International Journal of Forecasting 36 (2020) 1092–1113

Contents lists available at ScienceDirect

International Journal of Forecasting

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

Predicting bank insolvencies using machine learning techniques[☆]

Anastasios Petropoulos, Vasilis Siakoulis, Evangelos Stavroulakis, Nikolaos E. Vlachogiannakis [*]

Bank of Greece, 3 Amerikis, 10250 Athens, Greece

ARTICLE INFO

Keywords:
Bank's insolvencies
Forecasting
Random Forests
Support Vector Machines
Neural Networks
Conditional inference trees

ABSTRACT

Proactively monitoring and assessing the economic health of financial institutions has always been the cornerstone of supervisory authorities. In this work, we employ a series of modeling techniques to predict bank insolvencies on a sample of US-based financial institutions. Our empirical results indicate that the method of Random Forests (RF) has a superior out-of-sample and out-of-time predictive performance, with Neural Networks also performing almost equally well as RF in out-of-time samples. These conclusions are drawn not only by comparison with broadly used bank failure models, such as Logistic, but also by comparison with other advanced machine learning techniques. Furthermore, our results illustrate that in the CAMELS evaluation framework, metrics related to earnings and capital constitute the factors with higher marginal contribution to the prediction of bank failures. Finally, we assess the generalization of our model by providing a case study to a sample of major European banks.

© 2020 International Institute of Forecasters. Published by Elsevier B.V. All rights reserved.

1. Introduction – Motivation

Supervisory authorities are primarily concerned with protecting depositors' interests, via ensuring that financial institutions are able to survive under business as usual conditions and are sufficiently immune to any adverse market shocks. Hence, the comprehensive assessment of the current financial conditions of a bank as well as the evaluation of its future sustainability is the cornerstone of proactive banking supervision. To distinguish between strong and weak banks, supervisory authorities make use of early warning expert systems or/and statistical modeling techniques. The outcome of this analysis can drive the imposition of targeted regulatory measures. These measures can take the form of preemptive corrective actions addressing vulnerabilities of weaker banks and as a result increase their chances for sustainability. While, in specific cases of "likely to fail" banks, whose return to viability is rather considered irreversible, it will provide the necessary evidence to the supervisory authorities in order to take even more drastic actions. Essentially, supervisory actions serve in retaining depositors' confidence to the financial system, so that any domino effects that can even trigger a potential systemic crisis are precluded.

Between 1934 and 2014, there were 4069 banks in the United States that failed or received financial assistance from FDIC. More precisely, according to the FDIC records 3483 banks failed or were assisted by the Central Bank from 1980 to 2014 following the deregulation of the US banking system in 1980's, notwithstanding the considerable efforts made by supervisory authorities in identifying vulnerable financial institutions.

As a response to the numerous defaults of credit institutions during the recent global financial crisis, the Basel

[☆] The views expressed in this paper are those of the authors and not necessarily those of Bank of Greece.

[*] Corresponding author.
E-mail addresses: apetropoulos@bankofgreece.gr (A. Petropoulos), VSiakoulis@bankofgreece.gr (V. Siakoulis), EStavroulakis@bankofgreece.gr (E. Stavroulakis), nvlachogiannakis@bankofgreece.gr (N.E. Vlachogiannakis).

https://doi.org/10.1016/j.ijforecast.2019.11.005
0169-2070/© 2020 International Institute of Forecasters. Published by Elsevier B.V. All rights reserved.

Committee on Banking Supervision (BCBS) has introduced an updated set of regulations, known as the Basel III accord.[1] This set of rules aims at further strengthening the monitoring processes of supervised entities, so that the effectiveness of banking supervision is enhanced. However, it seems that the compliance with an even more extended set of minimum regulatory standards or/and the close monitoring by supervisory authorities of the evolution of a bank's risk indicators, should not be assessed on a standalone basis. It is essential that all risk drivers and relevant information should be combined into a single measure, representing each bank's financial strength. Reflecting in a single score a bank's overall risk level have been proved to be a difficult task due to the big bulk of information that is currently collected by supervisory authorities. In the absence of strong analytical and data filtering tools, this oversupply of information could even mislead regulators during the decision-making process. Hence, supervisory authorities should utilize robust aggregation methodologies, which could result in the efficient calculation of a survival probability for each financial institution as well as its classification into different riskiness classes.

In the last decades, various statistical methodologies have been exploited to aggregate bank specific information into a single figure in order to distinguish between solvent and insolvent financial institutions. These classification methods range from simple Discriminant analysis (Altman, 1968; Cox & Wang, 2014; Kočišová & Mišanková, 2014) and Logit/Probit regressions (Ohlson, 1980; Cole & Wu, 2017), to advanced machine learning techniques such as Support Vector Machines (Boyacioglu, Yakup, & Baykan, 2009; Chen & Jen, 2006), conditional inference trees and Neural Networks (Messai & Gallali, 2015; Ravi & Pramodh, 2008). At the same time, other novel modeling approaches such as Random Forests (RF) (Breiman, 2001) have not been employed to date in the problem of assessing bank failures, regardless of these models being really popular for modeling classification problems in recent years.

However, no academic study exists that thoroughly assesses simultaneously all, or at least some, of the abovementioned methodologies on a common dataset in order to determine in a concrete way their relative forecasting performance. In this work, addressing the aforementioned gap in the current literature, we employ a series of performance statistics to assess the explanatory power of six modeling techniques in predicting bank insolvencies, including Logistic Regression (LR), Linear Discriminant analysis (LDA), Random Forests (RF), Support Vector Machines (SVM), Neural Networks (NN) and Random Forests of Conditional Inference Trees (CRF). The model evaluation measures utilized in this analysis are tailored to assess model performance on imbalanced samples, like all datasets used in related studies to ours. Model performance is assessed based on in-sample, out-of-sample, and out-of-time scenarios. We deem that our comprehensive analysis, which is coupled with an extended and robust validation process across all developed models, provides significant findings regarding the selection of the "optimal" method for identifying bank failures.

Another critical component in predicting bank insolvencies, apart from selecting the appropriate modeling technique, is the universe of explanatory variables to be analyzed. Although a series of studies employs macroeconomic determinants to develop early warning systems for bank failures (Betz, Oprica, Peltonen, & Sarlin, 2014; Mayes & Stremmel, 2012; Rebel & White, 2012), recent empirical evidence suggests that the financial condition of individual banks is a key driver in distinguishing their performance during the recent financial crisis (Berger & Bouwman, 2013; Vazquez & Pablo, 2015). Moreover, supervisory authorities are interested in the bank-specific weaknesses that may drive banks to insolvency, so that they are able to address them via the specification of targeted remedial actions in each particular case. In this study, following the same philosophy, we use an extended dataset of bank-specific variables to differentiate between failing and non-failing financial institutions. Specifically, we test the explanatory power of more than 40 variables that can be broadly classified under the categories prescribed by CAMELS (Messai & Gallali, 2015; Cole & Wu, 2017), along with their lags up to 8 quarters and their transformed changes in various time intervals, amounting to a total number of 660 covariates examined. In short, the high number of independent variables investigated along with the state-of-the-art methods used for variable selection and model setup, envisages capturing the bulk of any available bank specific information under the problem of distinguishing between solvent and insolvent banks.

There is a big debate in the current literature regarding the superiority of certain indicators in predicting bank failures. Mayes and Stremmel (2012) claim that a simple leverage ratio (unweighted) is a better predictor than the capital adequacy ratio (risk weighted), while others Cole and Wu (2017) have identified that those related to capital adequacy, liquidity and asset quality are the most important predictors of bank failures. In an attempt to offer more evidence on this inconclusive aspect in the current literature, we rank the predictors used across all the models developed based on their marginal contribution. Our results indicate that metrics related to capital and earnings constitute the factors with the highest marginal contribution in predicting bank failures.

To sum up, our modeling approach is balanced between capturing the determinants that strongly affect the health of a financial institution, while at the same time developing an early warning system to predict bank failures (i.e., via assessing our result in out-of-sample and out-of-time validation). The modeling framework that we implement captures temporal dependencies in a bank's financial indicators. At the same time, it explores up to 2 years of lagged observations, which are assumed to carry all the necessary information to describe and predict the financial soundness of a bank.

The structure of the remaining part of this study is organized as follows. In Section 2, we focus on the related literature review on bank failure prediction. Section 3 describes the data collection and processing. In Section 4,

[1] The new proposals address risks not covered in the existing regulatory framework, by introducing stricter criteria for the quality of capital, a binding leverage ratio as well as two indicators to capture liquidity risk.

we provide details regarding the estimation process of the developed alternative models employed. In Section 5, we compare across methodologies providing important insights on variable importance whereas we also provide a case study by applying our selected technique in a sample of European banks and benchmark it relative to Moody's credit ratings. Thus, we assess not only its applicability but also its generalization capacity. Finally, in the concluding Section 6, we summarize the performance superiority of the proposed methodology, we identify any potential weaknesses and limitations, while we also discuss areas for future research extensions.

2. Literature review

There is an extensive literature on the various methods and analyses performed, regarding bank default prediction. Demyanyk and Iftekhar (2010) provide a summary of various papers focusing on analyzing, forecasting, and providing remedial actions for potential financial crises or bank defaults. For completeness purposes, Appendix A contains a summary of the existing literature that deals with the problem of forecasting bank failures, by outlining the statistical techniques along with the dataset employed.

A large group of literature related to bank failure prediction focuses on the set of supervisory CAMELS indicators. This is the acronym for Capital, Asset Quality, Management, Earnings, Liquidity, and Sensitivity to market risk indicators, which are typically used by investors and regulators to assess the soundness of a financial institution. Several empirical studies also combine CAMELS with additional indicators (Cole & White, 2011, Altunbas, Manganelli, & Marques-Ibanez, 2011; Betz et al., 2014; Chiaramonte, Croci, & Poli, 2015; Cox & Wang, 2014; Lall, 2014; Poghosyan & Čihák, 2009; Wanke, Azad, Barros, & Hadi-Vencheh, 2015) in an attempt to increase the explanatory power of their models. However, there is inconclusive evidence on which variables are important in predicting bank insolvencies. Poghosyan and Čihák (2009) show that indicators related to capitalization, asset quality and profitability can effectively identify weak banks. Berger and Bouwman (2013) showed that capital, had a positive impact on the survival probabilities and market shares of small banks. While, Mayes and Stremmel (2012) indicated that the leverage ratio out-performs risk-weighted capital ratios. In an attempt to receive a concluding answer on what are the variables that lead banks to default, this study incorporates a wide range of CAMEL-related variables, along with various transformations, so as to identify the ones with the higher explanatory power and provide a ranking among them.

At the same time, the modeling techniques used for predicting bank insolvencies range from simple logistic regressions or their extensions (Audrino, Kostrov, & Ortega, 2018), up to advanced machine learning techniques. Halling and Hayden (2006) introduced a two-step survival time procedure that combines a multiperiod logit model and a survival time model, while Kolari, Glennon, Shin, and Caputo (2002) introduced the parametric approach of trait recognition to develop early warning systems. Advanced modelling techniques have also been used by Gogas, Papadimitriou, and Agrapetidou (2018), who apply Support Vector Machines for tackling the problem of bank failure prediction. Messai and Gallali (2015) by applying discriminant analysis, logistic regression and artificial intelligence methods pointed that the neural network method performed better than the other models. Finally, a comparison of artificial intelligence methods was introduced in Ekinci and Halil (2016). However, all those studies do not simultaneously benchmark the performance both of the standard modeling techniques (e.g. Logit, LDA) and of the more advanced ones (e.g. SVM, NN etc.) for predicting bank failures, and, to the best of our knowledge, none of them apply Random Forests. In short, although there do exist many studies that benchmark the performance of multiple modeling techniques for corporate defaults forecasting, we are not aware of any other study that performs a parallel analysis of all these modeling techniques (i.e., two standard and 4 advanced) in the prediction of banking failures, which exhibit the unique property of a low default portfolio.

The approach outlined in this paper offers a significant advantage over most of the existing literature, as the assessment of modeling options sufficiently covers most of the available and applicable statistical methods in predicting bank distress, whereas we include a wide range of explanatory variables. Namely, the following models are included in our analysis: Logistic Regression (LogR), Linear Discriminant Analysis (LDA), Random Forests (RF), Support Vector Machines (SVMs), Neural Networks (NNs) and Random Forest of Conditional Inference Trees (CRF). In addition, over 660 variables are assessed in the basis of their potential predictive value. In this way we can address simultaneously many of the topics occurring in the literature either concerning the final model implemented or relatively to the variables' selection process. We also utilize a robust assessment methodology to evaluate the performance of each model. In doing so, we include out-of-sample and out-of-time validation samples as well as various discriminatory and accuracy tests. Finally, we extend the dataset used in the vast majority of previous studies (which use development samples that marginally reach 2010) by employing a dataset reaching up to Q4 2014.

3. Data collection and processing

We have collected information on non-failed, failed, and assisted entities from the database of the Federal Deposit Insurance Corporation (FDIC), an independent agency created by the US Congress in order to maintain the stability and the public confidence in the financial system. The collected information is related to all US banks, while the adopted definition of a default event in this dataset includes all bank failures and assistance transactions of all FDIC-insured institutions. Under the proposed framework, each entity is categorized either as solvent or as insolvent based on the indicators provided by FDIC.

The dataset covers the 2008–2014 period; a 7-year period with quarterly information resulting in dataset with more than 175,000 records. The selected time period seems to approximate a full economic cycle, in terms of the Default Rate evolution. Fig. 1 shows the number of records included in each observation quarter and the corresponding default rate. It is clear that the default rates are significantly increased in the first half of sample, compared to the second half. Specifically, the Default Rates follow an increasing trend in the 2008–2009 period, where they peak at 2.5% in the third quarter of 2009. Thereafter, they follow a decreasing trend. The default rates seem to have flattened out around 2013, further decreasing during 2014, reaching 0.1% in the fourth quarter of 2014.

The dataset was split into three parts (Fig. 2). An in-sample dataset (Full in sample) that comprises the data pertaining to the 80% of the examined companies over the observation period 2008–2012 (randomly stratified[2]) amounting to 101.641 observations. An out-of-sample dataset, including the remaining 20% of the observations for the period 2008–2012 (randomly stratified) amounting to 25.252 observations, and an out-of-time dataset that spans over the 2013–2014 observation period reaching 48.756 observations. In all cases, the dependent variable is a binary indicator that takes the value of one in case there is a default event, while it takes the value of zero otherwise.

The model development process was performed in a shorter dataset named "Short in sample", which is derived from the "Full in sample" by randomly excluding 90% of the solvent banks while keeping all the insolvent banks, amounting to 11.573 observations. This is done primarily to account for the low number of defaults observed during the in sample period. Therefore, we artificially increased the bad to good mix of the dataset used for development so as to reach a 10% proportion of insolvent banks. Depending on the model type under consideration, we further equally split in certain cases our "Short in sample" dataset into a training set and into a validation set. This is especially true for RFs and NN, in which the training sample is used to train the candidate model, while the validation set is used for selecting the best parameter setup.

In particular, we have opted for using a subsample of the solvent banks in order to train our models (short in sample), following Japkowicz (2000) and Altini (2015). Under-sampling is a method usually used for dealing with imbalanced datasets in building classification models, as significantly imbalanced datasets can be a challenge to train hyper parameters on various statistical techniques. In this specific problem, failed banks account for a significant small proportion of the total population. Therefore, by using a less imbalanced train sample, the learning process of various statistical algorithms was more efficient. At the same time, in order to ensure that overfitting is not an issue in Random Forests and Neural Networks, we used the "train" part of the dataset for development purposes, and the "validation" part of the dataset for assessing the performance of the fitted model. The validation dataset provides an unbiased evaluation of a model's fit on the train dataset, while tuning the model's hyper parameters (e.g., the number of hidden units in a neural network). In particular, training stops when the error on the validation dataset increases, as this is a sign of overfitting to the train dataset. Most importantly, we have assessed the performance of our models using an out-of-time validation sample, which reflects the proportion of failing banks as it appears in this period (i.e., 48.619 Good, 137 Bad). The results of the out-of-time validation suggest the good performance of our models in highly imbalanced datasets, even if these models are originally developed in the "short in sample". So, in case overfitting had been an issue in our models, we would expect that the out-of-time validation results would have indicated a really poor model performance, which is not the case.

To sum up, after developing our models in the "short in sample" dataset, we assess their performance results under three different validation samples. The first, being the "Full in sample", is used to evaluate the generalization capacity of our models in a population with less frequent default events than the ones observed in the development sample (short in-sample). The second is the "Out-of-sample" that is used to assess the performance of each model across banks during the same period. While, under the third "Out-of-time" datasets the performance of each model is evaluated during a future time period. The spectrum of the validation sets used in assessing the performance of our models ensures, to the greatest extend possible, that no material factors have been disregarded due to a potentially biased selection of the "short in sample". Essentially, the models would have substantially underperformed in the validation samples (and even more in the out-of-time validation sample), in case any material factor driving banks insolvencies had been ignored. This is particularly true as the validation samples include all the banks at the proportions they appear in the population. Hence, as this is not implied by our results, it is not that probable to have omitted important variables.

In developing our model specifications, we examine an extended set of variables that follow under the classification categories of CAMELS (i.e. Capital, Asset Quality, Management, Earnings, Liquidity, and Sensitivity to market risk). Specifically, the following independent variables are tested:

- Capital adequacy (C):
  i. Equity capital to assets (EQ_ASS)
  ii. Core capital (leverage) ratio (LEV)
  iii. Tier 1 risk-based capital ratio (TIER1)
  iv. Total risk-based capital ratio (CAR)
  v. Common equity tier 1 capital ratio (CET1)
- Asset quality (A):
  i. Loan and lease loss provision to assets (PROV_ASS)
  ii. Net charge-offs to loans (CHOF_LOAN)
  iii. Credit loss provision to net charge-offs (PROV_CHOF)

[2] The number of solvent and insolvent banks is selected for each quarter such that the same default rate for the quarter is retained.

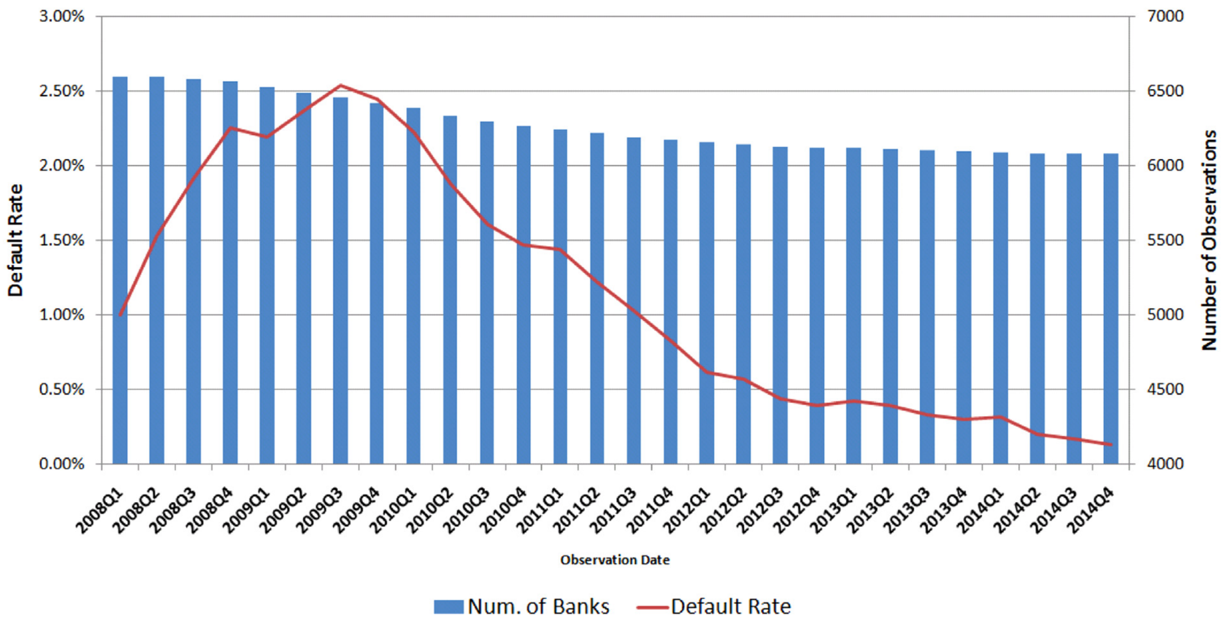


Fig. 1. Bank defaults in the USA. Historical overview for the period 2008–2014.
Source: FDIC.

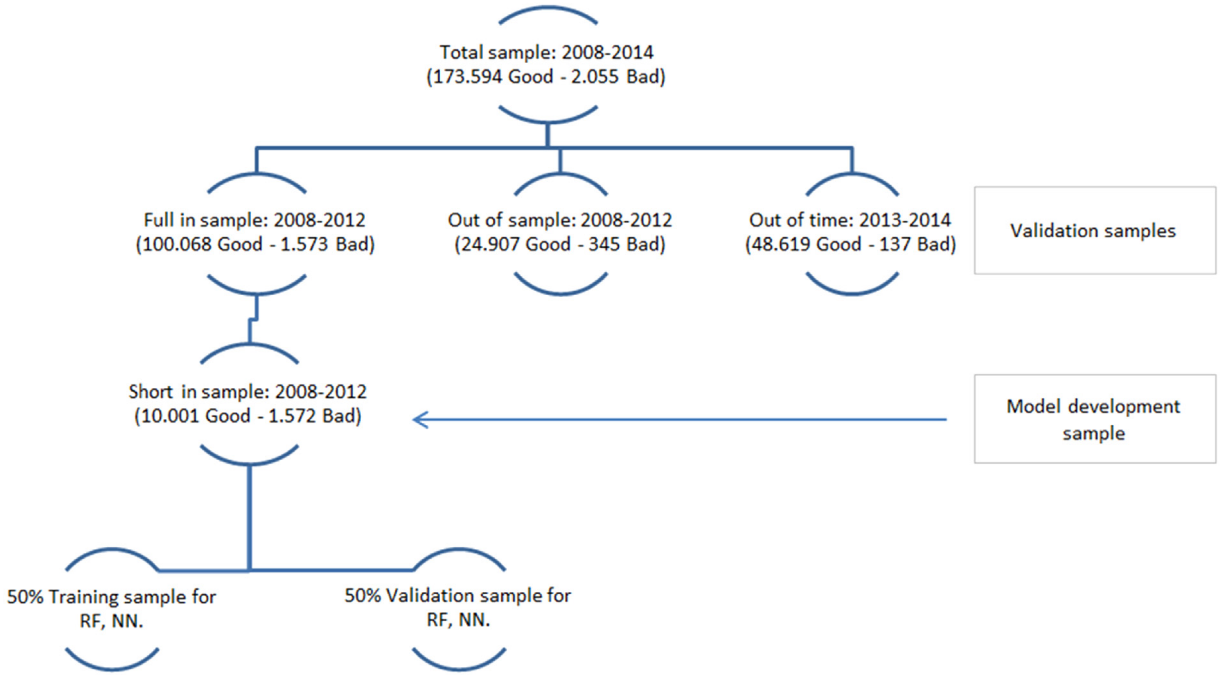

Total sample: 2008-2014
(173.594 Good - 2.055 Bad)
Full in sample: 2008-2012
(100.068 Good - 1.573 Bad)
Out of sample: 2008-2012
(24.907 Good - 345 Bad)
Out of time: 2013-2014
(48.619 Good - 137 Bad)
Validation samples
Short in sample: 2008-2012
(10.001 Good - 1.572 Bad)
Model development sample
50% Training sample for RF, NN.
50% Validation sample for RF, NN.

Fig. 2. Model development and Validation samples.

iv. Assets per employee ($millions) (ASS_EMP)
v. Earning assets to total assets ratio (EASS_ASS)
vi. Loss allowance to loans (LOSS_LOAN)
vii. Loan loss allowance to noncurrent loans (LOSS_NPL)
viii. Noncurrent assets plus other real estate owned to assets (NCASS_ORE)
ix. Noncurrent loans to loans (NPL)
x. Average total assets (ASSET)
xi. Average earning assets (EASSET)
xii. Average equity (EQUITY)
xiii. Average total loans (LOAN)
xiv. Net loans and leases (LNLSNET)
xv. Loan loss allowance (LNATRES)

xvi. Restructured Loans & leases (RSLNLTOT)
xvii. Assets past due 30–89 days (P3ASSET)
xviii. Restructuring ratio (RESTR)
xix. Provisions to loans (PROVTL)
xx. Provision to assets (PROVTA)

- Management capability (M):

  i. Noninterest income to average assets (NFI_ASS)
  ii. Noninterest expense to average assets (EXP_ASS)
  iii. Net operating income to assets (NOI_ASS)
  iv. Earnings coverage of net charge-offs (EAR_CHOF)
  v. Efficiency ratio (EFF)
  vi. Cash dividends to net income (ytd only) (DIV_INC)

- Earnings (E):

  i. Yield on earning assets (YEA)
  ii. Cost of funding earning assets (CFEA)
  iii. Net interest margin (NIM)
  iv. Return on assets (ROA)
  v. Pretax return on assets (PTR_ASS)
  vi. Return on Equity (ROE)
  vii. Retained earnings to average equity (ytd only) (RE_EQ)

- Liquidity (L):

  i. Net loans and leases to total assets (NLOAN_ASS)
  ii. Net loans and leases to deposits (NLOAN_DEP)
  iii. Net loans and leases to core deposits (NLOAN_CDEP)
  iv. Total domestic deposits to total assets (DDEP_ASS)
  v. Volatile Liabilities (V_LIAB)

- Market risk (S):

  i. Asset Fair Value (AFV)

In addition to the abovementioned variables, we have also explored whether the business model/sector, according to the classification provided by FDIC, has any statistical power in predicting bank insolvencies. Furthermore, we created a dummy indicating whether in the last period a significant bank filed for bankruptcy. This variable can be considered as a systemic shock indicator, which can potentially capture any contagion effects among banks.

To derive more representative drivers to train our models, we experimented with various transformations of the aforementioned ratios. Below in brackets [], we present the naming convention pertaining to each type of transformation. Specifically, we proceeded by executing the following sequential steps:

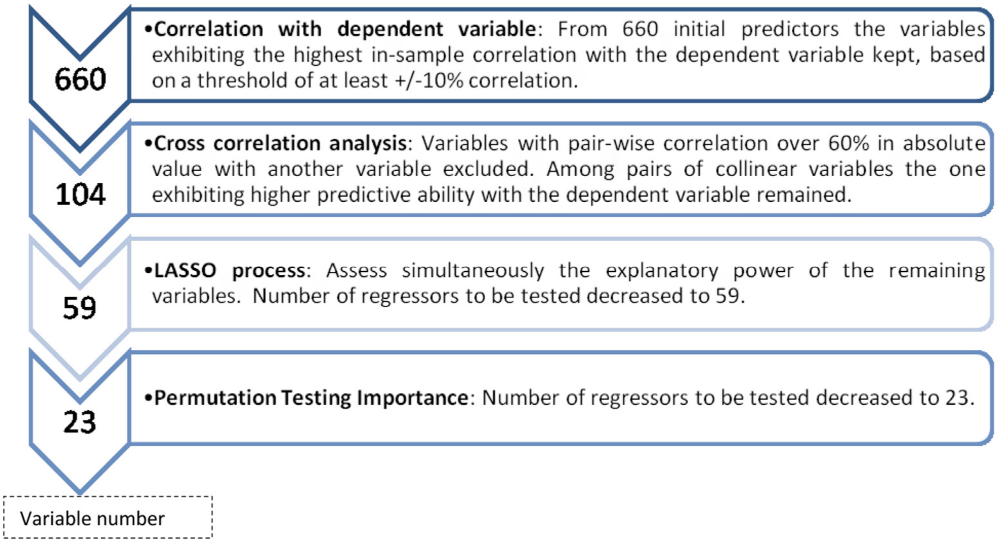
(i) We applied a series of simple log transformations [log] on the indicators referring to amounts/ currency units (e.g. Assets, Equity, Net loans and leases, etc.)
(ii) We calculated lagged variables on a quarterly basis for each indicator [lag1, lag2,…., lag8] starting from 1 up to 8 quarters (i.e. 2 years)
(iii) We computed the quarter-over-quarter change (first difference) [d] for every indicator referring to ratios or log amounts and the quarter-over-quarter percentage change (relative difference) [PCT] for every indicator referring to amounts.
(iv) All variables in the dataset were floored and capped based on the 1st and 99th percentile of each variable respectively.
(v) For a number of selected regressors, we calculated a series of "distance from sector" [DFS] indicators for each quarter. These indicators aim to capture the relative performance of a bank relative to its peers. Specifically:

  a. The sector was approximated by the "Asset Concentration Hierarchy" variable, which is defined as an indicator of the institution's primary specialization in terms of assets concentrations. It includes the following categories:

    i. international specialization,
    ii. agricultural,
    iii. credit card,
    iv. commercial lending,
    v. mortgage lending,
    vi. consumer lending specialization,
    vii. other specialized less than $1 billion,
    viii. all other less than $1 billion and
    ix. all other more than $1 billion.

  b. For each one of the selected regressors, the mean value for each category of the sector proxy was calculated for each quarter.
  c. The "Distance from Sector" was calculated as the difference between the mean and the underlying value of each regressor of the same quarter.

The variable generation process led to a set of almost 660 predictors as potential candidates for our modeling procedures. Our decision to explore various transformations and lags of all regressors was motivated by the current literature that has inconclusive evidence on how many lags to include, and whether using levels or deltas (e.g. first differences) is preferable in predicting bank insolvencies. That is, our choice to create as many regressors as possible was driven by our intention not to miss any important information, which can be conveyed either in a different lag or in a different transformation (e.g., the delta of a regressor and not its level itself). The so-obtained set of time-series was narrowed down in four consecutive stages (Fig. 3). In particular, the following steps serve in reducing the dimensionality space of the candidate variables, so that the core models are developed on a subsample of variables that includes the most relevant ones for the purposes of the analysis. The natural way to do so is to begin with univariate analysis. That is, we investigate for any linear relations by assessing the



•**Correlation with dependent variable**: From 660 initial predictors the variables exhibiting the highest in-sample correlation with the dependent variable kept, based on a threshold of at least +/-10% correlation.

•**Cross correlation analysis**: Variables with pair-wise correlation over 60% in absolute value with another variable excluded. Among pairs of collinear variables the one exhibiting higher predictive ability with the dependent variable remained.

•**LASSO process**: Assess simultaneously the explanatory power of the remaining variables. Number of regressors to be tested decreased to 59.

•**Permutation Testing Importance**: Number of regressors to be tested decreased to 23.

Fig. 3. Flowchart illustrating the variable selection process.

pair-wise correlation of each variable with the dependent variable. Then, considering the large number of explanatory variables in our dataset, we proceed by excluding the ones that convey, to a large extent, the same type of information in interpreting the dependent variable. Thus, we attain our objective regarding the development of a series of parsimonious models. In the next step, we jointly assess the explanatory power of the remaining variables, in a nonlinear set up, by utilizing a variable selection algorithm to isolate the ones the explanatory power of which fully compensates any extra ‘‘noise’’ created from their introduction in the model.

i. Initially, we kept the variables exhibiting the highest in-sample correlation with the modeled (binary) dependent variable, i.e., the categorization of banks as solvent or insolvent at the end of the observation period. Specifically, a threshold of at least +/-10% correlation with the default flag variable was applied to narrow down the extended set of independent variables.
ii. Then, a cross correlation matrix analysis was performed. In particular, an explanatory variable exhibiting pair-wise correlation higher than 60% in absolute value with another explanatory variable was excluded, as it is considered to offer the same qualitative information. Among the pair of variables with significant correlation, the one exhibiting higher predictive ability with the dependent variable remained in the sample. The threshold was set to 60% to account for the high number of lagged and distance from sector variables of the same regressor, which in some sense capture the same information from a different perspective (i.e. across time or relative to the peers/sector respectively). The only exception under this stage was that Leverage (LEV) and Capital Adequacy Ratio (CAR) variables were both kept, as additional analysis was performed in a later stage to assess their standalone importance.
iii. In the third stage, we used LASSO (Least Absolute Shrinkage and Selection Operator) to assess simultaneously the explanatory power of the 104 variables remained under step (ii). We used the Elastic Net parameterization with an alpha of 50% and a 10-fold cross validation. The regressors to be included in our model were selected based on their capacity to minimize the Root Mean Square Error (RMSE). After applying LASSO, the number of regressors to be tested decreased to 59. It should be noted that as we keep on gradually increasing the value of λ, the number of regressors selected in the final model specification decreases well below 59. Not surprisingly, after certain step-wise increases of λ, LASSO ends up to the same set of variables as the ones selected under step iv.
iv. Finally, we omitted the variables whose absence do not lead to a statistically significant loss in accuracy (as measured by the Root Mean Squared Error) based on a permutation type statistic. In this framework, if a predictor is important, then assigning other values for that predictor, permuting this predictor's values over the dataset, should have a negative influence on overall model prediction. In other words, using the same model to predict from data that is the same except for this variable, should give worse predictions. In doing so, we had to select a modeling specification for calculating the RMSE under the original and the permutated data so as to measure the relative loss. We resorted to Random Forest as it can handle a large number of input variables without any correlation restrictions. From this process, the final number of regressors to be tested decreased to 23.

To sum up, to ensure that we keep only the unique and valuable information on the final dataset, our variable selection process starts with a univariate analysis of each regressor with the dependent variable, so that regressors not conveying any important information are removed. Then, we remove regressors that are highly correlated between them, that is, convey the same information for the dependent variable. In the next step, we apply LASSO to select only the variables that have the capacity to minimize the Root Mean Square Error (RMSE).

At this point, it is critical to ensure that the final system of variables remains robust for different model specifications. Specifically, it is not the use of the RF algorithm in the final step of the variable selection process that contributed to the RF showing superior performance. To this end, we extend our analysis by also applying a Stepwise Logit (SL) and a Genetic Algorithm (GA) model specification. The underlying idea of the GA is to generate some random possible variable sets (called population) and then combine the best solutions in an iterative process. In the first step random selections of variables are performed (first generation) and picking up the fittest individuals in each generation (i.e., the variable sets providing the highest ROC) we create sequentially cross-overs of variables till we reach the best solutions in an iterative process spanning a number of generations. As the number of generations increases, more variables are dropped leading to narrower fittest variable sets. We employed a series of trials for optimizing the number of generations based on a ROC criterion ending up to an optimal number of 30 generations.

Specifically, the variable selection process entails the assessment of (i) linear combination of regressors via Stepwise Logit, (ii) tree based multivariate model via RF, and (iii) non-linear model specification via Genetic Algorithm. The vast majority of the variables selected by each and every algorithm, presented in detail in Appendix B, remains robust to different model specifications. In particular, Stepwise Logit (SL) ends up with 22 variables (out of which 20 are also considered by RF), Genetic Algorithm (GA) ends up with 12 variables (out of which 11 are also considered by RF), while there are no variables selected by both SL and GA, and not by RF. Therefore, the superior performance of RF, does not depend on the algorithm used under the variable selection process. That is, our main conclusions regarding the predictive performance of each model seem not to be biased against the variable selection algorithm used.

4. Model development

In our effort to approach the issue of predicting bank insolvency, we apply most of the available and applicable statistical methods in predicting bank distress whereas we include a wide range of explanatory variables. Namely, the following models are included in our analysis: Logistic Regression (LogR), Linear Discriminant Analysis (LDA), Random Forests (RF), Support Vector Machines (SVMs), Neural Networks (NNs), and Random Forest of Conditional Inference Trees (CRF).

4.1. Logistic regression (LogR)

Logistic regression is an approach broadly employed for building corporate rating systems and retail scorecards due to its parsimonious structure. It was first used by Ohlson (1980) to predict corporate bankruptcy based on publicly available financial data. Logistic regression models determine the relative importance of coefficients in classifying debtors into two distinct classes based on their credit risk (i.e. good or bad obligors). To account for nonlinearities and relaxing the normality assumption, a sigmoid likelihood function is typically used (Kamstra, Kennedy, & Suan, 2001).

We implemented logistic regression in R by using the glm function that performs optimization through Iteratively Reweighted Least Squares. To reduce the number of parameters and obtain more intuitive results, we performed a stepwise selection process. In each step, we dropped variables with p-values more than 15% and we re-estimated the model. To avoid any multicollinearity issues, we used only the Leverage Ratio (LEV), while we excluded the Capital Adequacy Ratio (CAR) on the basis of Akaike Information Criterion.

4.2. Linear discriminant analysis (LDA)

Linear discriminant analysis (LDA) is a method to find a linear combination of features that characterizes or separates two or more classes of objects or events. The main assumptions are that the modeled independent variables are normally distributed and that the groups of modeled objects (e.g. good and bad obligors) exhibit homoscedasticity. LDA is broadly used for credit scoring. For instance, the popular Z-Score algorithm proposed by Altman (1968) is based on LDA to build a rating system for predicting corporate bankruptcies. In particular, he estimated a linear discriminant function using a series of financial ratios, which covered the areas of liquidity, profitability, leverage, solvency and turnover, so as to estimate credit quality.

The normality and homoscedasticity assumptions are hardly ever the case in real-world scenarios, thus, being the main drawbacks of this approach. As such, this method cannot effectively capture nonlinear relationships among the modeled variables, which is crucial for the performance of a credit rating system. We implemented this approach in R using the MASS R package, while we restricted our model to the selected variables from the logistic regression to reduce the parameters' dimension and avoid multicollinearity issues.

4.3. Random forests (RF)

Random Forests (RF) is a popular method for modeling classification problems. Since its inception (Breiman, 2001), RFs has gained significant ground and is frequently used in many machine learning applications across various fields of the academic community.

Random forests' basic philosophy is based on combining three concepts: (i) classification or regression decisions trees, ii) bootstrap aggregation or bagging and (iii) random subspaces. Its structure follows a divide-and-conquer approach used to capture nonlinearity in the data and perform pattern recognition. Its core principle is that a group of "weak learners" combined, can form a "strong predictor" model.

The outline of the algorithm is the following: Let's assume that under a supervise setup that we want to model the dataset D which is composed by a series of features denoted by X_i-X_N, where X_i belongs to $\mathbb{R}^d$ space and Y is the dependent variable. The dependent variable can either be continuous, in case we have a regression problem, or binary, in case we investigate a classification problem. Let's also denote B the number of decisions trees the algorithm will generate. This group of trees forms the Forest. The randomness is attributed in two steps of the algorithm, which basically lead to the generation of random trees.

So for i=1 to B the algorithm performs the following steps:

Using bootstrap, a random subsample is selected from D denoted D_i. Then a tree T_i is generated on D_i such that in each node of the tree (or each split) a random subset of feature or explanatory variables is selected and considers splitting only on these features using the CART criterion. So if the number of original features is denoted by N, we select a random subset of them m in each split for each random tree i.e. m < N. Thus the construction of trees is performed on a random subspace of features and a random sample of D. After constructing the random trees, prediction is performed using the bagging method in the following way. Each input is entered through each decision tree in the forest and produces a forecast. Then, all predictions of each tree are aggregated either as a (weighted) average or majority vote, depending on whether the underlying problem is a regression or a classification, to produce a global forecast. Random forests usually avoid overfitting due to the aforementioned bagging process and the random spaces procedure embedded in the algorithm and provide strong generalization efficacy.

In this work, we build a statistical framework to classify financial institutions in two categories, that is, solvent and insolvent. Thus, the model setup employed is random forests for binary classification. In the initial run, the 59 candidate variables were used as input for supervised learning in the short in-sample dataset. To build the Random Forests the randomForest package in R statistical software was employed.

For construction, the predictive ability of RFs increases as the inter tree correlation decreases. Thus, a large number of predictors can provide increased generalization capacity, like in our case where the predictors used were initially 23. Furthermore, performance of random forests depends strongly on m, the number of parameters to be used in each split for each node creation. If m is relatively low, then both inter tree correlation and strength of individual tree decreases. Thus, it is critical for the overall performance of random forests to find the optimal of m through a tuning algorithm. Breiman (2001) in his original work suggests three possible values for (m) of the following form: $\frac{1}{2}\sqrt{\mathrm{m}}$, $\sqrt{\mathrm{m}}$, and $2\sqrt{\mathrm{m}}$, where m equals 23 in our case. In this work, we followed a grid tuning approach using a cross validation method. That is, we equally split the short in-sample dataset into two distinct samples, namely a training subsample and a cross validation subsample. The grid search was two dimensional taking a range of values for (m) and for the number of trees to generate. The random forests classifier was tested thoroughly on the cross validation dataset to avoid over-fitting and to increase generalization during the tuning process. One significant theoretical feature is that this method provides consistency in estimation as the number of trees increases (Denil, Matheson, & De Freitas, 2013). We employed the MSE metric for selecting the most efficient parameterization. In the optimized model, the number of trees is 650. Fig. 4 depicts the MSE error as the number of trees increases. It is evident that as the number of trees approaches 650, the MSE flattens.

In Fig. 5, we present for each financial indicator, its importance for classification.[3] The ranking is based on two criteria: Mean Square Error and Node Purity. The left part of chart, pertaining to the MSE, can be 'interpreted' as follows: if a predictor is important, then assigning other values for that predictor, permuting this predictor's values over the dataset, should have a negative influence on overall model prediction. In other words, using the same model to predict from data that is the same except for this variable, should give worse predictions. Thus, this chart compares MSE of the original dataset with the 'permuted' dataset. The values of the variables are scaled to be comparable across all variables. The right part of the chart presents node impurity. That is, at each split, we calculate how much this split reduces node impurity, calculated as the difference between Residual Sum of Squares (RSS) before and after the split. This is summed over all splits for that variable, over all trees. Overall, our results indicate that capital indicators such as Leverage Ratio and CAR exhibit high importance along with ROE, NPL and CFEA (Cost of funding earning assets).

In Fig. 6, the forest floor main effect plots of random forest are shown. These plots map the structure of bank failure prediction model on the basis of bank specific regulatory and financial characteristics. The plots are arranged according to variable importance, where X-axis shows variable values and Y-axis the corresponding cross validated feature contributions. The goodness-of-visualization is evaluated with leave-one-out k-nearest neighbor estimation (black line, R^2 values), and the graphical representation is based on the forestfloor package in R.

In particular, we present the charts for the variables that interact mostly, based on R-squared measure, with the dependent variable. The flatter the line the weaker is the relation between each regressor and the dependent variable. The parallel color gradients identify interactions between the regressors. The graphs point the nonlinear negative relation between capital measures, such as Leverage and Capital Adequacy Ratio, as well as Retained Earnings to Equity with the default intensity of US banks. Equity metrics as measured by ROE and ROE_DFS provide also significant interaction with bank default. Furthermore, high profitability reduces substantially the probability that a bank will fail. Finally, it seems that asset quality, as measures by the NPL ratio, plays a less significant role in predicting bank failures in comparison to capital and equity measures.

[3] The plot presents the 16 more significant variables out of the 23 included in the model.

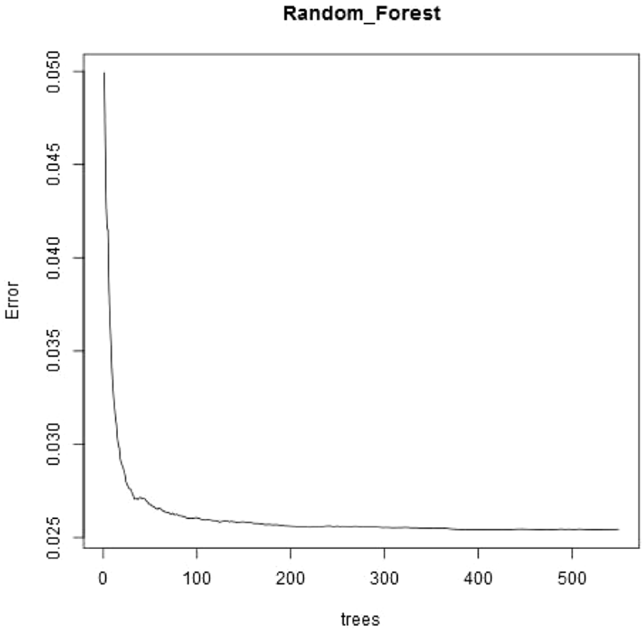


Fig. 4. Random Forests error relatively to the number of trees.

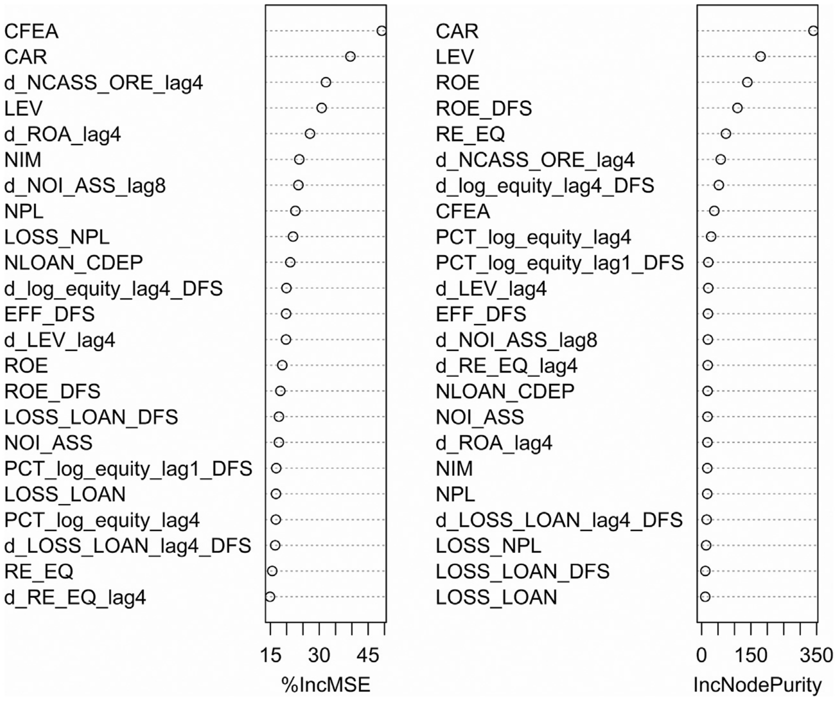


Fig. 5. Random Forests Variable Importance Plot.

4.4. Support vector machines (SVMs)

SVMs are a family of non-linear, large-margin binary classifiers. SVMs estimate a separating hyperplane that achieves maximum separability between the data of the two modeled classes (Vapnik & Vapnik, 1998). A significant number of studies point the usefulness of SVMs in credit rating systems (Harris, 2015; Huang, 2009), since they reduce the possibility of overfitting and alleviate the need of tedious cross-validation for the purpose of appropriate hyper parameter selection. The main drawbacks of

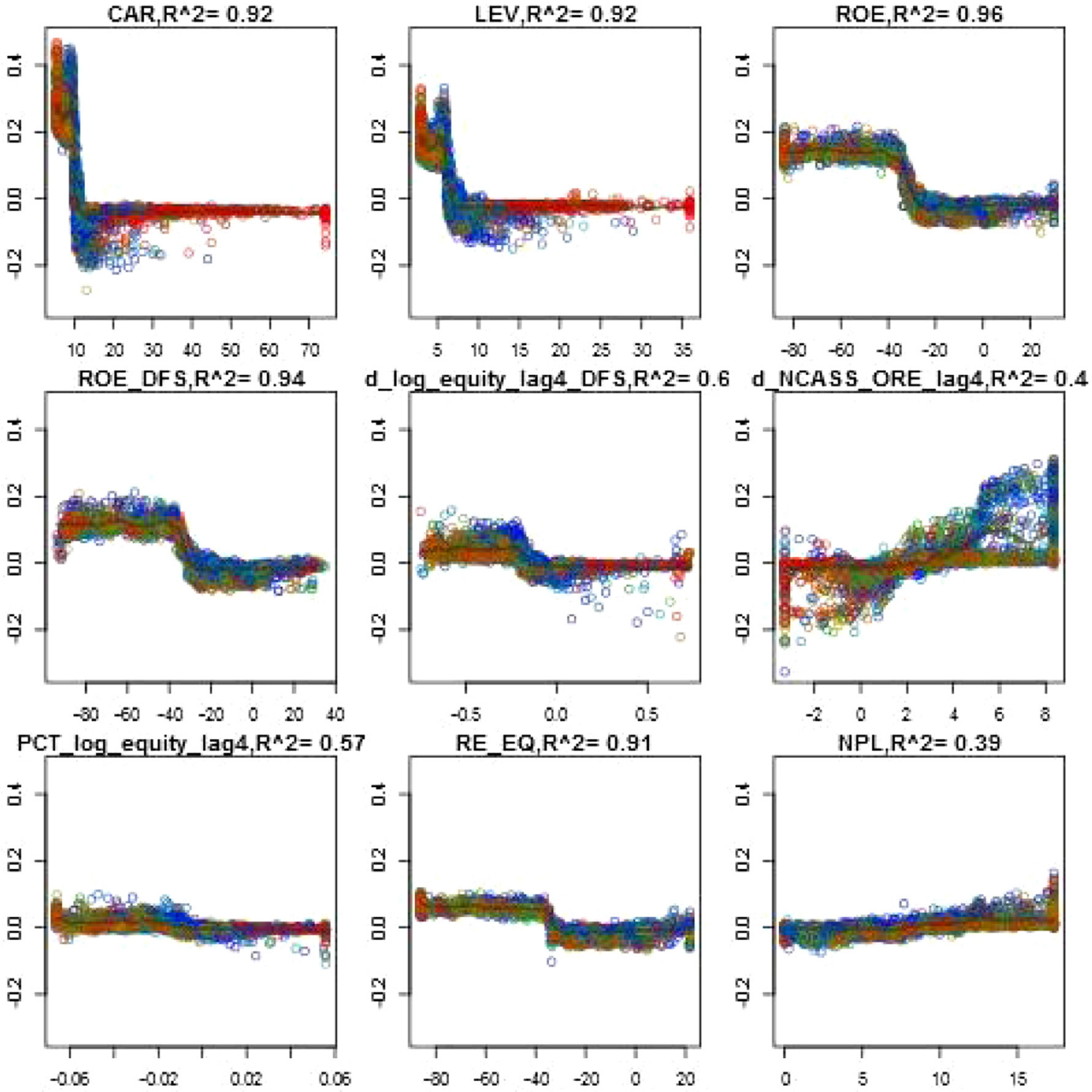


Fig. 6. Random Forests important variables effect.

SVMs stem from the fact that they constitute black-box models, thus limiting their potential of offering deeper intuition and visualization of the obtained results and inference procedure.

In this study, we evaluate soft-margin SVM classifiers using linear, radial basis function (RBF), polynomial, and sigmoid kernels, and retain the model configuration yielding optimal performance.

For selecting the proper kernel, we exploit the available validation set. To select the hyperparameters of the evaluated kernels as well as the cost hyperparameter of the SVM (related to the adopted soft margin), we resort to cross-validation. The candidate values of these hyperparameters are selected based on a grid-search algorithm (Vapnik & Vapnik, 1998). We implemented this model in R using the kernlab package along with the grid-search functionality included in the e1071 package (Tune routine). The SVM selected is of C classification type with a Radial Basis "Gaussian" kernel.

In short, to improve the performance of the support vector regression, we need to select the best parameters for the model. The process of choosing these parameters is called "hyper-parameter" optimization, or model selection. Fig. 7 presents the results of a grid search for different couples of cost (y-axis) and gamma (x-axis) for fine tuning the parameters of the SVM model.

A large misclassification cost parameter gives low bias, as it penalizes the cost of misclassification a lot. However, it leads to high variance, so that the algorithm is forced to explain the input data stricter and potentially overfit. Whereas, a small misclassification cost allows more bias and lower variance. Regarding gamma, when it is very small the model is too constrained and cannot capture

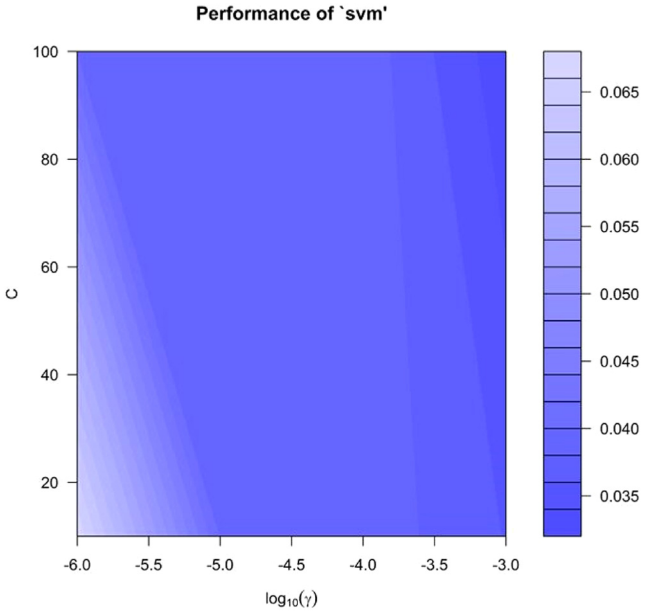


Fig. 7. Plot of the Parameter tuning for SVM. Sampling method: 10-fold cross validation.

the complexity of the data. In this case, two points can be classified the same, even if they are far from each other. On the other hand, a large gamma means that two points are classified the same, only if they are close to each other. We employed the RMSE metric in the validation sub-samples for selecting the most efficient parameterization. Based on Fig. 7, the parameters closer to the upper right region (darker part of the figure) lead to smaller RMSE and so those were employed in the optimization process of the SVM specification.

4.5. Neural networks (NN)

Neural networks is a well-known machine learning technique that is broadly used in credit rating classification problems. Classification problems are characterized by the availability of a big datasets, many explanatory variables, and the possibility of noise existence in the data. Experimental results offer evidence that neural networks can capture complex non-linear patterns in the data analyzed. Current literature offers numerous structural variations of Neural Networks depending on the number of layers, the flow of information and the algorithms used to train them.

The most often setup is composed by three layers. The input layer in which all candidate variables are imported as a high dimensional vector. The hidden layer where the information is transformed and processed forward to the output layer via non-linear functions such as sigmoid. The output layer in which the signal from individual neurons is aggregated to complete the supervised learning function.

To produce the benchmark neural network model, we trained on a train and a validation set, both belonging to the in sample dataset, various structures of multi-layer perceptron neural network (MLP). The structures investigated depended on the number of hidden layers, in our case 1–3, as well as the number of neurons in each layers. The latter number varied from 2 through 10, following the rule of thumb that each layer must be composed of fewer neurons than the previous one in the NN queue. In parallel for initializing the weights, we employed stochastic gradient descent algorithm which uses random weights in selecting a starting point for the search and in the progression of the search. Specifically, stochastic gradient descent requires that the weights of the network are initialized to small (close to zero) random values. Randomness is also used during the search process in the shuffling of the training dataset prior to each epoch, which in turn results in differences in the gradient estimate for each batch. We employed the accuracy metric (AUROC) in the validation sub-sample for selecting the most efficient parameterization.

The candidate neural network models were trained using the back propagation supervised learning algorithm. That is, each input along with the desired output fed into the model, while the weights at both the hidden and output layers are adjusted so that the actual output corresponds to the desired output using the gradient descent optimization method. The error between actual vs predicted values of the dependent variable decreases in every iteration of the algorithm. The iterative process stops when the error falls below a predefined threshold, in our case 0.01. The MSE of the performance of each NN on the validation sample was used to find the best candidate model. Through this process, the optimal NN that offered the best generalization capacity on the in sample dataset, while avoiding overfitting of the training data was selected. The best performer was a complete 2 layer back propagation Multilayer Perceptron (MLP) neural network with hidden neurons. To increase overall performance of the neural network the variables were transformed

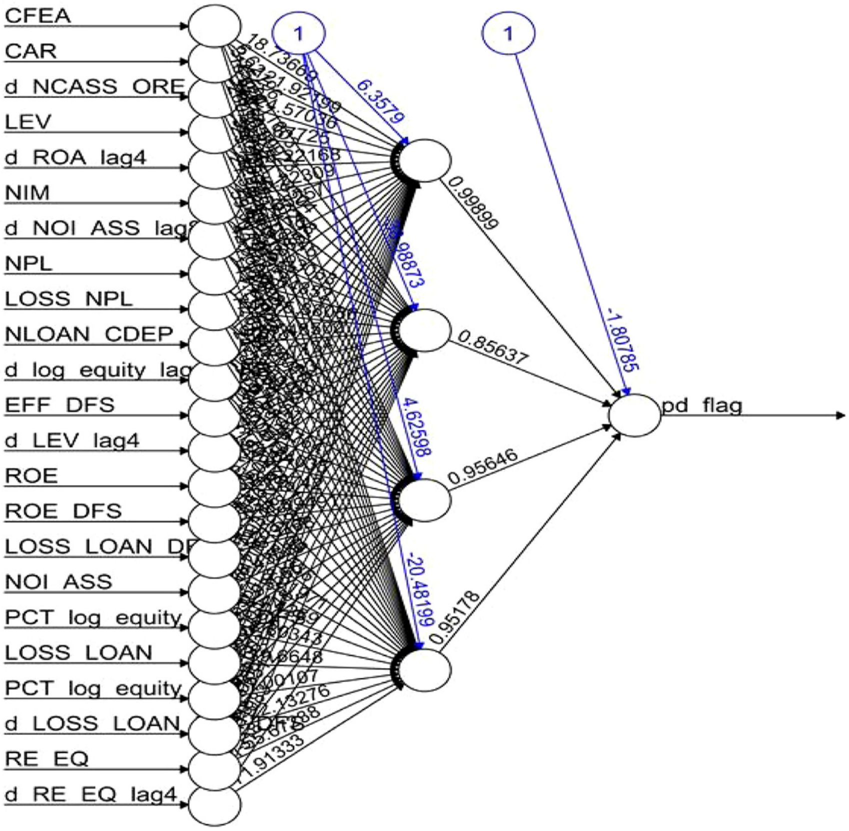


Fig. 8. Optimized Neural Network Depiction.

to take values in the continuous interval of [0,1]. Along with the different structures explored during the training process, further tuning was performed for various step sizes (learning rate), momentum values, the number of processing elements (nodes) in the hidden layer(s) and the maximum number of learning iterations (epochs) to avoid over-fitting (early stopping). The sigmoid was assumed as a process activation function for each node. Training and optimization of the neural networks were performed in R using the Neuralnet package.

Although neural networks are difficult to interpret and their training process can take longer than Random Forests, their performance provides a good benchmark to validate other methodologies. Fig. 8 depicts the structure of the optimized neural network. In particular, the input layer to the left side of the plot corresponds to the vector of explanatory variables used. Then, the hidden layer in which the data processing/transformation occurs follows in the middle of the plot. Finally, the output layer to the right part of the plot generates a prediction of the dependent variable.

4.6. *Conditional inference random forest (CRF)*

Random Forests comprising conditional Inference Trees take into account the distributional properties of the measures when distinguishing between a significant and an insignificant improvement in the information measure. More precisely, Conditional Inference Trees test the global null hypothesis of independence between any of the input variables and the response variable. If this hypothesis is not rejected, the algorithm stops. Otherwise, the algorithm selects the input variable with the strongest association to the response variable. This association is measured by a *p*-value, corresponding to a test for the partial null hypothesis of a single input variable and the response variable based on permutation tests. That is, by calculating all possible values of a test statistic under rearrangements of the labels on the observed data points. We implemented Conditional Inference Random Forest Trees using the party package in R, which is based on a unified framework for conditional inference, or permutation tests, developed by Strasser and Weber (1999). Hyperparameter tuning for the Conditional Inference Random Forest followed the same process described in Chapter 4.3, i.e., the work we followed a grid tuning approach using a cross validation method splitting equally the short in-sample dataset into two distinct samples, namely a training subsample and a cross validation subsample. The grid search was two dimensional taking a range of values for m (random subsamples of features) and for the number of trees to generate. We employed the accuracy metric (AUROC) in the validation sub-sample for selecting the most efficient parameterization.

In Fig. 9, we present the variance importance plot of Conditional Inference Random Forest, according to the

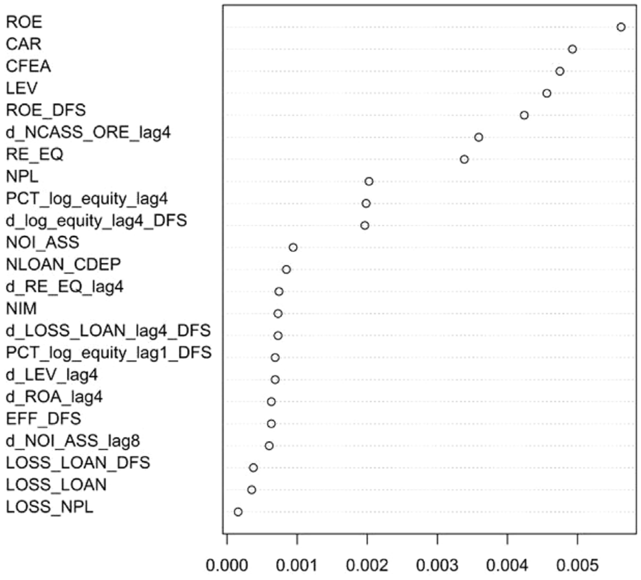


Fig. 9. Variance importance plot of Conditional Inference Random Forest (CRF).

significance of each variable in reducing MSE. Our results indicate that profitability indicators, such as Return on Equity (ROE) and Cost of Funding Earning Assets (CFEA), along with capital indicators, like Capital Adequacy Ratio (CAR) and Leverage Ratio (LEV), exhibit the highest importance in explaining the response variable.

5. Model validation

To assess the robustness of our approach, we perform a thorough validation procedure. More precisely, we report the performance results obtained from the experimental evaluation of our method, in terms of short in-sample fit, out-of-sample performance, out-of-time performance and in terms of evaluating the model's predictive ability on the full in-sample dataset.

5.1. Validation measures

Classification accuracy, as measured by the discriminatory power of a rating system, is the main criterion to assess the efficacy of each method and to select the most robust one. In this section, we present a series of metrics that are broadly used for quantitatively estimating the discriminatory power of each scoring model.

Considering that a bank failure is not as common as a corporate default, there is a predominance of solvent banks in our validation subsamples. That is, our dataset is strongly imbalanced, in the sense that it is not evenly split between low and high risk financial institutions. Imbalanced data learning is one of the most challenging problems in data mining. The skewed class distribution of such datasets may provide misleading classification accuracy based on common evaluation measures. We therefore used a PD cutoff point according to which we separate the predicted healthy and failed banks. After thoroughly examining different values for this parameter and based on the performance of the classification in the short in-sample dataset used for model development, we set the cut off criterion to be 50%. Translating sensitivity and specificity as the accuracy of positive (i.e. solvent) and negative (i.e. insolvent) cases respectively, we use a set of combined performance measures that aim to provide a more credible evaluation. As aptly noted by Bekkar, Kheliouane, and Taklit (2013), these measures overcome any issues of model performance misinterpretation, which can arise in cases where one uses traditional model performance metrics that are based on overall model accuracy. In particular, sensitivity and specificity are defined as follows:

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

where: TP = True Positive, the number of positive cases (i.e. solvent) that are correctly identified as positive,
TN = True Negative, the number of negative cases (i.e. insolvent) that are correctly identified as negative cases,
FN = False Negative, the number of positive cases (i.e. solvent) that are misclassified as negative cases (i.e. insolvent),
FP = False Positive, the number of negative cases (i.e. insolvent) that are incorrectly identified as positive cases (i.e. solvent).

More precisely we focus on the following measures

- **G-mean**: The geometric mean G-mean is the product of sensitivity and specificity. This metric indicates the balance between classification performances on the majority and minority class.

  $$G = \sqrt{sensitivity * specificity}$$

  A poor performance in prediction of the positive cases will lead to a low G-mean value, even if the negative cases are correctly classified from the algorithm.

- **LR-**: The negative likelihood ratio is the ratio between the probability of predicting a case as negative when it is actually positive, and the probability to predict a case as negative when it is truly negative.

  $$LR- = \frac{1 - sensitivity}{specificity}$$

  A lower negative likelihood ratio means better performance on the negative cases, which is the main point of interest in this study as we model bank failures.
- **DP**: Discriminant power is a measure that summarizes sensitivity and specificity.

  $$DP = \frac{\sqrt{3}}{\pi}\left[\log\left(\frac{sensitivity}{1 - sensitivity}\right) + \log\left(\frac{specificity}{1 - specificity}\right)\right]$$

  For DP values higher than 3 then the algorithm distinguishes well between positive and negative cases.
- **BA**: The balanced accuracy is the average of Sensitivity and Specificity. If the classifier performs equally well on either class, this term reduces to the conventional accuracy measure.

  $$BA = \frac{1}{2}\,(sensitivity + specificity)$$

  In contrast, if the conventional accuracy is high merely because the classifier takes advantage of good prediction on the majority class (i.e. dominant in terms of events, solvent banks in our case), the balanced accuracy will drop thus signaling any performance issues. That is, BA does not disregard the accuracy of the model in the minority class (i.e. insolvent banks in our case).
- **Youden's** γ: Youden's index is a linear transformation of the mean sensitivity and specificity therefore it is difficult to interpret.

  $$\gamma = sensitivity - (1 - specificity)$$

  As a general rule, a higher value of Youden's γ indicates better ability of the algorithm to avoid misclassifying banks.
- **WBA1**: Is a weighted balance accuracy measure that weights specificity more than sensitivity (75%/25%).
- **WBA2**: Is a weighted balance accuracy measure that weights sensitivity more than specificity (75%/25%).
- **AUC**: The area under the ROC[4] curve (Area Under Curve, AUC) is a summary indicator of the performance of a classifier into a single metric. The AUC can be estimated through various techniques, the most commonly used being the trapezoidal method. This is a geometrical method based on linear interpolation between each point on the ROC curve. The AUC of a classifier is equivalent to the probability that the classifier will rank a randomly chosen positive instance higher than a randomly chosen negative instance. In practice, the value of AUC varies between 0.5 and 1 with a value above 0.8 to denote a very good performance of the algorithm.

These measures are used to derive a full spectrum conclusion regarding the classification power of each model relative to the others. Even though there could be an amount of correlation between metrics, we employ them all to classify correctly the employed techniques based on their predictive performance even though in some cases the differences may be marginal. It should be noted that because there is a much higher financial stability risk when insolvent banks are not identified, rather than when solvent banks are wrongly classified as insolvent, we consider that the validation measures focusing on the proper identification of insolvent banks are the preferred ones for our concrete case. In particular, we use the results in the out-of-time sample to draw our conclusions for model performance, by simultaneously focusing on the following indicators: (i) LR-, a lower value of which indicates better performance on the negative cases, (ii) BA that does not disregard the accuracy of the model in the minority class (i.e. insolvent banks), and (iii) WBA1 that weights specificity (identification of insolvent banks) more than sensitivity (identification of solvent banks).

5.2. Validation results

Our original development sample contains 101.641 observations that can be divided into 100.068 solvent and 1573 insolvent cases, and we call it "Full in sample". The overbalanced nature of our dataset, which presents a preponderance of solvent banks (i.e. good cases), does not facilitate the training of complex techniques. To this end, we created a new training sample (called "Short in sample"), including randomly chosen 10% of the good cases and all the bad cases. So, the final training sample used to develop our models contains 10.001 good cases and 1.572 bad cases, reaching 11.573 observations in total. For the purpose of fine tuning the parameters of the random forests and neural networks specifications, we further equally divide the short in-sample dataset into training and validation sub-samples (50% each). In short, the term "Short in sample" refers to the more balanced dataset, while the term "Full in sample" refers to the sample that includes all the good cases. As already mentioned, the "Out-of-sample" dataset refers to the 20% randomly selected observations covering the years 2008–2012. Finally, the "Out-of-time sample" refers to the data for the years 2013–2014.

In terms of performance metrics in the short in-sample, we noted in Table 1 that Neural Networks and Random Forests provide the best fit, while Logit and LDA are underperforming across all performance metrics. It should be noted that the overperformance of Neural Networks in the in-sample dataset, may have been the results of a potential over fitting, being the usual pitfall of using Neural Networks. To this end, we base our core conclusions on the predictive performance of each model on the out-of-sample and out-of-time datasets. When examining the out-of-sample (Table 2) performance, RFs are again the best across almost all performance measures, while

[4] Receiver Operating Characteristic curve.

Table 1
Short in sample performance metrics.

	Logit	LDA	RF	SVM	NN	CRF
AUROC	0.980	0.973	0.989	0.981	0.984	0.991
G-mean	0.898	0.884	0.921	0.898	0.923	0.914
LR-	0.183	0.209	0.139	0.184	0.137	0.156
DP	3.116	2.971	3.255	3.181	3.356	3.312
BA	0.902	0.889	0.923	0.902	0.925	0.916
Youden	0.804	0.778	0.846	0.804	0.851	0.833
WBA1	0.943	0.936	0.953	0.944	0.955	0.951
WBA2	0.861	0.842	0.893	0.860	0.895	0.881

Table 2
Out-of-sample performance metrics.

	Logit	LDA	RF	SVM	NN	CRF
AUROC	0.990	0.983	0.990	0.992	0.980	0.989
G-mean	0.919	0.905	0.934	0.916	0.922	0.907
LR-	0.144	0.169	0.113	0.150	0.130	0.165
DP	3.239	3.099	3.352	3.268	3.051	3.147
BA	0.921	0.908	0.935	0.919	0.923	0.910
Youden	0.842	0.816	0.871	0.837	0.847	0.821
WBA1	0.952	0.945	0.959	0.952	0.948	0.947
WBA2	0.890	0.871	0.912	0.886	0.898	0.874

Table 3
Out-of-time performance metrics.

	Logit	LDA	RF	SVM	NN	CRF
AUROC	0.990	0.974	0.976	0.993	0.990	0.965
G-mean	0.741	0.824	0.862	0.819	0.862	0.838
LR-	0.452	0.321	0.255	0.329	0.255	0.296
DP	3.684	3.590	3.793	3.804	3.722	3.668
BA	0.774	0.839	0.871	0.835	0.871	0.851
Youden	0.548	0.677	0.743	0.670	0.742	0.702
WBA1	0.886	0.918	0.934	0.916	0.934	0.924
WBA2	0.662	0.759	0.809	0.754	0.809	0.778

Table 4
Full in sample performance metrics.

	Logit	LDA	RF	SVM	NN	CRF
AUROC	0.980	0.973	0.998	0.981	0.981	0.990
G-mean	0.898	0.884	0.992	0.897	0.926	0.914
LR-	0.184	0.209	0.000	0.185	0.125	0.153
DP	3.079	2.960	Inf	3.115	3.124	3.202
BA	0.901	0.889	0.992	0.901	0.927	0.916
Youden	0.803	0.777	0.984	0.802	0.854	0.832
WBA1	0.942	0.935	0.988	0.943	0.951	0.950
WBA2	0.860	0.842	0.996	0.859	0.903	0.883

logistic regression seems also to be an adequate tool for assessing bank failure probability as it is ranked second. Regarding out-of-time performance, presented in Table 3, Random Forests and Neural Networks provide again the best fit, with the former method exhibiting marginally better performance in 5 criteria and better performance in 1 criterion relative to the latter. Logistic regression performs poorly in the out-of-time period, as it shows the worst performance in 6 out of 8 criteria. Finally, when assessing the discriminatory power of our specifications in the "Full in-sample", Random Forests is the dominant methodology. That is, in Table 4, we note that Random Forests outperform across all performance metrics.

Summarizing the results in all samples, it is evident that the proposed RF rating system exhibits higher discriminatory power than all the considered benchmark models when taking into account the skewness of the data. More importantly, the obtained performance is more stable and more consistent across all test samples, resulting in lower performance variability. Another interesting finding stemming from our results is that NN perform relatively well in the "in-sample" and "out-of-time" samples. We point though that the non-anticipated failure of a bank may come at a much higher cost for the economy environment relative to a corporate default. In the former case, depositors could start concern themselves about the safety of their savings, banks may face liquidity problems generated by deposit outflows; thus, when banks cut off business lines, the business activity faces a slowdown and generally the economic environment is destabilized. It is therefore imperative for supervisory purposes to achieve the maximum possible accuracy when setting an Early Warning System for bank failures.

To further illustrate the higher discriminatory power of the proposed statistical model, we present in Fig. 10 the corresponding ROC curves corresponding to the four datasets analyzed. Receiver operating characteristic curve, or ROC curve, illustrates the performance of a binary classifier system as its discrimination threshold is varied. The curve is created by plotting the true positive rate against the false positive rate at various threshold settings. It shows the tradeoff between sensitivity and specificity, as any increase in sensitivity will be accompanied by a decrease in specificity. The closer the curve follows the left-hand border and then the top border of the ROC space, the more accurate is the modeling approach. The ROC curve across all samples is approaching the perfect classification line, thus supporting the high degree of efficacy and generalization of the proposed RFs rating system.

5.3. Bootstrapping (stability)

To further assess the stability of the developed RFs-based rating system, we perform a bootstrapping approach on the joint dataset consisting of the out-of-sample and out-of-time parts. Specifically, we generate random samples with replacement from the above-mentioned dataset with a balanced mix between good and bad banks (i.e. 50% and 50% respectively) to estimate all the discriminatory power statistics as described in Section 5.2. The experiment is performed with 10.000 repetitions. Then, for each one of the performance measures, we construct confidence intervals to assess its stability as well as the existence of any bias in the prediction of the proposed RFs model. The results, reported in Table 5, denote that the RFs' performance is stable. In particular, each performance metric is distributed in a narrow range around the whole sample performance metric. Hence, there is strong evidence for the generalization capacity of RFs, regardless of the composition mix between insolvent and solvent financial institutions. In addition, our empirical results support the efficacy of the model to capture possible outliers and nonlinear behaviors in the underlying sample, without significant deterioration in its ability to

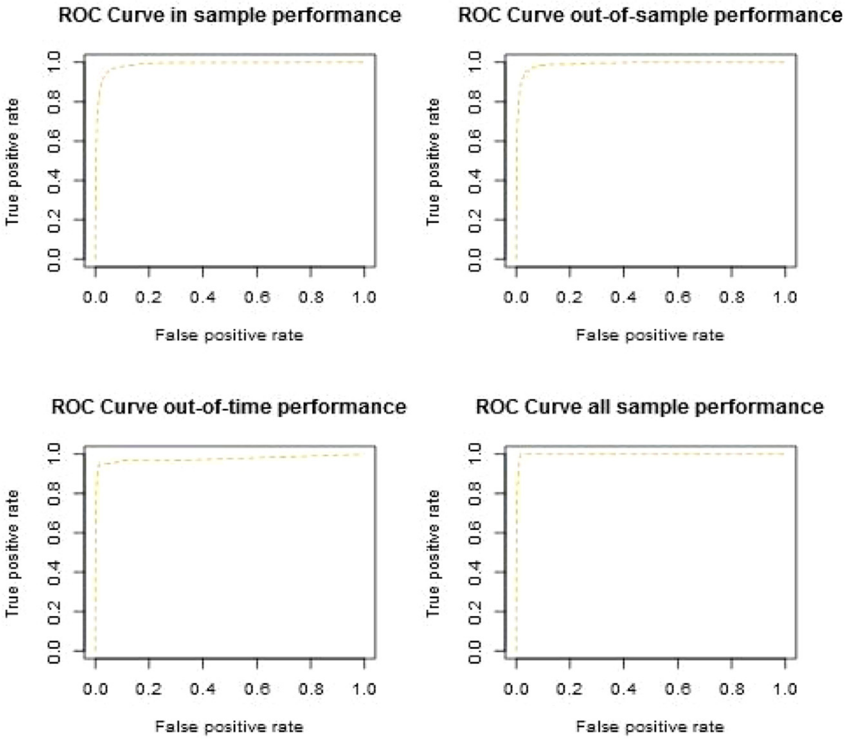


Fig. 10. ROC curve performance evaluation of Random Forests.

Table 5
Performance measure stability of RFs.

	Mean	−CI 99%	+CI 99%
AUROC	0.9886	0.9886	0.9887
G-mean	0.9183	0.9181	0.9184
LR	0.1511	0.1509	0.1514
DP	3.6619	3.6538	3.6701
BA	0.9210	0.9209	0.9212
Youden	0.8421	0.8418	0.8423
WBA1	0.9565	0.9564	0.9566
WBA2	0.8856	0.8854	0.8857

discriminate. In short, the bootstrapping exercise verifies the stability of RFs across different types of samples.

5.4. Variable importance

There is a big debate in the current literature regarding the level of significance of the regressors used in predicting bank failures under the CAMELS framework. Variables related to capital, asset quality and earnings most of the times are significant in a typical CAMELS' based model (Poghosyan & Čihák, 2009). Liquidity-related variables are also sometimes included as significant indicators in various models (Cole & Wu, 2017; Mayes & Stremmel, 2012), while indicators related to Management and Sensitivity to Market appear to be less significant in predicting bank insolvencies (Betz et al., 2014; Mayes & Stremmel, 2012). However, there is neither a unanimous conclusion on the significance of certain indicators across studies, nor all statistically significant indicators retain their importance up to the default event. For example, Cole and White (2011) show that the Equity to Assets ratio losses its predictive power when move back more than two year prior the default date. Betz et al. (2014) showed that Reserves to Impaired assets ratio and RoE are not statistically important at all.

To obtain further insights on the importance of each explanatory variable in predicting bank insolvencies under the CAMELS framework, we use all the benchmark models developed in this study to perform a comparative analysis. The aim is to produce a ranking among all explanatory variables used as inputs in each one of the statistical models developed. The results of this analysis could provide important feedback in an expert judgment approach, in which the weighting is done using qualitative criteria. We have not included in this comparative analysis all variables selected through the variable reduction space process described in Chapter 3, but we focus mainly on those variables that are more commonly used by supervisory authorities under the CAMELS assessment framework. The final ranking was produced by applying the "leave one out" method. That is, we trained each model by excluding one candidate variable at a time, and then, we measured the performance of the resulting model. This approach was applied uniformly across all models, and the results are summarized in Table 6.

We assessed the relative importance of each variable based on its marginal contribution to AUROC metric.[5].

[5] For Random Forests, the ranking is based on the incMSE% variable importance plot

Table 6
Covariate importance ranking per model.

	Logit	LDA	RF	SVM	NN	CRF	Average score
log(equity)(−4)%	4	3	11	3	7	6	5,7
d(LEV)(−4)	3	5	8	5	6	8	5,8
LOSS_LOAN_DFS	7	6	10	1	1	10	5,8
d(NCASS_ORE)(−4)	9	10	2	8	4	4	6,2
d(ROA)(−4)	11	8	4	6	8	9	7,7
LEV	2	2	3	10	5	3	4,2
NLOAN_CDEP	6	9	7	2	3	7	5,7
NPL	8	7	5	9	11	5	7,5
LOSS_NPL	10	11	6	11	9	11	9,7
ROE	5	4	9	4	10	1	5,5
CFEA	1	1	1	7	2	2	2,3

(1: Highest importance, 11: Lowest importance).

Table 7
Dominance of capital adequacy ratio vs leverage ratio per model.

	CAR	LEV
Logit		✓
LDA		✓
RF	✓	
SVM	✓	
NN	✓	
CRF	✓	

(✓: indicates the dominant ratio for each model).

Specifically, we excluded each variable, in turn, from each model and we measured the loss in AUROC for each specification. We ranked first the variables that led to the largest loss in AUROC metric. We can note in Table 6 that for most models, Cost of Funding Earnings Assets (CFEA) and leverage ratio (LEV) are leading indicators in bank failure forecasting. Indeed, CFEA is by far the most important indicator across all models, as it is on average ranked in position 2.3. Apart from CFEA, additional earnings related indicators such as Return on Equity (ROE) are also important determinants. On the other hand, Loan loss allowance to noncurrent loans (LOSS_NPL) and Noncurrent loans to loans (NPL) appear to be the ones with the lower importance across all models, as they are ranked in position 9.7 and 7.5 respectively. Furthermore, Liquidity risk as measured by the Net Loans to Core Deposits (NLOAN_CDEP) and Asset Quality as measured by the distance from the sector of Loss allowance to loans (LOSS_LOAN_DFS) have increased significance in the SVM and NN models. Thus, the results of the variable importance analysis suggest that profitability (CFEA) and capital (LEV) indicators are the most important drivers across all models.

There is inconclusive evidence in the current literature regarding the superiority of certain indicators falling under the category of Capital assessment in predicting bank failures. On the one hand, Mayes and Stremmel (2012) claim that a simple leverage ratio (unweighted) is a better predictor than capital adequacy ratio (risk weighted). On the other hand, Cole and Wu (2017) have identified that those related to capital adequacy are the most important predictors of bank failures. To this end, we utilize the models developed in this study to further explore the discriminatory power of a simple leverage (LEV) relative to a risk-weighted capital adequacy ratio (CAR). This analysis will equip us with a deeper insight on the regulatory aspect of these indicators. The comparison is made based on AUROC performance metric for Logit, LDA, SVM, NN and based on MSE% variable importance plot for RF and CRF. The results of this comparison are summarized in Table 7. Capital adequacy ratio outperforms leverage ratio when entered as covariate in more complex models, such as Random Forests, Support Vector Machines, Neural Network and Random Forest of Conditional Inference Trees. Whereas, in simpler models such as logistic regression and LDA, the leverage ratio is the dominant covariate. In short, our analysis implies that the importance of the one indicator relative to the other is purely model driven. That is, any conclusions are strongly related to the sophistication of the underlying model used to predict bank failures.

For illustration purposes, we present in Table 8 the two different Logit models, corresponding to the inclusion of LEV and CAR variables respectively. According to the AIC and BIC information criteria, the Logit model that incorporates the Leverage Ratio (LEV) has better fit.

5.5. Implementation on European banks dataset

To test the generalization of our approach, we apply the selected Random Forests technique in the European banking system. Essentially, we make use of the Random Forests specification in creating an Early Warning System of bank failures in Europe. Random Forests has been utilized in the area of credit risk attempting to model the underlying dynamics that drive a company to default on its obligations. Specifically, Yeh, Chi, and Lin (2014) combined random forests and rough set theory to address the problem of prediction of a firm's ability to remain a going concern. Wu, Hu, and Huang (2014) constructed a corporate credit rating prediction model by using RFs to evaluate financial variables. Finally, Random Forests can be proven as a really useful regulatory tool for monitoring the stability of the financial system. To this end, Alessi and Detken (2014) proposed RFs to form an early warning system for macroprudential purposes, by identifying excessive credit growth and leverage that could potentially jeopardize the stability of a banking system.

This is a strong test for classification purposes as European banking system is characterized by significant disparity in financial institutions driven by country macroeconomic specificities. More specifically, we employ the selected Random Forests specification for calculating the default Probability for 173 European banks based on year end-2015[6] accounting and regulatory data[7]. To benchmark our results, we mapped our PDs to rating classes based on lower bound PD thresholds described in 2016 Moody's rating methodology document.

[6] We did not take into account variables based on lag differences greater than 4 (pre 2014 data) and we also excluded the variable related to retained earnings to equity ratio, as it was not available in a quarterly basis. Because this variable is ranked last in the Random Forests variable importance plot, we do not expect any bias in our results.

[7] Source: SNL

Table 8
Comparing Logistic regression models with different capital-related ratios.

Coefficients:	Logit (LEV)	Logit (CAR)
(Intercept)	-2.696^{***}	-2.809^{***}
	(0.349)	(0.374)
CFEA	1.008^{***}	0.999^{***}
	(0.0077)	(0.076)
ROE	-0.017^{***}	-0.021^{***}
	(0.003)	(0.003)
LOSS_NPL	−0.0004	−0.0001
	(0.000)	(0.000)
NPL	0.068^{***}	0.077^{***}
	(0.018)	(0.018)
NLOAN_CDEP	0.007^{***}	0.002
	(0.002)	(0.002)
LEV	-0.449^{***}	
	(0.032)	
d_ROA_lag4	-0.069^{*}	−0.058
	(0.038)	(0.037)
d_NCASS_ORE_lag4	0.137^{***}	0.110^{***}
	(0.026)	(0.026)
LOSS_LOAN_DFS	0.147^{**}	0.160^{***}
	(0.057)	(0.057)
d_LEV_lag4	-0.096^{***}	-0.074^{***}
	(0.030)	(0.028)
PCT_log_equity_lag4	-19.902^{***}	-23.342^{***}
	(3.289)	(3.227)
CAR		-0.246^{***}
		(0.021)
AIC	2457.652	2518.936
BIC	2545.650	2606.934
Log Likelihood	−1216.826	−1247.468
Deviance	2433.652	2494.936
Num. obs.	11307	11307

$^{***}p < 0.01$.
$^{**}p < 0.05$.
$^{*}p < 0.1$.

Table 9
Classifying High-Risk banks by European countries based on RFs credit rating system.

	High risk banks	Banks in sample
AT	0	5
BA	0	2
BE	1	2
BG	1	2
CH	0	17
CY	0	2
CZ	0	1
DE	2	8
DK	0	15
ES	1	8
FI	0	2
FR	3	11
GB	0	10
GE	0	1
GR	3	5
HR	2	3
HU	2	2
IE	0	3
IT	7	15
LI	0	1
MD	1	1
MK	0	2
MT	0	3
NO	0	12
PL	0	11
PT	1	2
RO	1	2
RS	0	1
RU	5	5
SE	1	3
SK	0	4
TR	7	9
UA	3	3
Total	**41**	**173**

We evaluated the concordance of our ranking with the respective Moody's ranking[8] by calculating Kendal's tau, Spearman's rho and the classical Fisher correlation coefficient. Seeing that Moody's ratings take into account the sovereign rating of a bank's resident country, we adapted our ranking for sovereign rating in a similar way as described in Moody's respective document.[9] Our credit rating scale has 67% Spearman's Rho, 59% correlation and 47% Kendal's Tau with the Moody's Rating system, thus, verifying the high positive concordance. In Table 9 the number of High-Risk banks is shown by country. A bank is defined as High Risk when its Probability of Default, as calculated by the RFs specification, is larger than 25%.

Focusing on Eurozone, we note that countries experiencing prolonged macroeconomic deterioration, which has eroded local banks' capital and increased non-performing exposures show the highest relative number of "High Risk banks". Specifically, in Greece (GR) 3 out of 5 banks and in Italy (IT) 7 out of 15 banks are classified as "High Risk". On the other side, our results confirm that stronger Eurozone economies are accompanied by resilient banking systems, so that in Germany (DE), France (FR), Austria (AT), Finland (FI) and Belgium (BE) only 6 out of 28 banks are classified as "High Risk". Finally, countries regaining competitiveness such as Ireland (IE) and Spain (ES) also exhibit relatively low levels of risky banks, that is, 0 out 3 and 1 out of 8 respectively.

Outside Eurozone, strong economies such as Switzerland (CH), Norway (NO), Denmark (DK), Sweden (SE) and United Kingdom (GB) exhibit close to zero levels of "High Risk Banks". On the contrary Ukraine (UA) that suffered from a military conflict, Turkey (TR) who lost 13% of its GDP in the last 3 years and Russia (RU) present a large proportion of "High Risk" Banks (7 out of 9 in Turkey, 3 out of 3 in Ukraine and 5 out of 5 in Russia).

We also point the zero number of "High Risk Banks" in Poland (PL) and Slovakia (SK), whereas we also gain insight on the fragile Portugal (PT) banking sector. We finally remain cautious on the results in Eastern European countries (Bosnia-Herzegovina (BA), Bulgaria (BG), Czech Republic (CZ), Georgia (GE), Croatia (HR), Hungary (HU), Moldova (MD), FYROM (MK), Romania (RO) and Serbia (RS)) and small countries such as Cyprus (CY), Malta (MT) and Lichtenstein (LT), for which our sample contains a limited number of banks.

6. Conclusions and future work

In this paper, we propose a holistic approach, ranging from the selection of the most significant bank specific

[8] Moody's rating was available for 95 banks out of 173 of our European banks sample.

[9] p.31 https://www.moodys.com/research/Banks--PBC_186998

Table A.1
Literature review summary table.

Author	Title	Year published	Dataset	Statistical technique
Altunbas, Manganelli, Marques-Ibanez	Bank Risk during the Great Recession: Do business models matter?	2012	Global sample of 16 countries. Quarterly data from 2003:q4 to 2007:q3	Probit regression, Quantile Regression
Berger, Bouwman	How does capital affect tbank performance during financial crises ?	2013	1984-2010, quarterly	Logit Regression for survival probability OLS for market share
Betz, Oprica, Peltonen, Sarlin	Predicting Distress in European Banks	2013	546 banks (most EU countries) assets>1bn 2000Q1-2013Q2	Recursive Logit model and benchmark Logit model
Chiaramonte	Should we trust the Z-score? Evidence from the European Banking Industry	2015	European banks from 12 countries over the period 2001–2011 (Banscope)	probit and complementary log–log models hazard rate model
Cihak, Poghosian	Distress in European Banks: An Analysis Based on a New Data Set	2009	European banks 1996–2007	Logit
Cole , Wu	Hazard versus probit in predicting U.S. bank failures: a regulatory perspective over two crises	2014	FDIC 1984–1992	Static Probit and time varying hazard model
Cox, Wang	Predicting the US bank failure: A discriminant analysis	2014	USA Bank failures 2007 to 2010 FDIC Quarterly	Linear and Quadratic Discriminant Analysis
De Young, Torna	Nontraditional banking activities and bank failures during the financial crisis	2013	2008Q3-2010Q4 lag quarters FDIC,assets < 100bn (plus other exclusions)	multi-period Logit model (hazard)
Demyanyk, Hasan	Financial crises and bank failures: a review of prediction methods	2009		
Ekinci, Halil	Forecasting Bank Failure: Base Learners, Ensembles and Hybrid Ensembles	2016	Turkey: 37 privately owned commercial banks operating in Turkey between 1997 and 2001. 17 out of the 37 banks faced with financial failure because of 1998 Asian and 2001 financial crises.	Logistic, J48 and Voted Perceptron, Random Subspaces, Bagging, Hybrid
Halling, Hayden	Bank Failure Prediction: A Two-Step Survival Time Approach	2008	Austrian banks 1995 - 2002	Two-Step Survival Time
Kočišová, Mišanková	Discriminant analysis as a tool for forecasting company's financial health	2014		
Kolari, Glennon, Shin, Caputo	Predicting large US commercial bank failures	2002	1989-1992 US banks defaults (1989 defaults Development) assets > 250 mil	Trait recognition model, Logit model
Lall	Factors affecting U.S. Banking Performance: Evidence From the 2007-2013 Financial Crisis	2014	2007-2013 Quarterly Call Report, Federal reserve bank of Chichago	GLS (to overcome heteroskedsticity issue in panel data)
Mayes, Stremmel	The effectiveness of capital adequacy measures in predicting bank distress	2012	Quarterly data set of FDIC insured US banks from 1992 to 2012 710.000 obs	Logit ,Discrete survival time analysis
Messai , Gallali	Financial Leading Indicators of Banking Distress: A Micro Prudential Approach - Evidence from Europe	2015	618 European Banks, 18 countries, 2007–2011	Discriminant Analysis, Logistic regression, neural networks
Ng - Roychowdhury	Do Loan Loss Reserves Behave like Capital? Evidence from Recent Bank Failures	2014	FDIC 2001–2010	Logit and hazard regression
Rebel, White	Deja vu all over again: The causes of U.S. commercial bank failures this time around	2011	FDIC 2004–2008	Logit
Wanke	Predicting performance in ASEAN banks: an integrated fuzzy MCDM–neural network approach	2015	88 Association of Southeast Asian Nations banks from 2010 to 2013,	Ntegrated fuzzy MCDM–neural network approach

indicators, which can predict its survival probability, to the choice of the appropriate machine learning technique that aggregates all critical information into a single score. Our empirical results indicate that the method of Random Forests (RF) has a superior out-of-sample and out-of-time predictive performance, with Neural Networks also performing almost equally well as RF in out of time samples.

The main contributions of this empirical study and its stark differences from other studies in the related literature of bank failures can be summarized in four layers. First and foremost, the extensive exploration of the appropriate statistical technique to address this problem by implementing six broadly used and state of the art modeling approaches. Second, the robust validation approach that we implement to test the efficacy of each modeling technique, which includes both out-of-sample and out-of-time validation. Third, the performance measures that we use to assess each model are appropriate for imbalanced datasets, like the bank failures dataset we use in this empirical study. Last but not least, the examination of an extended set of candidate explanatory variables that cover the full spectrum of a bank's financial state, both along time and cross sectionally.

Summarizing our experimental results, Random Forests consistently outperforms a series of benchmark approaches like Logistic Regression, Linear Discriminant Analysis, Support Vector Machines, Neural Networks, and Random Forest of Conditional Inference Trees, almost across all metrics broadly used for assessing the discriminatory power under an imbalance dataset. The only notable exception is NN that perform almost equally well as RF in the out of time sample. Furthermore, we estimated the predictive variance for each performance assessment measure by employing bootstrapping. Our analysis provides strong evidence for the model's increased stability and capacity to retain the high-performance levels observed in the in-sample dataset, when evaluation is

Table A.2
Selection of regressors by different algorithms.

Variable name\model	Random Forests (RF)	Stepwise Logit (SL)	Genetic Algorithm (GA)
ROE	X	X	X
CAR	X	X	X
CFEA	X	X	X
LEV	X	X	
ROE_DFS	X	X	X
d_NCASS_ORE_lag4	X	X	X
RE_EQ	X	X	
NPL	X	X	
PCT_log_equity_lag4_DFS	X	X	
NOI_ASS	X	X	
NLOAN_CDEP	X	X	
d_RE_EQ_lag4	X	X	X
NIM	X	X	
d_LOSS_LOAN_lag4_DFS	X		X
PCT_log_equity_lag1_DFS	X	X	
d_LEV_lag4	X	X	
d_ROA_lag4	X	X	X
EFF_DFS	X		X
d_NOI_ASS_lag8	X	X	
LOSS_LOAN_DFS	X	X	
LOSS_LOAN	X	X	X
LOSS_NPL	X	X	
d_log_equity_lag4_DFS	X		X
PTR_ASS			X
NFI_ASS		X	
EQ_ASS		X	
Total	**23**	**22**	**12**

performed using out-of-sample and out-of-time datasets. This performance consistency implies a much stronger generalization capacity compared to the state-of-the-art models, which renders our approach much more attractive to researchers and practitioners working in real-world financial institutions. Indeed, they are mainly interested in the generalization capacity of their systems, rather than in their in-sample performance. Furthermore, our results illustrate that in the CAMELS evaluation framework, Earnings and Capital metrics constitute the factors with the higher marginal contribution in the prediction of bank failures. Finally, we test the performance of the proposed Random Forest technique on the European Banking system, in order to further explore its generalization capacity. In particular, we classified the European banks on a credit rating scale based on their riskiness as derived by our RF specification. The produced ranking was benchmarked against Moody's rating scale to validate its performance. This way, we provided additional evidence for the robustness and stability of the proposed RF model even in datasets derived from different jurisdiction.

One aspect that this work does not consider is whether allowing for our model to account for macroeconomic variables can improve prediction performance. Such an approach though could be explored in multiple business cycle setup in order to capture the variability in the state of the whole banking system. Finally, we note that in our approach, we have postulated a Random Forests model based only on US Banks performance data and exploited its capacity on European Banks. In the future, we aim to perform our analysis on enriched dataset composing by multiple jurisdictions in order to build a global rating system for banks. Nevertheless, the results of this analysis provide valuable information to policy-makers and regulators to assess the health of the financial system based on the individual status of each participant and develop policy responses.

Appendix A

See Table A.1.

Appendix B

See Table A.2.

References

Alessi, L., & Detken, C. (2014). Identifying excessive credit growth and leverage. ECB working paper no. 1723.

Altini, M. (2015). Dealing with imbalanced data: undersampling, oversampling and proper cross-validation. https://www.marcoaltini.com/blog/dealing-with-imbalanced-data-undersampling-oversampling-and-proper-cross-validation.

Altman, E. (1968). Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *The Journal of Finance*, *23*(4), 589–609.

Altunbas, Y., Manganelli, S., & Marques-Ibanez, D. (2011). Bank risk during the great recession: Do business models matter? ECB working paper no. 1394.

Audrino, F., Kostrov, A., & Ortega, J.-P. (2018). Extending the logit model with Midas aggregation: The case of US bank failures. SSRN.

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

Berger, A., & Bouwman, H. (2013). How does capital affect bank performance during financial crises? *Journal of Financial Economic*, *109*(1), 146–176.

Betz, F., Oprica, S., Peltonen, T., & Sarlin, P. (2014). Predicting distress in European banks. *Journal of Banking & Finance*, *45*, 225–241.

Boyacioglu, M., Yakup, K., & Baykan, O. (2009). Predicting bank financial failures using neural networks, support vector machines and multivariate statistical methods: A comparative analysis in the sample of savings deposit insurance fund (SDIF) transferred banks in Turkey. *Expert Systems with Applications*, *36*(2), 3355–3366.

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

Chen, W., & Jen, Y. (2006). A study of Taiwan's issuer credit rating systems using support vector machines. *Expert Systems with Applications*, *30*(3), 427–435.

Chiaramonte, L., Croci, E., & Poli, F. (2015). Should we trust the Z-score? Evidence from the European Banking Industry. *Global Finance Journal*, *28*, 111–131.

Cole, Rebel A., & White, Lawrence J. (2011). Déjà vu all over again: The causes of US commercial bank failures this time around. *Journal of Financial Services Research*, *42*, 5–29.

Cole, Rebel A., & Wu, Qiongbing (2017). *Hazard versus Probit in Predicting U.S. Bank Failures: A Regulatory Perspective over Two Crises*.

Cox, R., & Wang, G. (2014). Predicting the US bank failure: A discriminant analysis. *Economic Analysis and Policy*, *44*(2), 202–211.

Demyanyk, Y., & Iftekhar, H. (2010). Financial crises and bank failures: A review of prediction methods. *Omega*, *38*(5), 315–324.

Denil, M., Matheson, D., & De Freitas, N. (2013). Consistency of online random forests. *ICML*, *28*(3), 1256–1264.

Ekinci, A., & Halil, I. (2016). Forecasting bank failure: Base learners, ensembles and hybrid ensembles. *Computational Economics*, 1–10.

Gogas, P., Papadimitriou, T., & Agrapetidou, A. (2018). Forecasting bank failures and stress testing: A machine learning approach. *International Journal of Forecasting*, *34*(3), 440–455.

Halling, M., & Hayden, E. (2006). Bank failure prediction: a two-step survival time approach. IFC bulletin no. 28.

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

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

Japkowicz, N. (2000). The class imbalance problem: Significance and strategies. In *Proceedings of the 200 international conference on artificial intelligence*, Special track on inductive learning.

Kamstra, M., Kennedy, p., & Suan, T. K. (2001). Combining bond rating forecasts using logit. *Financial Review*, *36*(2), 75–96.

Kolari, J., Glennon, D., Shin, H., & Caputo, M. (2002). Predicting large US commercial bank failures. *Journal of Economics and Business*, *54*(4), 361–387.

Kočišová, K., & Mišanková, M. (2014). Discriminant analysis as a tool for forecasting company's financial health. *Procedia -Social and Behavioral Sciences*, *110*, 1148–1157.

Lall, P. (2014). Factors affecting US banking performance: Evidence from the 2007–2013 financial crisis. *International Journal*, *3*(6).

Mayes, D., & Stremmel, H. (2012). The effectiveness of capital adequacy measures in predicting bank distress. *SUERF Studies*.

Messai, A. S., & Gallali, M. I. (2015). Financial leading indicators of banking distress: A micro prudential approach-evidence from Europe. *Asian Social Science*, *11*(21), 78.

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

Poghosyan, T., & Čihák, M. (2009). Distress in European banks: An analysis based on a new dataset. IMF working papers. (pp. 1–37).

Ravi, V., & Pramodh, C. (2008). Threshold accepting trained principal component neural network and feature subset selection: Application to bankruptcy prediction in banks. *Applied Soft Computing*, *8*(4), 1539–1548.

Rebel, A. C., & White, L. (2012). Déjà vu all over again: The causes of US commercial bank failures this time around. *Journal of Financial Services Research*, *42*(1–2), 5–29.

Strasser, H., & Weber, C. (1999). On the asymptotic theory of permutation statistics. *Mathematical Methods of Statistics*, *8*, 220–250.

Vapnik, V. N., & Vapnik, V. (1998). *Statistical Learning Theory (Vol. 1)*. New York: Wiley.

Vazquez, F., & Pablo, F. (2015). Bank funding structures and risk: Evidence from the global financial crisis. *Journal of Banking & Finance*, *61*, 1–14.

Wanke, P., Azad, A. K., Barros, C. P., & Hadi-Vencheh, A. (2015). A predicting performance in ASEAN banks: an integrated fuzzy MCDM–neural network approach. *Expert Systems*.

Wu, Hsu-Che, Hu, Ya-Han, & Huang, Yen-Hao (2014). Two-stage credit rating prediction using machine learning techniques. *Kybernetes*, *43*(7), 1098–1113.

Yeh, C., Chi, D., & Lin, Y. (2014). Going-concern prediction using hybrid random forests and rough set approach. *Information Sciences*, *254*, 98–110.