5

Stress-testing applications of Machine Learning Models

Jorge A. Chan-Lau

International Monetary Fund

This chapter will discuss stress-testing applications of machine-learning methods against the background of forecasting accuracy, interpretability of results, and the ability to capture the adaptive behaviour of firms and households facing structural breaks in the economic and business environment in which they operate. Adequate forecasting accuracy in stress tests can be difficult to accomplish given our imperfect knowledge about macro-financial linkages and their impact on financial firms' profitability, liquidity and soundness. The chapter will therefore argue that machine learning offers a viable option for improving forecast accuracy due to the models' ability to capture nonlinear effects between the scenario variables and the risk factors driving the soundness of a financial firm. Stress-testing applications are reviewed, highlighting the advantages of machine-learning models over more standard econometric-based stress-test models.

The traditional methods used for stress testing originated in statistics. In the words of Larry Wasserman of Carnegie Mellon: "Statistics emphasises formal statistical inference (confidence intervals, hypothesis tests, optimal estimators) in low dimensional problems. Machine learning emphasises high dimensional prediction problems." The accuracy gains of machine-learning models come at the expense of model interpretability as decision-makers perceive machine-learning models to be mysterious black boxes. This perception, unjustified in some cases, could impair the models' effectiveness for guiding strategic decisions at the firm level, and policy actions and recommendations at regulatory agencies. The chapter will briefly explain why such prejudice is not justified for simpler machine-learning models, and how advances in interpretable machine learning could facilitate communicating results to senior decision-makers, and contribute to their broader and faster adoption.

To complicate stress-testing practice further, the dynamic nature of financial linkages suggests that, to paraphrase Tolstoy's heroine Anna Karenina, every unhappy financial stress episode is unhappy in its own way. Machine-learning models' ability to forecast events accurately is much impaired by the presence of structural breaks not experienced in the past, a deficiency shared with statistical and econometric models. One driver of these breaks is that firms and households react to changes in their operating environment through adaptation and innovation. The chapter will conclude by providing new directions towards incorporating adaptive behaviour that, when combined with machine-learning methods, could greatly improve stress-testing practices.

THE NEED OF FORECASTING ACCURACY IN STRESS TESTS

Forecasting accuracy is paramount in stress tests, whether they be conducted internally by the firm or externally by the authorities to satisfy supervisory and macroprudential requirements. The numerical outcomes of internal tests guide and influence a firm's strategic decisions related to resource and capital allocation, and serve to ensure the overall strategy conforms to the firm's risk appetite. Strategic decisions need to be firmly grounded on the best numerical estimate from the tests. Were the results too conservative, the firm would be missing substantial profit opportunities; too optimistic, and the firm would be facing risks it has not foreseen adequately; too unbalanced in their risk identification across sectors, and the firm's capital allocation would be uneven, overcapitalised in certain business lines and undercapitalised in others.

The most common stress tests rely on the evaluation of the impact of large shocks on the response to single risk factors responsible for, among other items, the performance of loan and trading portfolios as well as the liquidity and funding conditions facing the firm. The likelihood of the occurrence of the shock is roughly determined and framed within a scenario narrative using expert judgement.[1] Comprehensive macro-financial stress tests, which have become the *de facto* tool for assessing the vulnerability of a financial firm in the aftermath of the global financial crisis, specify multi-year stress scenarios. When calibrated, these scenarios may require the specification of numerical values for a large number of economic and financial variables.

The narrative has a qualitative nature, and in many instances the likelihood of the scenario is categorised into broad categories such as “high”, “medium” and “low”. These broad categories do not necessarily have a numerical probability assigned to them. For example, Table 5.1 shows some of the downside risk scenarios the International Monetary Fund (IMF) used in its analysis of the Chinese and US economy under its 2018 Article IV Consultations; and one risk considered in the adverse scenario formulated by the European Systemic Risk Board for the 2018 EU-wide stress test.

For the stress-test team, articulating these scenarios into shocks to the risk factors could be difficult. The first task is to translate these scenarios into quantitative shocks to certain key variables, such as real GDP growth, short- and long-term interest rates, equity

Table 5.1 IMF: Selected adverse scenarios in China and the US

Likelihood	**Scenarios**
China, medium likelihood	Uncoordinated tightening: In the near term, uncoordinated financial and local government regulatory action could have unintended consequences that trigger disorderly repricing of corporate and local government financing vehicle credit risks. Reduced liquidity and higher rollover risk could lead to an excessively sharp acceleration of defaults, and a sharp tightening of local government spending. High impact.
US, medium likelihood	Significant economic slowdown: As the current recovery ages and vulnerabilities build up, the risks of a sharper-than-expected slowdown increase. The proximate causes could be a fiscal contraction associated with the eventual planned withdrawal of the tax stimulus or market fears of overheating.
EU (no likelihood specified)	Adverse feedback loop between weak bank profitability and low nominal growth amid structural challenges in the EU banking sector, leading to the following financial and economic shocks: investment and consumption demand shocks in EU countries; residential and commercial property price shocks in EU countries; and EU-wide uniform shock to interbank money market rates due to higher credit risk of the banking sector.

Source: IMF (2018a, 2018b); ESRB (2018).

prices and exchange rates.[2] Once the numerical parameters of the scenarios are set, the second task is to translate them into shocks to the relevant risk factors. This is not an easy task. By way of illustration, a bank examining mortgage defaults and valuations under the stress-test scenario conditions needs to predict risk factors that are borrower-specific, such as income, debt-to-income, creditworthiness and employment status; and loan-specific, such as market value in geographic locations, changes in loan-to-value ratios and potential refinancing rates.

The problem stress testers face when translating scenarios into shocks can be decomposed into two connected problems: feature selection (ie, choosing which scenario variables are useful as predictors of the risk factors) and model selection (ie, choosing the best predictive model, conditional on the scenarios). Financial supervisory agencies and systemic risk oversight bodies mandate regulated firms to conduct periodic stress tests. The outcome of these regulatory stress tests, if short from specified minimum thresholds, may force remedial actions on the failing firms. In the US, for instance, firms that fail to meet the targets of the annual Comprehensive Capital Analysis and Review (CCAR) need to modify their planned capital actions and improve their internal risk management processes.[3]

Further complicating a regulatory stress-test exercise is the inherent tension between the internal analysis of the regulated firms and the tests conducted by the supervisory authority on its own. The internal and external outcomes may diverge due to the use of different modelling methodologies applied to dissimilar datasets in terms of time coverage and granularity, and different assumptions on how the scenario variables interact with the relevant risk factors for each business line of the firm.[4] The tension can be attributed to the problem of model selection given our still imperfect knowledge of the interaction between the real economy and the financial system through the multiple feedback loops linking them together, the so-called macro-financial linkages (see Claessens and Kose, 2018). Traditional empirical methods long used by economists and financial practitioners could face severe limitations when attempting to uncover the highly nonlinear relationship between variables bearing on the soundness of a financial firm, such as funding access and the default risk of loans, and changes in the economic and business environment due to adverse macro-financial shocks.

It is argued below that machine-learning models could be very useful in the feature selection and model selection stages of a stress-testing exercise. Typically, the estimation of a machine-learning model simultaneously addresses the feature selection and model selection stages. Since these models have been developed with a view towards efficient computer implementation, they could be integrated into automated real-time stress testing and risk analysis platforms.[5]

FEATURE SELECTION AND MODEL SELECTION CHALLENGES IN STRESS TESTS

The large number of primary variables specified in stress scenarios poses a challenge for the design, selection and estimation of forecasting models. In the case of the US CCAR, restricting models to use only contemporaneous values of the variables and ruling out interactions lead to 228 possible linear models for explaining a single dependent variable – say, the default rate of a specific asset class. Model selection intractability worsens as the number of potential models grows linearly with the number of variables to forecast. Moreover, in the context of macro stress tests, the dimensionality of the data may exceed the length of the sample size, raising concerns about model overfitting. In these tests, data is typically available at an annual frequency and may be available only for a few years. In this case, least squares cannot yield unique coefficient estimates and some method is necessary to reduce the number of covariates included in the model.

Given the large set of potential covariates, multicollinearity is likely to pose problems. This situation justifies selecting a reduced subset of variables for forecasting purposes. Arguably, expert judgement could help to identify the covariates relevant for the forecasting exercise, facilitating the model selection process. However, in the context of a relatively complex firm (eg, a commercial bank), expert judgement may not compensate for the lack of specialised and detailed knowledge on the firm's operations and exposures, including on its main counterparties. Expert judgement, especially if exercised by outsiders to the firm, may create blind spots and foster an unjustified sense of security.

Dimension reduction techniques offer a formal approach for dealing with the curse of dimensionality. Commonly used techniques are factor analysis and principal component analysis, which construct factors and components as linear combinations of the covariates. Typically, only a reduced number of factors and components suffice to explain the variability of the data. For instance, three principal components are enough to explain the term structure of government yield curves. While useful, there is an important caveat with dimension reduction techniques: it is somewhat difficult to associate an economic meaning to a factor or a principal component. Hence, it may be difficult to specify the path of a factor under a given scenario or understand what a negative shock to the factor is.

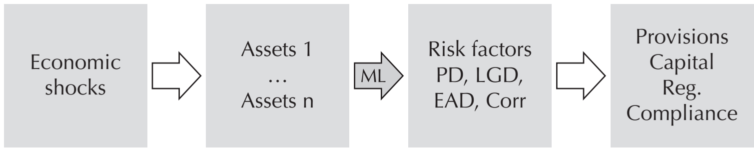
Expert judgement and dimension reduction techniques lead to a lower number of potential covariates ahead of the model selection stage. This may not be desirable from an economic perspective. Financial and non-financial corporations are becoming increasingly global, and domestic and international factors affect their operations, profitability and solvency. Reducing the number of covariates prior to selecting the forecasting model may not reveal what the drivers of a firm's performance are. Arguably, it may be preferable to address the curse of dimensionality at the model selection level rather than at the covariate level. Machine-learning methods excel at this task.[6]

MACHINE-LEARNING METHODS[7]

Machine learning encompasses a number of techniques for identifying patterns and relationships in the data, making it suitable for forecasting and simplifying the model selection process.[8] In the realm of machine learning, forecasting models falls under the category of supervised learning, in which we are interested in finding a rule (eg, an econometric model) that maps the set of covariates or inputs into an output (eg, the ratio of non-performing loans in an asset category). Linear regression models are just one subset of supervised learning models, and are not discussed here since they are already familiar to economists and analysts involved in stress tests, and are covered exhaustively in econometric textbooks (ie, Greene, 2017).

Different machine-learning methods are available for

Figure 5.1 Solvency stress test



forecasting, with increased flexibility offset by increased difficulty to interpret the results. More flexible methods are better at capturing patterns in the data, but simpler, easier to interpret methods are better suited for understanding and communicating results. Among the methods, in decreasing order of interpretability, we have subset selection, lasso regressions, least squares, generalised additive models, tree-based methods, support vector machines, methods combining different base learning methods such as bagging and boosting, and deep learning.

To assess the role forecasting plays in stress tests, let's focus on a simplified solvency stress test as depicted in Figure 5.1. The stress test starts with the specification of economic shocks that impact the credit quality of the assets held by the firm, and which ultimately lead to a re-evaluation of the provisions and capital the firm needs to hold.

A quantitative assessment of the deterioration in the assets' creditworthiness requires evaluating the impact of the shocks on the risk factors determining the loss distribution of the credit portfolio: the probability of default (PD), the loss given default (LGD), the exposure at default (EAD) and the default correlation. Once the risk factors are determined, it is straightforward to calculate the necessary provisions and capital to meet both internal and regulatory requirements. Machine-learning methods, which excel as prediction tools, are the building block that maps the scenario shocks onto the risk parameters. In an automated or semi-automated stress-test system, there would be a scenario generation model. Once the scenarios are agreed upon, a machine-learning model would gather data on the firm's exposures and create the prediction models for the relevant risk factors. A final module would take the risk factors input to calculate losses and determine whether the buffers are adequate.

Machine-learning methods do not require the modeller to specify what predictors (ie, covariates or variables) the machine-learning model should include. The models use search algorithms to find the combination of variables that yield the best predictive model using cross-validation errors to discriminate against competing models. Cross-validation requires the sample to be randomly partitioned into k groups, referred to as folds. The machine-learning algorithm fits the model using k-1 folds (ie, the training sample) and reserves the remaining fold (ie, the test sample) to calculate the mean squared error (MSE). The process is repeated exactly k times, and the model error is set equal to the average MSE (or other criterion set by the modeller) obtained when the k folds serve as the testing sample.

To illustrate conceptually the search methods of machine-learning algorithms, let's focus on the best subset selection algorithm. It starts with a null model containing no predictors. For a given number of predictors, n, it fits all possible model combinations containing exactly n predictors and picks the best one among them. Afterwards, it picks the single best model among all best n predictor models based on their cross-validated prediction error, information criteria or adjusted R-squared. Best subset selection, hence, reduces the number of predictors used in the equation.

Another machine-learning method that is being increasingly used in economic and finance applications is the lasso (least absolute shrinkage and selection operator) method introduced by Tibshirani (1996).[9] The lasso estimates a linear regression but restricts the number of predictors included in the regression by adding a penalty proportional to the absolute value of the coefficients of the predictors. The penalty forces a substantial number of coefficients to shrink to zero, especially if the predictors are correlated. The lasso method tries different values for the penalty weight, and yields the linear regression with the lower cross-validated MSE. Practitioners prefer to not use the minimum weight model but rather one that does not exceed the minimum error by more than one standard deviation and uses a lower number of predictors.

More complex models can capture nonlinearities better at the expense of interpretability. One exception is the simple regression tree. The algorithm requires dividing the predictor space into a set of possible values, and to assign any observation falling into the region the average predicted value of the training observation that fell in the same region. The tree is relatively simple to understand, as in each node there is a threshold value and one needs to compare whether the value of a predictor is either lower than, or greater than, or equal to the threshold. While easy to explain, a simple regression tree seldom yields good predictions and it is necessary to combine several of them through different methods, including random forests, boosting and bagging. This chapter does not discuss these and more complex models in detail, but refers the reader to the references cited initially in this section. The discussion now turns to a non-exhaustive review of machine-learning applications relevant for stress testing.

SOME MACHINE-LEARNING APPLICATIONS RELEVANT FOR STRESS TESTING

This section will review academic work on machine-learning applications that could serve to enhance stress-testing practices. The applications are listed in increasing order of the mathematical sophistication and complexity of the model used. An early application of machine learning in default prediction, which is an essential element in stress testing, is Perdeiy (2009). The author used stepwise selection and lasso methods, and found that including non-traditional financial ratios could improve the accuracy ratio of bankruptcy prediction models for US firms. The study looked at 100 classical accounting and market ratios, and enriched the predictor set by including trends of the traditional ratios and industry averages of the ratios, as well as technical accounting ratios. The lasso model outperformed stepwise selection, albeit by only a small margin.

Chan-Lau (2017b) uses the lasso and the relaxed lasso to model the probabilities of default in 10 different industrial sectors in South Africa, which are then fed into a stress-testing model.[10] The sectors included are basic materials, communications, consumer cyclicals, consumer non-cyclicals, diversified industries, energy, financials, industrials, technology and utilities.[11] In this application, the number of potential covariates, p, was equal to 90 excluding the intercept, and the number of observations, n, was at most 100, a situation well suited for lasso estimation. With p approximately equal to *n*, ordinary least squares estimates would exhibit high sensitivity to outliers. Pre-selecting the variables and choosing the number of lags was not feasible as it would have required a detailed knowledge of the drivers of default in different sectors.[12]

Kapinos and Mitnik (2015) used the lasso variable selection method in combination with principal component analysis to identify the macroeconomic drivers of banking variables, namely pre-provisions net revenue and net charge-offs on all loans and leases. The lasso selected the predictors more relevant for explaining the banking variables, and served to identify the balance-sheet and income statement factors responsible for the heterogeneous response among banks to macroeconomic shocks. Afterwards, the principal components associated with those predictors served as the explanatory variables in stress test models.

Kupiec (2017) examined the performance of alternative stress-test modelling approaches using as stress scenario the 2008 financial crisis, and evaluates the forecasting performance of these approaches by comparing their forecasts against actual bank performance. The response variable used was quarterly bank income before tax and extraordinary items, an important determinant of a bank's capital adequacy. The models used different level of disaggregation and a different set of predictors, and were calibrated using two possible methods. The first was the standard model selection criterion used in econometrics and based on adjusted R-squared statistic. This calibration method mimicked the Federal Reserve Capital and Loss Assessment under Stress Scenarios (CLASS) model approach (Hirtle *et al*, 2015). The second method was the lasso selection algorithm. Kupiec found that the traditional econometric approach produced inaccurate forecasts despite its good in-sample performance. Lasso-based models improved substantially over them. The findings suggest caution is warranted when supervisory model outcomes differ from those delivered by the firms' internal models.

Jacobs (2018) proposed using the multivariate adaptive regression splines (MARS) model to stress test credit risk. Using Federal Reserve filing data and macroeconomic data released by the regulatory authorities in the context of the CCAR exercise,[13] it was found that a MARS model outperformed a traditional vector autoregression (VAR) based econometric model in forecasting losses in three loan portfolio categories: commercial and industrial; commercial real estate; and consumer credit.[14] In the last two portfolios, the MARS outperformed the VAR model by a wide margin.

Barboza *et al* (2017) examined the performance of several machine-learning models, including support vector machines, bagging, boosting and random forests, to predict bankruptcy one year ahead of the event for North American firms during the period 1985–2013. They compared the performance of these models against discriminant analysis, logistic regression and neural networks. Relative to the latter set of models, the machine-learning models recorded a 10% gain in accuracy. For example, random forest had an accuracy of 87% compared with that of logistic regression (69%) and discriminant analysis (50%).

Mortgage loan defaults, which were one of the root causes of the global financial crisis, are difficult to model since their performance depends on many variables and their complex interaction with economic and financial variables. Practitioners often relied on logistic regression to model mortgage defaults. Bagherpour (2018) applied several machine-learning algorithms to a large dataset comprising over 20 million quarterly loan observations from Fannie Mae and Freddie Mac covering the period 2001–16. The machine-learning algorithms analysed were K-nearest neighbours, random forest, support vector machines and factorisation machines. All the models outperformed the logistic regression model, albeit only by a relatively small margin. From an interpretability perspective, the logistic regression may be a better option since it is easier to communicate the results.

Deep-learning methods are also rapidly making their way into the analytical toolkit available to stress testers. Sirignano *et al* (2018) developed a deep-learning model of multi-period mortgage risk trained using a dataset of origination and monthly performance records for over 120 million mortgages originated across the US between 1995 and 2014. The model uncovered the nonlinear influence on borrowers' behaviour of a large set of predictors, including macroeconomic variables and local economic conditions, and the interaction between the predictors. The authors suggested using these methods to rate and hedge mortgage-backed securities portfolios.

THE INTERPRETABILITY OF MACHINE-LEARNING OUTCOMES

The bias–variance trade-off formalises the tension between interpretability and flexibility. Given a forecast method or learning algorithm, its expected MSE can be decomposed into its squared bias (ie, errors due to erroneous assumptions underlying the method), its variance (ie, errors due to the sensitivity of the method to noise in the calibrating dataset) and the variance of the residual term. The more flexible the method the lower its bias, since it can approximate better the true relationship existent in the data. However, increased flexibility increases the variance of the method since it attempts to fit not only true data points but also the unavoidable noise present in the dataset.

Since stress tests are an input for formulating business strategy or guiding policy, the tests' results and conclusions need to be communicated to different constituencies, including senior decision-makers, each with a different grasp and understanding of the technical details underlying the assumptions and analytical methods. This situation places a premium on interpretability rather than flexibility since it is somewhat difficult to build persuasive arguments based on black-box analysis. Hence, whenever possible, linear models are preferred. Since best subset selection models and lasso regressions are framed within the familiar regression model, they are relatively simple to explain. The geometric intuition behind support vector machines, which generalise the concept of separating hyperplanes and discriminant analysis, is somewhat easy to grasp although the concept of kernels could prove problematic to explain to those not familiar with linear algebra.

For better or worse, the emphasis on prediction rather than casual inference has also contributed to the perception that machine-learning models are black boxes, which makes it difficult to trust the models and their outputs. More importantly, for decision-making purposes it is not enough to know what a model predicts, but also why the model predicts it. The machine-learning academic community has been aware of this situation, and there is a growing recognition that human interpretation of machine-learning models is necessary.[15] One promising way to make machine-learning models output accessible is by the use of model-agnostic interpretability methods. These allow the stress-testing team to use its preferred machine-learning model without having to worry about a model-specific interpretation. Model-agnostic methods include partial dependence plots, individual conditional expectation, feature interaction, feature importance, global surrogate models, local surrogate models and Shapley value explanations (see Molnar, 2018).

Among model-agnostic methods, the use of surrogate models could be the most appropriate in stress testing. The method requires fitting an interpretable model – eg, logistic regression, linear regression or decision tree – to the predictions of the machine-learning model. If the surrogate model replicates the output of the machine-learning model relatively well, the interpretation of the machine-learning model outcome can be built on our intuitive understanding of the surrogate model. Surrogate models can be estimated at the global level (ie, including all sample observations) or locally, by fitting different models in subsamples.[16]

MACHINE-LEARNING MODELS AND AGENT-BASED MODELS: THE WAY FORWARD IN STRESS TESTING?

Machine-learning models outperform traditional econometric models in prediction tasks, which makes them suitable for stress testing given how important forecasting accuracy is for capital allocation decisions and regulatory actions. One shortcoming of these methods is that prediction gains come at the expense of neglecting casual inference. We expect this to be less of a problem going forward since substantial efforts are under way to combine machine-learning methods and causal inference (see Athey, 2019).

Powerful as the machine-learning models are, they share the same weakness as traditional econometric models used in stress testing: their calibration (ie, training) depends on past historical data. As long as past periods of distress are representative of what we can expect in the future, the models would produce reliable predictions. This may not necessarily hold true whenever there are regime changes or structural breaks in business and operating environment affecting firms' behaviour. Take, as an example, bank regulation, which has experienced substantial changes since the global financial crisis. As banks adapt their behaviour and business strategies under a new set of constraints, would calibrations based on past data capture their response to shocks adequately?

To capture the behaviour of economic agents – households and firms – researchers in academia, central banking and government organisations are increasingly turning to agent-based models (ABMs). These models follow a bottom-up approach and introduce adaptive behaviour at the agent level. Aggregate behaviour emerges from the interaction between agents, and is not specified *a priori*. When properly calibrated, ABMs can serve as artificial economies where what-if simulations can be conducted, including stress tests.[17] The quality of the stress tests conducted in an ABM framework depends on how well the models are calibrated, which is not a trivial task. This is an area where machine-learning methods could prove effective and help us move the frontier of stress testing further, allowing us to capture strategic interactions and feedback loops explicitly.[18]

The views in this chapter are those of the author and do not necessarily represent the views of the IMF, its Executive Board, or IMF management. The chapter benefited from comments by the editors, Akhtar Siddique and David Lynch, and from previous work with J.-C. Duan, L. Lin and W. Miao. The author retains sole responsibility for any errors or omissions.

1 This is common practice in banks (see, for instance, Deutsche Bank, 2016, and BCBS, 2017).

2 In central banks and other official agencies, the most common approach is to use structural macroeconometric models, time series econometric models or a set of regression equations for each risk factor. Kitamura *et al* (2014) and Banco de México, 2013, *Financial System Report*, México, D.F. describe stress-test scenario design and calibration at the central banks of Japan and Mexico, respectively. In the banking sector, one survey indicates scenario calibration is mainly based on expert judgement (BCBS, 2017).

3 See Zhang (2013) for an in-depth review of CCAR; Board of Governors of the Federal Reserve, 2018, "Comprehensive Capital Analysis and Review 2018: Assessment Frameworks and Results," Washington, D.C. for a description of the 2018 CCAR; and Siddique and Hasan (2019) for a comprehensive review of stress testing.

4 See the detailed discussion in Kupiec (2017), whose numerical results suggest that the class of regulatory models used by the Federal Reserve may yield inaccurate forecasts and penalise banks' models that deliver different results.

5 Chan-Lau and Li (2015) describe a machine-learning based stress-testing model developed for the South Africa Reserve Bank. Duan *et al* (2015) set a benchmark analytical model for scenario analysis useful for stress testing. Regularisation methods such as lasso are used to forecast firm-specific predictors of the probability of default of publicly listed firms.

6 An alternative approach to selecting sparse models is to use aggregate forecasts of multiple models (Claeskens and Hjort, 2008). An extensive exploratory analysis suggests this may not be feasible in high dimensional problems (Guan et al, 2014).

7 The chapter does not describe these methods in detail. Chakraborty and Joseph (2017) provide a non-technical broad overview of the methods; good introductory and comprehensive textbook expositions of the main machine learning algorithms, in increasing order of difficulty are James, G., D. Witten, T. Hastie, and R. Tibshirani, 2013, *An Introduction to Statistical Learning with Applications in R* (New York: Springer) and Mohri *et al* (2012).

8 Or, in machine-learning parlance, forecasting can be referred to as predictive analytics or predictive modelling.

9 For applications in economics, see Fan, Lv and Qi (2011). For a brief overview of lasso applications in economics, finance and financial networks in Chan-Lau (2017b).

10 In the relaxed lasso, once the lasso algorithm has selected the relevant predictors they are used in a simple linear regression model. This allows calculating confidence intervals (see Meinshausen, 2007).

11 The monthly median PD series data came from the CRI database maintained by the Risk Management Institute at the National University of Singapore (Duan and Van Laere, 2012).

12 The set of predictors comprised 13 domestic primary variables, including the exchange rate *vis-à-vis* the US dollar, the nominal effective exchange rate, the domestic policy rate, the consumer price index, the real GDP growth rate, the unemployment rate, the total amount of credit in the economy, the money market rate, the three-month Treasury bill rate, the bank deposit rate, the bank lending rate and the 10-year Treasury bond rate. The domestic variables are complemented by five international variables: the US real GDP, the China real GDP, the US policy rate, the US consumer price index and a commodity price index.

13 The analysis included 15 domestic economic and financial variables, 12 international macroeconomic variables and four country blocks.

14 VaR-based models of credit risk have been a workhorse in macro stress tests conducted by central banks and multilateral organisations such as the IMF and the World Bank (see Virolainen, 2004, for details on these models).

15 See Doshi-Velez and Kim (2017); Miller (2017); Molnar (20189) for a textbook treatment.

16 Assessing the results, however, requires a careful assessment of the quality of the approximation. For instance, one widely used method is local interpretable model-agnostic explanations (LIME, see Ribeiro *et al*, 2016). The reliability of the method depends on whether a linear model can approximate local behaviour well. In the case of small, local regions exhibiting high nonlinearity, the local approximation could fail.

17 See, among others, Bookstaber *et al* (2014) and Chan-Lau (2017a). Aymanns *et al* (2018) discuss in detail how these models could be applied in stress testing.

18 See Furtado (2017); van der Hoog (2017); Lamperti *et al* (2017).

REFERENCES

Athey, S., 2019, "The Impact of Machine Learning in Economics", in A. Agrawal, J. Gans and A. Goldfarb (Eds), *The Economics of Artificial Intelligence: An Agenda* (Chicago, IL: The University of Chicago Press).

Aymanns, C., J. D. Farmer, A. Kleinnijenhuis and T. Wetzer, 2018, "Models of Financial Stability and their Application in Stress Tests", INET Oxford Working Paper No. 2018–6, University of Oxford.

Bagherpour, A., 2018, "Predicting Mortgage Loan Default with Machine Learning Methods", mimeo, University of California, Riverside.

Barboza, F., H. Kimura and E. Altman, 2017, "Machine Learning Models and Bankruptcy Prediction", *Expert Systems with Applications,* 83, pp 405–17.

BCBS, 2017, "Supervisory and Bank Stress Testing: Range of Practices", Bank for International Settlements.

Bookstaber, R., M. Paddrik and B. Tivnan, 2014, "An Agent-based Model for Financial Vulnerability", OFR Working Paper 14–05, US Treasury, Office of Financial Research, Washington, D.C.

Chakraborty, C. and A. Joseph, 2017, "Machine Learning at Central Banks", Staff Working Paper No. 674, Bank of England.

Chan-Lau, J. A., 2017a, "ABBA: An Agent-based Model of the Banking System", IMF WP/17/136, International Monetary Fund.

Chan-Lau, J. A., 2017b, "Lasso Regressions and Forecasting Models in Applied Stress Testing", IMF WP/17/108, International Monetary Fund.

Chan-Lau, J. A. and L. Lin, 2015, "A Machine-learning Stress Testing Integrated Tool: User's Guide", mimeo, South Africa Reserve Bank.

Claeskens, G. and N. Hjort, 2008, *Model Selection and Model Averaging* (New York, NY: Cambridge University Press).

Claessens, S. and M. A. Kose, 2018, "Frontiers of Macrofinancial Linkages", BIS Papers No. 95, Bank for International Settlements.

Deutsche Bank, 2016, "Annual Report".

Doshi-Velez, F. and B. Kim, 2017, "Towards a Rigorous Science of Interpretable Machine Learning", mimeo, Harvard University and Google Brain.

Duan, J.-C. and E. van Laere, 2012, "A Public Good Approach to Credit Ratings – From Concept to Reality", *Journal of Banking and Finance,* 36, pp 3,239–47.

Duan, J.-C., W. Miao and J. A. Chan-Lau, 2015, "BuDA: A Bottom-up Default Analysis Platform", mimeo, National University of Singapore.

European Systemic Risk Board, 2018, "Adverse Macro-financial Scenario for the 2018 EU-wide Banking Sector Stress Test".

Fan, J., J. Lv and L. Qi, 2011, "Sparse High-dimensional Models in Economics", *Annual Review of Economics,* 3, pp 291–317.

Furtado, B. A., 2017, "Machine Learning Simulates Agent-based Model", mimeo, Institute for Applied Economics Research, Brazil.

Greene, W. H., 2017. *Econometric Analysis (8e)* (Upper Saddle River, NJ: Prentice Hall).

Guan, Z., Q. Luo, G. Wang and B. Wu, 2014, "Satellite Models in Stress Tests: Automatic Model Fitting", Credit Practicum Report, Risk Management Institute, National University of Singapore.

Hirtle, B., A. Kovner, J. Vickery and M. Bhanot, 2015, "Assessing Financial Stability: The Capital and Loss Assessment Under Stress Scenarios (CLASS) Model", Federal Reserve Bank of New York Staff Report No. 663.

van der Hoog, S., 2017, "Deep Learning in (and of) Agent-based Models: A Prospectus", mimeo, Computational Economics Group, Bielefeld University.

International Monetary Fund, 2018a, "United States: Staff Report for the 2018 Article IV Consultation", IMF Country Report No. 18/207, Washington, D.C.

International Monetary Fund, 2018b, "People's Republic of China: Staff Report for the 2018 Article IV Consultation", IMF Country Report No. 18/240, Washington, D.C.

Jacobs, M., 2018, "The Validation of Machine Learning Models for Stress Testing Credit Risks", mimeo, Accenture Consulting.

Kapinos, P. and O. Mitnik, 2015, "A Top-down Approach to the Stress Testing of Banks", mimeo, The Federal Deposit Insurance Corporation.

Kitamura, T., S. Kojima, K. Nakamura, K. Takahashi and I. Takei, 2014, "Macro Stress Testing at the Bank of Japan", BOJ Reports and Research Papers.

Kupiec, P. H., 2017, "Inside the Black-box: The Accuracy of Alternative Stress Test Models", American Enterprise Institute Economics Working Paper 2017–03.

Lamperti, F., A. Roventini and A. Sani, 2017, "Agent-based Model Calibration Using Machine Learning Surrogates", mimeo, Institute of Economics, Scuola Superiore Sant' Anna and OFCE-Sciences Po.

Meinshausen, N., 2007, "Lasso with Relaxation", *Computational Statistics and Data Analysis,* 52(1), pp 374–93.

Miller, T., 2017, "Explanation in Artificial Intelligence: Insights from the Social Sciences", mimeo, University of Melbourne.

Mohri, M., A. Rostamizadeh and A. Talwalkar, 2012, *Foundations of Machine Learning* (Cambridge, MA: MIT Press).

Molnar, C., 2019, *Interpretable Machine Learning* (lulu.com).

Perdeiy, V., 2009, "Bankruptcy Prediction Revisited: Non-traditional Ratios and Lasso Selection", mimeo, Deutsche Bank Group.

Ribeiro, M., S. Singh and C. Guestrin, 2016, "Why Should I Trust You? Explaining the Predictions of Any Classifier", arvix.org.

Siddique, A. and I. Hasan, 2019, *Stress Testing: Approaches, Methods and Applications*, 2nd ed (London: Risk Books).

Sirignano, J., A. Sadhwani and K. Giesecke, 2018, "Deep Learning for Mortgage Risk", mimeo, University of Illinois, Urbana-Champaign, Google Brain and Stanford University.

Tibshirani, R., 1996, "Regression Shrinkage and Selection Via the Lasso", *Journal of the Royal Statistical Society*, *Series B (Methodological)*, 58(1), pp 267–88.

Virolainen, K., 2004, "Macro Stress Testing with a Macroeconomic Credit Risk Model for Finland", Bank of Finland Discussion Paper No. 18/2004.

Zhang, J. (Ed). 2013, *CCAR and Beyond: Capital Assessment, Stress Testing and Applications* (London: Risk Books).