https://economix.fr/

What do bankruptcy prediction models tell us about banking regulation? Evidence from statistical and learning approaches

Document de Travail
Working Paper
2021-2

Pierre Durand
Gaëtan Le Quang

EconomiX - UMR7235
Université Paris Nanterre
Bâtiment G - Maurice Allais, 200, Avenue de la République
92001 Nanterre cedex

Email : secretariat@economix.fr

Université Paris Nanterre

What do bankruptcy prediction models tell us about banking regulation? Evidence from statistical and learning approaches

Pierre Durand* Gaëtan Le Quang*

October 27, 2020

Abstract

Prudential regulation is supposed to strengthen financial stability and banks' resilience to new economic shocks. We tackle this issue by evaluating the impact of leverage, capital, and liquidity ratios on banks default probability. To this aim, we use logistic regression, random forest classification, and artificial neural networks applied on the United-States and European samples over the 2000-2018 period. Our results are based on 4707 banks in the US and 3529 banks in Europe, among which 454 and 205 defaults respectively. We show that, in the US sample, capital and equity ratios have strong negative impact on default probability. Liquidity ratio has a positive effect which can be justified by the low returns associated with liquid assets. Overall, our investigation suggests that fewer prudential rules and higher leverage ratio should reinforce the banking system's resilience. Because of the lack of official failed banks list in Europe, our findings on this sample are more delicate to interpret.

Keywords: Banking regulation ; Capital requirements ; Basel III ; Logistic ; Statistical learning classification ; Bankruptcy prediction models.
JEL classification : C44, G21, G28

Acknowledgements: The authors would like to thank their colleagues from the EconomiX research centre. They are particularly grateful to Laurence Scialom and Valérie Mignon for their guidance and advice.

*EconomiX – CNRS, University of Paris Nanterre

Corresponding author: pierre(.)alexis(.)durand(@)hotmail(.)fr

1 Introduction

Bankruptcy prediction is a constantly growing field of research. Starting with the seminal paper by Altman (1968) that identified several key ratios to predict firms' bankruptcy, the literature has evolved both by diversifying the type of firms considered and by proposing increasingly complex methods. The main interest associated with developing good bankruptcy prediction models is to offer a way to monitor the soundness of a given firm in real time. Another interest is to provide a ground to regulatory constraints. This is of particular interest as far as banks are concerned.

Banking regulators indeed need to know what are the main predictors of banks' default to design rules meant to prevent such default from happening. The purpose of this paper is to discuss current banking regulation in the light of what bankruptcy prediction models tell us about the main determinants of banks' failure. To do so, we resort both (i) to a standard statistical approach by estimating a logistic regression (logit) on data covering both US and European banks, and (ii) to more sophisticated intelligent approaches by presenting results coming from random forest classifications (RF) and from artificial neural networks (ANN). Overall, we find that capital outperforms other balance sheet variables in predicting bankruptcy. In addition, the complex Basel capital ratio does not outperform the simple leverage ratio in predicting banks' default, which forces to question the rationale behind the former. As for liquid assets holding, our models suggest that banks that hold a great amount of liquid assets go more frequently bankrupt than banks investing in less liquid assets. Our models perform well on the US database, but exhibit low performances on European data.

Banking regulation is currently implemented through several rules whose main purpose is to ensure both the soundness of the banking system as a whole (macro-prudential rules) and of each bank individually (micro-prudential rules). Banking regulation was traditionally implemented through a capital ratio whose purpose was to ensure a loss-absorbing capacity on the liability side of the balance sheet. To better take into account the risk taken on the asset side of the balance sheet, this ratio has progressively evolved toward a risk-based ratio, meaning that capital requirements are computed as a function of the risk-weighted assets (RWA). Given the complexity of banks' activities, computing the RWA is not trivial and regulators often lack information or expertise to assess the risk associated with each individual bank. As a consequence, under certain conditions, banks are allowed to resort to internal models, through what the regulatory framework referred to as the advanced internal ratings-based approach (A-IRB), to assess the risk associated with their portfolio of assets. Such internal computations of RWA have however been shown to underestimate the risk associated with banks' activities (Mariathasan and Merrouche, 2014).

After the 2007-2008 crisis, banking regulators added liquidity ratios to the risk-weighted capital ratio. The necessity of liquidity regulation is grounded on the illiquidity spirals that materialized during the crisis and led to the collapse of the banking system (Brunnermeier and Pedersen, 2009). Liquidity regulation has been implemented through two different rules: the liquidity coverage ratio (LCR) and the net stable funding ratio (NSFR). The LCR states that banks need to hold enough high quality liquid assets (HQLA) to withstand a liquidity crisis lasting 30 days. The NSFR states that banks' illiquid assets need to be funded through stable funding instruments. Having a closer look at the two ratios, we notice that they are in fact redundant (Bolton et al., 2019). Instead of two ratios, liquidity regulation would thus be better off defining only one ratio. Which ratio should then be ruled out and which should remain? We believe that the perspective that should be adopted is that of the NSFR. Our models indeed suggest that liquid assets holding could actually increase the probability that banks go bankrupt. If liquid assets allow banks to face short-term liquidity needs, they are nonetheless often associated with low returns that could explain why in some cases banks that hold a great proportion of their asset portfolio in liquid assets go more frequently bankrupt than other banks.

From this quick overview of our results and of banking regulation, we formulate policy recommendations. Specifically, we think that banking regulation would be better off focusing on equity to ensure the soundness of the banking system. As can indeed be theoretically shown (see Appendix A), equity outperforms liquid assets in preventing a bank from defaulting even when the return associated with those assets is not lower than the return demanded by the creditors of the bank (i.e. even when liquid assets holding is assumed to have a negative impact on the probability of default). In addition, we argue that the simple leverage ratio should be preferred to the sophisticated Basel one. Our results indeed suggest that the latter does not outperform the former at predicting banks' failure, while this latter is far more difficult to compute than this former. Equity should thus be preferred to more complex definitions of capital. Moreover, as shown in Durand and Le Quang (2020), increasing equity requirements has a positive impact on banks' profitability when measured as the ROA. Since the ROA is by far the main predictor of bankruptcy, increasing equity requirements would probably lower the occurrence of defaults through two channels: the direct channel of capital (increase in the loss-absorbing capacity of the liability side of the balance sheet), and the indirect channel of the ROA (increase in the return associated with the asset side of the balance sheet).

This paper is in line with the literature through the use and comparison of traditional and more recent classification methods. It is quite conventional, on this type of issue, to propose a comparison of the models and their respective performances. We have therefore come here to expand the literature, by basing ourselves on some of the models identified as the most efficient for this subject, and by proposing to focus on the role of banking regulation in determining bank default. The study of the European case also constitutes an innovation in the literature, since the emphasis is generally done on the US case. Finally, our investigation gives keys to understand regulatory efficiency and complexity issue.

The rest of the paper is organized as follows. The next section reviews the literature on bankruptcy prediction models. Section 3 offers some details on the models we use. Section 4 describes our database. Section 5 presents the main results. Robustness checks are provided in section 6, and section 7 concludes.

2 Literature review

The main challenge associated with bankruptcy prediction is that, by definition, bankruptcies are very rare events. Datasets are thus severely imbalanced with one class (that of bankrupted banks) far less represented than the other (that of non-bankrupted banks). There are several ways to deal with imbalanced datasets: either under-sampling or over-sampling (or mixing the two). Under-sampling aims at reducing the size of the majority class to match that of the minority class. It therefore has the inconvenient to delete potentially interesting information, but is in general less computationally demanding than over-sampling. Over-sampling consists in balancing class distribution by replicating items in the minority class, either by exactly replicating some randomly selected items found in the minority class (Random Oversampling With Replication – ROWR (Zhou, 2013)) or by creating new items through the Synthetic Minority Oversampling Technique (SMOTE) proposed by Chawla et al. (2002). If under-sampling could sometimes be preferred to over-sampling when the dataset is weakly imbalanced (Zhou, 2013), there is a consensus in the literature that SMOTE is the best option for severely imbalanced datasets (Chawla et al., 2002; García et al., 2012; Zhou, 2013; Haixiang et al., 2017). Given our database, we therefore resort to SMOTE to balance our dataset.

Once the dataset re-sampled, the bankruptcy prediction problem consists in a simple classification problem. Such a problem can be solved either by resorting to a statistical approach or to an intelligent approach (Ravi Kumar and Ravi, 2007). Statistical methods include well-known logistic regressions and are widely used to deal with classification problems, including bankruptcy prediction for firms (Ohlson, 1980; Jones and Hensher, 2004) and for banks (Martin, 1977; Kolari et al., 2002). Intelligent methods consist in machine learning techniques such as neural network or random forest. Specifically, neural network is largely used in the bankruptcy prediction literature (Ravi Kumar and Ravi, 2007) and is often shown to perform better than logistic regressions (Tam and Kiang, 1990; Tam, 1991; Salchenberger et al., 1992). Fewer papers resort to random forest regressions to predict firms' failures (Zoričák et al., 2020).

The literature on bankruptcy prediction has reached a consensus around several financial ratios that are considered as the main determinants of defaults. Those ratios are the rationale behind the computation of the widely used Z-score (Altman, 1968; Altman et al., 1977). Capital adequacy, Assets quality, Management, Earnings, Liquidity, and Sensitivity (CAMELS)[1] ratings

[1]CAMELS constitutes the six factors used by regulatory authorities to classify financial institutions in function of their quality.

are also based on the main results found in the literature on bankruptcy prediction. Ravi Kumar and Ravi (2007) provide an exhaustive review of the variables found as predictors of banks' bankruptcy in papers published from 1968 to 2005.

3 Methodology

As discussed above, in order to assess the importance of capital and equity ratios on banks' default probability, we rely on three classification methods. For all of those approaches, the objective is to estimate the function f, on which we have no *a priori*, that defines the true model: $P(y = \{0, 1\}|X = x) = f(x) + \epsilon$, where $P(y = \{0, 1\}|X = x)$ is the probability that y, the explained variable, equals 1 or 0, y takes the value 1 at time $t-1$ for banks that fail in t and 0 otherwise, x refers to the explanatory variables and ϵ designates the error term.

The major issue in our empirical approach is the sparsity of the Y matrix. To avoid this problem we use the Synthetic Minority Oversampling Technique (SMOTE). We describe this procedure and give an overview of our methodologies and interpretation techniques. For the sake of clarity, we give only brief insight on our methodology, and refer the reader to Appendix C for more details.

Synthetic Minority Oversampling Technique (SMOTE)

As said earlier, we consider methods associated with extreme rare events to tackle our deeply imbalanced database: SMOTE (Chawla et al., 2002). As a robustness check and for transparency matters, we give class weight results in Section 6.

SMOTE uses the k nearest neighbors of all minority class examples to synthesize new minority class instances: synthetic observations are created on the line between the existing ones. The recourse to nearest neighbors ensures to replicate the distribution of the original data. In order to avoid over-fitting issues, the SMOTE procedure is only applied on the training sample. The test sample remains imbalanced.

Logistic regression

Logistic regression model (Hastie et al., 2009) comes from the wish to assess the probability of classes as a linear function of explanatory variables while respecting that the sum of probabilities equals 1. In a binary class model, it takes the following form:

$$log\frac{P(y=0|X=x)}{P(y=1|X=x)} = \beta_0 + \beta_1^T x \tag{1}$$

where β_0 is the intercept included in the model and β_1^T stands for the vector of parameters.

After rearrangement, we obtain:

$$\begin{cases} P(y=1|X=x) = \frac{1}{1+exp(\beta_0+\beta_1^T x)} \\ P(y=0|X=x) = 1 - \frac{1}{1+exp(\beta_0+\beta_1^T x)} \end{cases}$$

It is widely accepted that in a matter of logit interpretation, one has to refer to the odds ratio ($OR(x_j)$) which is calculated as the exponential of the coefficient associated with an explanatory variable. It is interpreted as follows: all things being equal, an increase of one unit in variable j induces a change in the probability of class 1 by a factor of $OR(x_j)$.

Random forest (RF)

Random forest classification (Breiman, 2001) consists in averaging a more or less large number of decision trees. A tree is built *via* recursive binary partition of the explanatory variables, or features, spaced into M final regions. At every step, or node, the best splitting point is computed for each feature and the model retains the variable that minimizes the loss function. Once the tree is built, the estimated probability $\hat{p}_{1m}$ of default in region m is given by the proportion of default in the region:

$$\hat{p}_{1m} = \frac{1}{N_m} \sum_{x_i \in R_m} I(y_i = 1) \tag{2}$$

where N_m is the cardinal of region m, m is the region with $m \in [\![1; M]\!]$, $I(y_i = 1)$ is the function that scores 1 if y_i equals 1 and 0 otherwise.

In order to avoid over fitting issues, but to improve the out-of-sample prediction of the model, there are two parameters to optimize at the trees' level and one at the forest's one: the number of final observations per leave (*i.e.*, per final partition space), the number of splits (*i.e.*, the depth of trees), and the number of trees in the forest. To set those hyperparameters, we run numerous estimations and selected those that give the best out-of-sample score.[2]

Artificial neural network (ANN)

Artificial neural networks (McCulloch and Pitts, 1943; Hastie et al., 2009) model links between features and explained variables, or label, through the application and composition of non-linear functions. For complexity matters, we recourse here to the most widely used neural network, called the single hidden layer back-propagation network. It means that we use only one hidden layer between the inputs and the output:

[2]In this particular context, the score we use is the true positive rate: the number of banks identified as default banks (true positive) over the sum of true positive, and the number of banks identified as not having defaulted while they have. This ratio is called sensitivity.

$$
\begin{aligned}
Z_m &= \sigma(\alpha_{0h} + \alpha_h^T X),\ h \in [\![1, H]\!] \\
T_k &= \beta_{0k} + \beta_k^T Z,\ k \in \{0, 1\} \\
f_k(X) &= g_k(T),\ k \in \{0, 1\}
\end{aligned}
\tag{3}
$$

where $\sigma(.)$ is the simoid function given by $\sigma(v) = \frac{1}{1+e^{-v}}$, H is the number of hidden units in the hidden layer, and $g_k(.)$ is the softmax function given by $g_k(T) = \frac{e^{T_k}}{\sum_{l \in \{0,1\}} e^{T_l}}$. Z_m are called hidden units and form the hidden layer because they are not directly observed. ANN have three hyperparameters to be set: the number of hidden layers, the number of hidden units and the batch size.[3] As for RF, to determine those parameters, we select those that maximize the proportion of default banks identified as so among all of those.

Interpretation

An important part of the empirical strategy resides in our capacity to evaluate models' performance, features' significance (or importance), and features' marginal impact on the estimated probability of default. Our objective is to obtain efficient models in their ability to identify default. Then, we look into the role played by equity and capital ratios in the determination of default's probability. Finally, we assess the marginal impact of those ratios on the output. To this aim, we use several performance scores and interpretation tools.

In order to assess the performance of our models, we rely on the confusion matrix (Hand, David, 2012), which is widely used in classification studies. In binary classification problems, confusion matrix gives four elements: the number of true positive (TP, failed banks identified as failed banks), the number of true negative (TN, unfailed banks identified as so), the number of false positive (FP, unfailed banks identified as failed ones), and the number of false negative (FN, failed banks identified as unfailed ones). From those quantities, we can compute some performance scores:[4]

- Mean score of the model: $\frac{TP+TN}{TP+TN+FP+FN}$
- True positive rate (TPR, also called sensitivity or recall): $\frac{TP}{TP+FN}$.
- True negative rate (TNR, or specificity): $\frac{TN}{TN+FP}$
- Positive predictive value (PPV, or precision): $\frac{TP}{TP+FP}$

Since the TPR gives the proportion of failed banks identified as so among all failed banks, this is the ratio[5] we are looking to maximize in the hyperparameterization of our model. Indeed,

[3] It corresponds to the number of observations took to fit the model at each iteration.
[4] All those scores are computed in and out-of-sample. We favor the out-of-sample score.
[5] We maximize the TPR calculated with out-of-sample data.

it is far more important for us to identify default banks, even if it increases the false negative rate. We believe that the cost of identifying a bank as not defaulting while it is, is greater than identifying a bank as defaulting while it is not.

In order to evaluate the performance of our models, we rely on conventional measures in binary classification:

- The receiver operating characteristic (ROC), that plots the true positive rate against the false positive rate for different levels of differentiation threshold. This curve allows us to calculate the area under the ROC curve (AUROC) that gives the probability that the classifier ranks a positive randomly selected instance higher than a negative one. Therefore, the AUCROC should be maximized.

- The precision recall curve (PR) that plots the true positive rate against the positive predictive value for a set of different thresholds. The area under the PR (AUPR) should also be maximized even if it has not intuitive interpretation as the AUROC.

The next step is to assess the statistical significance, or importance, of independent variables. Logistic regression being a parametric model, it provides a Z-score, a p-value and a confidence interval. It is not the case for RF and ANN classifiers. Therefore, we resort to interpretable machine learning tools in order to assess features' importance in determining the output. The independent variables' importance in random forests is assessed given by a generalization of Breiman et al. (1984)'s calculation of relevance in classification trees that measures the improvement made in each node of a tree for each predictor.

However, this measure is specific to RF and decision trees, so we rely on permutation feature importance (Breiman, 2001) for artificial neural networks. This measure attributes to each independent variable a score allowing to order them in function of their importance in determining the output. For a given variable j, it is computed as follows:

1. We calculate the model's score S

2. Then, the variable j is shuffled N times

3. The mean of scores s_n^j given by the model using the shuffled variable is then calculated:

$$S_{mean}^j = \frac{1}{N} \sum_{n \in [\![1,N]\!]} s_n^j$$

4. The importance of the variable j is given by the difference $Imp_j = S - S_{mean}^j$. Therefore, the larger this difference, the more important we can consider the variable to be in the predictive capacity of the model.

Since this permutation feature importance can be assessed for any model that gives prediction, we also computed it for the logistic regression.

As for the marginal impact of features on default probability for non-parametric RF and ANN models, we rely on two quantitative input influence measures: Partial Dependence Plots (PDP, Friedman (2000), Hastie et al. (2009)) that assess the variations of the output when making one feature varying, and Accumulated Local Effect (ALE, Datta et al. (2016)) that is based on the same logic as PDPs but being computed on variables definition's space and supposed to take into account the potential correlations between independent variables.

4 Data and descriptive statistics

4.1 Data

Our sample consists in US and European[6] bank's balance sheet variables on the 2000-2018 period extract from FitchConnect database. In order to capture the specific information of variables, we checked their correlation with banks' size measured by total assets. We therefore reported size to variables highly correlated with it. In order to attenuate outliers' effect, we applied a log transformation to variables displaying extreme values away from the mean by several tens of times the standard deviation.

After data treatment for missing values, we managed to keep 24 variables, 4707 banks in US among which 454 have defaulted and 3529 European banks among which 205 defaults. Table 1 displays the evolution of the number of banks per year in the two samples.

[6]We consider 12 European countries: Austria, Belgium, Cayman Islands, Denmark, France, Germany, Ireland, Italy, Luxembourg, Netherlands, Spain, United Kingdom. We selected those countries based on the number of defaults per country.

Table 1 – Evolution of observations and defaults per year

Year	US			Europe		
	Nb. obs.	Nb. defaults	% default	Nb. obs.	Nb. defaults	% default
2000	3866	0	0.0	353	8	2.27
2001	3961	1	0.03	865	9	1.04
2002	4029	1	0.02	275	2	0.73
2003	4067	3	0.07	272	7	2.57
2004	4076	0	0.0	257	3	1.17
2005	4305	0	0.0	185	0	0.0
2006	4337	1	0.02	197	2	1.02
2007	4435	18	0.41	728	16	2.2
2008	4516	115	2.55	1275	4	0.31
2009	4465	138	3.09	1409	10	0.71
2010	4361	80	1.83	1470	18	1.22
2011	4291	42	0.98	1545	20	1.29
2012	4226	21	0.5	1754	15	0.86
2013	4207	13	0.31	1777	31	1.74
2014	4194	7	0.17	2209	45	2.04
2015	4198	5	0.12	2234	13	0.58
2016	4191	6	0.14	1716	2	0.12
2017	4196	0	0.0	1659	0	0.0
2018	4197	3	0.07	1376	0	0.0

Source: Authors' calculations. The number of defaults refers to the following year.

Failed banks are identified using the FDIC list of failed banks[7] for the US sample. There is no such official list for European banks. Therefore, we used the FitchConnect variable identifying closed banks, withdrawing those that are closed because of merger or acquisition. We believe that taking into account balance sheet variables of the year of default to identify default is not relevant for two reasons: (i) it might be too easy to classify failed from unfailed banks since it is likely that some variables take abnormal values during the year of default, and (ii) it is far more interesting to be able to predict default at least one year before its occurrence. In particular, we aim at assessing the role played by regulatory requirements on the probability of default. Therefore, we consider that the incentive to increase certain balance sheet variables is effective when it improves a bank's resilience or, in other words, when it reduces the probability of default. For those reasons, the default dummy has been shifted by one year before its occurrence.

[7]The US Federal Deposit Insurance Corporation offers a public list of US banks that have failed since October 1, 2000.

Some comments on our European sample should be made. As can be seen, the evolution of the number of banks between 2000 and 2008 does not seem to reflect reality. For the sake of comparison with US results, we keep all the period in the results we present in Section 5.2. However, we check for a potential data selection bias, running the models for the European sample on the sub-period 2008-2018 and display the results of this robustness check in Section 6.

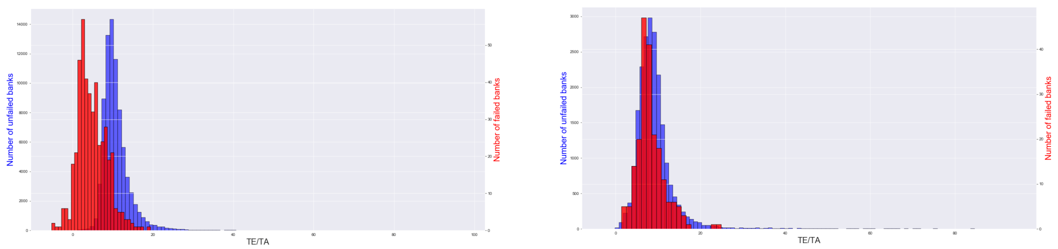
4.2 Descriptive statistics

Equity and capital distribution: failed versus unfailed banks

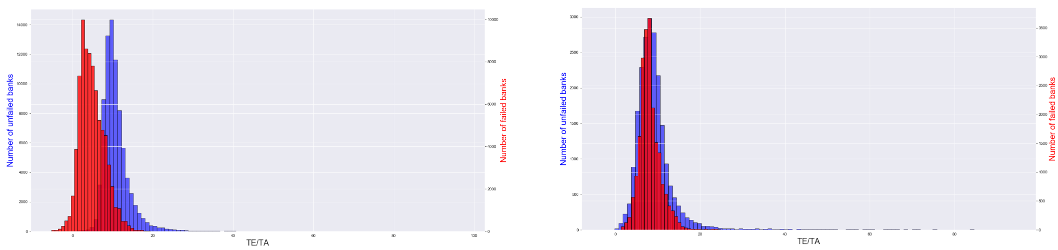
The goal of our models is to classify failed banks aside from unfailed banks. Therefore, we first look into our main variables' distribution separating default from no default in order to reveal eventual differences. Figure 1 shows TE (Total Equity)/TA (Total Assets) distribution for US and European banks.

Figure 1 – TE/TA distribution - US versus Europe

(a) Before SMOTE



(b) After SMOTE



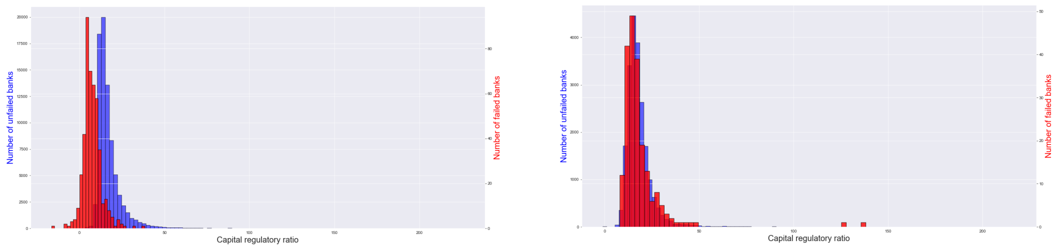
Source: Authors' calculations. Total equity over total assets distribution before and after applying SMOTE on data. The US sample is displayed on the right, the European one on the left.

As can be seen, in the case of the US sample, failed banks are characterized by a lower leverage ratio in the year before default than unfailed banks. The same remark cannot be made regarding the European sample: the distribution of TE/TA is indeed quite similar for failed and unfailed banks. Therefore, we can expect that TE/TA will be a relevant determinant of US banks' default probability, but not for European banks. Besides we should expect a negative impact of this variable on default probability. We can also remark that the distribution seems to keep its characteristics after SMOTE application on data.

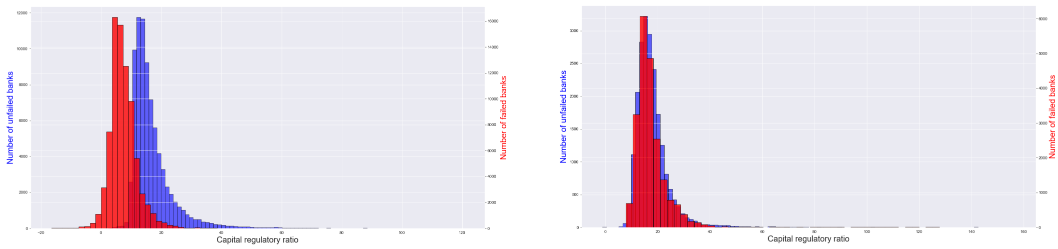
The same remarks can be made on regulatory capital ratio's distribution, as shown in Figure 2: (i) it appears to be determinant in the US sample, with negative impact on default's probability, (ii) it has almost the same distribution for failed and unfailed banks in the case of European banks, and (iii) SMOTE application to data does not interfere with variables' distributions.

Figure 2 – Regulatory capital ratio distribution - US versus Europe

(a) Before SMOTE



(b) After SMOTE



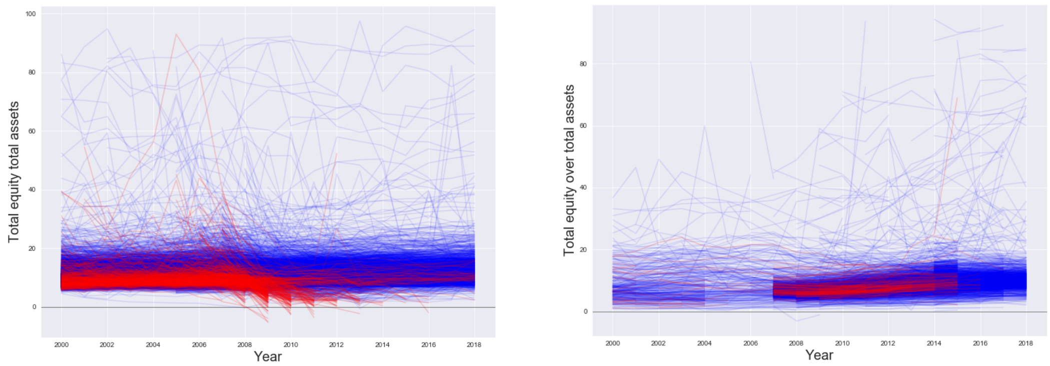
Source: Authors' calculations. Regulatory capital ratio distribution before and after applying SMOTE on data. The US sample is displayed on the right, the European one on the left.

Equity and capital evolution: failed versus unfailed banks

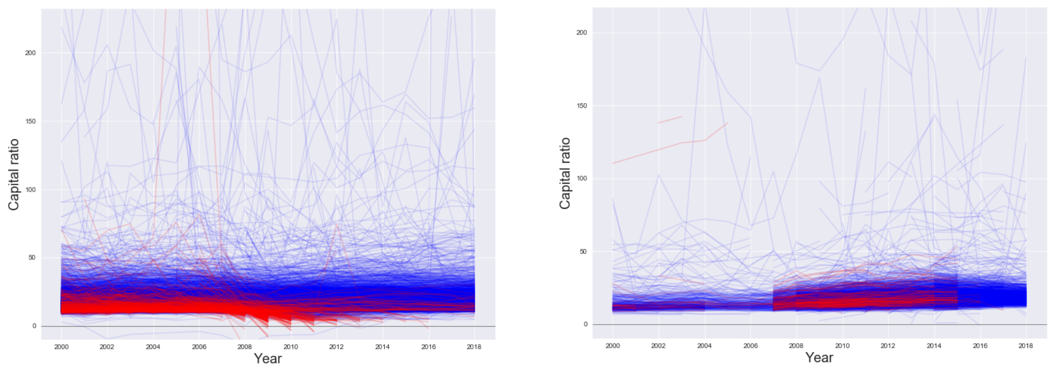
In order to control for potential dynamic effect from independent variables on default probability, we look into our main variables' evolution in Figure 3.

Figure 3 – Equity and capital ratios evolution in time - US versus Europe

(a) TE/A



(b) Regulatory capital ratio



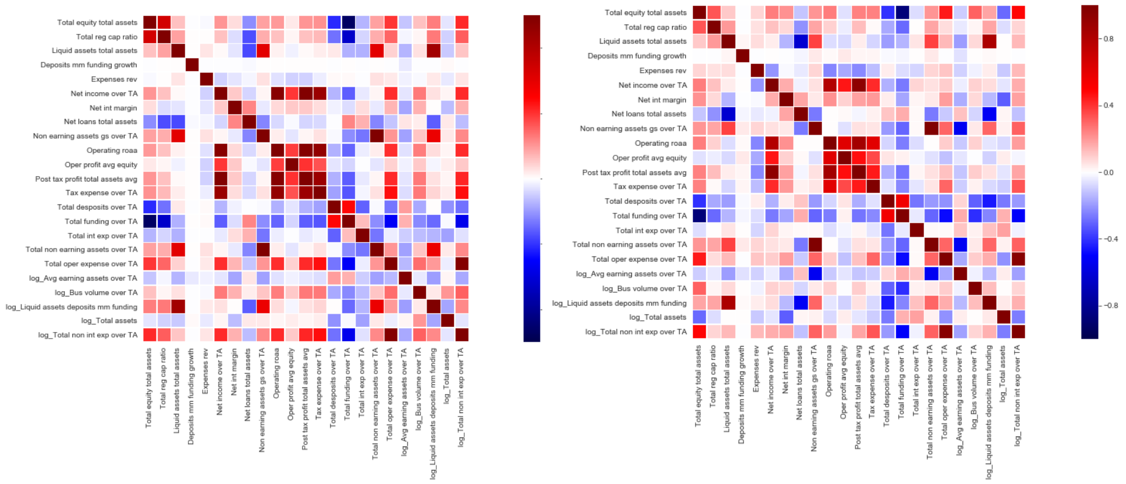
Source: Authors' calculations. TE/TA and regulatory capital ratio distributions. US sample is displayed on the right, European one on the left. Unfailed banks are displayed in blue, failed ones in red.

As can be seen, at least for equity and capital ratios in US, there is a decrease in the two to three years prior to the default. This dynamic, once again, is not observed for European banks. Those observations confirm those we made based on variables' distributions, and show that taking into account temporal dynamic can help to better capture the different balance sheet's characteristics between failed and unfailed banks. To account for this remark, we show the results of models when including time dimension in a robustness check in Section 6.

Variables correlations

The logistic regression does not handle multicolinearity, and Partial Dependence Plots (PDPs) can be biased when independent variables are highly correlated with each other. Figure 4 shows correlation hitmaps for both samples.

Figure 4 – Correlation hitmap - US versus Europe



Source: Authors' calculations. The US sample is displayed on the right, the European one on the left.

As can be seen, some variables display quite important correlation coefficients with each other. To avoid any bias in our estimations we remove the variables showing high correlation with multiple other features and that could potentially contain quite similar information.[8] Numerous variables remain in our models. This could especially be problematic for the logistic regression that is not built to handle important number of features. As it will be discussed thereafter, we removed variables associated with explosive coefficients from the logistic regression. Regarding PDPs, the use of Accumulated Local Effects (ALE) should ensure the stability of our results.

5 Results

5.1 US banks

We begin by presenting results for US banks. To do so, we first present the performance of our different models at correctly sorting banks. We then rank variables (features) according to their importance in predicting banks' default. We finally inquire the impact of each significant feature on the probability of default.

5.1.1 Models' performance

To study the performance of our models, we resort to the performance measures presented in the methodological section. We focus in particular on the true positive rate (TPR). Recall

[8]Precisely, we removed three variables: Net income over Total Assets, Operating profit avg equity, Post tax profit total assets avg.

that this rate measures the proportion of bankrupted banks that models identified correctly as bankrupted. Table 2 presents the value of different performance measures for our three models. We notice that all three models perform well in predicting out-of-sample defaults, with ANN performing better than both Logit and RF. However, ANN seems to under-perform when it comes to identifying non-bankrupted banks. In general, the different performance measures presented indicate that all three models perform well in classifying banks.

Table 2 – Models' performance - US

Scores	Logit		RF		ANN	
	In-sample	Out-of-sample	In-sample	Out-of-sample	In-sample	Out-of-sample
Score	90.67	94.95	93.35	96.59	90.50	87.63
TPR	86.36	85.37	89.95	86.18	92.99	91.06
TNR	95.0	95.0	96.75	96.64	88.03	87.62
AUROC	95.55	93.15	98.26	96.57	96.70	94.65
AUPR	96.47	46.54	98.40	41.61	96.94	31.98

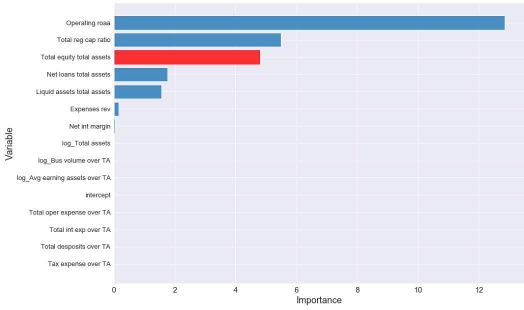
Source: Authors' calculations. All scores are defined in Section 5 and displayed in %. In red, the out-of-sample rate of failed banks identified as so.

5.1.2 Features' importance

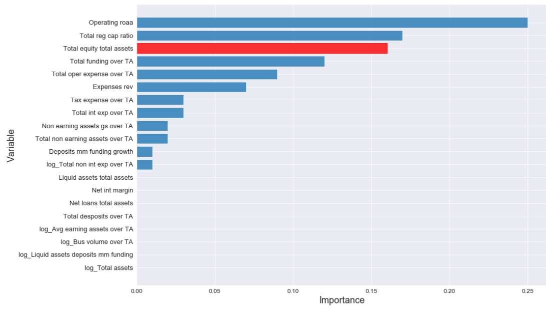
The next step in our study consists in determining which variables (features) impact the most the probability of bankruptcy. To do so, we compute the relative significance of the different features, resorting to the computation of the relative importance of features for the RF model (Hastie et al., 2009) and to that of the permutation feature importance for the ANN model (Breiman, 2001). Even if the importance of features is not calculated in the same way for our three models, such calculation allows in each case to rank features according to their importance and thus to gain insight on the main predictors of default. Figure 5 presents variables' relative importance for our three models.

Figure 5 – Variables' relative importance - US

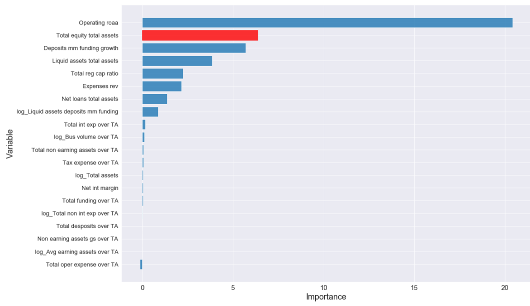
(a) Logistic regression



(b) Random Forest



(c) Artificial Neural Networks



Source: Authors' calculations. In red the importance of TE/TA ratio.

We notice that the three models exhibit more or less similar rankings. Operating Return On Average Assets (ROAA) always ranks first, which is not surprising and in line with the literature. Equity over total assets (TE/TA) always ranks among the top three predictors of banks' default. Total regulatory capital also ranks among the main predictors of default. In line with the simple theoretical model presented in Appendix A, we notice that capital is a stronger predictor of bankruptcy than the proportion of liquid assets held.[9]

5.1.3 Features' impact on default probability

Now that we have exhibited which variables are the main predictors of banks' failure, we have to wonder what is the impact of those variables on the probability of default. From a regulatory perspective it is indeed of the utmost importance to know on which variables focusing to design proper rules. We first begin by presenting the results drawn from the logit model. Results are presented in Table 3.

[9]The only situation where this is not the case is in the ANN model where total regulatory capital is less important than liquid assets over total assets. However, even in this case, equity over total assets is more important than the latter ratio.

Table 3 – Logistic regression - US

Variables	Odds Ratio	Coefficient p-values
Total equity total assets	-0.217***	0.000
Total reg cap ratio	-0.118***	0.000
Liquid assets total assets	0.079***	0.000
Expenses rev	-0.001***	0.000
Net int margin	0.748***	0.000
Net loans total assets	0.036***	0.000
Operating roaa	-0.626***	0.000
Tax expense over TA	-1.000***	0.000
Total desposits over TA	-0.481***	0.000
Total int exp over TA	2.5e+19***	0.000
Total oper expense over TA	-1.000***	0.000
log_Avg earning assets over TA	345.334***	0.000
log_Bus volume over TA	-0.984***	0.000
log_Total assets	-0.127***	0.000
intercept	1.539**	0.023
Nb. of observations	111502	
Nb. of banks (before SMOTE)	3138	
Nb. of defaults (before SMOTE)	331	

Source: Authors' calculations. Odds ratio are calculated as the exponential of estimated coefficients. To ease the reading, we have subtracted 1 from the OR.

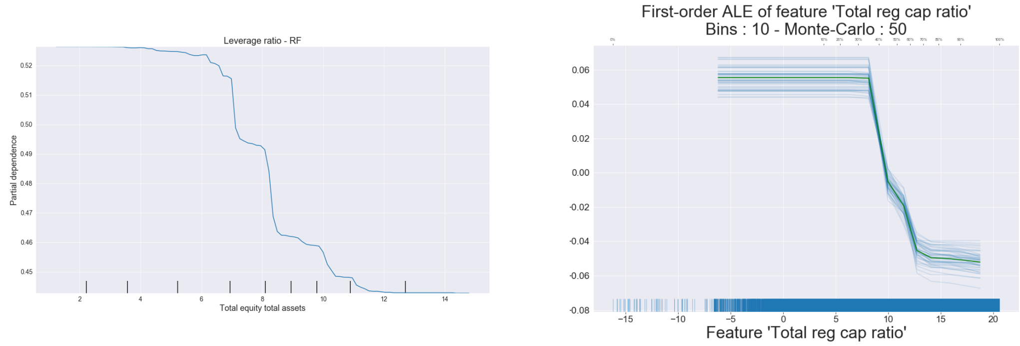
We notice that both total equity over total assets and total regulatory capital have a negative impact on the probability of default. More precisely, the total equity over total assets ratio has a stronger negative impact than total regulatory capital. This suggests that a simple constraint on the leverage ratio would perform better than a sophisticated capital ratio in preventing banks from defaulting. Surprisingly, the impact of liquid assets holding on the probability of default is positive, suggesting that the more banks hold liquid assets, the more they are likely to go bankrupt. This seems in contradiction with the simple theoretical model presented in Appendix A. However, in this model, we assume that liquid assets yield the same return as that paid to depositors. This assumption was meant to simplify the interpretation of the results, but has the consequence to ensure a positive impact of liquid assets holding on the probability of default.[10] On the contrary, when liquid assets pay less than what banks have to pay to their depositors, it is likely that liquid assets holding will have a positive impact on the probability of default.

[10]This is however not a problem *per se* since our model has the only purpose to provide some insight on the absolute value of the impact of capital and liquid assets holding on the probability of default.

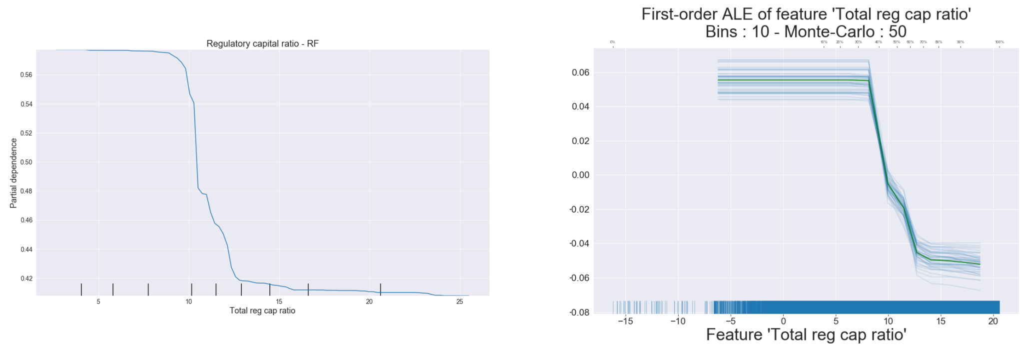
Tables 6 and 7 present results for, respectively, RF and ANN. In both cases, the results drawn from the logit model are confirmed: capital has a negative impact on the probability of default, while liquid assets holding has a positive effect.

Figure 6 – PDP and ALE - RF classifier - US

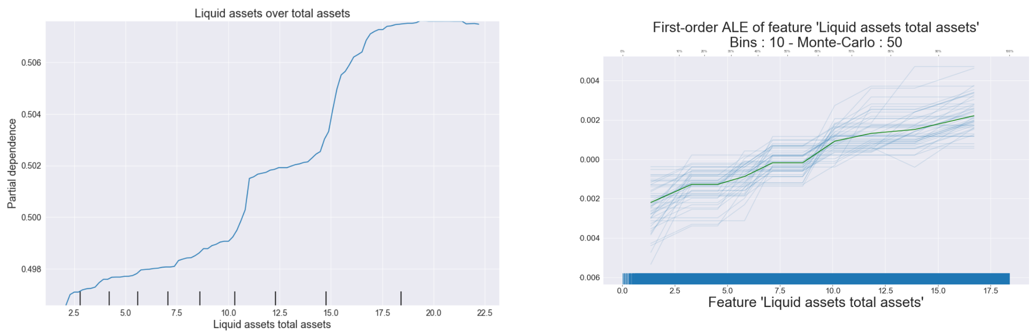
(a) Total equity over total assets



(b) Total regulatory capital



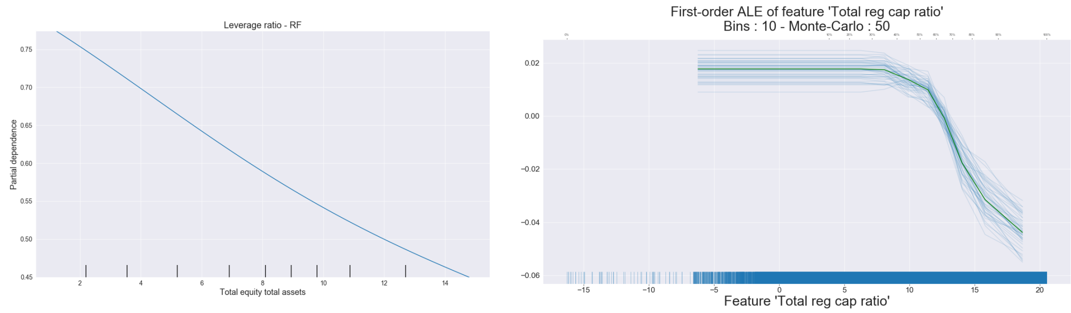
(c) Liquid assets over total assets



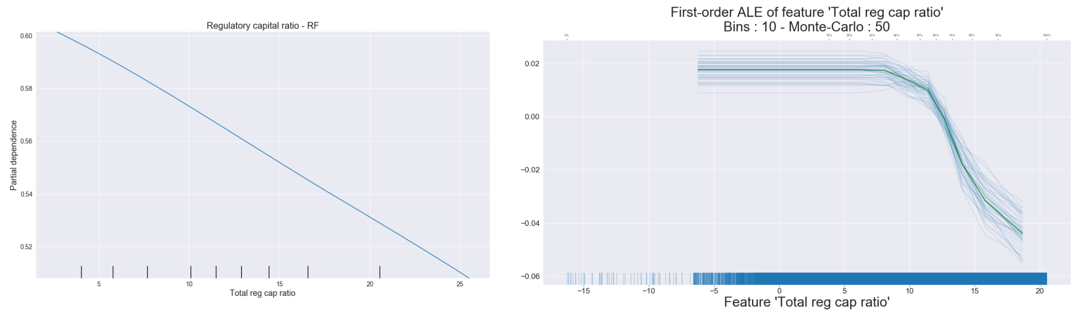
Source: Authors' calculations. PDPs on the left, ALEs on the right.

Figure 7 – PDP and ALE - ANN classifier - US

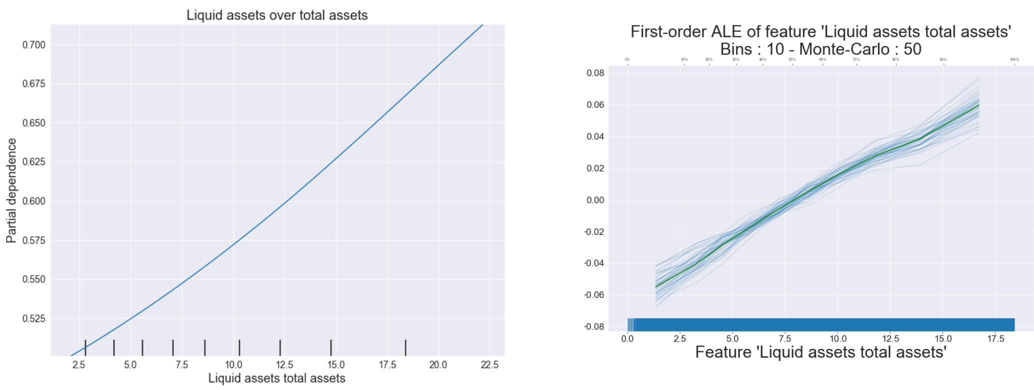
(a) Total equity over total assets



(b) Total regulatory capital



(c) Liquid assets over total assets



Source: Authors' calculations. PDPs on the left, ALEs on the right

In sum, capital has a negative impact on the probability of default, which confirms that it is indeed the main instrument through which banking regulation should intervene. In addition, it seems that the ratio equity over total assets is a stronger determinant of the probability of default than total regulatory capital. As for liquid assets holding, it counter-intuitively appears that it has a positive impact on the probability of default. This result can be explained by the low return associated which liquid assets.

5.2 European banks

We now turn to the results concerning European banks. As already stated in the data section, the main difficulty when it comes to inquiring the question of bankruptcy prediction in the European context is that there does not exist a unique database identifying banks' defaults as is the case for the US. We therefore identify banks' defaults directly in the Fitch Connect database.

Table 4 presents the performance of our three models in predicting banks' default. Here again we resort to different measures of performance, with a particular focus on the true positive rate (TPR). Results are far less convincing than those for US banks. We indeed notice that our models perform less well than for US banks. The performance of the three models is however consistent, with none performing very differently from the others. The best model here may be RF since it performs better than the others out-of-sample, event if it predicts correctly only a little more than half of the defaults.

Table 4 – Models' performance - Europe

Score	Logit		RF		ANN	
	In-sample	Out-of-sample	In-sample	Out-of-sample	In-sample	Out-of-sample
Score	67.24	62.87	86.07	74.68	69.74	50.06
TPR	70.92	53.01	96.69	51.81	83.97	65.06
TNR	63.56	63.0	75.46	74.98	55.52	54.93
AUROC	70.87	61.12	92.36	66.89	74.17	61.39
AUPR	62.99	1.9	89.93	2.26	65.85	1.94

Source: Authors' calculations. All scores are defined in Section 3 and displayed in %. In red, the out-of-sample rate of failed banks identified as so.

The difficulty to offer a satisfying prediction model of banks' default for European banks is certainly related to the lack of data concerning defaulting banks. As a consequence of the poor predictive power of our models, results on the importance and on the impact of features on the probability of default are hard to interpret. We nonetheless report those results in Appendix D for the sake of completeness.

6 Robustness

6.1 Taking time dynamic into account

The results displayed in Section 5 are based on first-order lagged variables so that the default is predicted one year before its occurrence. However, it is possible that probability of default is more influenced by balance sheet variables in dynamic than in static terms. This view is supported by the variables' evolution displayed in Section 4.2: at least for US failed banks, we can observe a drop in some variables in the two to three years preceding default.

In this robustness check, we intend to account for this potential dynamic effect. To do so, we fit the model using first difference variables. Table 7 in Appendix E.1 shows the performance scores of those estimations.

As can be seen, even if the results for Random Forest classification in the case of US banks are quite decent, most of the models using first difference variables display poor classification capacity. So, it is the balance sheet's state in the year before default that constitutes the main determinant of default.

6.2 Variables standardization

Standardization of independent variables is a usual procedure before implementing the models we use. It is supposed to decrease multicolinearity risks and ensure measurement units equivalence between features. However we decided to keep our variables as they are for three reasons: (i) we control for multicolinearity issues during our modeling process, (ii) considering that we only have balance sheet variables, we do not believe the measurement units differences to be strong, and (iii) interpretation is much easier when keeping features in their original units.

Nevertheless, as a robustness check, we look into our models' performances when variables have been standardized. Table 8 in Appendix E.2 gives the performance results for those models.

Results with standardized variables are similar to those obtained with untransformed variables. Particularly, the logistic regression performs slightly better with those variables. This finding was expected since standardization helps for multicolinearity treatment. Since performances are not deeply improved, we are comforted in the choice of displaying results with untransformed variables.

6.3 Alternative treatment of extreme rare events

As mentioned in Section 2, there are multiple ways to treat extreme rare events. We chose to proceed with the SMOTE procedure by comparing the estimation results with three other methods: the implementation of models on raw data, the use of class weighting, and the implementation of an anomaly detection methodology.[11]

[11] Results are available upon request

Models with no treatment of extreme rare events are unsurprisingly out-performed by all the others.

Class weight methodology is based on King and Zeng (2001) and consists in weighting the data, resulting in a weighted log-likelihood. Even if the results are better than the precedent approach, we found that SMOTE procedure produces slightly higher performances.

As a robustness check, we also implemented a model used in anomaly detection, namely the autoencoder (Olshausen and Field, 1996). Autoencoder is built on a similar structure as Artificial Neural Networks. It consists in dimensionality reduction (encoder) and input reconstruction (decoder). It is used for anomaly detection as follows: the model is trained only on "normal" cases (non default in our case). Then, the testing data, that includes default events, is passed through the model. The prediction error is supposed to increase importantly when a failed banks' input occurs. We can therefore create a variable that scores 1 when the error exceeds a certain threshold (default) and 0 the rest of the time. In our case, the results for the autoencoder are not satisfying enough to privilege this methodology to the other ones. Moreover, interpretability is far more complicated with this kind of deep learning methodology.

6.4 Reduced time dimension for the European sample

We noticed in Section 4.1 that there might be some issues in our data on European banks: the number of banks is very low until 2008 where it rises from 197 in 2006 to 1275 in 2008. In order to control for a potential data selection bias, we run our models on the sample going from 2008 to 2018. Table 9 in Appendix E.3 displays the performance results for those estimations.

Compared to our results on the full period, the overall performance of models is slightly improved but no one shows better capacity to identified default. The global scores' enhancement is mainly due to better non default identification. We believe that this is due to the fact that reducing time dimension, we remove important information on default banks occurring between 2000 and 2007.

6.5 Reduced sample for logistic model

As mentioned earlier, logistic regression does not support multicolinearity issues. To tackle this issue, we drop most correlated variables with each-other in the regressions presented in Section 5. However, we observed high coefficients values associated with some variables, which is characteristic of multicolinearity. As a robustness check, we run logistic regressions for both US and Europe, removing variables associated with explosive coefficients.

Results for those models are displayed in Tables 10 and 11 in Appendix E.4. We can see that models' performances and odds ratios' values are quite stable compared to those obtained with full features. Equity and capital ratios remain statistically significant and have negative influence on default probability. We can notice that the intercept takes high values in both models. This can be explained as follows: when all balance sheet variables are null, the probability of default equals 1.

7 Conclusion

In this paper, we tackle one of the most essential aspects of banking regulation: do prudential rules prevent banks from going bankrupt? Indeed, Basel III accords are supposed to strengthen financial stability both through macro- and micro-prudential perspectives. We focus our study on the latter one by looking at the efficiency and impact of some prudential ratios on banks' probability of default. To this aim, we rely on large databases of 4707 US banks and 3529 European ones, with respectively 454 and 205 observations of default, over the 2000-2018 period. Using SMOTE procedure to balance our data, we apply three different approaches to classify failed banks from the others: logistic regressions, random forest classifications, and artificial neural networks.

Our results on the US sample show high classification performances and identify three main determinants of bankruptcy probability: profitability as measured by operating ROAA, total regulatory capital ratio, and total equity over total assets ratio. Our findings also underline strong negative impact of equity over total assets and regulatory capital ratios on default probability. Turning to liquid assets over total assets ratio, even though its predictive power of default probability is found to be weak, its impact is surprisingly assessed as positive. We justify this result by the fact that liquid assets are likely to have lower returns than deposits. Therefore, this finding must be seen in the particular context of the period covered by our study: low interest rates since the crisis at the end of the 2000s.

Overall, our investigation suggests regulatory requirements to focus more on capital than on liquidity. Moreover, since equity over total assets and regulatory capital ratios seem to have similar impact on banks default probability, we believe that the actual regulatory agreements would gain in terms of complexity costs if focusing on leverage ratio. Besides, as shown in Durand and Le Quang (2020), equity ratio has positive impact on profitability as measured by ROAA. Therefore, prudential framework based on fewer rules but higher leverage ratio could also create a healthy dynamic between leverage, profitability and distance to default.

Our findings on the European sample are far less convincing. Since the quality of the models is not as great as for US banks, the interpretation of the results is much more delicate. The poor quality of our estimations on the European sample can be explained in two manners: (i) their is too much uncertainty in our data since there is no official list of failed banks in Europe as there is in US, and (ii) the differences between US and European banking system structures are so important that it implies an unequivocal opposition in their banks default determinants. We do not believe that European banks failure cannot be explained by balance sheet variables at all, so the first reason is the most probable.

The health crisis of 2020 related to Covid 19 creates great uncertainty about its economic repercussions, and there may be an opportunity to see whether some lessons from the 2007 financial crisis have been learned. Specifically, the next few years are likely to test the strength and relevance of Basel III regulatory agreements. Therefore, the after crisis period will be the occasion to test our hypothesis on a more efficient regulation when based on strong leverage ratios.

References

E. I. Altman. Financial ratios, dirscriminant analysis and the prediction of corporate bankruptcy. *The Journal of Finance*, 23(4):589–609, Sept. 1968. ISSN 00221082. 1, 3

E. I. Altman, R. G. Haldeman, and P. Narayanan. ZETATM analysis A new model to identify bankruptcy risk of corporations. *Journal of Banking & Finance*, 1(1):29–54, June 1977. ISSN 03784266. 3

P. Bolton, S. Cecchetti, J.-P. Danthine, and X. Vives. *Sound At Last? Assessing a Decade of Financial Regulation.* CEPR Press, 2019. 2

P. Bolton, M. Després, L. A. Pereira da Silva, F. Samama, and R. Svartzman. The green swan – Central banking and financial stabilit in the age of climate change. *Bank for International Settlements and Bank of France Working Paper*, 2020.

L. Breiman. Random Forests. *Machine Learning*, 45(1):5–32, Oct 2001. ISSN 1573-0565. 5, 7, 14

L. Breiman, J. Friedman, R. Olshen, and C. Stone. Classification and Regression Trees. 1984. 7

M. K. Brunnermeier and L. H. Pedersen. Market Liquidity and Funding Liquidity. *The Review of Financial Studies*, 22(6):2201–2238, 2009. ISSN 08939454, 14657368. 2

E. Campiglio. Beyond carbon pricing: The role of banking and monetary policy in financing the transition to a low-carbon economy. *Ecological Economics*, 121:220–230, 2016.

N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, 16:321–357, 2002. 3, 4, 28

A. Datta, S. Sen, and Y. Zick. Algorithmic Transparency via Quantitative Input Influence: Theory and Experiments with Learning Systems. pages 598–617, 05 2016. 8, 32

P. Durand and G. Le Quang. Banks to basics! Why banking regulation should focus on equity. *Working Paper EconomiX*, 2020-2, 2020. 2, 22

J. Friedman. Greedy Function Approximation: A Gradient Boosting Machine. *The Annals of Statistics*, 29, 11 2000. 8, 31

V. García, J. S. Sánchez, R. Martín-Félez, and R. A. Mollineda. Surrounding neighborhood-based SMOTE for learning from imbalanced data sets. *Progress in Artificial Intelligence*, 1 (4):347–362, Dec. 2012. ISSN 2192-6352, 2192-6360. 3

C. Goodhart. Problems of monetary management: The U.K. experience. *Papers in monetary economics Reserve Bank of Australia*, 1975.

T. M. Ha and H. Bunke. Off-line, handwritten numeral recognition by perturbation method. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(5):535–539, 1997. 28

G. Haixiang, L. Yijing, J. Shang, G. Mingyun, H. Yuanyue, and G. Bing. Learning from class-imbalanced data: Review of methods and applications. *Expert Systems with Applications*, 73: 220–239, May 2017. ISSN 09574174. 3

T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer series in statistics. Springer, 2009. ISBN 9780387848846. 4, 5, 8, 14, 29, 30, 31

Hand, David. Assessing the performance of classification methods. *International Statistical Review*, 80, 12 2012. 6

S. Jones and D. A. Hensher. Predicting Firm Financial Distress: A Mixed Logit Model. *The Accounting Review*, 79(4):1011–1038, 2004. ISSN 00014826. 3

G. King and L. Zeng. Logistic regression in rare events data. *Political Analysis*, 9(2):137–163, 2001. doi: 10.1093/oxfordjournals.pan.a004868. 21

J. Kolari, D. Glennon, H. Shin, and M. Caputo. Predicting large US commercial bank failures. *Journal of Economics and Business*, 54(4):361–387, 2002. 3

M. Mariathasan and O. Merrouche. The manipulation of basel risk-weights. *Journal of Financial Intermediation*, 23(3):300–321, 2014. 1

D. Martin. Early warning of bank failure. *Journal of Banking & Finance*, 1(3):249–276, Nov. 1977. ISSN 03784266. 3

W. McCulloch and W. Pitts. A logical calculus of the ideas immanent in nervous activity. *Bulletin of Mathematical Biophysics*, 5:115–133, 1943. 5, 30

J. A. Ohlson. Financial Ratios and the Probabilistic Prediction of Bankruptcy. *Journal of Accounting Research*, 18(1):109, 1980. ISSN 00218456. 3

B. Olshausen and D. Field. Emergence of Simple-Cell Receptive Field Properties by Learning a Sparse Code for Natural Images. *Nature*, 381:607–609, 1996. 21

P. Ravi Kumar and V. Ravi. Bankruptcy prediction in banks and firms via statistical and intelligent techniques – A review. *European Journal of Operational Research*, 180(1):1–28, July 2007. ISSN 03772217. 3, 4

L. M. Salchenberger, E. M. Cinar, and N. A. Lash. Neural Networks: A New Tool for Predicting Thrift Failures. *Decision Sciences*, 23(4):899–916, July 1992. ISSN 0011-7315, 1540-5915. 3

K. Tam. Neural network models and the prediction of bank bankruptcy. *Omega*, 19(5):429–445, Jan. 1991. ISSN 03050483. 3

K. Y. Tam and M. Kiang. Predicting bank failures: A neural network approach. *Applied Artificial Intelligence*, 4(4):265–282, Jan. 1990. ISSN 0883-9514, 1087-6545. 3

T. Xu, K. Hu, and S. Das, Udaibir. Bank Profitability and Financial Stability. *IMF Working Paper*, 2019.

L. Zhou. Performance of corporate bankruptcy prediction models on imbalanced dataset: The effect of sampling methods. *Knowledge-Based Systems*, 41:16–25, Mar. 2013. ISSN 09507051. 3

M. Zoričák, P. Gnip, P. Drotár, and V. Gazda. Bankruptcy prediction for small- and medium-sized companies using severely imbalanced datasets. *Economic Modelling*, 84:165–176, Jan. 2020. ISSN 02649993. 3

A Why capital dominates liquid assets in predicting banks' failure: a theoretical insight

Let us assume a bank that invests in a portfolio of assets funding through both capital (in proportion $1-y$) and bonds (in proportion y). This portfolio is made both of a riskless asset (in proportion x) and of a risky asset (in proportion $1-x$). The balance sheet of the bank is thus as follows:

Asset	Liability
Risky Asset $(1-x)$	Bonds $(1-y)$
Riskless Asset (x)	Capital (y)

The risky asset pays a random return that it as follows: it pays $\pi > 1$ with probability p and 0 with probability $1-p$. We assume that bondholders are paid the riskless return 1. The bank is thus solvent whenever the following inequality holds:

$$x + (1-x)\pi p \geq (1-y) \iff p \geq p^* \equiv \frac{1-y-x}{(1-x)\pi}. \tag{4}$$

p^* is therefore the threshold value of p such as when $p < p^*$ the bank ends up defaulting, and when $p \geq p^*$ the bank is solvent. More precisely, we have:

- when $y > 1-x$, we have $p^* < 0$, the bank is always solvent,
- when $y \leq 1-x$, the bank is solvent whenever $p \geq p^*$.

Let us differentiate p^* with respect to y:

$$\frac{\partial p^*}{\partial y} = -\frac{1}{(1-x)\pi} < 0. \tag{5}$$

Let us differentiate p^* with respect to x:

$$\frac{\partial p^*}{\partial x} = -\frac{y}{(1-x)^2\pi} < 0. \tag{6}$$

We notice that increasing capital always reduces more the probability of default than increasing liquid asset holding. We indeed have $\left|\frac{\partial p^*}{\partial y}\right| \geq \left|\frac{\partial p^*}{\partial x}\right| \iff y \leq 1-x$. When $y > 1-x$, we know that the bank is always solvent. In this very simplistic model, capital thus always dominates liquid assets as a regulatory tool to prevent bankruptcy.

B Data sources and definitions

Table 5 – Data sources and definitions

Data	Definition	Source
Total equity total assets	Ratio of total equity to total assets. This ratio is close to the leverage ratio as defined under Basel agreements.	FitchConnect
Total reg cap ratio	Total regulatory capital ratio as defined under Basel agreements. It is fixed to 8% of the risk weighted assets, plus a conservation buffer (2%).	FitchConnect
Liquid assets total assets	Liquid assets detained by the bank over its total assets	FitchConnect
Net loans total assets	Ratio of net loans to total assets.	FitchConnect
Deposits mm funding growth	Growth rate of deposits to money market funding.	FitchConnect
Expenses rev	Expenses over revenues ratio.	FitchConnect
Net int margin	Returns on invested funds. It is measured by the difference between the interests received and those paid, divided by the average invested assets.	FitchConnect
Non earning assets gs over TA	All assets that do not generate income over total assets.	FitchConnect
Operating roaa	Ratio of net income to average total assets. It measures the profitability of assets, meaning how a firm uses the resources it owns to generate profit. It refers to the returns on the assets purchased using each unit of money invested.	FitchConnect
Tax expense over TA	Expense for current and deferred tax for the period over total assets.	FitchConnect
Total desposits over TA	Total deposits over total assets.	FitchConnect
Total funding over TA	Total Deposits, Money Market and Short-term Funding + Total Long Term Funding + Derivatives + Trading Liabilities, all over total assets.	FitchConnect
Total int exp over TA	Ratio of total interest expense / Total assets.	FitchConnect

Table 5 – (continued)

Total non earning assets over TA	All assets that do not generate income, over total assets.	FitchConnect
Total oper expense over TA	Operating costs include administration costs such as staff costs, over total assets	FitchConnect
log Avg earning assets over TA	Logarithm of year assets that generate income, over total assets.	FitchConnect
log Total assets	Logarithm of total assets. It gives a proxy for banks' size.	FitchConnect
log Bus volume over TA	Logarithm Total Business Volume = Managed Securitized Assets Reported Off-Balance Sheet + Other off-balance sheet exposure to securitizations + Guarantees + Acceptances and documentary credits reported off-balance sheet + Committed Credit Lines + Other Contingent Liabilities + Total Assets. All over total assets.	FitchConnect
log Liquid assets deposits mm funding	Liquid assets as a deposit.	FitchConnect
log Total non int exp over TA	Non interest expenses over total assets.	FitchConnect

C Methodology

C.1 Synthetic Minority Over-sampling Technique (SMOTE)

Introduced by Chawla et al. (2002), Synthetic Minority Over-sampling Technique is inspired by Ha and Bunke (1997) and is designed to address both the issues associated with imbalanced data and the limitations of over-sampling with replacement. This technique is built in such a way that it replicates the initial data distribution. It works as follows:

- We focus on the minority class: $E_{min} = \{i \in [\![1, N]\!] | y_i = 1\}$, N being the number of banks, y is the dependent variable that scores 1s at time $t-1$ when a bank fails in t and 0 otherwise

- For all individual j in E_{min}, we take the difference between its features x_j and their k nearest neighbors: $diff(x_j, knn(x_j))$

- $diff(x_j, knn(x_j))$ is then multiplied by a random factor rd selected between 0 and 1

- $diff(x_j, knn(x_j)) \times rd$ constitutes a new synthetic observation in the minority class

This process is then repeated until the desired weights of classes are reached. As mentioned in Section 3, this procedure is only applied on the training data set. Therefore, we can measure its efficiency by a simple comparison between out-of-sample scores of classification models with and without SMOTE applied to data.

C.2 Decision tress and Random Forest (RF)

Random Forest classification is a supervised statistical learning methodology that performs well out-of-sample (Hastie et al., 2009), and allows to capture non-linearities and interactions between variables.

The main idea behind the RF method is to average a more or less large number of decision trees. A tree is built by partitioning the space of explanatory variables into regions, and then by predicting an output value in each final region. The M final regions (or leaves) of the tree $\{R_m, m \in [\![1, M]\!]\}$, are obtained *via* recursive binary partitions. At each split of features space, we choose the variable for which the split gives the best fit of the output variable (or label). Once the tree is built, the estimated probability $\hat{p}_{1m}$ of default in region m is given by the proportion of default in the region:

$$\hat{p}_{1m} = \frac{1}{N_m} \sum_{x_i \in R_m} I(y_i = 1)$$

where N_m is the cardinal of region m, m is the region with $m \in [\![1, M]\!]$, $I(y_i = 1)$ is the function that scores 1 if y_i equals 1 and 0 otherwise. Therefore, this method is a non-parametric estimation of the unknown function f. This function defines the true model: $P(y = \{0, 1\}|X = x) = f(x) + \epsilon$, where ϵ designates the error term.

The best splitting point is computed for all variables and the variable for which the splitting point gives the best minimization of the criterion is chosen. We use the Gini index impurity measure (criterion to minimize) given by (Hastie et al., 2009):

$$\sum_{k \neq k'} \hat{p}_{mk} \hat{p}_{mk'} = \sum_{k=0}^{1} \hat{p}_{mk}(1 - \hat{p}_{mk})$$

A second step in building a decision tree is to determine the maximum depth of a tree and the minimum number of observations in every leaves. Indeed shallow trees are likely to have poor prediction performance, and too deep trees might lead to overfitting issues and consequently bad out-of-sample forecasting. Following the same logic, a large number of observations per final region will predict poorly, while too little observations per leave are also subject to overfitting problems.

Thus the determination of those two parameters (depth and observations per leave) is crucial and can be done in various ways. In the context of a single tree, Hastie et al. (2009) propose

to rely on a cost complexity criterion that should be minimized. This procedure works on the fact that an increase in complexity (measured by the depth of the tree) that leads to overfitting the data and decreases the sum of squares is counterbalanced by an increase in a cost term that depends on the tree's depth. In the context of RF, this approach is quite demanding in terms of calculation: the criterion must be minimized for each tree. Another technique consists in making varying those two parameters in multiple RF classification estimations simultaneously and retain those that maximize the out-of-sample prediction performance.
The last parameter to establish is the number of trees in the forest. There are some debates on the optimal value for this parameter (and the very existence of an optimum). Hastie et al. (2009) suggest that the error of the model generally decreases and converges as the number of trees grows. From this perspective, the right number of trees corresponds to the moment where the error does not decrease below a certain threshold.

C.3 Artificial Neural Networks (ANN)

Artificial Neural Networks, introduced by McCulloch and Pitts (1943), also constitute a supervised statistical learning methodology that has gained attention in recent decades, in an increasing number of areas (Hastie et al., 2009). The general principle of ANN is to stem features T_k by linear combinations of the inputs Z_m and then predict output values $f_k(X)$ from a non-linear function $g_k(.)$ applied to those features. In a binary classification model, it gives:

$$\begin{aligned} Z_m &= \sigma(\alpha_{0h} + \alpha_h^T X),\, h \in [\![1, H]\!] \\ T_k &= \beta_{0k} + \beta_k^T Z,\, k \in \{0, 1\} \\ f_k(X) &= g_k(T),\, k \in \{0, 1\} \end{aligned}$$

where $\sigma(.)$ is the simoid function given by $\sigma(v) = \frac{1}{1+e^{-v}}$, H is the number of hidden units in the hidden layer, and $g_k(.)$ is the softmax function given by $g_k(T) = \frac{e^{T_k}}{\sum_{l \in \{0,1\}} e^{T_l}}$. We note the full set of parameters, or weights, θ:

$$\begin{aligned} &\{\alpha_{0,h}, \alpha_h; h \in [\![1, H]\!]\} \\ &\{\beta_{0,k}, \beta_k; k \in \{0, 1\}\} \end{aligned}$$

The error function to minimize is given by the cross-entropy measure:

$$R(\theta) = -\sum_{i=1}^{N} \sum_{k \in \{0,1\}} y_{ik} \log(f_k(x_i))$$

As specified by Hastie et al. (2009), the research for the global minimizer of $R(\theta)$ is likely to lead to overfitting issues. This is managed by either early stopping procedure or penalization term. The parameters of the model are estimated *via* gradient descent, and the gradient is computed using the back-propagation algorithm. This algorithm works as follows:

- Initial values for the weights are randomly chosen, generally close to zero
- The weights begin fixed, predicted value $\hat{f}_k(X)$ is computed
- This prediction allows to assess errors in the output δ_k and hidden layer s_m, that are used in the gradient computation
- The gradient is finally used to adjust the weights

ANN training must be done with some precautions regarding some aspects. Considerations on the initial values of weights must be done for two reasons: (i) too small values will lead the network to collapse into a linear model[12], and (ii) multiple values for the initial weights should be tested since $R(\theta)$ is non-convex and that the final solution can vary in function of those. As mentioned above, because of the important number of parameters, the search for a global minimum of R might lead to overfitting. To avoid this, a regularization (or penalization) term can be added to the error function that we seek to minimize. This hyperparameter can be optimized through multiple regression. Finally, it is worth mentioning that the final prediction of a network can depend on the fact that inputs ave been scaled or not. This can be controlled by comparing results on both models: with and without scaled features.

C.4 Partial Dependence Plots (PDP)

Partial Dependence Plots (Friedman, 2000; Hastie et al., 2009) belong to quantitative input influence techniques to visualize features' impact on labels in opaque models. A PDP provides a summary of the output dependence on the joint values of the inputs (Friedman, 2000; Hastie et al., 2009). Considering a subset of $l < p$ inputs $X_{S, S \subset \{1,2,\ldots,p\}}$ of $X^T = (X_1, \ldots, X_p)$, such that $f(X) = f(X_S, X_{S^C})$,[13] the partial dependence of f to X_S is given by:

$$f_S(X_S) = E_{X_{S^C}} f(X_S, X_{S^C}).$$

Note that this equation defines a measure of X_S effect on $f(X)$ after accounting for X_{S^c} effect. To calculate this impact in practice, we proceed as follows. We first assess Individual Conditional Effect (ICE), meaning the partial dependence of $f(X)$ on X_S when considering values of X_{S^c} for a given individual i:

$$ICE_i = \{\hat{f}(x_S^k, X_{i,S^C}), x_S^k \in [X_S^{min}, X_S^{max}]\}, \quad (7)$$

[12]Usually, we choose values close to zero. The network is then an approximately linear model and becomes more non-linear as the weights increase.
[13]S^C being the complementary of S: $S \cup S^C = \{1, 2, \ldots, p\}$.

where ICE_i is the ICE for the i-th individual, X_{i,S^C} refers to values of X_{S^C} of this individual, and x_S^k are the values of X_S that vary from its minimum to its maximum with a step k. This provides a set of points representing a plot of partial dependence of the explained label on the variables included in S for the i-th individual. In a second step, we average those plots for all the individuals, and we obtain the PDP.

C.5 Accumulated Local Effects (ALE)

One of the most important issues in PDPs is that they assume independence between the predictor for which the partial dependence is computed and the other one. Besides, making x_S^k vary across all the distribution of X_S creates a risk to overfit regions with almost no data. In order to overcome this issue, we rely on Accumulated Local Effect (ALE) (Datta et al., 2016) . ALE also proposes to calculate the marginal effect of X_S. The main differences with PDP can be summarized as follows: ALE is unbiased even when features are correlated, it marginalizes over probable combinations of features, and it is faster to compute. Technically, ALE bases its calculation on existing data intervals for explanatory variables. Moreover, ALE averages the changes of predictions, not the predictions themselves. Another significant difference with PDP is that ALE accumulates the local gradients over the range of features S, giving their effect on the predicted variable. Finally, ALE method is centred so that the average effect is zero.
In practice, ALE for one given feature is computed, dividing it into many intervals, and computing the differences in the predictions.[14] First, the uncentred effect is calculated:

$$\hat{\tilde{f}}_j(x) = \sum_{k=1}^{k_j(x)} \frac{1}{n_j(k)} \sum_{i:x_j^{(i)} \in N_j(k)} [f(z_{k,j}, x_{\setminus j}^{(i)}) - f(z_{k-1,j}, x_{\setminus j}^{(i)})],$$

where $\hat{\tilde{f}}_j(x)$ is the uncentred effect of the variable j. $f(z_{k,j}, x_{\setminus j}^{(i)})$ gives the prediction given by the model, and considers the i-th individual for features values excepted x_j that takes the value $z_{k,j}$. The z are the values taken by the variable X_j that has been distributed on a grid defined by a given step. The internal sum adds up the impacts of all individuals within an interval $(i : x_j^{(i)} \in N_j(k))$ that appears as a neighbourhood. This sum is weighted by the number of individuals $n_j(k)$ present in the k-th neighbourhood. Finally, we sum the average effect over all intervals.
Second, we center in order to obtain a null main effect:

$$\hat{f}_j(x) = \hat{\tilde{f}}_j(x) - \frac{1}{n} \sum_{i=1}^{n} \hat{\tilde{f}}_j(x^{(i)})$$

[14] This approximates the local gradients and allows us to compute ALE using RF classification, as well as ANN.

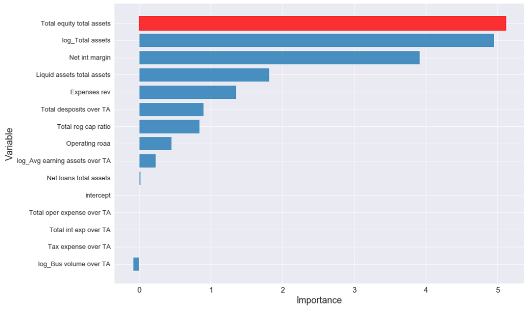
$\hat{f}_j(x)$ is interpreted as the main impact of the explanatory variable compared to the average prediction of the data.

D Results for Europeans banks

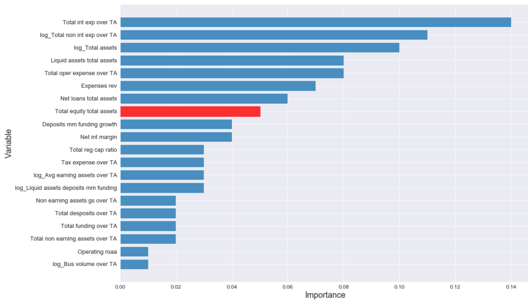
D.1 Features' importance

Figure 8 – Variables relative importance - Europe

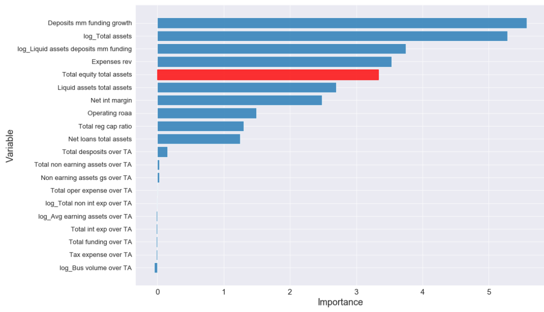
(a) Logistic regression



(b) Random Forest



(c) Artificial Neural Networks



Source: Authors' calculations.

D.2 Features' impact on default probability

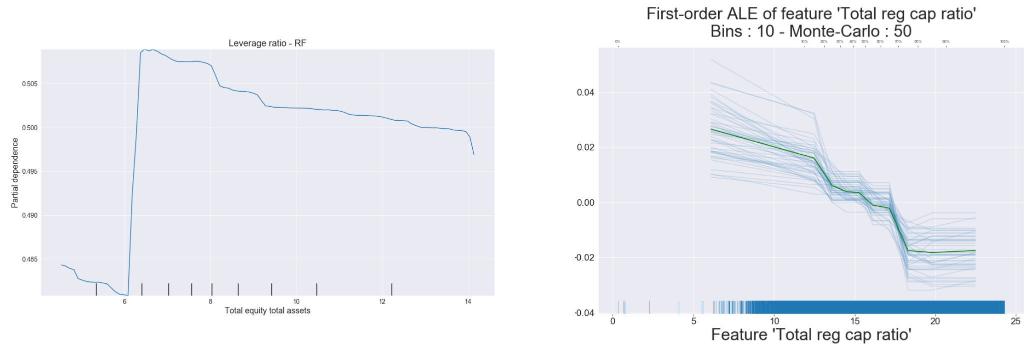
Table 6 – Logistic regression - US

Variables	Odds Ratio	Coefficient p-values
Total equity total assets	-0.157***	0.000
Total reg cap ratio	-0.051***	0.000
Liquid assets total assets	0.016***	0.000
Expenses rev	0.003***	0.000
Net int margin	0.232***	0.000
Net loans total assets	-0.011***	0.000
Operating roaa	-0.192***	0.000
Tax expense over TA	2.3e+57***	0.000
Total desposits over TA	-0.464***	0.000
Total int exp over TA	5.370	0.160
Total oper expense over TA	17.781***	0.001
log_Avg earning assets over TA	2114.767***	0.000
log_Bus volume over TA	-0.889***	0.000
log_Total assets	-0.375***	0.000
intercept	4029.465***	0.000
Nb. of observations	111502	
Nb. of banks (before SMOTE)	3138	
Nb. of defaults (before SMOTE)	331	

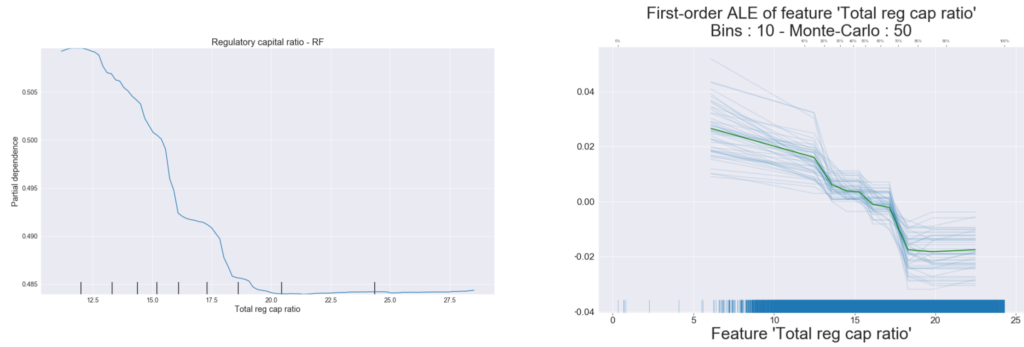
Source: Authors' calculations. Odds ratio are calculated as the exponential of estimated coefficients. To ease the reading, we have subtracted 1 from the OR.

Figure 9 – PDP and ALE - RF classifier - Europe

(a) Total equity over total assets



(b) Total regulatory capital

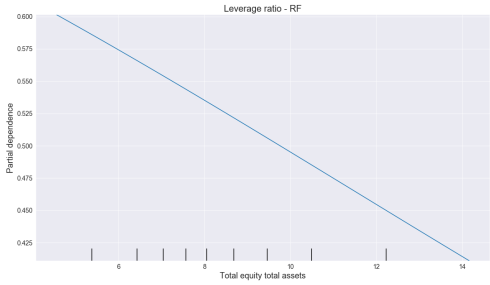


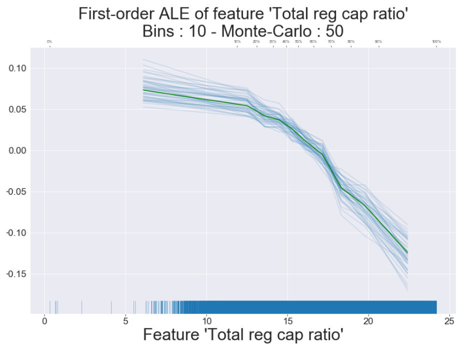
(c) Liquid assets over total assets

Source: Authors' calculations. PDPs on the left, ALEs on the right.

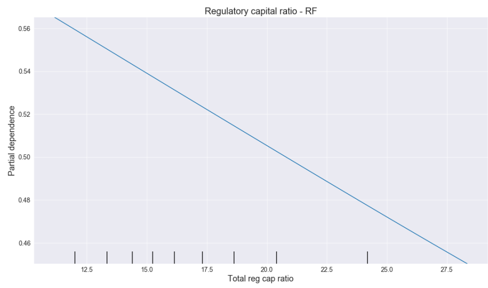
Figure 10 – PDP and ALE - ANN classifier - Europe

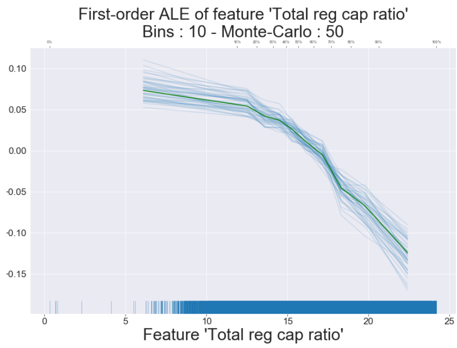
(a) Total equity over total assets



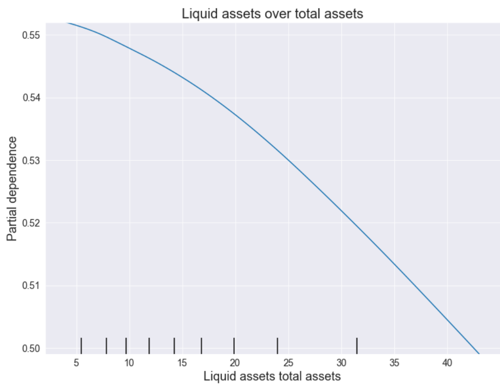


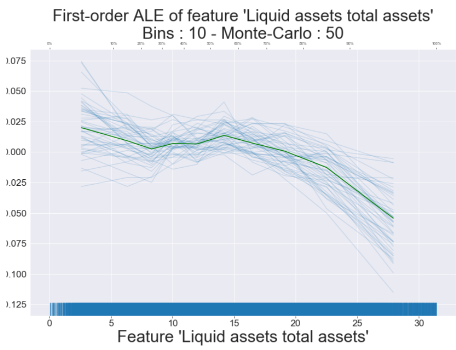
(b) Total regulatory capital





(c) Liquid assets over total assets





Source: Authors' calculations. PDPs on the left, ALEs on the right

E Robustness outputs

E.1 Models with first difference

Table 7 – Dynamic models' performance - US versus Europe

Scores	Logit		RF		ANN	
	In-sample	Out-of-sample	In-sample	Out-of-sample	In-sample	Out-of-sample
US						
Score	72.43	84.16	91.38	95.50	64.03	49.94
TPR	60.57	61.34	86.76	83.19	77.56	78.15
TNR	84.3	84.29	96.01	95.57	50.19	49.83
AUROC	72.18	69.74	97.13	96.16	76.58	77.32
AUPR	80.28	24.82	97.33	29.34	83.31	10.93
Europe						
Score	62.02	67.23	95.59	89.38	53.55	48.38
TPR	66.25	55.32	99.66	17.02	58.62	46.81
TNR	54.83	55.11	91.54	90.02	48.5	48.4
AUROC	63.73	54.95	99.44	61.63	52.67	44.43
AUPR	62.79	0.92	99.38	1.2	49.71	0.71

Source: Authors' calculations. All scores are defined in Section 3 and displayed in %. In red, the out-of-sample rate of failed banks identified as so.

E.2 Standardized variables

Table 8 – Models with standardized variables performance - US versus Europe

Scores	Logit		RF		ANN	
	In-sample	Out-of-sample	In-sample	Out-of-sample	In-sample	Out-of-sample
US						
Score	90.41	98.90	93.40	92.51	96.97	78.11
TPR	86.91	89.43	90.05	86.99	97.44	83.74
TNR	93.91	93.86	96.77	96.68	96.51	96.29
AUROC	96.49	80.44	98.25	92.46	99.49	70.80
AUPR	96.89	24.43	98.40	11.53	99.39	32.87
Europe						
Score	68.75	63.01	82.06	71.77	84.36	48.38
TPR	73.74	55.42	91.53	53.01	90.12	40.96
TNR	63.76	63.11	72.6	72.02	78.61	78.15
AUROC	72.82	63.09	89.49	41.56	91.45	56.10
AUPR	65.93	2.16	87.21	2.26	89.27	11.11

Source: Authors' calculations. All scores are defined in Section 3 and displayed in %. In red, the out-of-sample rate of failed banks identified as so.

E.3 Reduced time dimension for Europe

Table 9 – Models' performance on the 2008-2018 period - Europe

Score	Logit		RF		ANN	
	In-sample	Out-of-sample	In-sample	Out-of-sample	In-sample	Out-of-sample
Score	67.28	65.57	86.64	78.88	71.25	59.40
TPR	68.72	54.76	94.47	45.24	83.71	61.9
TNR	65.89	65.66	78.83	79.15	58.79	59.39
AUROC	72.30	65.47	94.59	70.60	77.11	65.32
AUPR	65.08	1.15	93.04	1.43	70.64	1.27

Source: Authors' calculations. All scores are defined in Section 3 and displayed in %. In red, the out-of-sample rate of failed banks identified as so.

E.4 Reduced number of independent variables for the logistic regression

Table 10 – Logistic regression in reduced number of features - US versus Europe

Scores	Logit	
	In-sample	Out-of-sample
US		
Score	90.67	94.95
TPR	86.36	85.37
TNR	95.0	95.0
AUROC	95.55	93.14
AUPR	96.47	46.53
Europe		
Score	66.73	62.44
TPR	70.44	53.01
TNR	63.03	62.56
AUROC	70.22	62.27
AUPR	60.85	1.87

Source: Authors' calculations. All scores are defined in Section 3 and displayed in %. In red, the out-of-sample rate of failed banks identified as so.

Table 11 – Logistic regression - US versus Europe

	US		Europe	
Variables	Odds Ratio	Coef. p-values	Odds Ratio	Coef. p-values
Total equity total assets	-0.239***	0.000	-0.138***	0.000
Total reg cap ratio	-0.111***	0.000	-0.053***	0.000
Liquid assets total assets	0.064***	0.000	0.010***	0.000
Expenses rev	-0.001***	0.000	0.000	0.421
Net int margin	0.259***	0.000	0.202***	0.000
Net loans total assets	0.047***	0.000	-0.012***	0.000
Operating roaa	-0.543***	0.000	-0.078***	0.000
Total desposits over TA	-0.901***	0.000	-0.417***	0.001
Total int exp over TA	-	-	1.286	0.494
log_Bus volume over TA	-0.997***	0.000	-0.798***	0.000
log_Total assets	-0.180***	0.000	-0.376***	0.000
intercept	9182.38***	0.000	650047.95***	0.000

Source: Authors' calculations. Odds ratio are calculated as the exponential of estimated coefficients. To ease the reading, we have subtracted 1 from the OR.