

Machine Learning for Financial Stability



Lucia Alessi and Roberto Savona

Abstract What we learned from the global financial crisis is that to get information about the underlying financial risk dynamics, we need to fully understand the complex, nonlinear, time-varying, and multidimensional nature of the data. A strand of literature has shown that machine learning approaches can make more accurate data-driven predictions than standard empirical models, thus providing more and more timely information about the building up of financial risks. Advanced machine learning techniques provide several advantages over empirical models traditionally used to monitor and predict financial developments. First, they are able to deal with high-dimensional datasets. Second, machine learning algorithms allow to deal with unbalanced datasets and retain all of the information available. Third, these methods are purely data driven. All of these characteristics contribute to their often better predictive performance. However, as “black box” models, they are still much underutilized in financial stability, a field where interpretability and accountability are crucial.

1 Introduction

What we learned from the global financial crisis is that to get information about the underlying financial risk dynamics, we need to fully understand the complex, nonlinear, time-varying, and multidimensional nature of the data. A strand of literature has shown that machine learning approaches can make more accurate data-driven predictions than standard empirical models, thus providing more and more timely information about the building up of financial risks.

L. Alessi (✉)

European Commission - Joint Research Centre, Ispra (VA), Italy
e-mail: lucia.alessi@ec.europa.eu

R. Savona

University of Brescia, Brescia, Italy
e-mail: roberto.savona@unibs.it

© The Author(s) 2021

S. Consoli et al. (eds.), *Data Science for Economics and Finance*,
https://doi.org/10.1007/978-3-030-66891-4_4

Advanced machine learning techniques provide several advantages over empirical models traditionally used to monitor and predict financial developments. First, they are able to deal with high-dimensional datasets, which is often the case in economics and finance. In fact, the information set of economic agents, be it central banks or financial market participants, comprises hundreds of indicators, which should ideally all be taken into account. Looking at the financial sphere more closely, as also mentioned by [25] and [9], banks should use, and are in fact using, advanced data technologies to ensure they are able to identify and address new sources of risks by processing large volumes of data. Financial supervisors should also use machine learning and advanced data analytics (so-called *suptech*) to increase their efficiency and effectiveness in dealing with large amounts of information. Second, and contrary to standard econometric models, machine learning algorithms allow to deal with unbalanced datasets, hence retaining all of the information available. In the era of big data, one might think that losing observations, i.e., information, is not anymore a capital sin as it used to be decades ago. Hence, one could afford cleaning the dataset from problematic observations to obtain, e.g., a balanced panel, given the large amount of observations one starts with. On the contrary, large datasets require even more flexible models, as they almost invariably feature large amounts of missing values or unpopulated fields, “ragged” edges, mixed frequencies or irregular periodic patterns, and all sorts of issues that standard techniques are not able to handle. Third, these methods are purely data-driven, as they do not require making *ex ante* crucial modelling choices. For example, standard econometric techniques require selecting a restricted number of variables, as the models cannot handle too many predictors. Factor models, which allow handling large datasets, still require the econometrician to set the number of the underlying driving forces. Another crucial assumption, often not emphasized, relates to the linearity of the relevant relations. While standard econometric models require the econometrician to explicitly control for nonlinearities and interactions, whose existence she should know or hypothesize *a priori*, machine learning methods are designed to address these types of dynamics directly. All of these characteristics contribute to their often better predictive performance.

Thanks to these characteristics, machine learning techniques are also more robust in handling the fitting versus forecasting trade-off, which is reminiscent of the so-called “forecasting versus policy dilemma” [21], to indicate the separation between models used for forecasting and models used for policymaking. Presumably, having a model that overfits in-sample when past data could be noisy leads to the retention of variables that are spuriously significant, which produces severe deficiencies in forecasting. The noise could also affect the dependent variable when the definition of “crisis event” is unclear or when, notwithstanding a clear and accepted definition of crisis, the event itself is misclassified due to a sort of noisy transmission of the informational set used to classify that event. Machine learning gives an opportunity to overcome this problem.

While offering several advantages, however, machine learning techniques also suffer from some shortcomings. The most important one, and probably the main reason why these models are still far from dominating in the economic and financial

literature, is that they are “black box” models. Indeed, while the modeler can surely control inputs, and obtain generally accurate outputs, she is not really able to explain the reasons behind the specific result yielded by the algorithm. In this context, it becomes very difficult, if not impossible, to build a story that would help users make sense of the results. In economics and finance, however, this aspect is at least as important as the ability to make accurate predictions.

Machine learning approaches are used in several very diverse disciplines, from chemometrics to geology. With some years delay, the potential of data mining and machine learning is also becoming apparent in the economics and finance profession. Focusing on the financial stability literature, some papers have appeared in relatively recent years, which use machine learning techniques for an improved predictive performance. Indeed, one of the areas where machine learning techniques have been more successful in finance is the construction of early warning models and the prediction of financial crises. This chapter focuses on the two supervised machine learning approaches becoming increasingly popular in the finance profession, i.e., decision trees and sparse models, including regularization-based approaches. After explaining how these algorithms work, this chapter offers an overview of the literature using these models to predict financial crises.

The chapter is structured as follows. The next section presents an overview of the main machine learning approaches. Section 3 explains how decision tree ensembles work, describing the most popular approaches. Section 4 deals with sparse models, in particular the LASSO, as well as related alternatives, and the Bayesian approach. Section 5 discusses the use of machine learning as a tool for financial stability policy. Section 6 provides an overview of papers that have used these methods to assess the probability of financial crises. Section 7 concludes and offers suggestions for further research.

2 Overview of Machine Learning Approaches

Machine learning pertains to the algorithmic modeling culture [17], for which data predictions are assumed to be the output of a partly unknowable system, in which a set of variables act as inputs. The objective is to find a rule (algorithm) that operates on inputs in order to predict or classify units more effectively without any a priori belief about the relationships between variables. The common feature of machine learning approaches is that algorithms are realized to learn from data with minimal human intervention. The typical taxonomy used to categorize machine learning algorithms is based on their learning approach, and clusters them in supervised and unsupervised learning methods.¹

¹See [7] for details on this classification and a comprehensive discussion on the relevance of the recent machine learning literature for economics and econometrics.

Supervised machine learning focuses on the problem of predicting a response variable, y , given a set of predictors, x . The goal of such algorithms is to make good out-of-sample predictions, rather than estimating the structural relationship between y and x . Technically, these algorithms are based on the *cross-validation* procedure. This latter involves the repeated rotation of subsamples of the entire dataset, whereby the analysis is performed on one subsample (the *training set*), then the output is tested on the other subset(s) (the *test set*). Such a rotational estimation procedure is conceived with the aim of improving the out-of-sample predictability (accuracy), while avoiding problems of overfitting and selection bias, this one induced by the distortion resulting from collecting nonrandomized samples.

Supervised machine learning methods include the following classes of algorithms:

- Decision tree algorithms. Decision trees are decision methods based on the actual attributes observed in the explored dataset. The objective is to derive a series of rules of thumb, visualized in a tree structure, which drive from observations to conclusions expressed as predicted values/attributes. When the response variable is continuous, decision trees are named regression trees. When instead the response variable is categorical, we have classification trees. The most popular algorithm in this class is CART (Classification and Regression Trees) introduced in [18].² CART partitions the space of predictors x in a series of homogeneous and disjoint regions with respect to the response variable y , whose nature defines the tree as classification (when y is a categorical variable) or regression (when y is a continuous variable) tree.
- Ensemble algorithms. Tree ensembles are extensions of single trees based on the concept of prediction averaging. They aim at providing more accurate predictions than those obtained with a single tree. The most popular ensemble methods are the following: Boosting, Bootstrapped Aggregation (Bagging), AdaBoost, Gradient Boosting Machines (GBM), Gradient Boosted Regression Trees (GBRT). and Random Forest.
- Instance-based algorithms. These methods generate classification predictions using only specific instances and finding the best match through a similarity measure between subsamples. A list of the most used algorithms includes the following: k-Nearest Neighbor (kNN), Learning Vector Quantization (LVQ), Locally Weighted Learning (LWL), and Support Vector Machines (SVM). SVM, in particular, are particularly flexible as they are used both for classification and regression analysis. They are an extension of the vector classifier, which provides clusters of observations identified through partitioning the space via linear boundaries. In addition, SVM provide nonlinear classifications by mapping their inputs into high-dimensional feature spaces through nonlinear boundaries. In more technical terms, SVM are based on a hyperplane (or a set of hyperplanes— which in a two-dimensional space are just lines) in a high/infinite-dimensional

²Less popular decision trees algorithms are: Chi-squared Automatic Interaction Detection (CHAID), Iterative Dichotomiser 3 (ID3); C4.5 and C5.0.

space. Hyperplane(s) are used to partition the space into classes and are optimally defined by assessing distances between pairs of data points in different classes. These distances are based on a kernel, i.e., a similarity function over pairs of data points.

- **Regularization algorithms.** Regularization-based models offer alternative fitting procedures to the least square method, leading to better prediction ability. The standard linear model is commonly used to describe the relationship between the y and a set of x_1, x_2, \dots, x_p variables. Ridge regression, Least Absolute Shrinkage and Selection Operator (LASSO), and Elastic Net are all based on detecting the optimal constraint on parameter estimations in order to discard redundant covariates and select those variables that most contribute to better predict the dependent variable out-of-sample.
- **Bayesian algorithms.** These methods apply Bayes Theorem for both classification and regression problems. The most popular Bayesian algorithms are: Naive Bayes, Gaussian Naive Bayes, Multinomial Naive Bayes, Averaged One-Dependence Estimators (AODE), Bayesian Belief Network (BBN), and Bayesian Network (BN).
- **Supervised Artificial Neural Networks.** Artificial neural networks (ANN) are models conceived to mimic the learning mechanisms of the human brain—specifically, supervised ANN run by receiving inputs, which activate “neurons” and ultimately lead to an output. The error between the estimation output and the target is used to adjust the weights used to connect the neurons, hence minimizing the estimation error.

Unsupervised machine learning applies in contexts where we explore only x without having a response variable. The goal of this type of algorithm is to understand the inner structure of x , in terms of relationships between variables, homogeneous clustering, and dimensional reduction. The approach involves pattern recognition using all available variables, with the aim of identifying intrinsic groupings, and subsequently assigning a label to each data point. Unsupervised machine learning includes clusters and networks.

The first class of algorithms pertains to clustering, in which the goal is, given a set of observations on features, to partition the feature space into homogeneous/natural subspaces. Cluster detection is useful when we wish to estimate parsimonious models conditional to homogeneous subspaces, or simply when the goal is to detect natural clusters based on the joint distribution of the covariates.

Networks are the second major class of unsupervised approaches, where the goal is to estimate the joint distribution of the x variables. Network approaches can be split in two subcategories: traditional networks and Unsupervised Artificial Neural Networks (U-ANN). Networks are a flexible approach that gained popularity in complex settings, where extremely large number of features have to be disentangled and connected in order to understand inner links and time/spatial dynamics. Finally, Unsupervised Artificial Neural Networks (U-ANN) are used when dealing with unlabeled data sets. Different from Supervised Artificial Neural Networks, here the objective is to find patterns in the data and build a new model based on a smaller set

of relevant features, which can represent well enough the information in the data.³ Self-Organizing Maps (SOM), e.g., are a popular U-ANN-based approach which provides a topographic organization of the data, with nearby locations in the map representing inputs with similar properties.

3 Tree Ensembles

This section provides a brief overview of the main tree ensemble techniques, starting from the basics, i.e., the construction of an individual decision tree. We start from CART, originally proposed by [18]. This seminal paper has spurred a literature reaching increasingly high levels of complexity and accuracy: among the most used ensemble approaches, one can cite as examples bootstrap aggregation (Bagging, [15]), boosting methods such as Adaptive Boosting (AdaBoost, [29]), Gradient Boosting [30], and [31], Multiple Additive Regression Trees (MART, [32]), as well as Random Forest [16].⁴ Note, however, that some of the ensemble methods we describe below are not limited to CART and can be used in a general classification and regression context.

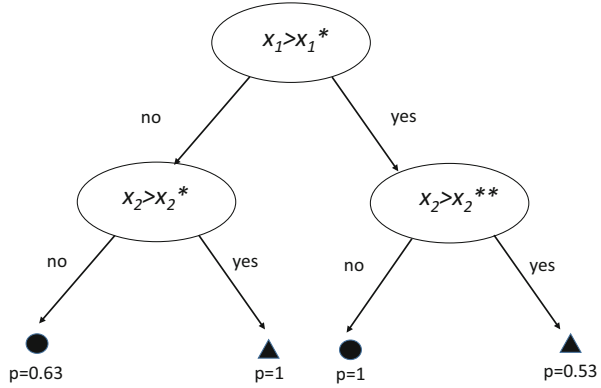
We only present the most well-known algorithms, as the aim of this section is not to provide a comprehensive overview of the relevant statistical literature. Indeed, many other statistical techniques have been proposed in the literature, that are similar to the ones we describe, and improve over the original proposed models in some respects. The objective of this section is to explain the main ideas at the root of the methods, in nontechnical terms.

Tree ensemble algorithms are generally characterized by a very good predictive accuracy, often better than that of the most widely used regression models in economics and finance, and contrary to the latter, are very flexible in handling problematic datasets. However, the main issue with tree ensemble learning models is that they are perceived as black boxes. As a matter of fact, it is ultimately not possible to explain what a particular result is due to. To make a comparison with a popular model in economics and finance, while in regression analysis one knows the contribution of each regressor to the predicted value, in tree ensembles one is not able to map a particular predicted value to one or more key determinants. In policymaking, this is often seen as a serious drawback.

³On supervised and unsupervised neural networks see [57].

⁴See [56] for a review of the relevant literature.

Fig. 1 Example of binary tree structure



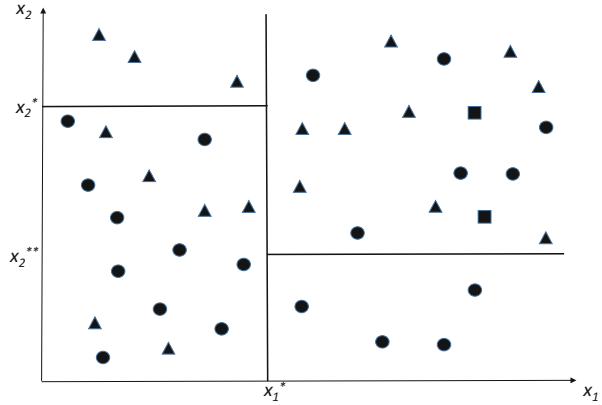
3.1 Decision Trees

Decision trees are nonparametric models constructed by recursively partitioning a dataset through its predictor variables, with the objective of optimally predicting a response variable. The response variable can be continuous (for regression trees) or categorical (for classification trees). The output of the predictive model is a tree structure like the one shown in Fig. 1. CART are binary trees, with one *root node*, only two branches departing from each *parent node*, each entering into a *child node*, and multiple *terminal nodes* (or “leaves”). There can also be nonbinary decision trees, where more than two branches can attach to a node, as, e.g., those based on Chi-square automatic interaction detection (CHAID, [43]). The tree in Fig. 1 has been developed to classify observations, which can be circles, triangles, or squares. The classification is based on two features, or predictors, x_1 and x_2 . In order to classify an observation, starting from the root node, one needs to check whether the value of feature x_1 for this observation is higher or lower than a particular threshold x^* . Next, the value of feature x_2 becomes relevant.⁵ Based on this, the tree will eventually classify the observation as either a circle or a triangle. In the case of the tree in Fig. 1, for some terminal nodes the probability attached to the outcome is 100%, while for some other terminal nodes, it is lower. Notice that this simple tree is not able to correctly classify squares, as a much deeper tree would be needed for that. In other words, more splits would be needed to identify a partition of the space where observations are more likely to be squares than anything else. The reason will become clearer looking at the way the tree is grown.

Figure 2 explains how the tree has been constructed, starting from a sample of circles, triangles, and squares. For each predictor, the procedure starts by considering all the possible binary splits obtained from the sample as a whole. In our example, where we only have two predictors, this implies considering all

⁵Notice that this is not necessarily the case, as the same variable can be relevant in the tree at consecutive nodes.

Fig. 2 Recursive partitioning



possible values for x_1 and x_2 . For each possible split, the relevant impurity measure of the child nodes is calculated. The impurity of a node can be measured by the Mean Squared Error (MSE), in the case of regression trees, or the Gini index, for classification trees, or information entropy. In our case, the impurity measure will be based on the number of circles and triangles in each subspace associated with each split. The best split is the value for a specific predictor, which attains the maximum reduction in node impurity. In other words, the algorithm selects the predictor and the associated threshold value which split the sample into the two purest subsamples. In the case of classification trees, e.g., the aim is to obtain child nodes which ideally only contain observations belonging to one class, in which case the Gini index corresponds to zero. Looking at Fig. 2, the first best split corresponds to the threshold value x_1^* . Looking at the two subspaces identified by this split, the best split for $x_1 < x_1^*$ is x_2^* , which identifies a pure node for $x_2 > x_2^*$. The best split for $x_1 > x_1^*$ is x_2^{**} , which identifies a pure node for $x_2 < x_2^{**}$. The procedure is run for each predictor at each split and could theoretically continue until each terminal node is pure. However, to avoid overfitting, normally a stopping rule is imposed, which, e.g., requires a minimum size for terminal nodes. Alternatively, one can expect “prune” large trees, by iteratively merging two adjoining terminal nodes.⁶

Decision trees are powerful algorithms that present many advantages. For example, in terms of data preparation, one does not need to clean the dataset from missing values or outliers, as they are both handled by the algorithm, nor does one need to normalize the data. Moreover, once the tree structure is built, the model output can be operationalized also by the nontechnical user, who will simply need to assess her observation of interest against the tree. However, they also suffer from one major shortcoming, i.e., the tree structure is often not robust to small variations in the data. This is due to the fact that the tree algorithm is recursive, hence a different split at any level of the structure is likely to yield different splits at any lower level. In

⁶See [38] for technical details, including specific model choice rules.

extreme cases, even a small change in the value of one predictor for one observation could generate a different split.

3.2 *Random Forest*

Tree ensembles have been proposed to improve the robustness of predictions realized through single models. They are collections of single trees, each one grown on a subsample of observations. In particular, tree ensembles involve drawing subsets of are collections of regression trees, where each tree is grown on a subsample of observations. In particular, tree ensembles involve drawing subsets of observations with replacement, i.e., Bootstrapping and Aggregating (also referred to as “BAGGING”) the predictions from a multitude of trees. The Random Forest [16] is one of the most popular ensemble learning inference procedures. The Random Forest algorithm involves the following steps:

1. Selecting a random subsample of the observations⁷
2. Selecting a random small subset of the predictors⁸
3. Growing a single tree based on this restricted sample
4. Repeating the first two steps thousands of times⁹

Predictions for out-of-sample observations are based on the predictions from all the trees in the Forest.

On top of yielding a good predictive performance, the Random Forest allows to identify the key predictors. To do so, the predictive performance of each tree in the ensemble needs to be measured. This is done based on how good each tree is at correctly classifying or estimating the data that are not used to grow it, namely the so-called out-of-bag (OOB) observations. In practice, it implies computing the MSE or the Gini impurity index for each tree. To assess the importance of a predictor variable, one has to look at how it impacts on predictions in terms of MSE or Gini index reduction. To do so, one needs to check whether the predictive performance worsens by randomly permuting the values of a specific predictor in the OOB data. If the predictor does not bring a significant contribution in predicting the outcome variable, it should not make a difference if its values are randomly permuted before the predictions are generated. Hence, one can derive the importance of each predictor by checking to what extent the impurity measure worsens.

⁷It is common practice to use 60% of the total observations (see [38]).

⁸The number of the selected predictors is generally around to the square root of the total number of predictors, while [16] tests with one variable at a time and with a number of features equal to the first integer less than $\log_2 M + 1$, where M is the total number of features.

⁹The accuracy of the Random Forest algorithm is heuristically proven to converge with around 3000 trees (see [38]).

3.3 *Tree Boosting*

Another approach to the construction of Tree ensembles is Tree Boosting. Boosting means creating a strong prediction model based on a multitude of weak prediction models, which could be, e.g., CARTs. Adaptive Boosting (AdaBoost, [29]) is one of the first and the most popular boosting methods, used for classification problems. It is called *adaptive* because the trees are built iteratively, with each consecutive tree increasing the predictive accuracy over the previous one. The simplest AdaBoost algorithm works as follows:

1. Start with growing a tree with just one split.¹⁰
2. Consider the misclassified observations and assign them a higher weight compared to the correctly classified ones.
3. Grow another tree on all the (weighted) observations.
4. Update the weights.
5. Repeat 3–4 until overfitting starts to occur. This will be reflected in a worsening of the predictive performance of the tree ensemble on out-of-sample data.

Later, the Gradient Boosting algorithm was proposed as a generalization of AdaBoost [30]. The simplest example would involve the following steps:

1. Grow a regression tree with a few attributes.¹¹
2. Compute the prediction residuals from this tree and the resulting mean squared error.¹²
3. On the residuals, grow another tree which optimizes the mean squared error.
4. Repeat 1–3 until overfitting starts to occur.

To avoid overfitting, it has been proposed to include an element of randomness. In particular, in Stochastic Gradient Boosting [31], each consecutive simple tree is grown on the residuals from the previous trees, but based only on a subset of the full data set. In practice, each tree is built on a different subsample, similarly to the Random Forest.

3.4 *CRAGGING*

The approaches described above are designed for independent and identically distributed (i.i.d.) observations. However, often this is not the case in economics and finance. Often, the data has a panel structure, e.g., owing to a set of variables

¹⁰Freund and Schapire [29] do not use CART and also propose two more complex algorithms, where the trees are grown by using more than one attribute.

¹¹Typically between 4 and 8, see [38].

¹²More generally, one can use other loss functions than the mean squared error, such as the mean absolute error.

being collected for several countries. In this case, observations are not independent; hence there is information in the data that can be exploited to improve the predictive performance of the algorithm. To this aim, the CRAGGING (CRoss-validation AGGregatING) algorithm has been developed as a generalization of regression trees [66]. In the case of a panel comprising a set of variables for a number of countries observed through time, the CRAGGING algorithm works as follows:

1. Randomly partition the whole sample into subsets of equal size. The number of subsets needs to be smaller than the number of countries.
2. One of the subsets is reserved for testing, while the others are used to train the algorithm. From the training set, one country is removed and a regression tree is grown and pruned.
3. The test set is used to compute predictions based on the tree.
4. The country is reinserted in the training set and steps 2–3 are repeated for all the countries.
5. A cross-validation procedure is run over the test set to obtain a tree which minimizes prediction errors. Hence, CRAGGING combines two types of cross-validation, namely, the leave-one-unit-out cross-validation, in which the units are removed one at a time from the training set and then perturbed, and the usual cross-validation on the test sets, run to minimize the prediction error out-of-sample. (see [66] for details).
6. Steps 1–5 are repeated thousands of times and predictions from the thousands of trees are aggregated by computing the arithmetic average of those predictions.
7. As a final step, a regression tree is estimated on the predictions' average (computed at step 6) using the same set of the original covariates.

This algorithm eventually yields one single tree, thereby retaining the interpretability of the model. At the same time, its predictions are based on an ensemble of trees, which increases its predictive accuracy and stability.

4 Regularization, Shrinkage, and Sparsity

In the era of Big Data, standard regression models increasingly face the “curse of dimensionality.” This relates to the fact that they can only include a relatively small number of regressors. Too many regressors would lead to overfitting and unstable estimates. However, often we have a large number of predictors, or candidate predictors. For example, this is the case for policymakers in economics and finance, who base their decisions on a wide information set, including hundreds of macroeconomic and macrofinancial data through time. Still, they can ultimately only consider a limited amount of information; hence variable selection becomes crucial.

Sparse models offer a solution to deal with a large number of predictor variables. In these models, regressors are many but relevant coefficients are few. The Least Absolute Shrinkage and Selection Operator (LASSO), introduced by [58] and

popularized by [64], is one of the most used models in this literature. Also in this case, from this seminal work an immense statistical literature has developed with increasingly sophisticated LASSO-based models. Bayesian shrinkage is another way to achieve sparsity, very much used, e.g., in empirical macroeconomics, when variables are often highly collinear. Instead of yielding a point estimate for the model parameters, it yields a probability distribution, hence incorporating the uncertainty surrounding the estimates. In the same spirit, Bayesian Model Averaging is becoming popular also in finance to account for model uncertainty.

4.1 Regularization

Regularization is a supervised learning strategy that overcomes this problem. It reduces the complexity of the model by *shrinking* the parameters toward some value. In practice, it penalizes more complex models in favor of more parsimonious ones. The Least Absolute Shrinkage and Selection Operator (LASSO, [58] and [64]), increasingly popular in economics and finance, uses L1 regularization. In practice, it limits the size of the regression coefficients by imposing a penalty equal to the absolute value of the magnitude of the coefficients. This implies shrinking the smallest coefficients to zero, which is removing some regressors altogether. For this reason, the LASSO is used as a variable selection method, allowing to identify key predictors from a pool of several candidate ones. A tuning parameter λ in the penalty function controls the level of shrinkage: for $\lambda = 0$ we have the OLS solution, while for increasing values of λ more and more coefficients are set to zero, thus yielding a sparse model.

Ridge regression involves L2 regularization, as it uses the squared magnitude of the coefficients as penalty term in the loss function. This type of regularization does not shrink parameters to zero. Also in this case, a crucial modeling choice relates to the value of the tuning parameter λ in the penalty function.

The Elastic Net has been proposed as an improvement over the LASSO [38], and combines the penalty from LASSO with that of the Ridge regression. The Elastic Net seems to be more efficient than the LASSO, while maintaining a similar sparsity of representation, in two cases. The first one is when the number of predictor variables is larger than the number of observations: in this case, the LASSO tends to select at most all the variables before it saturates. The second case is when there is a set of regressors whose pairwise correlations are high: in this case, the LASSO tends to select only one predictor at random from the group.¹³

The Adaptive LASSO [68] is an alternative model also proposed to improve over the LASSO, by allowing for different penalization factors of the regression coefficients. By doing so, the Adaptive LASSO addresses potential weaknesses of

¹³See [69].

the classical LASSO under some conditions, such as the tendency to select inactive predictors, or over-shrinking the coefficients associated with correct predictors.

4.2 *Bayesian Learning*

In Bayesian learning, shrinkage is defined in terms of a parameter's prior probability distribution, which reflects the modeler's beliefs.¹⁴ In the case of Bayesian linear regression, in particular, the prior probability distribution for the coefficients may reflect how certain one is about some coefficients being zero, i.e., about the associated regressors being unimportant. The posterior probability of a given parameter is derived based on both the prior and the information that is contained in the data. In practice, estimating a linear regression using a Bayesian approach involves the following steps:

1. Assume a prior probability distribution for the dependent variable, the coefficients, and the variance of the error term.
2. Specify the likelihood function, which is defined as the probability of the data given the parameters.
3. Derive the posterior distribution, which is proportional to the likelihood times the prior.
4. If the likelihood is such that the posterior cannot be derived analytically, use sampling techniques such as Markov Chain Monte Carlo (MCMC) method to generate a large sample (typically based on thousands of draws) from the posterior distribution.
5. The predicted value for the dependent variable, as well as the associated highest posterior density interval (e.g., at the 95% level), are derived based on the coefficients' posterior distribution.

By yielding probability distributions for the coefficients instead of point estimates, Bayesian linear regression accounts for the uncertainty around model estimates. In the same spirit, Bayesian Model Averaging (BMA, [46]) adds one layer by considering the uncertainty around the model specification. In practice, it assumes a prior distribution over the set of all considered models, reflecting the modeler's beliefs about each model's accuracy in describing the data. In the context of linear regression, model selection amounts to selecting subsets of regressors from the set of all candidate variables. Based on the posterior probability associated with each model, which takes observed data into account, one is able to select and combine the best models for prediction purposes. Stochastic search algorithms help

¹⁴The book by [34] covers Bayesian inference from first principles to advanced approaches, including regression models.

reduce the dimension of the model space when the number of candidate regressors is not small.¹⁵

Finally, some approaches have more recently been proposed which link the LASSO-based literature with the Bayesian stream. This avenue was pioneered by the Bayesian LASSO [53], which connects the Bayesian and LASSO approaches by interpreting the LASSO estimates as Bayesian estimates, based on a particular prior distribution for the regression coefficients. As a Bayesian method, the Bayesian LASSO yields interval estimates for the LASSO coefficients. The Bayesian adaptive LASSO (BaLASSO, [47]) generalizes this approach by allowing for different parameters in the prior distributions of the regression coefficients. The Elastic Net has also been generalized in a Bayesian setting [40], providing an efficient algorithm to handle correlated variables in high-dimensional sparse models.

5 Critical Discussion on Machine Learning as a Tool for Financial Stability Policy

As discussed in [5], standard approaches are usually unable to fully understand the risk dynamics within financial systems in which structural relationships interact in nonlinear and state-contingent ways. And indeed, traditional models assume that risk dynamics, e.g., those eventually leading to banking or sovereign crises, can be reduced to common data models in which data are generated by independent draws from predictor variables, parameters, and random noise. Under these circumstances, the conclusions we can draw from these models are “about the models mechanism, and not about natures mechanism” [17]. To put the point into perspective, let us consider the goal of realizing a risk stratification for financial crisis prediction using regression trees. Here the objective should be based on identifying a series of “red flags” for potential observable predictors that help to detect an impending financial crisis through a collection of binary rules of thumb such as the value of a given predictor being larger or lower than a given threshold for a given observation. In doing this, we can realize a pragmatic rating system that can capture situations of different risk magnitudes, from low to extreme risk, whenever the values of the selected variables lead to risky terminal nodes. And the way in which such a risk stratification is carried out is, by itself, a guarantee to get the best risk mapping in terms of most important variables, optimal number of risk clusters (final nodes), and corresponding risk predictions (final nodes’ predictions). In fact, since the estimation process of the regression tree, as all machine learning algorithms, is

¹⁵With M candidate regressors, the number of possible models is equal to 2^M .

based on cross-validation,¹⁶ the rating system is validated by construction, as the risk partitions are realized in terms of maximum predictability.

However, machine learning algorithms also have limitations. The major caveat sits on the same connotation of data-driven algorithms. In a sense, machine learning approaches limit their knowledge to the data they process, no matter how and why those data lead to specific models. In statistical language, the point relates to the question about the underlying data-generation process. In more depth, machine learning is expected to say little about the causal effect between y and x , but rather it is conceived with the end to predict y using and selecting x . The issue is extremely relevant when exploring the underlying structure of the relationship and trying to make inference about the inner nature of the specific economic process under study. A clear example of how this problem materializes is in [52]. These authors make a repeated house-value prediction exercise on subsets of a sample from the American Housing Survey, by randomly partitioning the sample, next re-estimating the LASSO predictor. In doing so, they document that a variable used in one partition may be unused in another while maintaining a good prediction quality (the R^2 remains roughly constant from partition to partition). A similar instability is also present in regression trees. Indeed, since these models are sequential in nature and locally optimal at each node split, the final tree may not guarantee a global optimal solution, with small changes in the input data translating into large changes in the estimation results (the final tree).

Because of these issues, machine learning tools should then be used carefully.¹⁷ To overcome the limitations of machine learning techniques, one promising avenue is to use them in combination with existing model-based and theory-driven approaches.¹⁸ For example, [60] focus on sovereign debt crises prediction and explanation proposing a procedure that mixes a pure algorithmic perspective, without making assumptions about the data generating process, with a parametric approach (see Sect. 6.1). This mixed approach allows to bypass the problem of reliability of the predictive model, thanks to the use of an advanced machine learning technique. At the same time, it allows to estimate a distance-to-default, via a standard probit regression. By doing so, the empirical analysis is contextualized within a theoretical-based process similar to the Merton-based distance-to-default.

¹⁶The data are partitioned into subsets such that the analysis is initially performed on a single subset (the training sets), while the other subset(s) are retained for subsequent use in confirming and validating the initial analysis (the validation or testing sets).

¹⁷See [6] on the use of big data for policy.

¹⁸See [7] for an overview of recently proposed methods at the intersection of machine learning and econometrics.

6 Literature Overview

This section provides an overview of a growing literature, which applies the models described in the previous section—or more sophisticated versions—for financial stability purposes. This literature has developed in the last decade, with more advanced techniques being applied in finance only in recent years. This is the so-called second generation of Early Warning Models (EWM), developed after the global financial crisis. While the first generation of EWM, popular in the 1990s, was based on rather simple approaches such as the signaling approach, the second generation of EWM implement machine learning techniques, including tree-based approaches and parametric multiple-regime models. In Sect. 6.1 we will review papers using decision trees, while Sect. 6.2 deals with financial stability applications of sparse models.

6.1 *Decision Trees for Financial Stability*

There are several success stories on the use of decision trees to address financial stability issues. Several papers propose EWM for banking crises. One of the first papers applying classification trees in this field is [22], where the authors use a binary classification tree to analyze banking crises in 50 emerging markets and developed economies. The tree they grow identifies the conditions under which a banking crisis becomes likely, which include high inflation, low bank profitability, and highly dollarized bank deposits together with nominal depreciation or low bank liquidity. The beauty of this tool stands in the ease of use of the model, which also provides specific threshold values for the key variables. Based on the proposed tree, policymakers only need to monitor whether the relevant variables exceed the warning thresholds in a particular country. [50] also aim at detecting vulnerabilities that could lead to banking crises, focusing on emerging markets. They apply the CRAGGING approach to test 540 candidate predictors and identify two banking crisis’ “danger zones”: the first occurs when high interest rates on bank deposits interact with credit booms and capital flights; the second occurs when an investment boom is financed by a large rise in banks’ net foreign exposure. In a recent working paper by [33], the author uses the same CRAGGING algorithm to identify vulnerabilities to systemic banking crises, based on a sample of 15 European Union countries. He finds that high credit aggregates and a low market risk perception are amongst the key predictors. [1] also develop an early warning system for systemic banking crises, which focuses on the identification of unsustainable credit developments. They consider 30 predictor variables for all EU countries and apply the Random Forest approach, showing that it outperforms competing logit models out-of-sample. [63] also apply the Random Forest to assess vulnerabilities in the banking sector, including bank-level financial statements as predictor variables. [14] compare a set of machine learning techniques, also including trees and the Random

Forest, to network- and regression-based approaches, showing that machine learning models mostly outperform the logistic regression in out-of-sample predictions and forecasting. The authors also offer a narrative for the predictions of machine learning models, based on the decomposition of the predicted crisis probability for each observation into a sum of contributions from each predictor. [67] implements BAGGING and the Random Forest to measure the risk of banking crises, using a long-run sample for 17 countries. He finds that tree ensembles yield a significantly better predictive performance compared to the logit. [20] use AdaBoost to identify the buildup of systemic banking crises, based on a dataset comprising 100 advanced and emerging economies. They also find that machine learning algorithms can have a better predictive performance than logit models. [13] is the only work, to our knowledge, finding an out-of-sample outperformance of conventional logit models over machine learning techniques, including decision trees and the Random Forest.

Also for sovereign crises several EWM have been developed based on tree ensemble techniques. The abundant literature on sovereign crises has documented the high complexity and the multidimensional nature of sovereign default, which often lead to predictive models characterized by irrelevant theory and poor/questionable conclusions. One of the first papers exploring machine learning methods in this literature is [49]. The authors compare the logit and the CART approach, concluding that the latter overperforms the logit with 89% of the crises correctly predicted; however, it issues more false alarms. [48] also use CART to investigate the roots of sovereign debt crises, finding that they differ depending on whether the country faces public debt sustainability issues, illiquidity, or various macroeconomic risks. [60] propose a procedure that mixes the CRAGGING and the probit approach. In particular, in the first step CRAGGING is used to detect the most important risk indicators with the corresponding threshold, while in a second step a simple pooled probit is used to parametrize the distances to the thresholds identified in the first step (so-called “Multidimensional Distance to Collapse Point”). [61] again use CRAGGING, to predict sovereign crises based on a sample of emerging markets together with Greece, Ireland, Portugal, and Spain. They show that this approach outperforms competing models, including the logit, while balancing in-sample goodness of fit and out-of-sample predictive performance. More recently, [5] use a recursive partitioning strategy to detect specific European sovereign risk zones, based on key predictors, including macroeconomic fundamentals and a contagion measure, and relevant thresholds.

Finally, decision trees have been used also for the prediction of currency crises. [36] first apply this methodology on a sample of 42 countries, covering 52 currency crises. Based on the binary classification tree they grow on this data, they identify two different sets of key predictors for advanced and emerging economies, respectively. The root node, associated with an index measuring the quality of public sector governance, essentially splits the sample into these two subsamples. [28] implement a set of methodological approaches, including regression trees, in their empirical investigation of macroeconomic crises in emerging markets. This approach allows each regressor to have a different effect on the dependent variable for different ranges of values, identified by the tree splits, and is thus able to capture

nonlinear relationships and interactions. The regression tree analysis identifies three variables, namely, the ratio of external debt to GDP, the ratio of short-term external debt to reserve, and inflation, as the key predictors. [42] uses regression tree analysis to classify 96 currency crises in 20 countries, capturing the stylized characteristics of different types of crises. Finally, a recent paper using CART and the Random Forest to predict currency crises and banking crises is [41]. The authors identify the key predictors for each type of crisis, both in the short and in the long run, based on a sample of 36 industrialized economies, and show that different crises have different causes.

6.2 *Sparse Models for Financial Stability*

LASSO and Bayesian methods have so far been used in finance mostly for portfolio optimization. A vast literature starting with [8] uses a Bayesian approach to address the adverse effect due to the accumulation of estimation errors. The use of LASSO-based approaches to regularize the optimization problem, allowing for the stable construction of sparse portfolios, is far more recent (see, e.g., [19] and [24], among others).

Looking at financial stability applications of Bayesian techniques, [23] develop an early warning system where the dependent variable is an index of financial stress. They apply Bayesian Model Averaging to 30 candidate predictors, notably twice as many as those generally considered in the literature, and select the important ones by checking which predictors have the highest probability to be included in the most probable models. More recently, [55] investigate the determinants of the 2008 global financial crisis using a Bayesian hierarchical formulation that allows for the joint treatment of group and variable selection. Interestingly, the authors argue that the established results in the literature may be due to the use of different priors. [65] and [37] use Bayesian estimation to estimate the effects of the US subprime mortgage crisis. The first paper uses Bayesian panel data analysis for exploring its impact on the US stock market, while the latter uses time-varying Bayesian Vector AutoRegressions to estimate cross-asset contagion in the US financial market, using the subprime crisis as an exogenous shock.

Turning to the LASSO, not many authors have yet used this approach to predict financial crises. [45] use a logistic LASSO in combination with cross-validation to set the λ penalty parameter, and test their model in a real-time recursive out-of-sample exercise based on bank-level and macrofinancial data. The LASSO yields a parsimonious optimal early-warning model which contains the key risk-driver indicators and has good in-sample and out-of-sample signaling properties. More recently, [2] apply the LASSO in the context of sovereign crises prediction. In particular, they use it to identify the macro indicators that are relevant in explaining the cross-section of sovereign Credit Default Swaps (CDS) spreads in a recursive setting, thereby distilling time-varying market sensitivities to specific economic fundamentals. Based on these estimated sensitivities, the authors identify distinct

crisis regimes characterized by different dynamics. Finally, [39] conduct a horse race of conventional statistical methods and more recent machine learning methods, including a logit LASSO as well as classification trees and the Random Forest, as early-warning models. Out of a dozen competing approaches, tree-based algorithms place in the middle of the ranking, just above the naive Bayes approach and the LASSO, which in turn does better than the standard logit. However, when using a different performance metric, the naive Bayes and logit outperform classification trees, and the standard logit slightly outperforms the logit LASSO.

6.3 *Unsupervised Learning for Financial Stability*

Networks have been extensively applied in financial stability. This stream of literature is based on the notion that the financial system is ultimately a complex system, whose characteristics determining its resilience, robustness, and stability can be studied by means of traditional network approaches (see [12] for a discussion). In particular, network models have been successfully used to model contagion (see the seminal work by [3], as well as [35] for a review of the literature on contagion in financial networks)¹⁹ and measure systemic risk (see, e.g., [11]). The literature applying network theory started to grow exponentially in the aftermath of the global financial crisis. DebtRank [10], e.g., is one of the first approaches put forward to identify systemically important nodes in a financial network. This work contributed to the debate on too-big-to-fail financial institutions in the USA by emphasizing that too-central-to-fail institutions deserve at least as much attention.²⁰ [51] explore the properties of the global banking network by modelling 184 countries as nodes of the network, linked through cross-border lending flows, using data over the 1978–2009 period. By today, countless papers use increasingly complex network approaches to make sense of the structure of the financial system. The tools they offer aim at enabling policymakers to monitor the evolution of the financial system and detect vulnerabilities, before a trigger event precipitates the whole system into a crisis state. Among the most recent ones, one may cite, e.g., [62], who study the type of systemic risk arising in a situation where it is impossible to decide which banks are in default.

Turning to artificial neural networks, while supervised ones have been used in a few works as early warning models for financial crises ([26] on sovereign debt crises, [27] and [54] on currency crises), unsupervised ones are even less common in the financial stability literature. In fact, we are only aware of one work, [59], using self-organizing maps. In particular, the authors develop a Self-Organizing Financial Stability Map where countries can be located based on whether they are

¹⁹Amini et al. [4], among others, also use financial networks to study contagion.

²⁰On the issue of centrality, see also [44] who built a network based on co-movements in Credit Default Swaps (CDS) of major US and European banks.

in a pre-crisis, crisis, post-crisis, or tranquil state. They also show that this tool performs better than or equally well as a logit model in classifying in-sample data and predicting the global financial crisis out-of-sample.

7 Conclusions

Forecasting financial crises is essential to provide warnings to be used in preventing impending abnormalities, and taking action with a sufficient lead time to implement adequate policy measures. The global financial crisis that started with the Lehman collapse in 2008 and the subsequent Eurozone sovereign debt crisis over the years 2010–2013 have both profoundly changed economic thinking around machine learning. The ability to discover complex and nonlinear relationships, not fully biased by a priori theory/beliefs, has contributed to getting rid of the skepticism around machine learning. Ample evidence proved indeed the inconsistency of traditional models in predicting financial crisis, and the need to explore data-driven approaches. However, we should be aware about what machine learning can and cannot do, and how to handle these algorithms alone and/or in conjunction with common traditional approaches to make financial crisis predictions more statistically robust and theoretically consistent. Also, it would be important to work on improving the interpretability of the models, as there is a strong need to understand how decisions on financial stability issues are being made.

References

1. Alessi, L., & Detken, C. (2018). Identifying excessive credit growth and leverage. *Journal of Financial Stability*, 35, 215–225.
2. Alessi, L., Balduzzi, P., & Savona, R. (2019). *Anatomy of a Sovereign Debt Crisis: CDS Spreads and Real-Time Macroeconomic Data*. Brussels: European Commission—Joint Research Centre. Working Paper No. 2019-03.
3. Allen, F., & Gale, D. (2000). Financial contagion. *Journal of Political Economy* 108(1), 1–33.
4. Amini, H., Cont, R., & Minca, A. (2016). Resilience to contagion in financial networks. *Mathematical Finance*, 26, 329–365.
5. Arakelian, V., Dellaportas, P., Savona, R., & Vezzoli, M. (2019). Sovereign risk zones in Europe during and after the debt crisis. *Quantitative Finance*, 19(6), 961–980.
6. Athey, S. (2017). Beyond prediction: Using big data for policy problems. *Science*, 355(6324), 483–485.
7. Athey, S., & Imbens, G. W. (2019). Machine learning methods economists should know about. *Annual Review of Economics*, 11(1), 685–725.
8. Barry, C. B. (1974). Portfolio analysis under uncertain means, variances, and covariances. *Journal of Finance*, 29, 515–22.
9. Basel Committee on Banking Supervision (2018). *Implications of fintech developments for banks and bank supervisors, consultative document*. Accessed 12 March 2021. <https://www.bis.org/bcbs/publ/d415.htm>.

10. Battiston, S., Puliga, M., Kaushik, R., Tasca, P., & Caldarelli, G. (2012). DebtRank: Too Central to Fail? Financial Networks, the FED and Systemic Risk. *Scientific Reports*, 2, 541–541.
11. Battiston, S., Caldarelli, G., May, R. M., Roukny, T., & Stiglitz, J. E. (2016). The price of complexity in financial networks. *Proceedings of the National Academy of Sciences*, 113(36), 10031–10036.
12. Battiston, S., Farmer, J. D., Flache, A., Garlaschelli, D., Haldane, A. G., Heesterbeek, H., Hommes, C., Jaeger, C., May, R., Scheffer, M. (2016). Complexity theory and financial regulation. *Science*, 351(6275), 818–819.
13. Beutel, J., List, S., & von Schweinitz, G. (2018). *An evaluation of early warning models for systemic banking crises: Does machine learning improve predictions?*. Deutsche Bundesbank Discussion Paper Series, No. 48.
14. Bluwstein, K., Buckmann, M., Joseph, A., Kang, M., Kapadia, S., & Simsek, O. (2020). *Credit growth, the yield curve and financial crisis prediction: Evidence from a machine learning approach*. Bank of England Staff Working Paper No. 848.
15. Breiman, L. (1996). Bagging predictors. *Machine Learning*, 24, 123–140.
16. Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5–32.
17. Breiman, L. (2001). Statistical Modelling: The Two Cultures. *Statistical Science*, 16(3), 199–215.
18. Breiman, L., Friedman, J., Olshen, R., & Stone, C. (1984). *Classification and regression trees*. Monterey, CA: Wadsworth and Brooks.
19. Brodie, J., Daubechies, I., De Mol, C., Giannone, D., & Loris, I. (2009). Sparse and stable Markowitz portfolios, *Proceedings of the National Academy of Sciences*, 106, 12267–12272.
20. Casabianca, E. J., Catalano, M., Forni, L., Giarda, E., & Passeri, S. (2019). An early warning system for banking crises: From regression-based analysis to machine learning techniques. In *Marco Fanno Working Papers 235, Dipartimento di Scienze Economiche “Marco Fanno”*.
21. Clements, M., & Hendry, D. (1998). *Forecasting economic time series*. Cambridge: Cambridge University.
22. Duttagupta, R., & Cashin, P. (2011). Anatomy of banking crises in developing and emerging market countries. *Journal of International Money and Finance*, 30(2), 354–376.
23. Eidenberger, J., Sigmund, M., Neudorfer, B., & Stein, I. (2014). *What Predicts Financial (In)Stability? A Bayesian Approach*. Bundesbank Discussion Paper No. 36/2014.
24. Fan, J., Zhang, J., & Yu, K. (2012). Vast portfolio selection with gross-exposure constraints. *Journal of the American Statistical Association*, 107, 592–606.
25. Financial Stability Board (2017). *Artificial intelligence and machine learning in financial services—Market developments and financial stability implications*. FSB Report, P011117. <https://www.fsb.org/wp-content/uploads/P011117.pdf>.
26. Fioramanti, M. (2008). Predicting sovereign debt crises using artificial neural networks: A comparative approach. *Journal of Financial Stability*, 4(2), 149–164.
27. Franck, R., & Schmied, A. (2003). *Predicting currency crisis contagion from East Asia to Russia and Brazil: An artificial neural network approach*. AMCB Working Paper No 2/2003, Aharon Meir Center for Banking.
28. Frankel, J., & Wei, S. J. (2004). *Managing macroeconomic crises: Policy lessons*. NBER Working Paper 10907.
29. Freund, Y., & Schapire, R. E. (1996). Experiments with a new boosting algorithm. *Proceedings of the International Conference on Machine Learning (ICML 96)*, 96, 148–156.
30. Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29(5), 1189–1232.
31. Friedman, J. H. (2002). Stochastic gradient boosting. *Computational Statistics & Data Analysis*, 38(4), 367–378.
32. Friedman, J. H., & Meulman, J. J. (2003). Multiple additive regression trees with application in epidemiology. *Statistics in Medicine*, 22(9), 1365–1381.
33. Gabriele, C. (2019). *Learning from trees: A mixed approach to building early warning systems for systemic banking crises*. European Stability Mechanism Working Paper No. 40/2019.

34. Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian data analysis*. New York: CRC Press.
35. Glasserman, P., & Peyton Young, H. (2016). Contagion in financial networks. *Journal of Economic Literature*, 54(3), 779–831.
36. Gosh, S., & Gosh, A. (2002). *Structural vulnerabilities and currency crises*. IMF Working Paper 02/9.
37. Guidolin, M., Hansen, E., & Pedio, M. (2019). Cross-asset contagion in the financial crisis: A Bayesian time-varying parameter approach. *Journal of Financial Markets*, 45, 83–114.
38. Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning*, 2 edn. New York: Springer.
39. Holopainen, M., & Sarlin, P. (2017). Toward robust early-warning models: A horse race, ensembles and model uncertainty. *Quantitative Finance*, 17(12), 1933–1963.
40. Huang, A., Xu, S., & Cai, X. (2015). Empirical Bayesian elastic net for multiple quantitative trait locus mapping. *Heredity*, 114, 107–115.
41. Joy, M., Rusnk, M., Smdkova, K., & Vacek, B. (2017). Banking and currency crises: Differential diagnostics for developed countries. *International Journal of Finance & Economics*, 22(1), 44–67.
42. Kaminsky, G. L. (2006). Currency crises: Are they all the same. *Journal of International Money and Finance*, 25, 503–527.
43. Kass, G. V. (1980). An exploratory technique for investigating large quantities of categorical data. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 29(2), 119–127.
44. Kaushik, R., & Battiston, S. (2013). Credit default swaps drawup networks: Too interconnected to be stable? *PLoS ONE*, 8(7), e61815.
45. Lang, J. H., Peltonen, T. A., & Sarlin, P. (2018). *A framework for early-warning modeling with an application to banks*. Working Paper Series 2182, European Central Bank.
46. Leamer, E. E. (1978). *Specification searches*. New York: Wiley.
47. Leng, C., Nott, D., & Minh-Ngoc, T. (2014). Bayesian adaptive lasso. *Annals of the Institute of Statistical Mathematics*, 66(2), 221–244.
48. Manasse, P., & Roubini, N. (2009). Rules of thumb for Sovereign debt crises. *Journal of International Economics*, 78, 192–205.
49. Manasse, P., Roubini, N., & Schimmelpfennig, A. (2003). *Predicting Sovereign debt crises*. IMF Working Paper WP 03/221.
50. Manasse, P., Savona, R., & Vezzoli, M. (2016). Danger zones for banking crises in emerging markets. *International Journal of Finance & Economics*, 21, 360–381.
51. Minoiu, C., & Reyes, J. (2011). *A Network Analysis of Global Banking:1978–2009*. IMF Working Papers. 11.
52. Mullainathan, S., & Spiess, J. (2017). Machine learning: An applied econometric approach. *Journal of Economic Perspectives*, 31(2), 87–106.
53. Park, T., & Casella, G. (2008). The Bayesian Lasso. *Journal of the American Statistical Association*, 103(482), 681–686.
54. Peltonen, T. (2006). *Are emerging market currency crises predictable? A test*. No 571, Working Paper Series, European Central Bank.
55. Ray-Bing, C., Kuo-Jung, L., Yi-Chi, C., & Chi-Hsiang, C. (2017). On the determinants of the 2008 financial crisis: A Bayesian approach to the selection of groups and variables. *Studies in Nonlinear Dynamics & Econometrics*, 21(5), 17.
56. Ren, Y., Zhang, L., & Suganthan, P. N. (2016). Ensemble classification and regression-recent developments, applications and future directions. *IEEE Computational Intelligence Magazine*, 11(1), 41–53.
57. Rojas, R. (1996). *Neural networks a systematic introduction* (p. 101). Berlin: Springer.
58. Santosa, F., & Symes, W. W. (1986). Linear inversion of band-limited reflection seismograms. *SIAM Journal on Scientific and Statistical Computing*, 7(4), 1307–1330.
59. Sarlin, P., & Peltonen, T. (2011). Mapping the state of financial stability. *Journal of International Financial Markets Institutions and Money*, 26, 46–76.

60. Savona, R., & Vezzoli, M. (2012). Multidimensional Distance-To-Collapse Point and sovereign default prediction. *Intelligent Systems in Accounting, Finance and Management*, 19(4), 205–228.
61. Savona, R., & Vezzoli, M. (2015). Fitting and forecasting sovereign defaults using multiple risk signals. *Oxford Bulletin of Economics and Statistics*, 77(1), 66–92.
62. Schuldenzucker, S., Seuken, S., & Battiston, S. (2020). Default ambiguity: Credit default swaps create new systemic risks in financial networks. *Management Science*, 66(5), 1981–1998.
63. Tanaka, K., Kinkyo, T., & Hamori, S. (2016). Random forests-based early warning system for bank failures. *Economics Letters*, 148, 118–121.
64. Tibshirani, R. (1996). Regression Shrinkage and Selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1), 267–288.
65. Tsay, R., & Ando, T. (2012). Bayesian panel data analysis for exploring the impact of subprime financial crisis on the US stock market. *Computational Statistics & Data Analysis*, 56(11), 3345–3365.
66. Vezzoli, M., & Stone, C. J. (2007). CRAGGING. In *Book of short papers of CLADAG 2007, EUM*, pp. 363–366.
67. Ward, F. (2017). Spotting the danger zone: Forecasting financial crises with classification tree ensembles and many predictors. *Journal of Applied Econometrics*, 32(2), 359–378.
68. Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101(476), 1418–1429.
69. Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 67(2), 301–320.

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

