Finding all relevant features, instead of only the non-redundant ones, may be very useful when one is interested in understanding mechanisms related to the subject of interest, instead of merely building a black-box predictive model.

Specifically, the algorithm performs the following steps: First, it adds randomness to the given dataset by creating shuffled copies of all features (shadow features). Then, it fits a Random Forest on the extended dataset and evaluates the importance of each feature. At every iteration, it checks whether a real feature has a higher importance than the best of its shadow features, and constantly removes features which are deemed unimportant. The algorithm stops when all features are classified as important or are rejected as noise. In our study, we employ the Boruta Package, provided by the R programming language, to implement variable selection using both the 1-day and 20-day flag as dependent variable. This way, all features relevant to both dependent variables are selected based on error minimization for the fitted Random Forest models.

On the other hand, LASSO is a regression model that penalizes the number of model parameters in its objective function as a means of excluding irrelevant variables from the model. One of the most important features of LASSO is its ability to cope with high numbers of independent variables (features) relative to the available training observations, which is pertinent in the context of our study. We performed LASSO analysis using the GLMNET package in R. This offers a very fast way to select the best model configuration, using both cross-validation and the Bayesian Information Criterion (BIC).

Finally, the employed qualitative criteria-driven method consists in evaluating the individual correlation of each feature against the constructed dependent variables (occurrence of stock crash events).

We combine the so-obtained sets of rankings by applying weighted average scoring. Specifically, each of the three selection methods assigns each independent variable a score of 1 or 0, depending on whether it is selected or not. Then, we multiply each score by a weight, different for each method. We assign Boruta scores a slightly higher weight, due to its extensive analysis of the features in the dataset (variable importance, as determined in

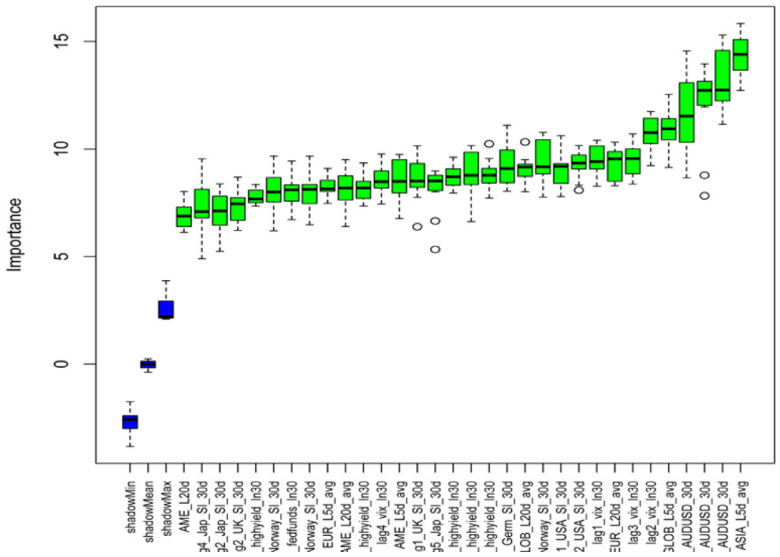

Fig. 4. Variable importance as determined in Boruta algorithm output.

Boruta algorithm output, is depicted in Fig. 4). Eventually, the final score obtained for each explanatory variable ranges in the interval from 0 to 4. To perform the final selection, we apply a cut-off score of 3, yielding a narrowed-down group of 131 candidate independent variables.

3.4. Fixing dataset imbalances

Due to the rather scarce nature of global stock crash events, a very small fraction of the derived binary dependent variables takes values equal to one (i.e., indicating a crash event) in our dataset. This is especially true when it comes to the 1-day ahead variable, where only around 1% of the available sample corresponds to crash events. This severe imbalance in the structure of our dataset may have adverse repercussions for the performance of the fitted machine learning models. Specifically, it is well-known that fitting machine learning classifiers to imbalanced datasets renders them biased against the minority class.

To stave off this possibility, in our work we investigated synthetic data generation. Specifically, we employed the ROSE (Random over Sampling Examples) package available in the R programming language. ROSE allows for generating balanced artificial datasets, by leveraging sampling methods and a smoothed bootstrap approach.

4. Model development

As already discussed, we employed the following methods: Logistic Regression –LogR (Ohlson, 1980), Random Forests -RF (Breiman, 2001), Support Vector Machines – SVMs (Vapnik & Vapnik, 1998), Neural Networks – NNs (Werbos, 1977), CART, Extreme Gradient Boosting – XGBoost and Deep Feedforward Networks -MXNET (LeCun et al., 2015). In the following, we provide an in-depth account of the development process and parameter tuning of each methodological framework.

To fit the considered models, we resorted to a well-established technique in machine learning, namely k-fold cross-validation, with the number of folds k set equal to five. K-fold cross-validation consists in iterating over a dataset k times. In each round, we split the dataset into k parts: one part is used for model selection (validation), and the remaining k-1 parts are merged into a single dataset used for model fitting. Model evaluation for our binary classification problem was performed on the grounds of the obtained accuracy ratio metric.

Table 3
Estimated logistic regression model: Dependent variable concerns stock crisis occurring within 20 days (Glob-20) (Notation: $^{***}<0.005$ p-value, $^{**}<0.010$ p-value).

	Estimate	Std. error	z value	Pr(> |z|)	
(Intercept)	−1.673	0.064	−26.125	0.000	***
ASIA_L5d_avg	3.019	0.974	3.100	0.002	**
EUR_L5d_avg	5.078	0.833	6.096	0.000	***
AME_L5d_avg	2.882	0.752	3.835	0.000	***
lag2_Jap_SI_30d	−1.874	0.683	−2.742	0.006	**
lag1_Norway_SI_30d	−1.391	0.303	−4.597	0.000	***
lag1_highyield_ln30	0.069	0.411	0.168	0.867	
lag1_fedfunds_ln30	−1.465	0.228	−6.426	0.000	***
lag1_CNYUSD_30d	−32.272	9.409	−3.430	0.001	***

4.1. Logistic regression (LogR)

Logistic regression is an approach broadly employed for building corporate rating systems and retail scorecards, due to its parsimonious structure. It was first used by Ohlson (1980) to predict corporate bankruptcy based on publicly available financial data. We implemented logistic regression in R, by using the glm function that performs model fitting through Iteratively Reweighed Least Squares. In order to reduce the number of fitted parameters, we developed the considered LogR model using only highly scoring features in the variable selection process described in Section 3.3; specifically, we only retained variables yielding a score equal to 4. Subsequently, we performed an additional stepwise selection process, whereby, on each step, we dropped variables for which the statistical significance test yields p-values more than 15%, and refitted the model on the remainder variables.

In Tables 3 and 4, we summarize some of the outcomes of our analysis. As we observe, LogR infers that lagged values of regional crises, in conjunction with falls in large stock markets, (USA, Japan) increase the probability of global crisis. In addition, it also infers that falling returns in Norwegian stock market, which is rather non –volatile, along with rising yields, falling federal fund rates, and falling Yuan, provide strong evidence of an upcoming financial crisis.

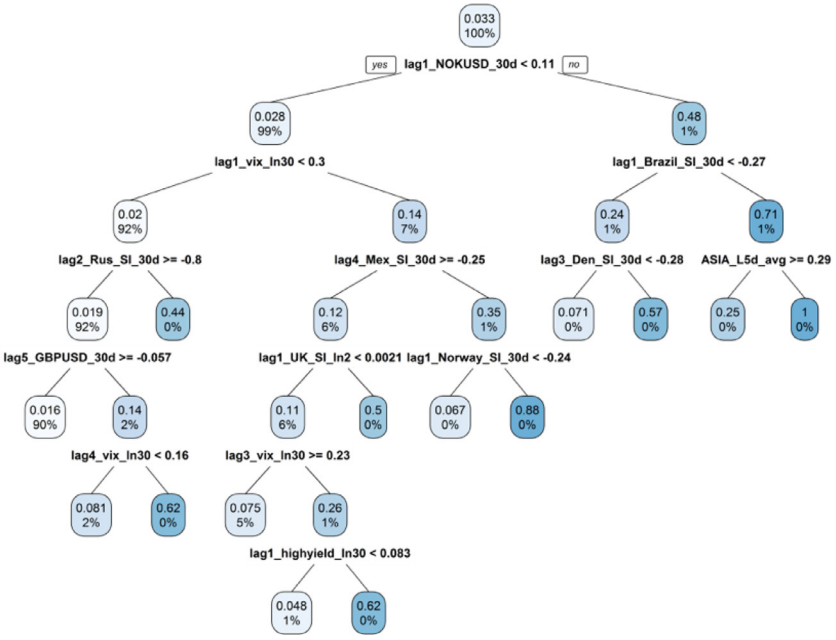

Fig. 5. CART: Dependent variable concerns a stock crisis occurring on the 1-day horizon (Glob-1).

Table 4
Estimated logistic regression model: Dependent variable concerns stock crisis occurring within one day (Glob-1) (Notation: ***<0.005 p-value, **<0.010 p-value).

	Estimate	Std. error	z value	Pr(> |z|)	
(Intercept)	−4.189	0.160	−26.206	0.000	***
ASIA_L5d_avg	−1.180	1.736	−0.680	0.497	
EUR_L5d_avg	0.683	1.604	0.426	0.670	
AME_L5d_avg	4.874	1.369	3.560	0.000	***
lag1_USA_SI_30d	4.440	2.395	1.854	0.064	
lag2_Jap_SI_30d	−3.606	1.830	−1.970	0.049	*
lag2_Germ_SI_30d	−1.635	1.792	−0.912	0.362	
lag1_Norway_SI_30d	−1.013	0.460	−2.203	0.028	*
lag1_highyield_ln30	1.433	0.908	1.578	0.115	
lag1_fedfunds_ln30	−0.679	0.301	−2.257	0.024	*
lag1_CNYUSD_30d	−22.350	19.430	−1.150	0.250	

4.2. Decision trees (CART)

Decision trees consist of a set of nodes, corresponding to the independent variables, and split conditions based on a hierarchical selection of the modeled independent variables. Two well-known algorithms in this field are the Chi-squared Automatic Interaction Detector (CHAID) (GV, 1978) and CART (Breiman, Friedman, Stone, & Ohlsen, 1984) techniques. Decision trees enjoy the advantage of offering simplicity and ease of visualization of the results. On the downside, they are prone to overfitting and a selection bias towards covariates with many possible splits. CART (Classification and Regression Trees) is a well-established algorithm for building decision trees (Breiman et al., 1984). In our case, we implemented CART using the rpart package of R, fed with all the variables stemming from the initial variable selection process. This is a reasonable strategy, since CART is robust to multicollinearity issues that may affect the selected set of independent variables, described in Section 3.3. An exhibition of the outcome of this process is provided in Figs. 5 and 6. Therein, we show the CART models fitted to the available dataset, pertaining to both the one-day dependent variable, and the 20-day dependent variable.

4.3. Random Forests (RF)

Random Forests (RF) (Breiman, 2000) has gained significant ground and is frequently used in many machine learning applications across various fields of the academic community. To build the considered Random Forests, we employed the randomForest package in R. The outline of the algorithm is the following: Let us assume a dataset D which is composed of a series of features denoted by $X_1 - X_N$, and the dependent variable Y. The dependent variable can either be continuous, in case we have a regression problem, or binary, in case we investigate a classification problem. Let us also denote as B the number of decision trees the algorithm is expected to generate. This group of decision trees forms the so-called Forest. Tree generation is randomly performed in an iterative fashion, as follows: On each iteration, a random subsample of the included features is selected from D by means of bootstrap; let us denote this as D_i. Then, a tree T_i is generated from D_i using the CART algorithm. Hence, each constructed tree contains a relatively limited number of features, say m < N. After constructing the random trees, prediction is performed using Bagging (Breiman, 1996). Random forests are usually robust to overfitting, since each forest is only presented with a subset of all the available features. On the other hand, the aforementioned bagging process facilitates strong generalization capacity.

By construction, the predictive ability of RFs increases as the inter-tree correlation decreases. Thus, a large number of predictors can provide increased generalization capacity, which is the case with our model. In addition, performance of random forests depends strongly on the number of parameters to be used in each split for node creation, m. If m is relatively low, then both the inter-tree correlation and the strength of each individual tree will decrease. Hence, it is critical for the overall performance of random forests to find the optimal value of m by means of a tuning algorithm. To this end, in this work we adopt a grid-search approach, using the validation set available on each one of the performed 5 folds of cross-validation. The employed grid search is two-dimensional, taking a range of values for both the number of splits, m, on each tree, and for the number of trees to generate, B.

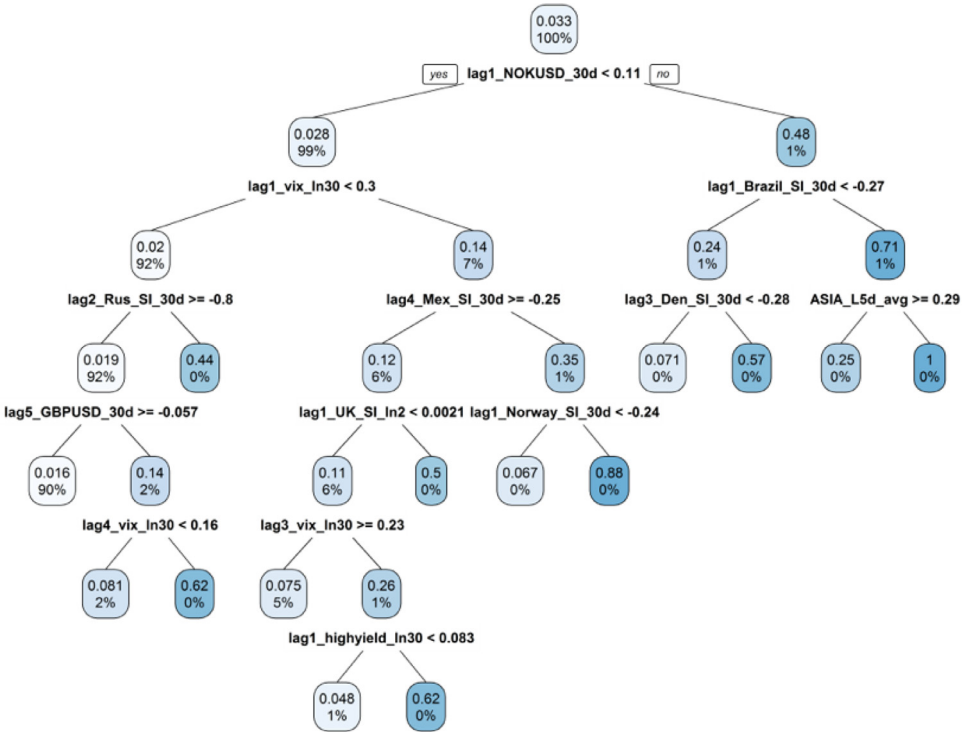

Fig. 6. CART: Dependent variable concerns a stock crisis occurring on the 20-day horizon (Glob-20).

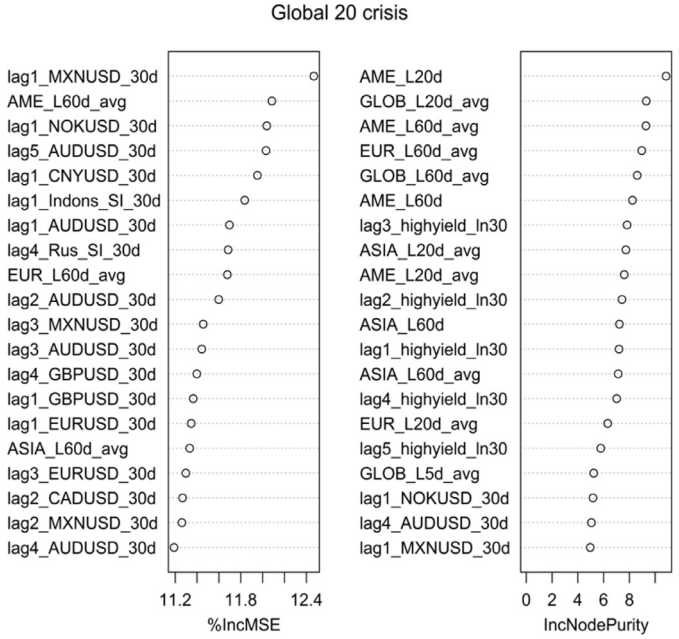

Fig. 7. Random forest variable importance plot. Dependent variable concerns a stock crisis occurring on the 20-day horizon (Glob-20) - Top 20 variables.

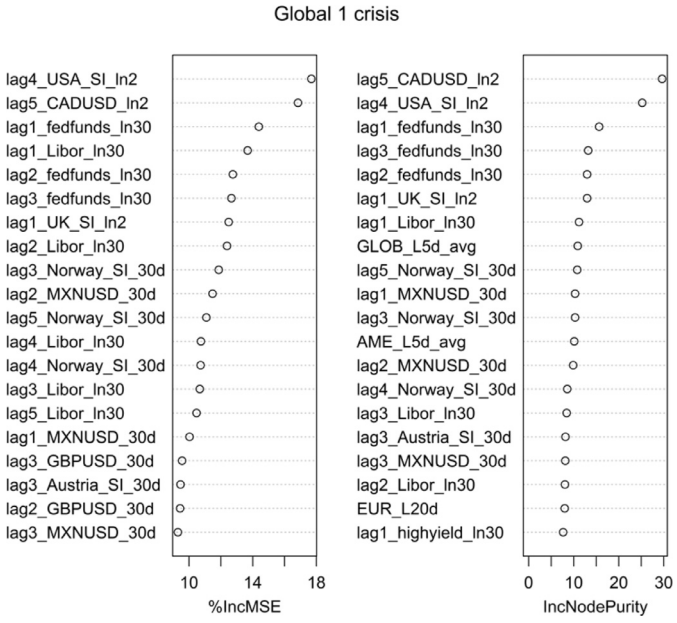

Fig. 8. Random forest variable importance plot. Dependent variable concerns a stock crisis occurring on the one-day horizon (Glob-1) - Top 20 variables.

In Figs. 7 and 8, we present for each financial indicator its importance for the classification outcome, as deemed by the RF algorithm. The obtained ranking is based on two criteria: Mean Square Error and Node Purity. The left part of the chart, pertaining to the MSE, can be interpreted as follows: if a predictor is important, then omitting it or distorting its value would have a negative influence on the overall predictive performance. On the basis of this rationale, this part of the chart compares the obtained MSE of the original dataset with the MSE of the distorted dataset. Note that, therein, the values of the variables are scaled for comparability purposes. The right part of the chart presents node impurity, calculated as the difference between the Residual Sum of Squares (RSS) before and after the split corresponding to each tree node. This is summed over all splits of a given variable, over all trees.

In summary, our illustrated results indicate that the lagged values of regional crises (especially in the 20-day horizon), US stock market crash variables, and disturbances in bond and currency markets are deemed strongly correlated with the probability of a global crisis. Federal Fund rates also are inferred to be an impor-

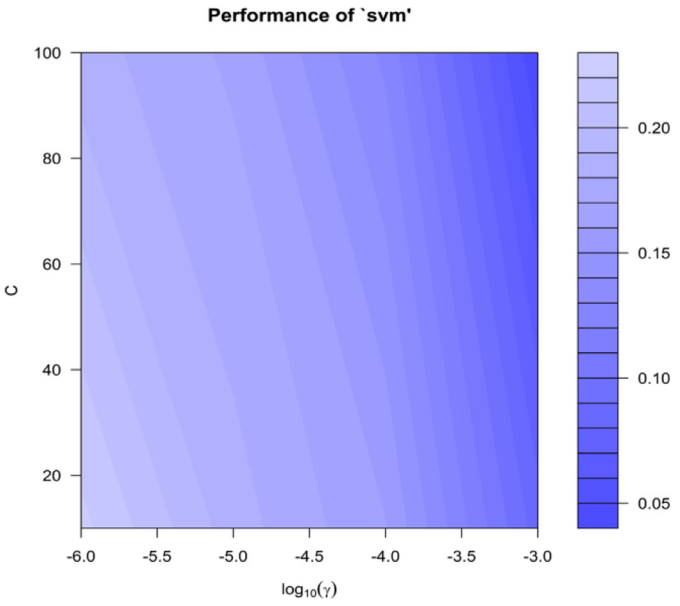

Fig. 9. Parameter tuning for the postulated SVM.

tant determinant. It is also interesting that global crises are found to be persistent, as shown from the high importance of lagged global crisis variables in reducing node impurity.

4.4. Support Vector Machines (SVMs)

SVMs (Vapnik & Vapnik, 1998) have been proven useful to credit rating systems in several studies (Harris, 2015; Huang, 2009). This is due to the fact that they reduce the possibility of overfitting and alleviate the need of tedious cross-validation for the purpose of appropriate hyper-parameter selection. The main drawbacks of SVMs stem from the fact that they constitute black-box models, thus limiting their potential of offering deeper intuition and visualization of the obtained results and inference procedure.

In this study, we evaluate soft-margin SVM classifiers using linear, radial basis function (RBF), polynomial, and sigmoid kernels, and retain the model configuration yielding optimal performance. For selecting the proper kernel, we exploit the available validation set on each fold. To select the hyperparameters of the evaluated kernels, as well as the cost hyperparameter, C, of the SVM (related to the adopted soft margin), we also resort to cross-validation. The candidate values of these hyperparameters are selected based on a grid-search algorithm (Vapnik & Vapnik, 1998). We implemented this model in R using the kernlab package, along with the grid-search functionality included in the e1071 package (Tune routine).

Fig. 9 presents the results of a grid search for different couples of cost, C, (y-axis) and γ (x-axis) hyperparameters of the employed SVM model. As we observe, a large cost parameter gives low bias, as it penalizes the cost of misclassification a lot. However, it leads to high variance, so that the algorithm is forced to explain the fitting data stricter, and potentially overfit. On the other hand, a small misclassification cost allows more bias and lower variance. Regarding γ, when it is very small the model is too constrained and cannot capture the complexity of the data. In this case, two points can be classified the same, even if they are far from each other. On the other hand, a large γ allows two points to be classified the same, only if they are close to each other. In a nutshell, based on Fig. 9 we deduce that parameters closer to the upper right region (darker part of the figure) lead to smaller RMSE in the validation sets; hence, these are selected in the postulated SVM specification.

4.5. Neural Networks (NN)

Neural Networks constitute a well-known machine learning technique that is broadly used in credit rating classification problems. The most usually considered setup is composed of three types of layers. The input layer, in which all candidate variables are imported as a high dimensional vector; one or more consecutive hidden layers, where the information is transformed into a lower-dimensional latent representation; and the output layer that generates predictions by making use of non-linear functions, such as a sigmoid. To develop the employed NN model, we investigated several selections regarding the number of hidden layers, in our case 1- 3, as well as the number of neurons on each layer. The latter number varied from 2 through 10, following the rule of thumb that each layer must be composed of fewer neurons than the previous one in the NN queue. The candidate neural network models were fitted using a standard gradient descent optimization method, namely the backpropagation algorithm Werbos (1977). In all cases, the used independent variables were transformed to take values in the continuous interval of [0,1]; such a normalization procedure is known to increase the reliability of the model fitting algorithm. In addition, in order to reduce the number of parameters, so as to avoid overfitting and obtain more intuitive and parsimonious networks, we only used the high-ranking independent variables obtained in Section 3.3 (scoring 4 in the entailed variable selection process).

We run backpropagation until the training set error fell below a predefined threshold, in our case 0.01, or until a maximum number of iterations was performed, to avoid over-fitting (early stopping; Srivastava, Hinton, Krizhevsky, Sutskever, & Salakhutdinov, 2014). The obtained performance on the validation set, on each k-fold cross-validation iteration, was used to select the best NN configuration. The best performer turned out to comprise two hidden layers. We used sigmoid activation functions. Training and optimization of the neural networks was performed in R using the Neuralnet package.

Although neural networks are difficult to interpret, and their training process can take longer than Random Forests, their performance provides a good benchmark to validate other methodologies. Figs. 10 and 11 depict the structure of the optimized neural networks. In particular, the input layer on the left-hand side of the plot corresponds to the vector of independent variables that we used. The hidden layers, where data processing/transformation takes place, follow in the middle of the plot. Finally, the output layer on the right-hand side of the plot generates predictions of the considered dependent variables.

4.6. Extreme Gradient Boosting (XGBoost)

One of the main novelties of this study is the application of the XGBoost (eXtreme Gradient Boosting) algorithm as a means of forecasting global stock market turbulence. XGBoost (Friedman, 1999) is an advanced implementation of gradient boosting, offering increased efficiency, accuracy and scalability over RFs and NNs. It supports fitting various kinds of objective functions, including regression, classification and ranking. XGBoost offers increased flexibility, since optimization is performed on an extended set of hyperparameters, while it fully supports online training, without the danger of catastrophic forgetting. We developed XGBoost in the context of our study by utilizing the XGBoost R package.

We performed an extensive (5-fold) cross-validation procedure to select a series of entailed hyperparameters, including the maximum depth of generated trees, the minimum leaf node size for performing a split, and the size of subsampling for building the classification trees and the variables considered in each split. We used the logistic objective function to train the model, due to the

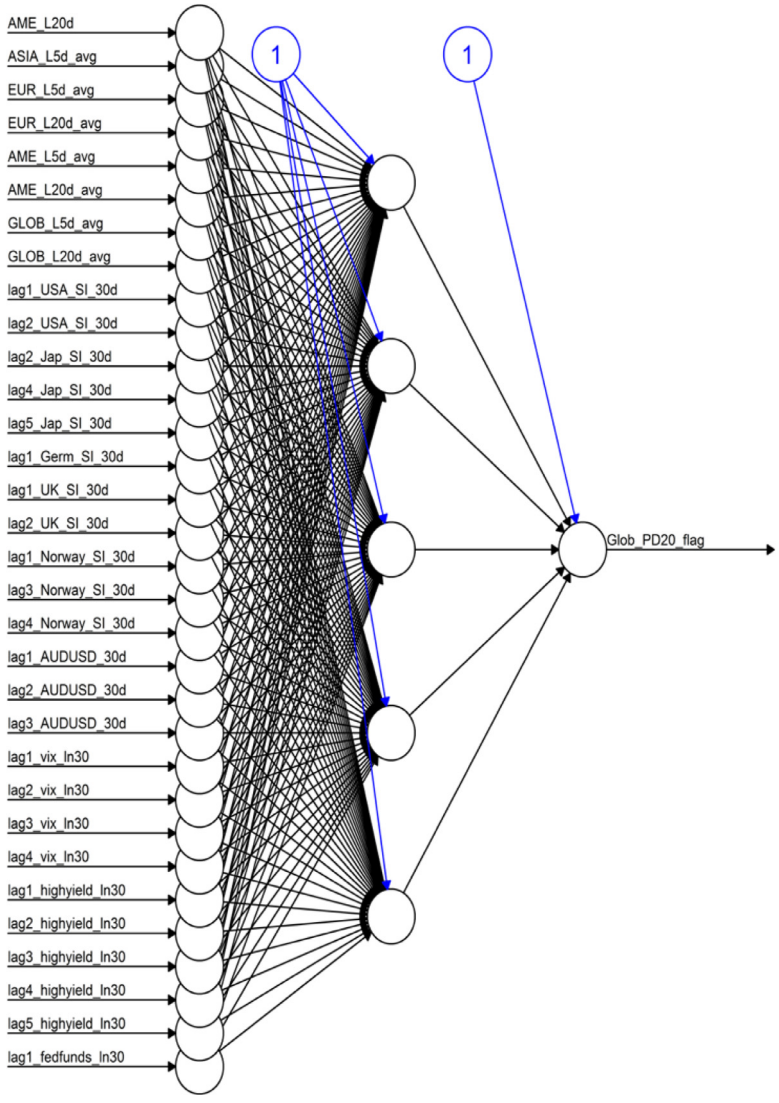

Fig. 10. Fitted NN: Dependent variable concerns a stock crisis occurring on the 20-day horizon (Glob-20).

binary nature of the dependent variable. We used the area under the curve (AUROC) metric for model selection in the context of cross-validation. The AUROC of a classifier is equivalent to the probability that the classifier will rank a randomly chosen positive instance higher than a randomly chosen negative instance. In practice, the value of AUROC varies between 0.5 and 1, with a value above 0.8 indicating a very good performance of the algorithm. To reduce overfitting tendencies, we tuned the γ hyperparameter; this controls model complexity by imposing the requirement that node splits should yield a minimum reduction in the loss function. We also tuned the α and λ hyperparameters, which perform regularization of model weights similar to LASSO.

In Figs. 12 and 13, we show the importance of each independent variable for the classification outcome, as deemed by the XGBoost algorithm. Similar to our analysis regarding the Random Forest algorithm, we show the persistence of crises on the 20-day horizon, and the transmission impact of regional crises both over the 1- and 20-day horizons. As we observe, disturbances in bond and currency markets increase the probability of global crisis, whereas the VIX index seems to be another important element in predicting future disturbances, especially on the 1-day horizon.

4.7. Deep learning feedforward network (MXNET)

Finally, we implemented a Deep Learning Network to address the problem of global crisis event forecasting. Deep learning has been an active field of research in the recent years, as it has achieved significant breakthroughs in the fields of computer vision and language understanding. However, its application in the field of finance is limited. Specifically, our paper constitutes the first work presented in the literature that considers application of deep learning to address this challenging task of crisis prediction.

Our approach consists in building a multi-layer perceptron using the MXNET package of R. We postulated modern deep models that are up to five hidden layers deep and comprise various numbers of neurons. Model selection using 5-fold cross-validation was performed by maximizing the area under the curve metric. We employed modern deep learning activation functions, namely Rectified Linear Unit (ReLU), as well as a strong regularization technique, namely the Dropout technique of Srivastava et al. (2014). The latter reduces the model' s overfitting tendencies by limiting the number of trained neurons on each algorithm iteration.

The final model structure, selected via 5-fold cross-validation, is outlined in Fig. 14. As we observe, in order to achieve the highest possible accuracy on this difficult financial analysis task, we

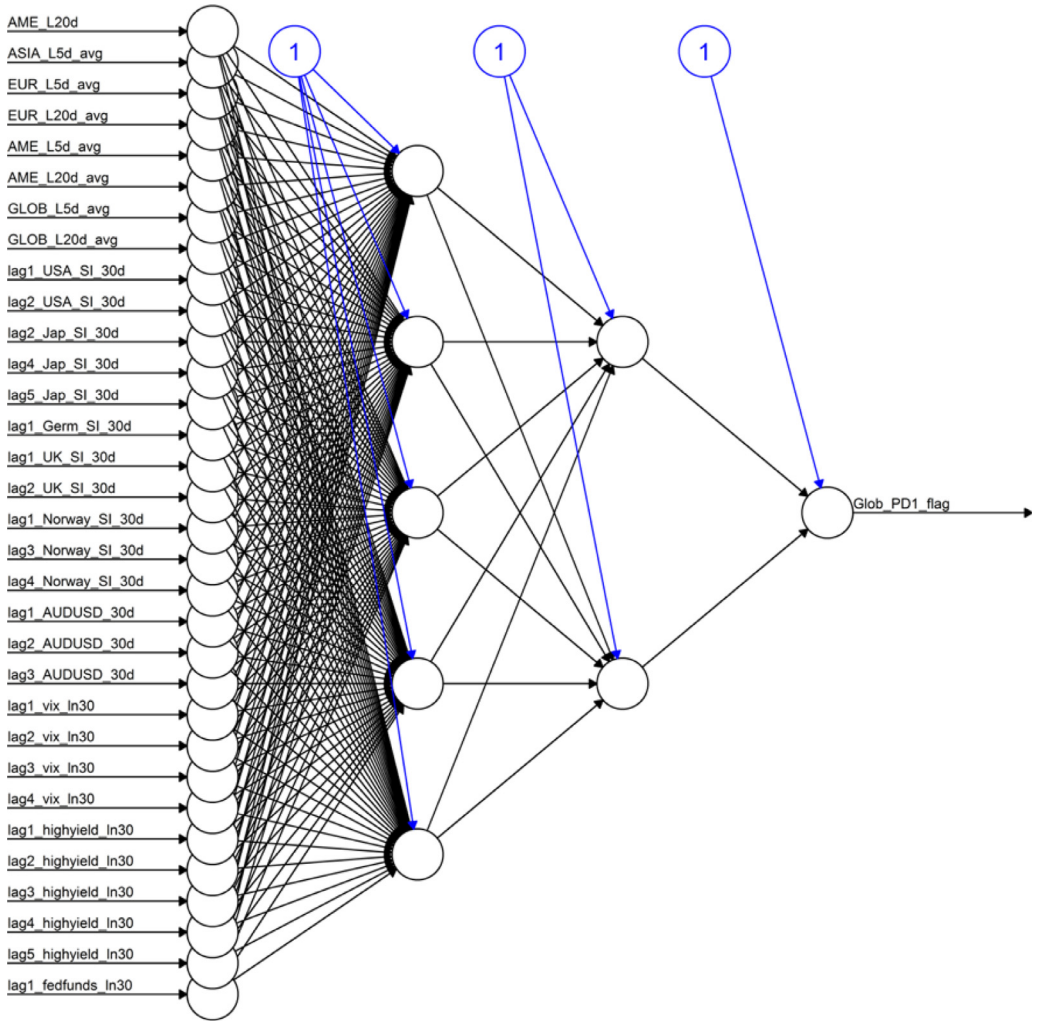

Fig. 11. Fitted NN: Dependent variable concerns a stock crisis occurring on the one-day horizon (Glob-1).

needed postulated deep models of increased complexity. Another point worth mentioning is that forecasting a crash event in the next day requires a slightly more complex architecture compared to forecasting over the next 20 days. Finally, in an effort to provide more transparency regarding what the fitted deep neural networks have managed to infer, we calculate variable importance using an iterative scoring method on the test sample. Specifically, for each feature in the test set, we set the rest of the variables to zero, present the so-constructed data to the fitted network, and estimate the resulting accuracy ratio. The higher the performance score obtained under this procedure, the higher the rating of the corresponding feature.

In Figs. 15 and 16, we show the relationships between the values of the 12 features highest ranked under our adopted procedure, and the corresponding probabilities of crash on the 1-day and 20-day horizon. It is evident that the postulated Deep Learning networks can effectively capture nonlinearities in the relationship between the input variables and the output variable. By examining these variable importance figures, we notice that the lagged variables counting regional and global crashes exhibit high predictive accuracy, similar to the case of Random Forest and XGboost algorithms. This autocorrelative result supports the hypothesis of persistence in the occurrence of stock market tail events.

4.8. Forecast combination

Due to the inherently challenging nature of forecasting stock market crisis events on the global scale, it is well-expected that all postulated models are rather weak learners. Some may obtain better modeling performance under specific observed patterns, others under different ones, but we do not expect any single one to capture best all the existing latent patterns. Hence, exploiting forecast combinations that allow us to assign different weights to each of the obtained predictions is expected to boost the overall obtainable predictive performance (Stock & Watson, 2004). Indeed, such an approach may lead to a decrease in forecast errors by aggregating across all statistical techniques and minimizing model misspecification.

In our work, we first perform simple forecast averaging of the three highest performing models in different ways. In the simplest case, we assign equal weights to the forecasts from each model (SIM). Further, we consider Ordinary Least Squares (OLS) forecast combination; this postulates a simple linear regression model over the individual model predictions, fitted on the available validation sets. We also perform Robust Regression (ROB), which performs the same procedure, but minimizes a different loss function, which is less sensitive to outliers. In addition, we effect constrained least squares (CLS), which also fits a linear model combining the individual predictions but minimizes the sum of squared errors under the restriction that the weights sum up

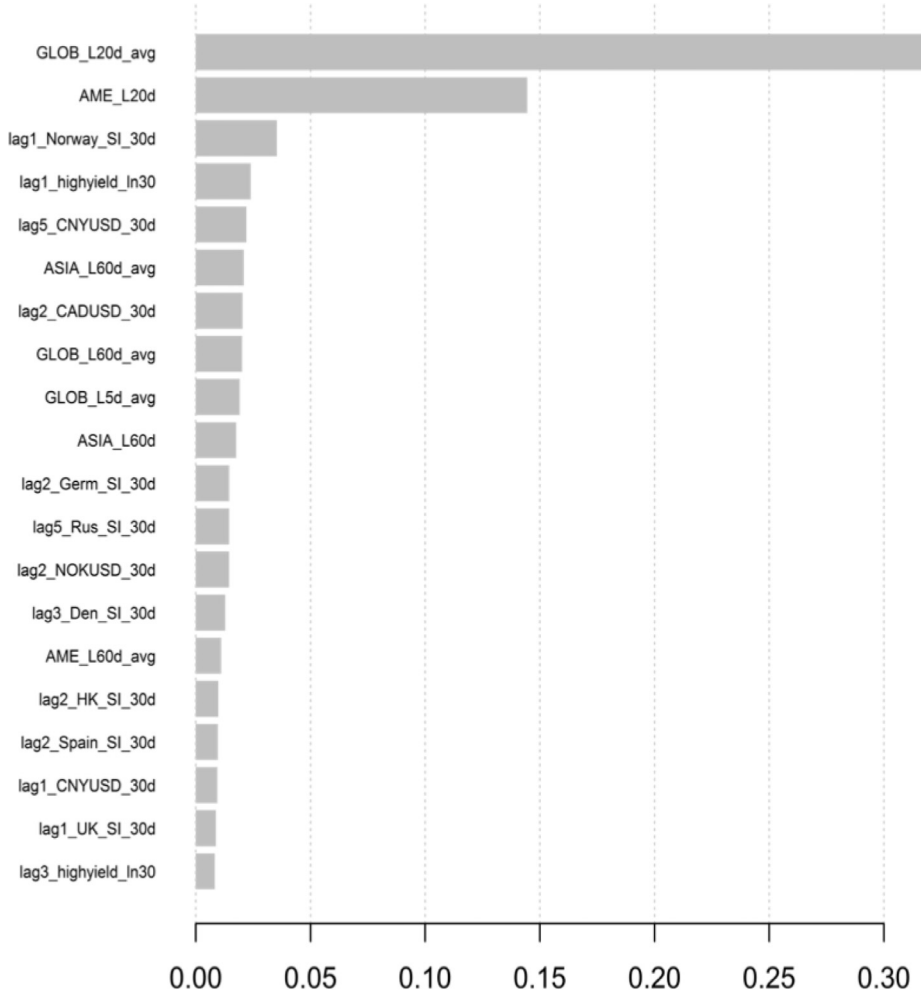

Fig. 12. XGBoost variable importance plot. Dependent variable concerns a stock crisis occurring on the 20-days horizon (Glob-20) - Top 20 variables.

to 1, and that the forecasts themselves are unbiased. Finally, we also employ a variance-based method (VAR), which computes the mean squared error and weights the forecasts according to their accuracy; accurate forecasts receive relatively higher weights. All these options are adopted from the ForecastCombinations package (Ravin, 2015) of the R programming language.

5. Experimental evaluation

In order to assess the robustness of our approach, we perform a thorough experimental evaluation procedure. More precisely, we report performance results obtained by evaluating our method over a long time-period comprising several crash events, specifically the period 2011–2017.

5.1. Validation measures

Classification accuracy, as measured by the discriminatory power of a rating system, is the main criterion to assess the efficacy of each method and to select the most robust one in terms of discriminatory power. In this section, we present a series of metrics that are broadly used for quantitatively estimating the discriminatory power of each scoring model. As usual, we employ the area under the curve metric, as well as the Kolmogorov Smirnov (KS) statistic as performance measures. However, an issue our work is confronted with concerns the considerably imbalanced nature of the available data, since the number of data points pertaining to crises constitute a very small fraction of the available dataset. As such, there is always the danger that the employed performance metrics may misinterpret the contribution of each class to the total accuracy score of the fitted models.

Bekkar, Kheliouane, and Taklit (2013) have proposed a series of measures that overcome any issues of model performance misinterpretation, when the distribution of data points among classes is considerably skewed, as is the case with our data. Specifically, we adopt the sensitivity and specificity metrics, which are defined as follows:

$$Sensitivity = \frac{TP}{TP+FN},\ Specificity = \frac{TN}{TN+FP} \tag{1}$$

where:

- TP = True Positive; the number of positive cases (i.e. crisis) that are correctly identified as positive.
- TN = True Negative; the number of negative cases (i.e. no crisis) that are correctly identified as negative cases.
- FN = False Negative; the number of positive cases (i.e. crisis) that are misclassified as negative cases (i.e. no crisis).
- FP = False Positive; the number of negative cases (i.e. no crisis) that are incorrectly identified as positive cases (i.e. crisis).

On this basis, we calculate a series of combined sensitivity – specificity evaluation measures; these are the following:

G-mean: The geometric mean (G-mean) is the product of sensitivity and specificity. This metric indicates the balance between

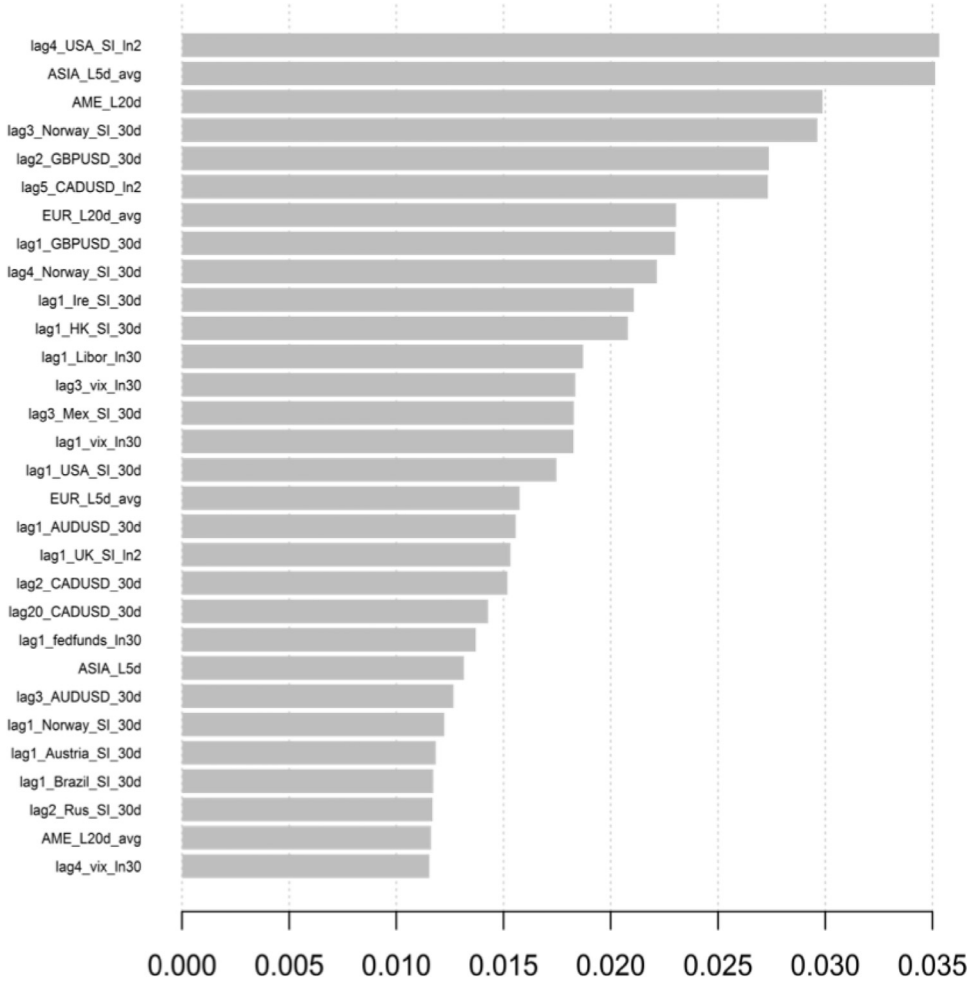

Fig. 13. XGBoost variable importance plot. Dependent variable concerns a stock crisis occurring on the one-day horizon (Glob-1) - Top 30 variables.

classification performances on the majority and minority class. We have

$$G = \sqrt{Sensitivity * Specificity} \tag{2}$$

Under this metric, poor performance in prediction of the positive cases will lead to a low G-mean value, even if the negative cases are correctly classified by the evaluated algorithm.

LP: The positive likelihood ratio represents the ratio between the probability of predicting an example as positive when it is actually positive, and the probability of predicting an example as positive when it is actually not positive. We have

$$LP = \frac{Sensitivity}{1 - Specificity} \tag{3}$$

LR: The negative likelihood ratio is the ratio between the probability of predicting a case as negative when it is actually positive, and the probability of predicting a case as negative when it is actually negative. It holds

$$LR = \frac{1 - Sensitivity}{Specificity} \tag{4}$$

Higher LP values and lower LR values mean better performance on data pertaining to the positive and negative classes, respectively.

DP: Discriminant power is a measure that summarizes sensitivity and specificity. It is defined as

$$DP = \frac{\sqrt{3}}{\pi}\left[\log\frac{Sensitivity}{1 - Sensitivity} + \log\frac{Specificity}{1 - Specificity}\right] \tag{5}$$

DP values higher than 3 indicate that the algorithm distinguishes well between positive and negative cases.

Youden' s γ: Youden' s index is a linear transformation of the mean sensitivity and specificity. It is defined as:

$$\gamma = Sensitivity - (1 - Specificity) \tag{6}$$

As a general rule, a higher value of Youden' s γ indicates better ability of the algorithm to avoid misclassification.

BA: The balanced accuracy metric is the average of Sensitivity and Specificity. In cases the evaluated classifier performs equally well on both classes, this term reduces to the conventional accuracy measure. We have:

$$BA = \frac{1}{2}(Sensitivity + Specificity) \tag{7}$$

In contrast, if the classifier works well only in the majority class (i.e. the class which is dominant in terms of events, no-crisis in our case), the balanced accuracy metric will plummet significantly, thus signaling severe performance issues. Hence, BA is a performance metric that takes into equal consideration the obtained accuracy of the evaluated model in the both the majority as well as the minority class (i.e., crisis events, in our case).

WBA: This is a weighted balance accuracy measure which weights sensitivity more than specificity (under the weighting scheme 75%-25%).

These metrics are used so as to derive a full-spectrum conclusion regarding the relative classification power of each model. Even though it is certain that the values of these distinct performance

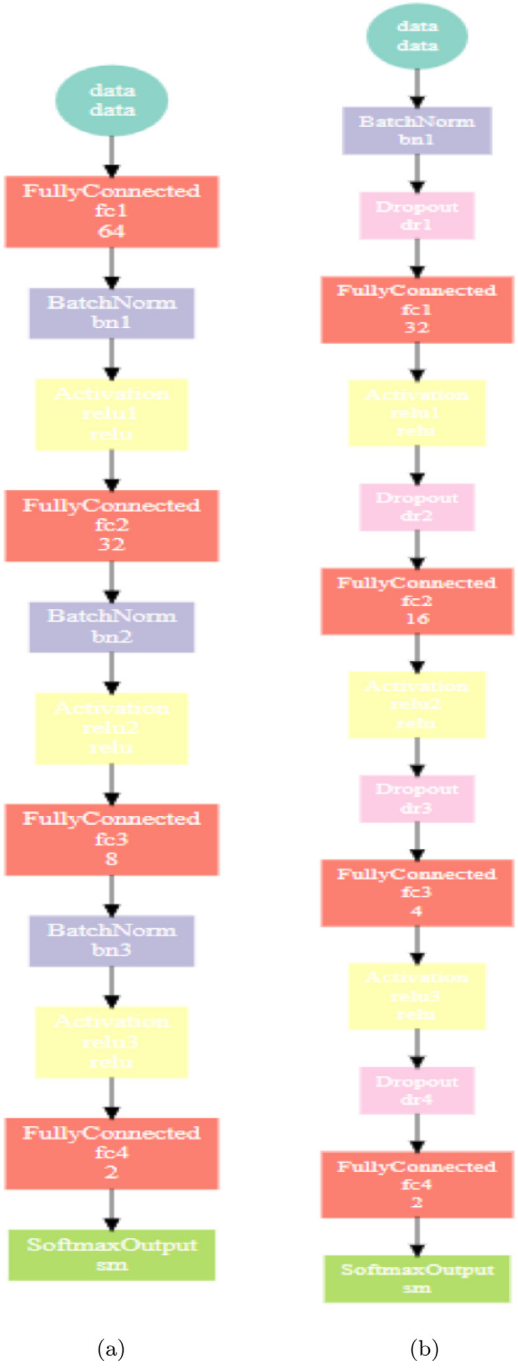

Fig. 14. Optimized deep neural network architectures. Right hand side: Dependent variable concerns a stock crisis occurring on the one-day horizon (Glob-1). Left hand side: Stock crisis occurring on the 20-day horizon (Glob-20).

metrics may indeed be correlated, we investigate them all to obtain a more holistic view on model performance. On the grounds of these outcomes, we eventually calculate an optimal stock market crash probability cut-off point for each fitted model, which obtains its optimal sensitivity and specificity measures. In these optimizations, we weight 10 times higher the cost of a generated false-negative compared to the cost of predicting a crisis that never happened (false-positive). This is reasonable, since the ultimate aim of this work is to create an advanced warning mechanism that produces correct signals as much as possible, with a special focus on diminishing false-negatives.

Table 5
Validation Measures – Dependent Variable concerns a stock crisis occurring on the 20-day horizon (Glob-20). AUROC (Area Under the Curve), KS (Kolmogorov – Smirnov), G-mean (Geometric Mean), LP (positive likelihood ratio), LR (negative likelihood ratio), DP (Discriminant power), Youden' s index, BA (Balanced Accuracy) and WBA (Weighted Balanced Accuracy).

Glob20	Logit	CART	RF	SVM	NN	XGBOOST	MXNET
AUROC	0.630	0.654	0.739	0.708	0.677	0.743	0.783
KS	0.278	0.284	0.322	0.367	0.296	0.398	0.441
G-mean	0.549	0.594	0.616	0.591	0.596	0.635	0.638
LP	3.116	3.474	3.858	2.947	3.485	4.153	4.083
LR	0.743	0.680	0.645	0.690	0.677	0.615	0.610
DP	0.790	0.899	0.987	0.800	0.904	1.053	1.048
Youden	0.229	0.284	0.316	0.267	0.286	0.343	0.346
BA	0.615	0.642	0.658	0.634	0.643	0.672	0.673
WBA	0.476	0.520	0.542	0.519	0.522	0.562	0.566

Table 6
Validation Measures – Dependent Variable concerns a stock crisis occurring on the 1-day horizon (Glob-1). AUROC (Area Under the Curve), KS (Kolmogorov – Smirnov), G-mean (Geometric Mean), LP (positive likelihood ratio), LR (negative likelihood ratio), DP (Discriminant power), Youden' s index, BA (Balanced Accuracy) and WBA (Weighted Balanced Accuracy).

Glob1	Logit	CART	RF	SVM	NN	XGBOOST	MXNET
AUROC	0.698	0.640	0.741	0.708	0.776	0.737	0.807
KS	0.363	0.301	0.424	0.398	0.483	0.438	0.516
G-mean	0.610	0.606	0.583	0.610	0.557	0.611	0.682
LP	4.114	3.669	3.664	4.114	3.428	4.237	5.142
LR	0.652	0.661	0.692	0.652	0.728	0.650	0.537
DP	1.016	0.945	0.919	1.016	0.854	1.034	1.246
Youden	0.313	0.301	0.276	0.313	0.244	0.316	0.417
BA	0.657	0.651	0.638	0.622	0.657	0.658	0.708
WBA	0.535	0.532	0.509	0.535	0.483	0.536	0.613

5.2. Quantitative results

Our original development sample contains 3300 observations; in the following, we refer it as the "Full in-sample" dataset. As already mentioned, we perform model development by resorting to 5-fold cross-validation. We evaluate the developed methods by using a dataset covering the years 2011–2017; in the following, we refer to this dataset as the "Out-of-time sample."

As we observe in Tables 5 and 6, the MXNet al.gorithm provides the best empirical performance on both horizons (1-day and 20-day). This is followed by the XGBoost methodology in the case of the 20-day horizon, and the Neural Network in the case of the 1-day horizon. Hence, MXNET deep neural networks offer significantly superior predictive accuracy both in the 1-day and 20-day forecasting setup on the out-of-time sample. Another remark is that, by moving from simple neural networks to deep networks, we are able to infer richer and subtler dynamics from the data, thus increasing our capacity in modeling nonlinearities and cross-correlations among financial market variables.

Summarizing the results across all metrics in the test sample, it is evident that the MXNET system exhibits higher discriminatory power compared to all the considered benchmark models. At this point, it is imperative to stress that a non-anticipated crash in the global stock markets may come at a much higher cost for the economy compared to a false-alarm. Hence, it is crucial for supervisory purposes to achieve the maximum possible accuracy in predicting imminent crises via a developed EWS for economic and financial crises.

Further, in Figs. 17 and 18 we present the ROC curves corresponding to the analyzed models. These curves are created by plotting the true positive rate against the false positive rate at various threshold settings. As such, they illustrate the obtained trade-offs between sensitivity and specificity, as any increase in sensitivity will be accompanied by a decrease in specificity. The closer the

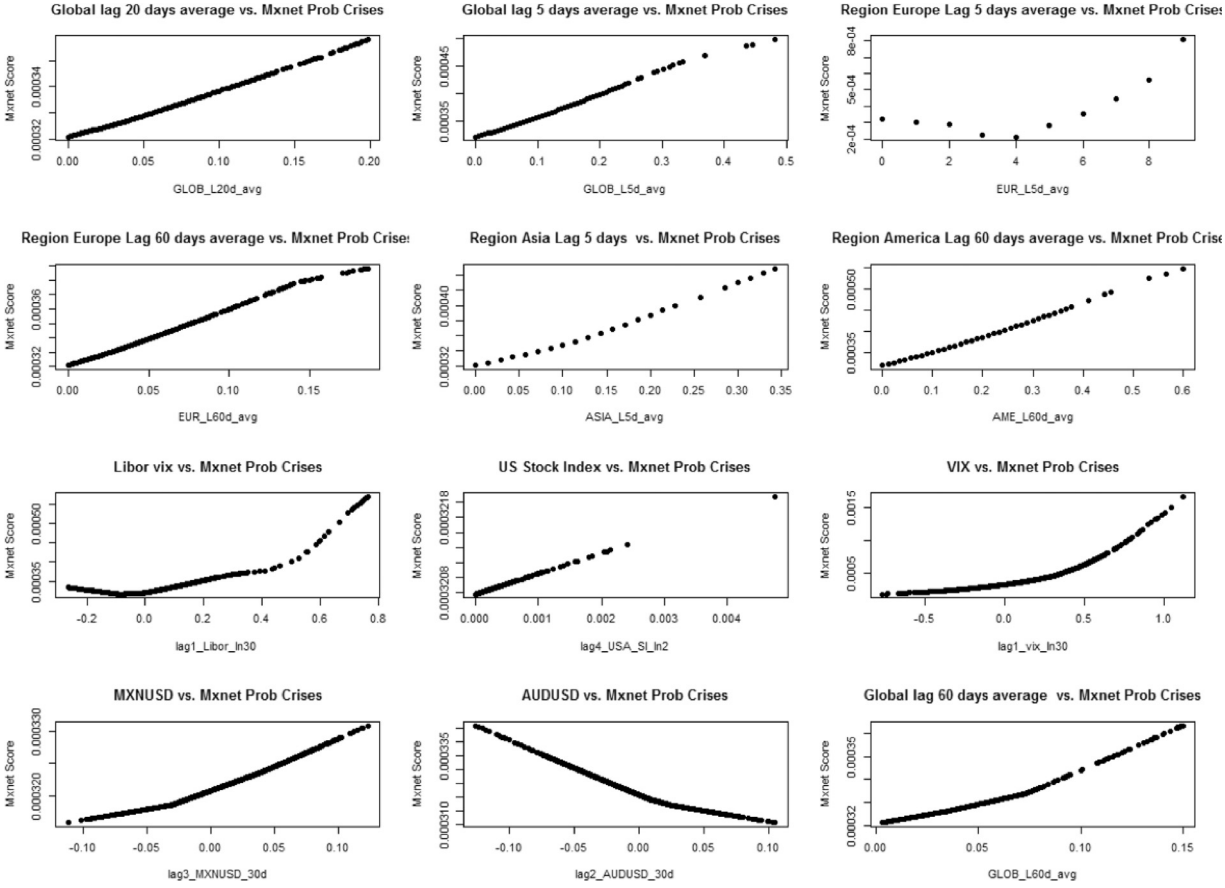

Fig. 15. MXNET variable importance plot: Dependent variable concerns a stock crisis occurring on the 20-day horizon (Glob-20).

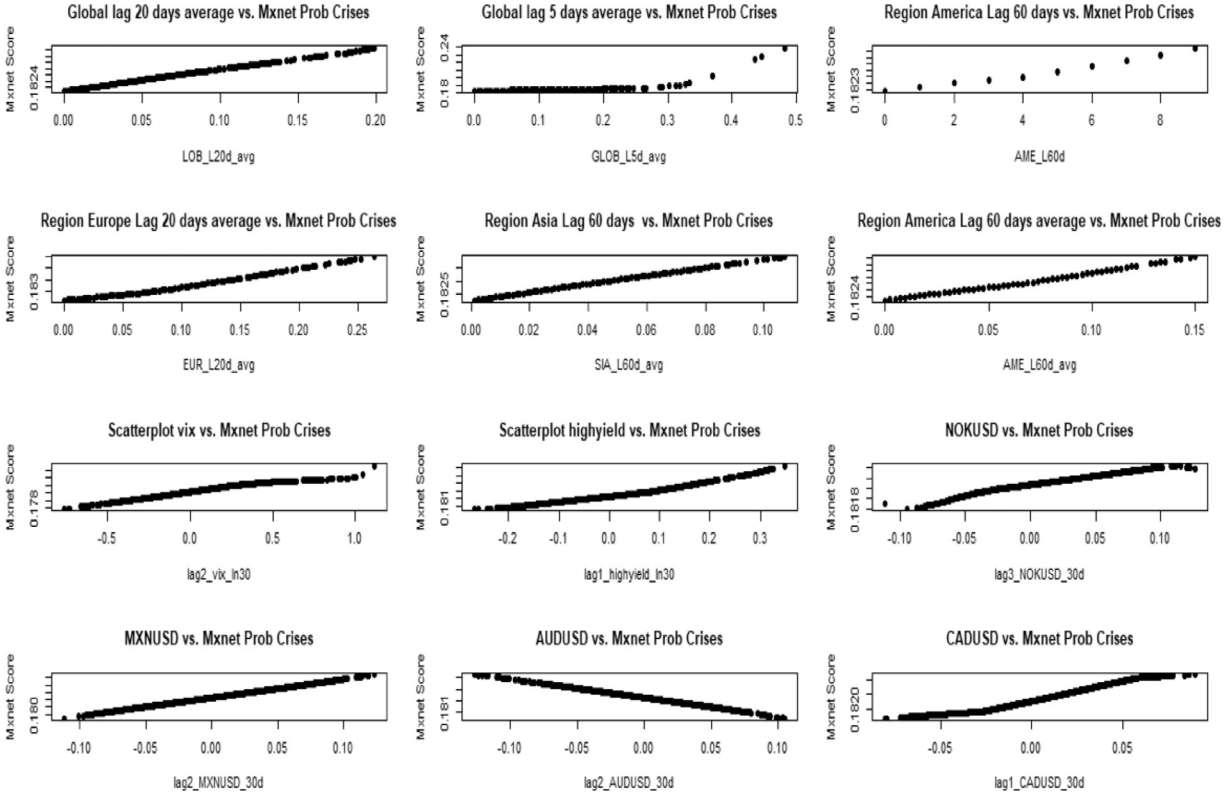

Fig. 16. MXNET variable importance plot: Dependent variable concerns a stock crisis occurring on the one-day horizon (Glob-1).

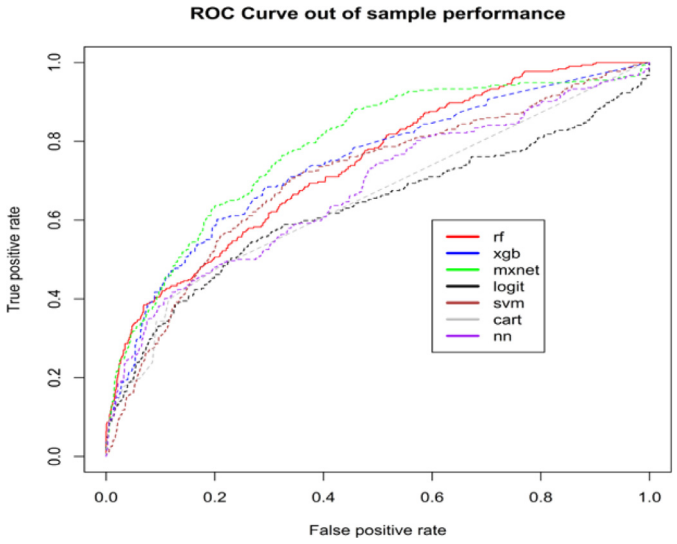

Fig. 17. ROC curve for forecasting a stock crisis occurring on the 20-day horizon (Glob-20).

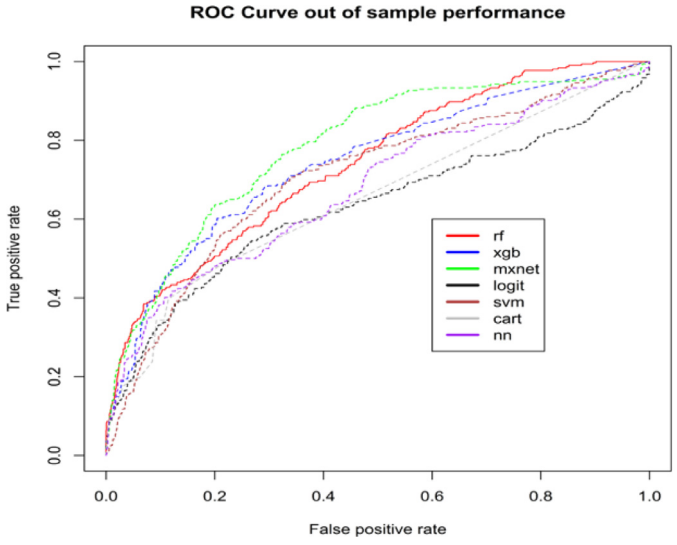

Fig. 18. ROC curve for forecasting a stock crisis occurring on the one-day horizon (Glob-1).

curve follows the left-hand border and then the top border of the ROC space, the more accurate the modeling approach. The corresponding ROC curve of deep neural networks is higher over all the considered competitors regarding both the explored dependent variables (pertaining to the one-day and 20-day horizons). Hence, we obtain yet another strong evidence supporting the high degree of efficacy and generalization capacity of the proposed deep learning system.

Finally, we also evaluate the proposed forecast combination techniques of the three best performing models, over each horizon. Therefore, we combine MXNET, Neural Network and Random Forest forecasts in the case of the one-day horizon, and MXNET, XGBoost and Random Forest forecasts for the 20-day horizon. The obtained results are depicted in Table 7. We observe that utilization of the CLS technique yields a minor improvement over MXNET in the one-day horizon. However, MXNET continues to be the top performer in the case of the 20-day setup. This finding provides yet another strong piece of evidence that deep learning techniques can offer an unprecedented level of accuracy compared to any technique currently used in the literature.

Table 7
Combination methods: AUROC (Area Under ROC) of Simple Averaging (Simple), Ordinary Least Squares (OLS), Robust regression (Robust), Variance-Based method (Var), Constrained Least Squares (CLS), and their comparison to MXNET. Glob_20 refers to the 20-day forecast horizon and Glob_1 to the one-day horizon.

	Simple	OLS	Robust	Var	CLS	MXNET
Glob_1 AUROC	0.772	0.813	0.643	0.783	0.813	0.807
Glob_20- AUROC	0.756	0.728	0.729	0.743	0.739	0.783

5.3. Accuracy of generated alarms

To conclude, we provide further insights into the possibility of using the developed methods as an early warning system for increasing awareness of an upcoming shock event. To this end, we infer for each trained model an optimal cutoff threshold of the predicted crisis probabilities. Predictive probabilities exceeding this threshold generate an alarm, either on the one-day or on the 20-day horizon. Specifically, we seek a threshold that minimizes the total weighted prediction loss, where the cost of not predicting a crisis (false-negative) is weighted 10 times more than the cost of a false-alarm, as we have already explained.

On this basis, we compute final classification performance tables for all the evaluated models. Our obtained results are depicted in Table 8. Concentrating on the best-performing model, namely the deep neural network, we deduce that by accepting a 10% false-alarm rate, we succeed in predicting half of the future global crashes in the stock markets. In addition, exploiting a forecast combination scheme allows to correctly predict a 60% of the days during which a stock market crisis occurred (1-day horizon). In the 20-day horizon case, the maximum hit rate obtained by means of the MXNET reached 46%.

6. Conclusions and future work

Financial crises forecasting is important for both practitioners and policymakers, since the in-depth comprehension of financial linkage breakdown after a crisis may substantially facilitate administration strategy selection and the development of contingency plans. In this paper, we proposed a full system that: (i) selects the most significant financial market indicators, which can be used to predict stock market tail events; (ii) chooses statistical machine learning techniques that are capable of inferring subtle and dynamic correlation patterns between the measured indicators and the occurrence probability of crisis events; and (iii) explores the efficacy of ensemble techniques, in an effort to improve overall classification accuracy. Our empirical results indicate that deep learning yields superior out-of-sample predictive performance, thus offering new exiting capabilities for practitioners and policy makers.

The main novel contribution of this empirical study to the literature of forecasting economic and financial crisis events is four-fold. First and foremost, we extensively explored relevant statistical machine learning techniques to address the problem at hand. Specifically, we implemented approaches that are popular in the field, as well as two new algorithms that yield state-of-the-art performance in other scientific fields, namely Deep Neural Networks and XGBoost. Second, we implemented model validation using latest market data, so as to test the efficacy of each modeling technique. Third, we used performance measures for model evaluation that are appropriate for imbalanced datasets, as is the case with datasets dealing with stock market crashes, which are scarce in nature. Last but not least, we offered a wide and in-depth examination of an extended set of explanatory variables that can be used to perform the predictive task at hand, which cover the full spectrum of major financial markets.

Table 8
Classification accuracy tables: Glob-20 refers to 20-day crisis forecast; Glob-1 refers to one-day horizons.

Glob 20					Glob 1				
Logit	pred				Logit	pred			
TRUE	0	1	Signal	Rate	TRUE	0	1	Signal	Rate
0	1573	190	False alarm	11%	0	1843	205	False alarm	10%
1	208	106	Hit rate	34%	1	17	12	Hit rate	41%
CART	pred				CART	pred			
TRUE	0	1	Signal	Rate	TRUE	0	1	Signal	Rate
0	1575	188	False alarm	11%	0	1817	231	False alarm	11%
1	206	108	Hit rate	34%	1	17	12	Hit rate	41%
RF	pred				RF	pred			
TRUE	0	1	Signal	Rate	TRUE	0	1	Signal	Rate
0	1568	195	False alarm	11%	0	1837	211	False alarm	10%
1	181	133	Hit rate	42%	1	18	11	Hit rate	38%
SVM	pred				SVM	pred			
TRUE	0	1	Signal	Rate	TRUE	1		Signal	Rate
0	1522	241	False alarm	14%	0	1843	205	False alarm	10%
1	187	127	Hit rate	40%	1	17	12	Hit rate	41%
NN	pred				NN	pred			
TRUE	0	1	Signal	Rate	TRUE	0	1	Signal	Rate
0	1561	202	False alarm	11%	0	1843	205	False alarm	10%
1	188	126	Hit rate	40%	1	19	10	Hit rate	34%
XGBoost	pred				XGBoost	pred			
TRUE	0	1	Signal	Rate	TRUE	0	1	Signal	Rate
0	1571	192	False alarm	11%	0	1849	199	False alarm	10%
1	173	141	Hit rate	45%	1	17	12	Hit rate	41%
MXNET	pred				MXNET	pred			
TRUE	0	1	Signal	Rate	TRUE	0	1	Signal	Rate
0	1565	198	False alarm	11%	0	1843	205	False alarm	10%
1	171	143	Hit rate	46%	1	14	15	Hit rate	52%
					CLS	pred			
					TRUE	0	1	Signal	Rate
					0	1842	206	False alarm	10%
					1	12	17	Hit rate	59%

Summarizing our experimental results, we have found that Deep Neural Networks built using the MXNET library consistently outperform the rest of the employed approaches. This has been the case across almost all metrics that are broadly used for assessing the discriminatory power of a binary classifier evaluated on imbalanced datasets. Further, we estimated the predictive capacity of an ensemble model built of the top-three performing models, and investigated the variance of each performance assessment measure. Our analysis provided strong evidence of increased classification accuracy and performance consistency, which implies a much stronger generalization capacity compared to state-of-the-art models. Hence, this finding renders our approach much more attractive to researchers and practitioners working in real-world financial institutions. Besides, the fact that traditional methods offer poor results signals that global financial dynamics are not only driven by volatility regimes or momentum effects; these can be effectively captured by traditional methods such as logistic regression. On the contrary, the employed nonlinear methods are better in capturing global shocks as well as isolated crash events; this provides strong evidence that our approach offers a good starting point for developing an early warning system.

From a qualitative perspective, one of the main conclusions of our work is that regional crashes in the last 20 and 5 days, as well as global crashes, increase the probability of reoccurrence of a crash event in the near term. This finding essentially implies a strong clustering behavior permeating such crisis events. This is supported by the inferred importance parameters assigned to the autoregressive variables included in the final model specification of all the considered statistical machine learning techniques. Additionally, these variable importance weights also indicate that there are interdependence effects among stock markets crashes, as well as cross-effects among stock, bond and currency markets. The aforementioned results may have important implications for policy makers, since they show that during crisis times economic decisions must take seriously into consideration potentially contagious repercussions in third markets. Of equal importance are the implications for asset management professionals, due to the fact that diversification benefits may cease to exist in turbulent periods.

In any case, since economic conditions continuously change, so do the potential causes of a financial breakdown and a consequent financial contagion. Hence, crisis events forecasting will remain an open research issue, with many more aspects and challenges to address. However, our work provides the first ever reported concrete empirical evidence that the constantly increasing computational capabilities, and the availability of large datasets that central banks have access to, create a fertile ground for leveraging the deep learning breakthrough to revolutionize financial forecasting processes.

An aspect this work has not considered concerns developing machine learning models that can be continuously retrained in a moving window (online learning) setup. Exploring the utility of an even more diverse set of machine learning algorithms and different deep learning architectures is also something that would be worth of investigation. For instance, the utilization of stronger regularization techniques, e.g. Partaourides and Chatzis (2017), may considerably enhance generalization capacity for the developed system. Finally, another possible way forward is the exploration of deep neural networks under a high-frequency data setup. This will allow for testing their ability to extract information in stock behavior of short-term nature, by filtering out noise embedded in such time-series data. The value of such novel developments remains to be examined in our future research endeavors.

References

Atsalakis, G., Protopapadakis, E., & Valavanis, K. (2016). Stock trend forecasting in turbulent market periods using neuro-fuzzy systems. *Operational Research, 16*(2), 245–269.

Babecký, J., Havranek, T., Mateju, J., Rusnak, M., Smidkova, K., & Vacisek, B. (2014). Banking, debt, and currency crises in developed countries: Stylized facts and early warning indicators. *Journal of Financial Stability, 15*, 1–17.

Bae, K., Karolyi, A., & Stulz, R. (2003). A new approach to measuring financial contagion. *Review of Financial Studies, 16*, 717–763.

Bagheria, A., Mohammadi, H., & Akbaric, M. P. (2014). Financial forecasting using ANFIS networks with quantum-behaved particle swarm optimization. *Expert Systems with Applications, 41*, 6235–6250.

Baig, T., & Goldfajn, I. (2000). *The russian default and the contagion to brazil*. Published in stijn claessens; kristin forbes. international financial contagion Kluwer Academic Publishers (pp. 268–299).

Barro, R., & Ursua, J. (2009). *Stock market crises and depressions*. NPER. working paper series. No 14760.

Bekkar, M., Kheliouane, H., & Taklit, A. (2013). Evaluation measures for models assessment over imbalanced data sets. *Journal of Information Engineering and Applications, 3*(10).

Bluedorn, J. C., Decressin, J., & Terrones, M. E. (2013). Do asset price drops foreshadow recessions?IMF Working Paper No. WP/13/203, International Monetary.

Breiman, L. (1996). Bagging predictors. *Machine Learning, 24*(2), 123–140.

Breiman, L. (2001). Random forests. *Machine learning, 45*(1), 5–32.

Breiman, L., Friedman, J., Stone, C., & Ohlsen, R. (1984). *Classification and regression trees*. CRC press.

Bussiére, & Matthieu (2013). In defense of early warning signals. Banque de France Working Paper No. 420.

Cervelló-Royo, R., Guijarroa, F., & Michniukab, K. (2015). Stock market trading rule based on pattern recognition and technical analysis: Forecasting the DJIA index with intraday data. *Expert Systems with Applications, 42*, 5963–5975.

Chiang, W. C., Enke, D., Wu, T., & Wang, R. (2016). An adaptive stock index trading decision support system. *Expert Systems with Applications, 59*, 195–207.

Christiansen, C., & Ranaldo, A. (2009). Extreme coexceedances in new EU member states stock markets. *Journal of Banking & Finance, 33*(6), 1048–1057.

Cuneyt, S., Oztekin, A., Ozkan, B., Serkan, G., & Erkam, G. (2014). Developing an early warning system to predict currency crises. *European Journal of Operational Research, 237*, 1095–1104.

Dabrowski, J. J., Beyers, C., & Villiers, J. P. (2016). Systemic banking crisis early warning systems using dynamic Bayesian networks. *Expert Systems with Applications, 62*, 225–242.

Dungey, M., & Martin, V. (2006). Unravelling financial market linkages during crises. *Journal of Applied Econometrics, 22*(1), 89–119.

Döpke, J., Fritsche, U., & Pierdzioch, C. (2017). Predicting recessions with boosted regression trees. *International Journal of Forecasting, 33*(4), 745–759.

Enke, D., & Thawornwong, S. (2005). The use of data mining and neural networks for forecasting stock market returns. *Expert Systems with Applications, 29*, 927–940.

Estrella, A., & Mishkin, F. (1996). *The yield curve as a predictor of US recessions. current issues in economic and finance*. Federal Reserve Bank of New York. 2(7).

Faranda, D., Flavio, M., Giachino, E., Vaienti, S., & Dubrulle, B. (2015). Early warnings indicators of financial crises via auto regressive moving average models. *Communications in Nonlinear Science and Numerical Simulation, 29*(1–3), 233–239.

Farmer, R. E. (2012). The stock market crash of 2008 caused the great recession: Theory and evidence. *Journal of Economic Dynamics and Control, 36*(5), 693–707.

Forbes, K., & Rigobon, R. (2002). No contagion, only interdependence: Measuring stock market comovements. *The Journal of Finance, 57*(5), 2223–2261.

Ghazali, R., Hussain, A. J., & Liatsis, P. (2011). Dynamic ridge polynomial neural network: Forecasting the univariate non-stationary and stationary trading signals. *Expert Systems with Applications, 38*, 3765–3776.

Harris, T. (2015). Credit scoring using the clustered support vector machine. *Expert Systems with Applications, 42*(2), 741–750.

Huang, S. C. (2009). Integrating nonlinear graph based dimensionality reduction schemes with SVMs for credit rating forecasting. *Expert Systems with Applications, 36*(4), 7515–7518.

Ibragimov, R., & Johan, W. (2007). The limits of diversification when losses may be large. *Journal of Banking and Finance, 31*(8), 2551–2569.

Kaminsky, G., Lizondo, S., & Reinhart, C. (1998). *Leading indicators of financial crises*. IMF Staff Papers. 45(1).

Kim, Y. H., & Chang, H. W. (2018). Forecasting the volatility of stock price index: A hybrid model integrating LSTM with multiple GARCH-type models. *Expert Systems with Applications, 103*, 25–37.

Kole, E., Koedijk, K., & Verbeek, M. (2006). Portfolio implications of systemic crises. *Journal of Banking and Finance, 30*, 2347–2369.

Kursa, M., & Rudnicki, W. (2010). Feature selection with the boruta package. *Journal of Statistical Software, 36*(11), 1–13.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature, 521*, 436–444.

Longin, F., & Solnik, B. (2001). Extreme correlation of international equity markets. *The journal of finance, 56*(2), 649–676.

Markwat, T., Kole, E., & Van Dijk, D. (2009). Contagion as a domino effect in global stock markets. *Journal of Banking & Finance, 33*(11), 1996–2012.

Ohlson, J. (1980). Financial ratios and the probabilistic prediction of bankruptcy. *Journal of Accounting Research, 18*, 109–131.

Oztekin, A., Kizilaslan, R., Freund, S., & Iseri, A. (2016). A data analytic approach to forecasting daily stock returns in an emerging market. *European Journal of Operational Research, 253*(3), 697–710.

Partaourides, H., & Chatzis, S. P. (2017). Deep network regularization via bayesian inference of synaptic connectivity. In J. Kim (Ed.), *Proceedings of the PAKDD* (pp. 30–41). Part I, LNAI 10234

Srivastava, N., Hinton, J., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: A simple way to prevent neural networks from overfitting,. *Journal of Machine Learning Research, 15*, 1929–1958.

Stock, J., & Watson, M. (2004). Combination forecasts of output growth in a seven—country data set. *Journal of Forecasting, 23*(6), 405–430.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B (methodological). Wiley, 58*(1), 267–288.

Vapnik, V., & Vapnik, V. (1998). *Statistical learning theory*. New York: Wiley.

Vašíč, B., Zingraoiva, D., Hoeberichts, M., Vermuelen, R., Smidkova, K., & De Haan, J. (2017). Leading indicators of financial stress: New evidence. *Journal of Financial Stability, 28*, 240–257.

Werbos, P. J. (1977). Advanced forecasting methods for global crisis warning and models of intelligence. *General Systems Yearbook, 22*, 25–38.

Zhong, X., & Enke, D. (2017). Forecasting daily stock market return using dimensionality reduction. *Expert Systems with Applications, 67*, 126–139.