To fail or not to fail? Assessing the microprudential efficiency of Basel III

Additional results and appendices

December 2025

This document provides additional results and robustness checks to support the main results presented in the paper, as well as its appendices. In particular, after reminding mains contributions of the paper, we give insights on the analysis of models' confusion matrices. Results are presented for GBC and HGBC, which are the models that perform best alongside with RF and Logit. Finally we display robustness checks in this document.

1 Contributions of the paper

1.1 Contribution of the paper

This article contributes to the above-presented literature in at least three respects:

- **Capital requirements.** We provide evidence that, to lower the most the probability of default, the best option is to set the leverage ratio at 10% and the risk-weighted ratio at 15%. Above these levels, no further impact on the probability of default is found. The values of the capital ratios found in this paper are consistent with those found by Karmakar (2016); Egan et al. (2017); Dagher et al. (2020); Mendicino et al. (2021).

- **Interaction between liquidity and capital.** We provide evidence that liquidity is not *per se* a strong driver of bank default. Consistently with Pérignon et al. (2018), we therefore question the idea that liquidity risk can act as a driver of financial instability independently of banks being under-capitalized. In fact, we notice that liquidity has a positive impact on the probability of default: failed banks are likely to exhibit more liquid asset portfolios than unfailed banks. In line with the evidence put forward by DeYoung et al. (2018), this may be because banks whose capital deteriorates are forced to sell their illiquid assets. As a consequence, banks whose situation deteriorates up to default might very well be more liquid than sound banks. Such evidence questions the implementation of the LCR in addition to capital requirements. In line with Thakor (2018), we therefore recommend to focus on capital requirements and to release the regulatory pressure put on liquidity. This would allow to reduce the complexity of banking regulation (Herring, 2018), to reduce the regulatory pressure exerted on safe assets (Caballero et al., 2017), and to prevent the shortening of banks' investment time horizon, which could prove detrimental to the funding of low-carbon sectors (Campiglio, 2016).

- **Methodology.** Among the seven models that are run, Logit, RF, GBC and HGBC perform the best. Linear SVC, MLP and KNN are, on the contrary, lagging behind.

2 Additional results: Determinants of bank default

2.1 Confusion matrices analysis

Let us have a look at the confusion matrices (Table 1) for the four models that perform best. Doing so, we notice that these models manage to properly identify the vast majority of the failed banks in the test sample: among the 121 failed banks, Logit identifies 100 defaults, HGBC, RF and GBC identify 98 defaults. What is however less convincing is the number of false positives (FP) : 1047 for Logit, 888 for RF, 848 for HGBC and 941 for GBC. Recall however that our dataset is made of bank-year observations. A failed bank is thus identified as a true positive only the year before the actual default occurred. As a consequence, at date $t-2$ this specific bank is identified as an unfailed one even if it might already exhibit the characteristics of a failed bank. To disentangle what the false positives displayed in Table 1 are made of, we try to identify whether or not they in fact consist in banks that at some point go bankrupt. Results are presented in Table 2. More precisely, we 1) identify all FP (Nb. of FP), 2) identify within these FP the actual number of banks keeping in mind that the same bank can be wrongly identified several times (Nb. of banks in FP), 3) look among these banks for those that actually go bankrupt at some point in the considered time period (Nb. of failed banks in FP), 4) compute the proportion these banks represent among all the banks at least once wrongly identified as failed ones (Prop. of failed banks in FP).

Table 1: Confusion matrices

Train	Logit		HGBC		RF		GBC	
	Pred. 0	Pred. 1	Pred. 0	Pred. 1	Pred. 0	Pred. 1	Pred. 0	Pred. 1
True 0	53353	2396	53988	1761	53832	1917	53660	2089
True 1	8609	47140	1779	53970	5055	50694	5163	50586
Test	Logit		HGBC		RF		GBC	
	Pred. 0	Pred. 1	Pred. 0	Pred. 1	Pred. 0	Pred. 1	Pred. 0	Pred. 1
True 0	22868	1047	23067	848	23027	888	22974	941
True 1	21	100	23	98	23	98	23	98

Source: Authors' calculations.

Having a look at Table 2, we notice that, depending on the considered model, between 14.61% and 18.75% of the false positives are actually previous observations of banks that do eventually go bankrupt. Even if the models make mistakes by identifying as failed banks some banks that are actually sound, a significant proportion of these mistakes concern banks that at some point do indeed go bankrupt.

Table 2: False positives (FP)

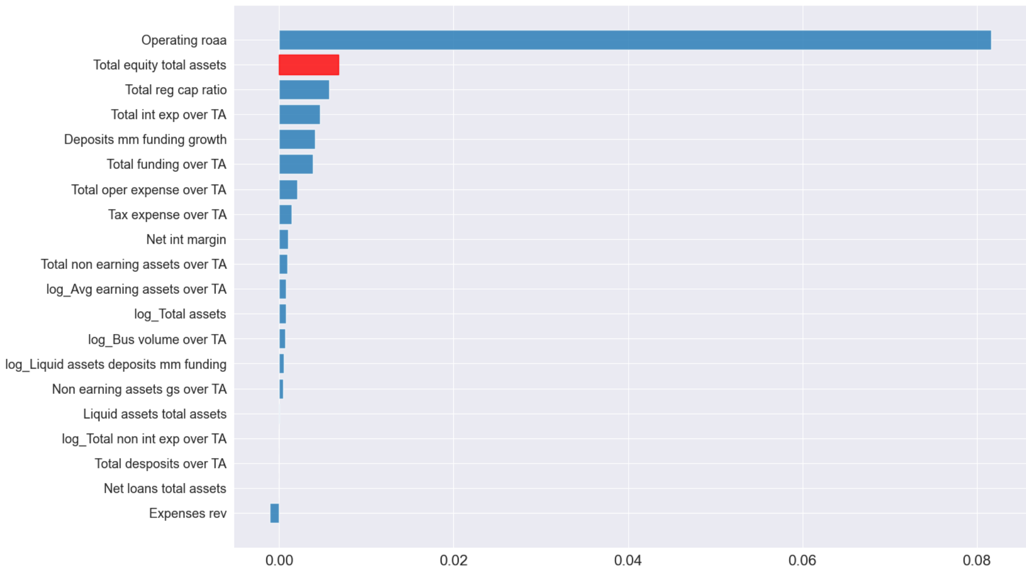
Model	Nb. of FP	Nb. of banks in FP	Nb. of failed banks in FP	Prop. of failed banks in FP (%)
Logit	1047	731	153	14.61
HGBC	848	650	159	18.75
RF	888	668	156	17.56
GBC	941	692	164	17.43

Source: Authors' calculations.

2.2 Permutation feature importance

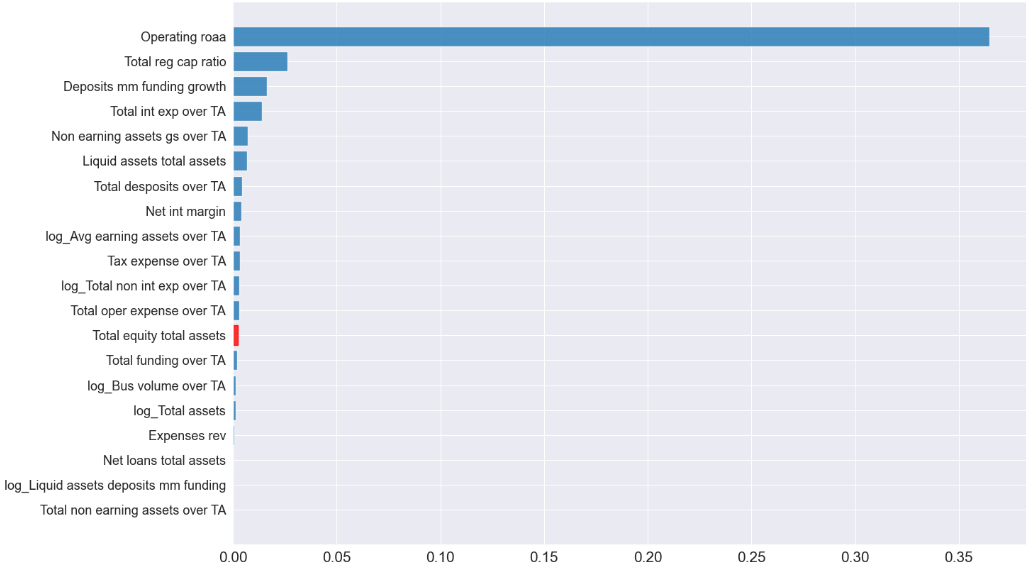
In the main document, total regulatory capital ratio (TRCR) and total equity over total assets (TE/TA) are identified as significant predictors of default in the RF model thanks to permutation feature importance. On the contrary, liquid assets over total assets (LA/TA) is not identified as a strong predictor. This result allows to conclude that capital (TE/TA and/or TRCR) is a greater predictor of default than liquidity. Results are here presented for GBC (Figure 1) and HGBC (Figure 2). The idea that capital is a stronger predictor of default than liquidity is corroborated by these figures: capital is always a stronger predictor of default than liquidity. We however notice that it is not possible to conclude, from these figures, which of TE/TA and TRCR is the strongest predictor.

Figure 1: Permutation feature importance (GBC)



Source: Authors' calculations.

Figure 2: Permutation feature importance (HGBC)

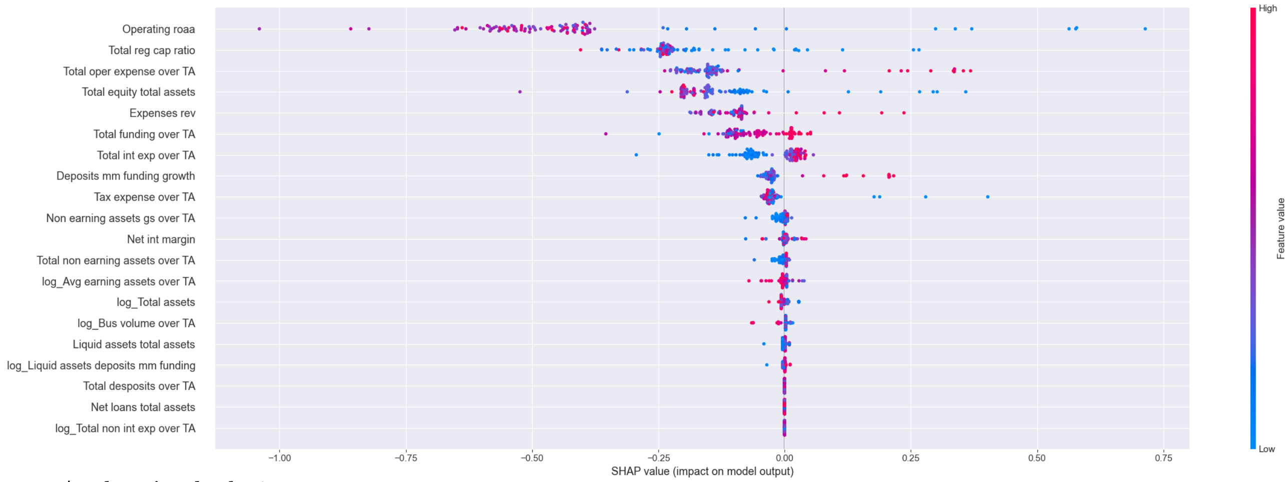


Source: Authors' calculations.

2.3 Shapley values

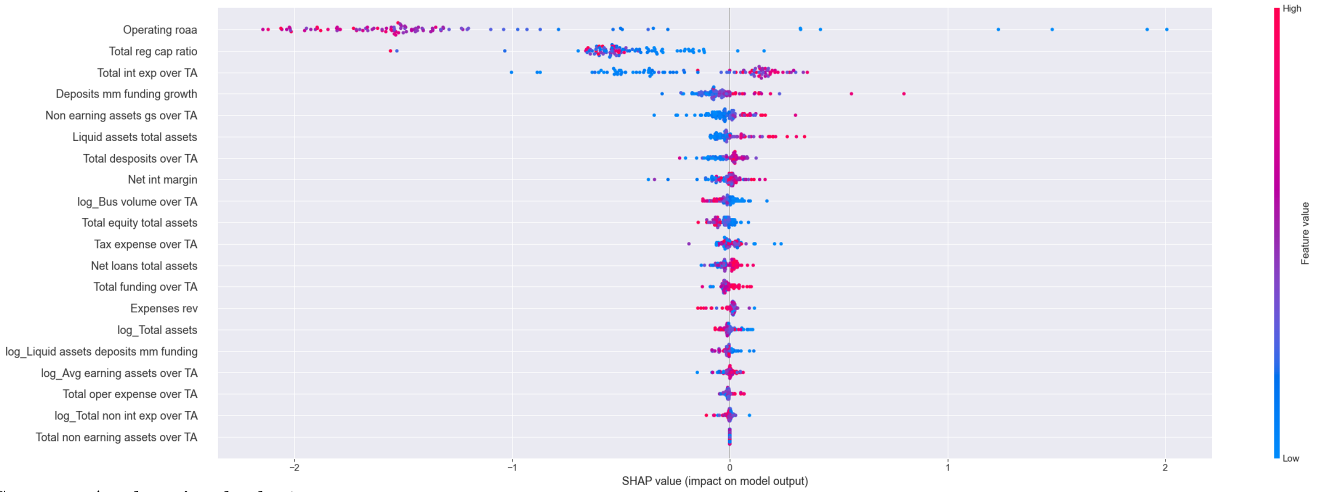
In the main document, Shapley values are used to assess the significance and the nature of the impact of the various predictors on the probability of default. In particular, from Shapley values, we conclude that both TRCR and TE/TA have a significant and negative impact on the probability of default, while the impact of LA/TA is less significant and ambiguous. Shapley values are here presented for GBC (Figure 3) and HGBC (Figure 4). These results are consistent with those of RF.

Figure 3: Shapley values (GBC)



Source: Authors' calculations.

Figure 4: Shapley values (HGBC)



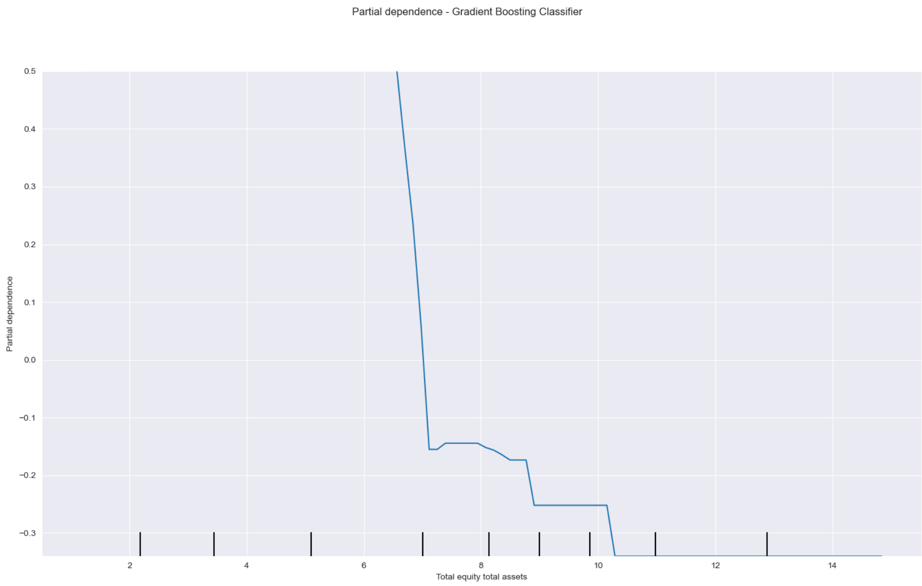
Source: Authors' calculations.

2.4 Capital ratio

2.4.1 Partial dependent plots (PDPs)

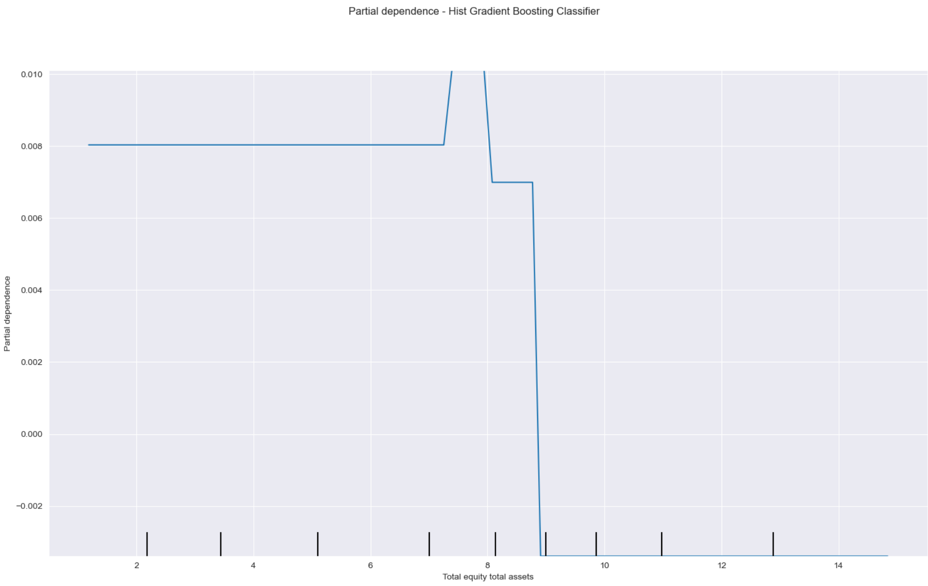
In the main document, PDPs are used to assess the nature of the impact of TE/TA and that of TRCR on the probability of default. Both predictors have a non-linear impact on the probability of default which allows to identify threshold values of TE/TA and TRCR such that below them the probability of default is large and above them the probability of default is low. Results are here presented for GBC (Figures 5 and 7) and HGBC (Figures 6 and 8). Figures 5 and 6 suggest that the impact of TE/TA on the the probability of default is indeed non-linear and that the threshold value of 10 is consistent. Similarly, Figures 7 and 8 corroborate the results presented in the main document concerning the impact of TRCR on the probability of default.

Figure 5: Partial Dependence Plots (PDPs) – TE/TA – GBC



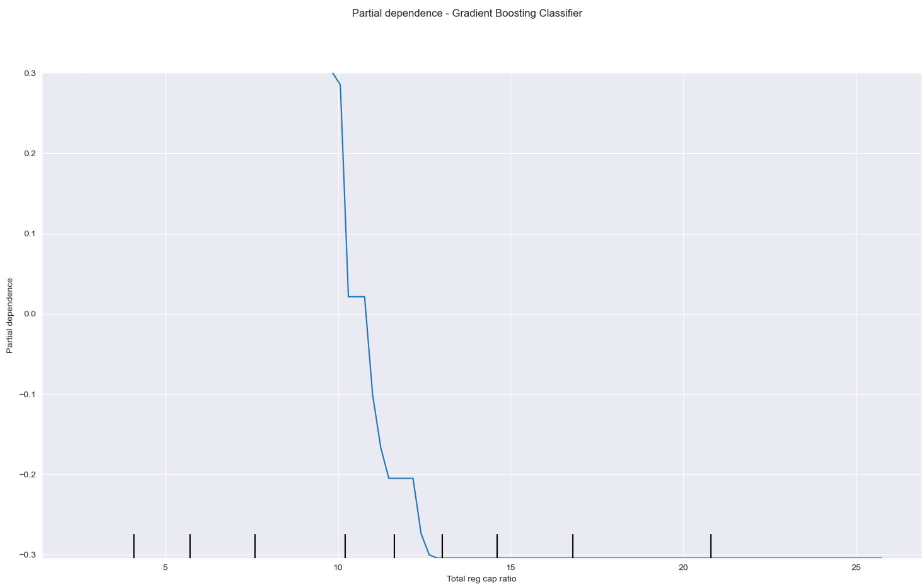
Source: Authors' calculations.

Figure 6: Partial Dependence Plots (PDPs) – TE/TA – HGBC



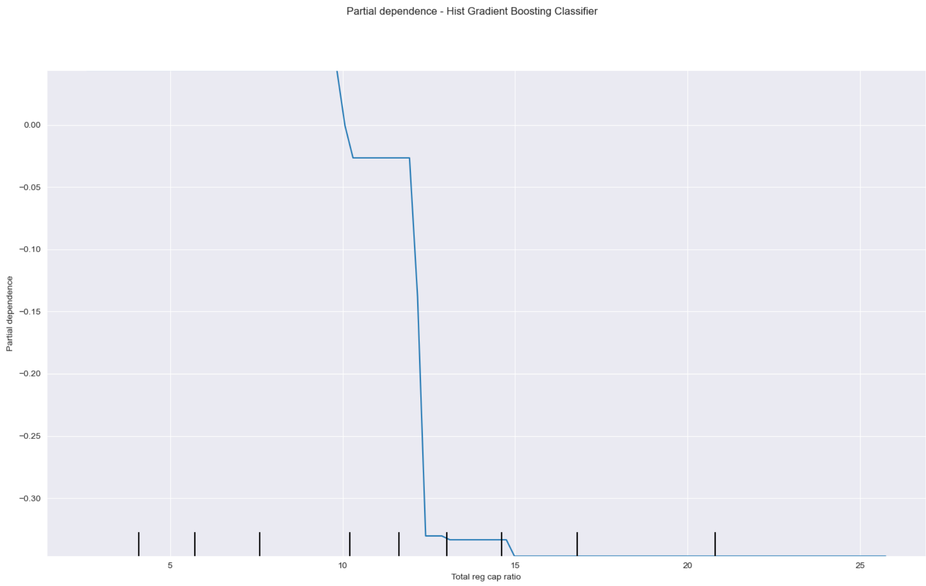
Source: Authors' calculations.

Figure 7: Partial Dependence Plots (PDPs) – TRCR – GBC



Source: Authors' calculations.

Figure 8: Partial Dependence Plots (PDPs) – TRCR – HGBC

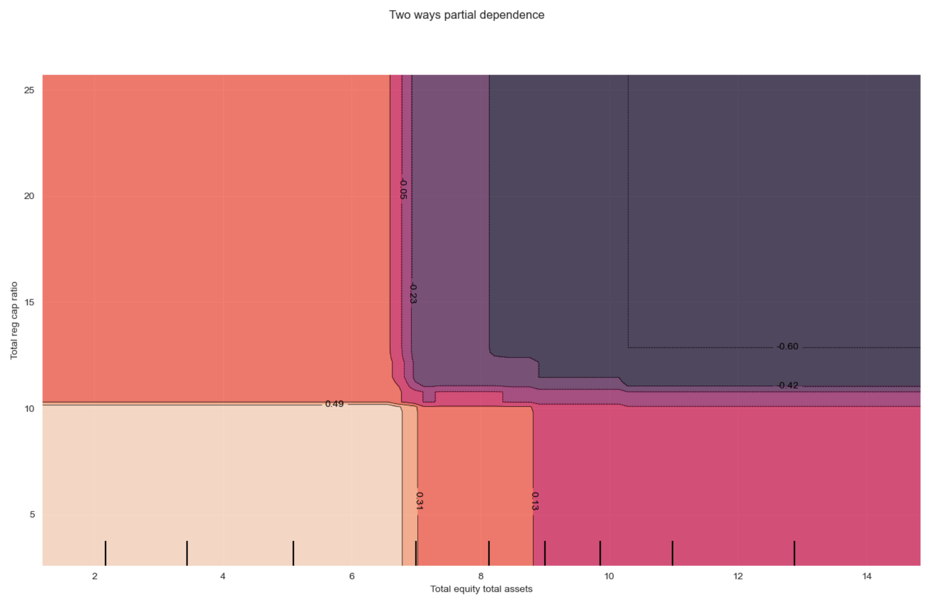


Source: Authors' calculations.

2.5 Two-way PDPs

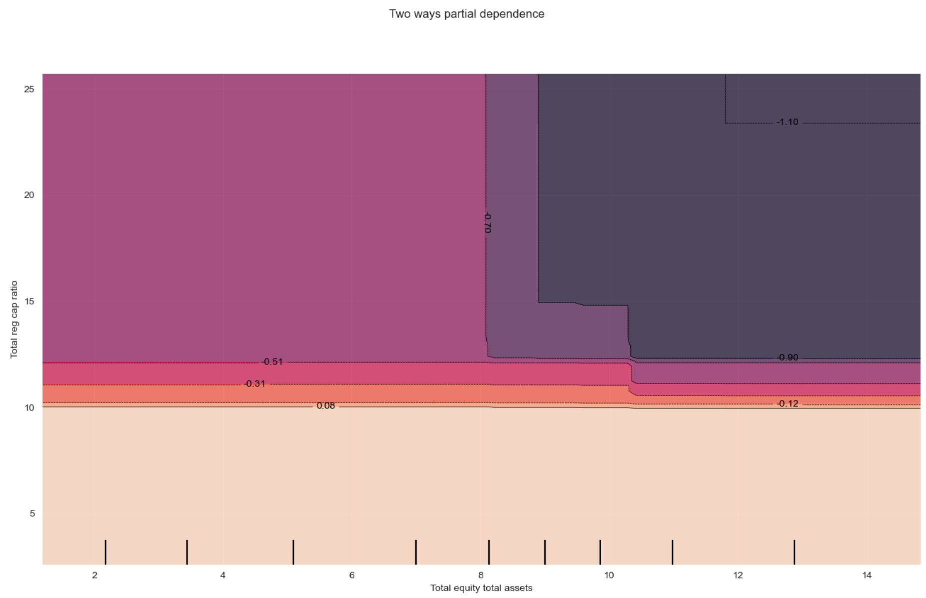
In the main document, two-way PDPs allow to study the impact of the interaction between TE/TA and TRCR on the probability of default. In particular, two-way PDPs allow to identify which values of these two variables allow to reach the space where the probability of default is the lowest. Two-way PDPs are here reported for GBC (Figure 9) and for HGBC (Figure 10). Consistently with the conclusion reached in the main document based on RF, we notice that the probability of default is the lowest when TRCR is above 15 and TE/TA above 10.

Figure 9: Two-way Partial Dependence Plots (PDPs) between TE/TA and TRCR – GBC



Source: Authors' calculations.

Figure 10: Two-way Partial Dependence Plots (PDPs) between TE/TA and TRCR – HGBC

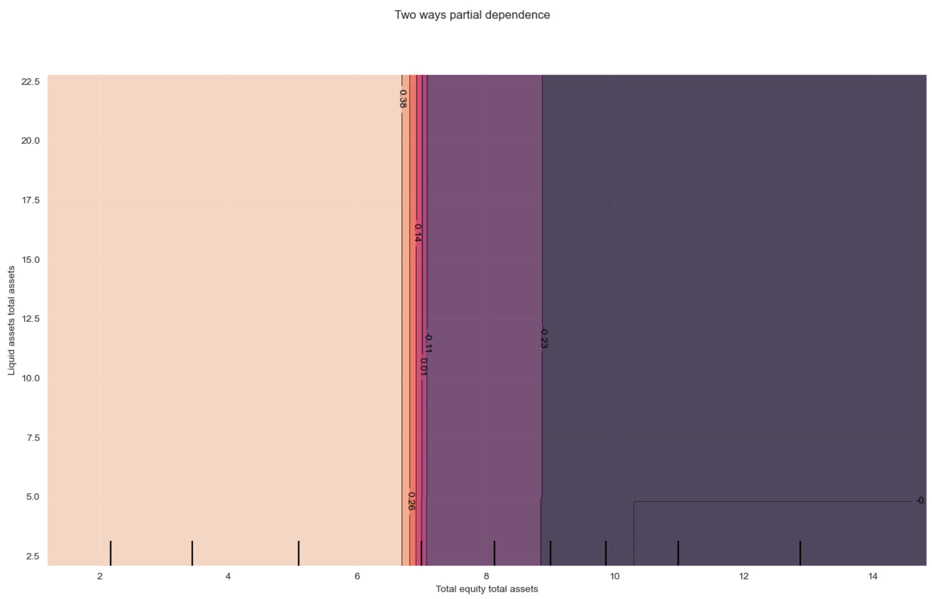


Source: Authors' calculations.

2.6 Interaction between liquidity and solvency risks

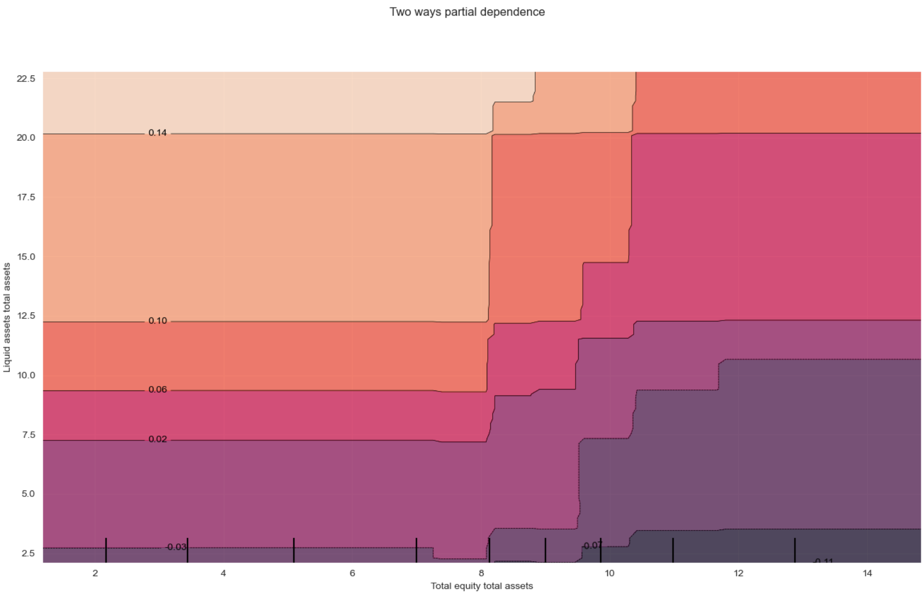
In the main document, we study the impact of the interaction between capital variables (TE/TA and TRCR) and LA/TA. We conclude that, most of the times, the impact of this interaction on the probability of default is driven by the variable accounting for capital. This idea is supported by results coming from GBC and HGBC. In particular, Figures 11, 13 and 14 unambiguously outline that the impact of the interaction between capital and liquidity on the probability of default is driven by the variable accounting for capital. Figure 12 qualifies this result. However, recalling that TE/TA is not a significant predictor of default according to HGBC (see Figures 2 and 4 above), interpretations coming from the reading of PDPs when TE/TA is considered need to be done cautiously.

Figure 11: Two-way Partial Dependence Plots (PDPs) – TE/TA and LA/TA – GBC



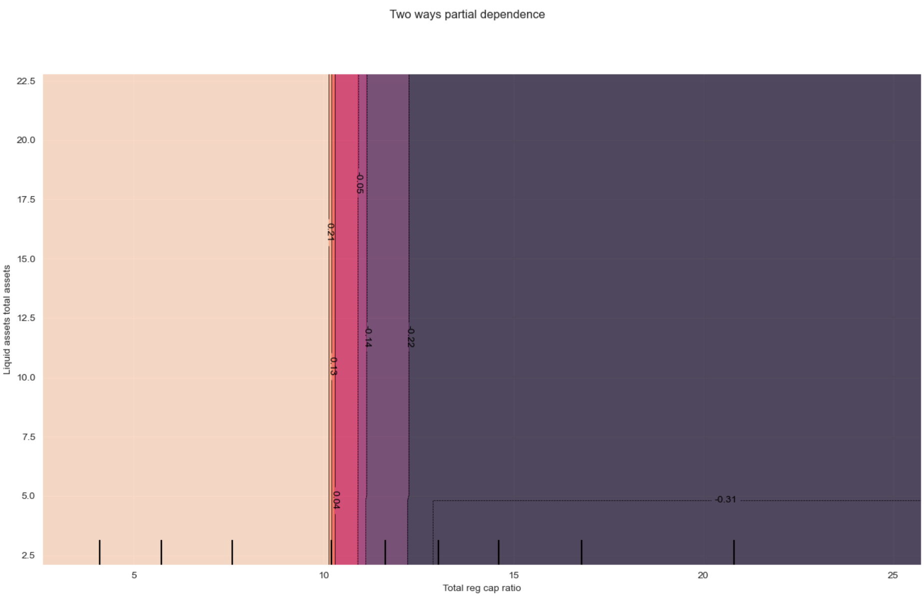
Source: Authors' calculations.

Figure 12: Two-way Partial Dependence Plots (PDPs) – TE/TA and LA/TA – HGBC



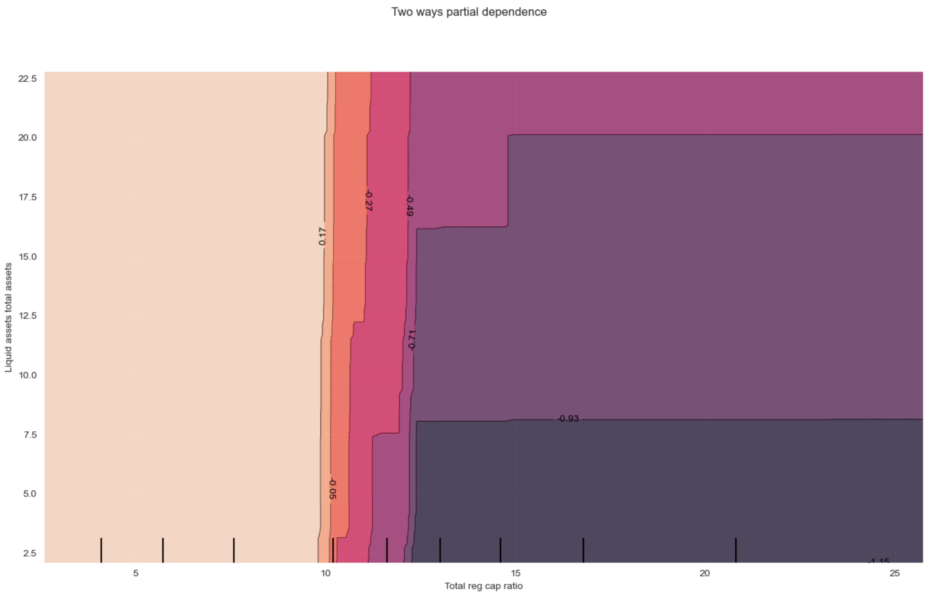
Source: Authors' calculations.

Figure 13: Two-way Partial Dependence Plots (PDPs) – TRCR and LA/TA – GBC



Source: Authors' calculations.

Figure 14: Two-way Partial Dependence Plots (PDPs) – TRCR and LA/TA – HGBC



Source: Authors' calculations.

3 Robustness

3.1 Results on the second database (2014-2024)

To check the robustness of our results accounting for recent bank failures (especially for those related to the 2023 US banking crisis), we run our models on a sample covering US commercial banks from 2014 to 2024. Data come from the Orbis database. Apart from covering recent events, this allows us to focus on the implementation period of Basel III. As in section **??**, results are only displayed for RF. In Figure 16 (see C.1), we report the permutation importance of each variable and the Shapley values. In line with the results presented in section **??**, we notice that the capital ratios (both total equity over total assets and the risk-weighted ratio) are great predictors of the default. On the contrary, the liquidity ratio has close to zero predictive power. Figure 17 (see C.1) displays the PDPs for capital ratios. In line with the results presented in section 2.4, we notice that the relationship between capital ratios and the probability of default is non-linear: the probability of default is high for values of the leverage ratio below 9% and low above, it is high for values of the risk-weighted ratio below 14% and low above. As a consequence, the values of the capital ratios identified when running our models on the main database in section 2.4 (i.e. 10% for TE/TA and 15% for TRCR) hold when running the models on the 2014-2024 database.

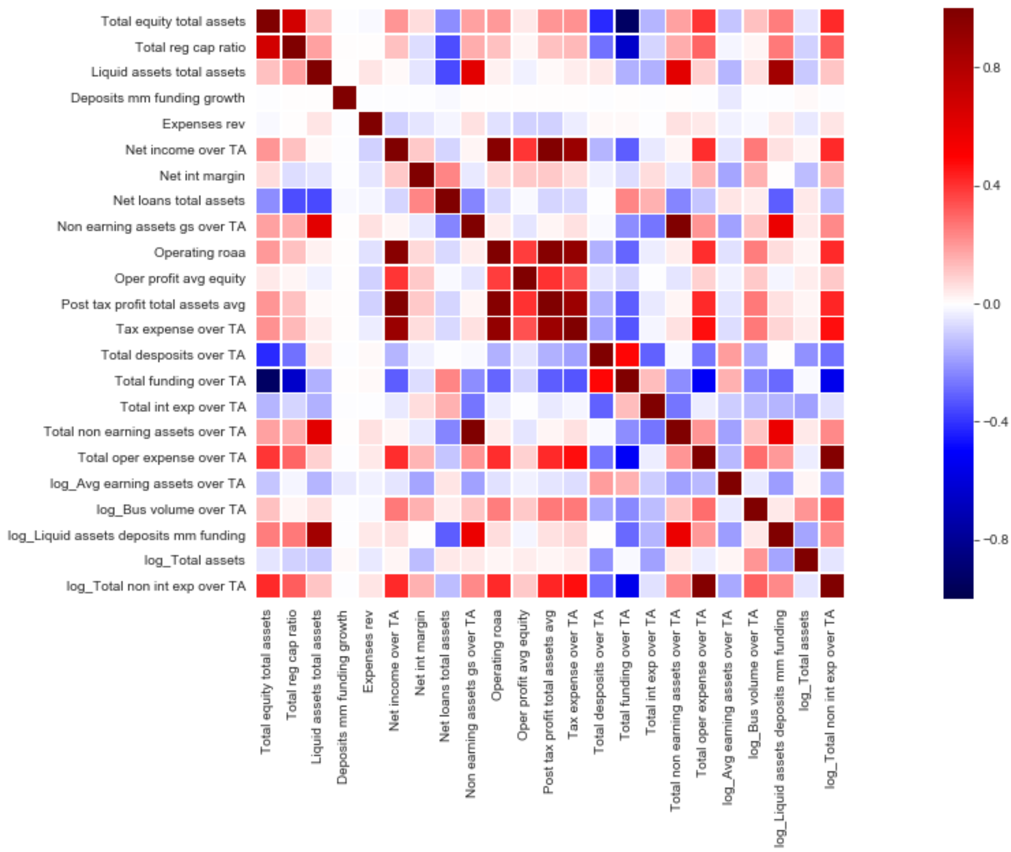
3.2 Fine-tuning the decision threshold?

To classify banks, models actually yield a probability: if this probability is superior (respectively inferior) to 0.5, the predicted class is 1 (respectively 0). Therefore, to increase the performance of a specific model, we could have fine tuned the decision threshold so as to choose the value that maximizes a given performance measure. Having a look at the distributions of the predicted probabilities for all the models (see Figure 18 in C.2) we notice that all the distributions exhibit two clear-cut mods: one around 0 and the other around 1. In other words, changing the value of the decision threshold would only marginally modify the performance of the models (it would actually mostly re-balance the proportion of false positives and that of false negatives). More precisely, having a look at the distribution for RF and differentiating between true (in blue) and false (in red) predictions, we notice that the probabilities having a value between 0.4 and 0.6 are equally distributed among true and false (see Figure 19 in C.2 for RF). This result outlines the fact that changing the decision threshold would not change much the global quality of the models.

3.3 Accumulated local effects (ALEs)

One of the most important issues associated with PDPs is that they assume that the predictor for which the partial dependence is computed and the other ones are independent. Figure 15 provides a correlation heatmap. As can be seen, some variables are strongly correlated with each other. To avoid any bias in our estimations, we remove those variables that are the most correlated. More precisely, the following variables have been removed: Net income over Total Assets, Operating profit avg equity, Post tax profit total assets avg. We are therefore left with 20 variables.

Figure 15: Correlation heatmap



Source: Authors' calculations.

As a consequence, the existence of strong linear correlations between some features may bias the computation of PDPs. In addition, making X_j vary across all its distribution creates the risk to overfit regions with almost no data. In order to take these issues into account, we compute Accumulated Local Effects (ALEs) (Datta et al., 2016) as a robustness check. By difference from PDPs, ALEs are unbiased even when features are correlated and they are computed over actual data intervals of the explanatory variables. ALEs are only reported for RF. They are displayed in C.3. Let us have a look at the different subplots presented in Figure 20.

Subplot (a): TRCR indeed has a negative impact on the probability of default. In addition, the two regimes identified in section 2.4 are clearly distinguishable: the probability of default is larger when TRCR $< 15\%$ and decrease above this threshold.

Subplot (b): TE/TA indeed has a negative impact on the probability of default.

In addition, the two regimes identified in section 2.4 are clearly distinguishable: the probability of default is larger when TE/TA < 10% and lower above this threshold.

Subplot (c): LA/TA indeed has a weakly positive impact on the probability of default.

3.4 Without over-sampling

Using the SMOTE procedure to re-balance the dataset allows to increase the performance of the models. However, since it consists in creating new instances in the minority class, it significantly modifies the information the classifiers find in the dataset. To check the robustness of the results presented in this paper, we therefore re-run the models without over-sampling the train sample. Results are presented for all the models in Table 5 in C.4. We do notice that the models underperform when the train sample is imbalanced. More specifically, they get lower scores on the train sample and over-fit the zero class and, as a consequence, fail to properly identify failed banks. Macro recalls are however not so low on both the train and test samples. This can be explained by the propensity of the models to over-predict zeros. KNN is a great illustration: all banks but one are classified as 0 (unfailed banks). In this case, the precisions for both zeros and ones reach 100%, while the recall for ones is only 1%. This is by the way a further argument in favor of using macro recall as the reference score to assess the performance of the models.

3.5 Standardized predictors

In order to control for the impact of (un)standardized features on our models' quality, we ran all hyperparameters pipelines allowing for standardization.[1] Our goal is to see if an increase appears in the models' macro recall, but we keep in mind that standardization does not facilitate for post-estimation interpretation. We find that this is the case for KNN (+ 8.48 % with RobustScaler), Linear SVC (+ 8.48 % with RobustScaler), MLP (+ 10.91 % with StandardScaler) and GBC (+ 2.75 % with RobustScaler). However, no standardization is optimal for Logit, RF and HGBC while potential increases imply a loss in interpretability. See C.5 for the confusion matrices in the test sample for all predicted years.

[1] We used "StandardScaler" and "RobustScaler" modules from Scikit-Learn as extra hyperparameters.

3.6 Prevision performance

Since our main database is limited to the 2000-2018 period because of data availability, we test our models' prevision performance using the latter years of our sample: 2015 to 2018. Precisely, we train our model on oversampled data going from 2000 to 2014 and then use 2015 data to predicted potential default in 2016. We do so for all of our test years untill the last training sample corresponds to the 2000-2017 period and the prediction uses 2018 data for potential defaults in 2019. Over the 14 banks that are going default among those 4 years, our models identify 13 of them. Among the false positives, we often find banks that end up defaulting. We are confident that our models correctly identify default at a given point in time and that, furthermore, the predicted probability serves as a leading indicator (sometimes more than one period in advance) of potential default.

References

R. J. Caballero, E. Farhi, and P.-O. Gourinchas. The Safe Assets Shortage Conundrum. *Journal of Economic Perspectives*, 31(3):29–46, Aug. 2017. 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, Jan. 2016. 2

J. Dagher, G. Dell'Ariccia, L. Laeven, M. L. Ratnovski, and M. H. Tong. Bank Capital: A Seawall Approach. *International Journal of Central Banking*, 16(2):249–291, Mar. 2020. 1

A. Datta, S. Sen, and Y. Zick. Algorithmic Transparency via Quantitative Input Influence: Theory and Experiments with Learning Systems. pages 598–617, 05 2016. 15

R. DeYoung, I. Distinguin, and A. Tarazi. The joint regulation of bank liquidity and bank capital. *Journal of Financial Intermediation*, 34:32–46, Apr. 2018. 2

M. Egan, A. Hortaçsu, and G. Matvos. Deposit Competition and Financial Fragility: Evidence from the US Banking Sector. *American Economic Review*, 107(1):169–216, Jan. 2017. 1

R. J. Herring. The Evolving Complexity of Capital Regulation. *Journal of Financial Services Research*, 53(2-3):183–205, June 2018. 2

S. Karmakar. Macroprudential regulation and macroeconomic activity. *Journal of Financial Stability*, 25:166–178, Aug. 2016. 1

C. Mendicino, K. Nikolov, J.-R. Ramirez, J. Suarez, and D. Supera. How much capital should banks hold? *Research Bulletin ECB*, 80, 2021. 1

C. Pérignon, D. Thesmar, and G. Vuillemey. Wholesale Funding Dry-Ups. *The Journal of Finance*, 73(2):575–617, 2018. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/jofi.12592. 1

A. V. Thakor. Post-crisis regulatory reform in banking: Address insolvency risk, not illiquidity! *Journal of Financial Stability*, 37:107–111, 2018. 2

A Data sources and definitions

Table 3: 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

Table 3: (continued)

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
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

B Hyperparameters

The list of the hyperparameters and their values are presented in Table 4. For each model, the optimal values of the hyperparameters are displayed in bold characters. The last column presents the mean effect (if any) of each hyperparameter (from the left of the list) on the out-of-sample macro recall score.

Table 4: Hyperparameters per model and macro recall

Model	Hyperparameter	value range	mean effect
LR	C	[0.01, 0.1, ***1***, 10, 100, 1000]	inv. U-shape
	penalty	[None, ***l1***, *l2*, ElasticNet]	inv. U-shape
	l1_ratio (if EN)	[0.1, 0.3, 0.5, 0.7, 0.9]	n.s.
	solver	[lbfgs, ***liblinear***, newton-cg, sag, saga]	decreasing
RF	n_estimators	[5, 10, 50, ***100***, 500, 1000]	inv. U-shape
	max_depth	[***5***, 10, 50, None]	decreasing
	max_features	[***5***, 10, sqrt, log2, None]	U-shape
	min_samples_split	[***5***, 10, 20]	increasing
GBC	learning_rate	[0.01, ***0.1***, 0.5]	decreasing
	n_estimators	[5, ***10***, 50,100, 500, 1000]	decreasing
	max_depth	[***5***, 10, 50, None]	decreasing
	max_features	[***5***, 10, sqrt, log2, None]	U-shape
	min_samples_split	[***5***, 10, 20]	decreasing
HGBC	learning_rate	[0.01, ***0.1***, 0.5]	decreasing
	max_iter*	[5, ***10***, 50,100, 500, 1000]	decreasing
	max_depth	[5, ***10***, 50, None]	inv. U-shape
	max_features	[***5***, 10, sqrt, log2, None]	U-shape
	l2_regularization	[***0***, 0.1]	decreasing
SVC	C	[0.1, ***1***, 10, 100, 1000]	inv. U-shape
	kernel	[***linear***, poly, rbf]	U-shape
	degree (if poly)	[***2***, 3]	decreasing
	gamma	[***0.0001***, 0.001]	decreasing
MLP	activation	[logistic, relu, ***tanh***]	inv. U-shape
	alpha	[***0.005***,0.01]	increasing

Table 4: (continued)

	max_iter	[10, ***50***, 100]	decreasing
	early_stopping	[False, ***True***]	increasing
	hidden_layers	39 combinations of [10, 100, 500] a.n.*	n.r.
	alpha	[***0.05***, 0.01]	increasing
KNN	n_neighbors	[4, 15, 25, 30, ***40***, 50]	increasing
	metric	[euclidean, ***manhattan***, minkowski]	inv. U-shape
	weights	[uniform, ***distance***]	increasing

Source: Authors' calculations.

* max_iter in HGBC corresponds to n_estimators in RF and GB.

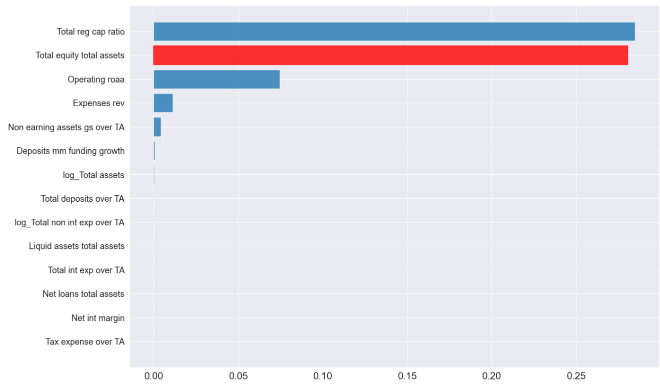
** The smallest possible MLP is chosen with a single layer of 10 artificial neurons.

C Robustness outputs

C.1 Results on the second database (2014-2024)

Figure 16: Variables' importance (second database)

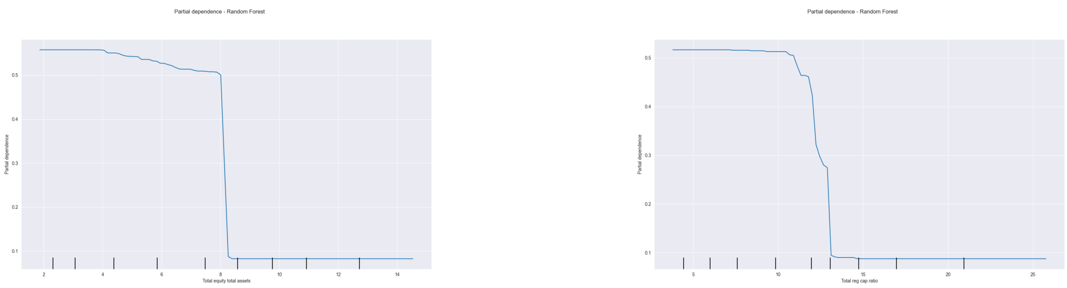
(a) Permutation feature importance



(b) Shapley values (second database)

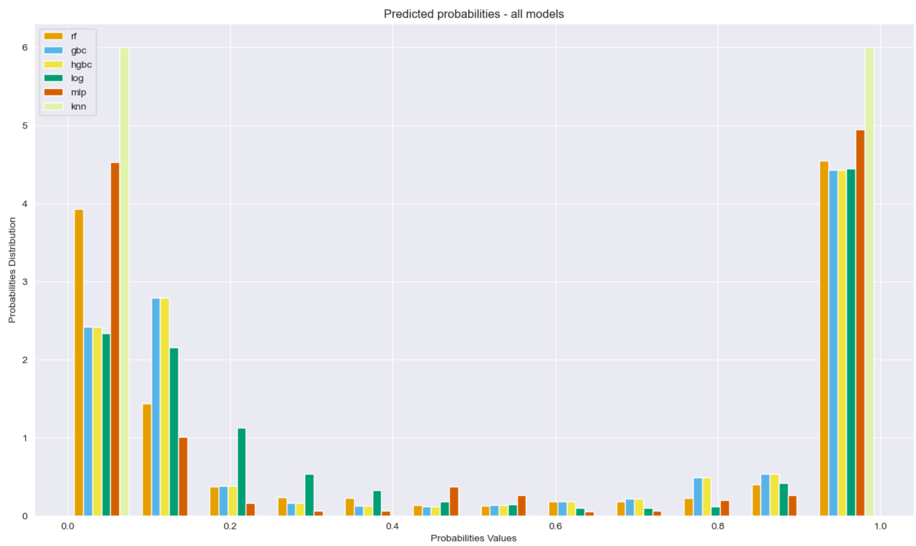
Figure 17: Partial dependence plots (second database)

(a) TE/TA (b) TRCR



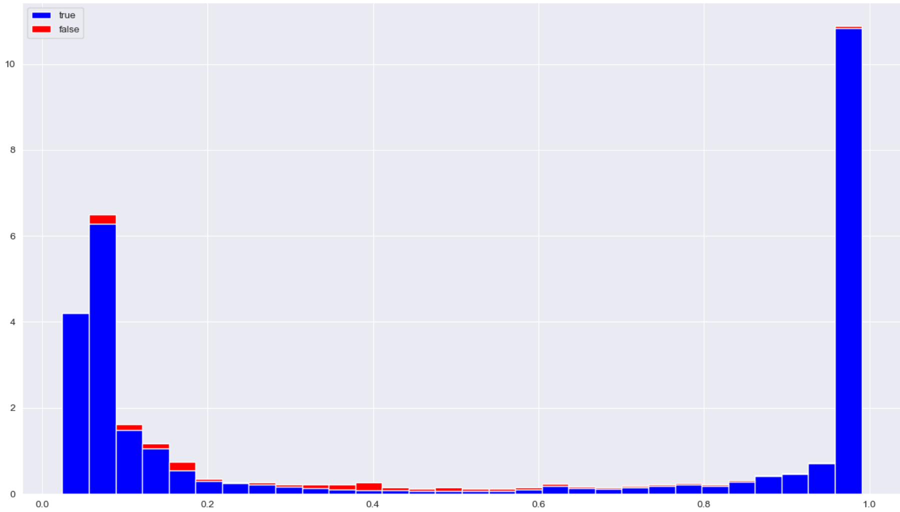
C.2 Fine-tuning the decision threshold?

Figure 18: Predicted probability distribution - all models



Source: Authors' calculations.

Figure 19: Predicted probability distribution - Random Forest

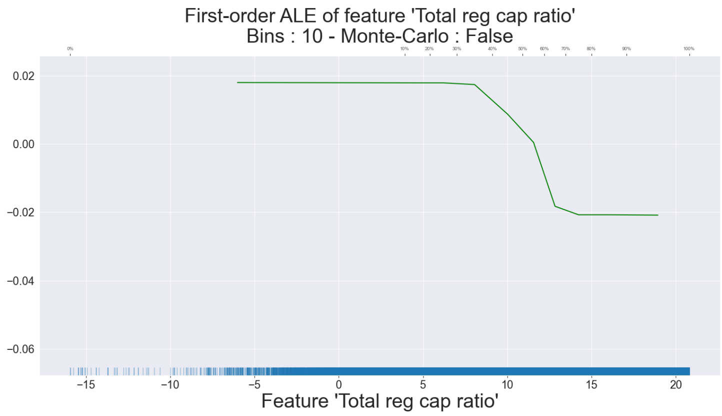


Source: Authors' calculations. Interpretation example: around 7% of the observations have a predicted probability around 0.05. Among those, more than 6% are well predicted (blue part) while 0.5% are wrongly predicted (red part).

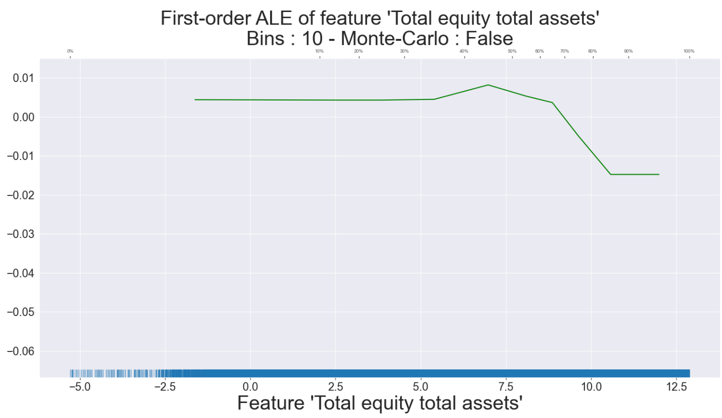
C.3 ALEs

Figure 20: ALEs - RF

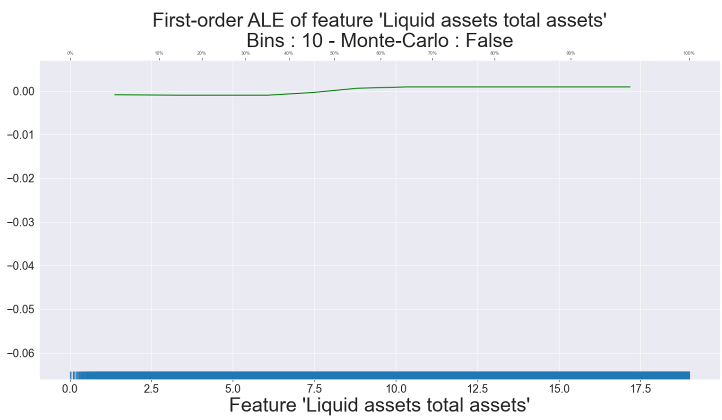
(a) TRCR



(b) TE/TA



(c) LA/TA



Source: Authors' calculations.

C.4 Without over-sampling

Table 5: Models' performance -

	Logit		RF		KNN		GBC	
	Train	Test	Train	Test	Train	Test	Train	Test
Recall	39	0	61	49	100	1	75	51
Macro recall (AUROC)	69	50	80	74	51	51	87	76

	HGBC		Linear SVC		MLP	
	Train	Test	Train	Test	Train	Test
Recall	55	13	31	44	44	46
Macro recall (AUROC)	77	57	65	71	72	73

Source: Authors' calculations.

C.5 Prevision performance - confusion matrices

Table 6: Year 2015 – Default 2016

		Prediction 0	**Prediction 1**
True class	**0**	4157	36
	1	0	5

Table 7: Year 2016 – Default 2017

		Prediction 0	**Prediction 1**
True class	**0**	4158	27
	1	1	5

Table 8: Year 2017 – Default 2018

		Prediction	
		0	**1**
True class	**0**	4174	22
	1	0	0

Table 9: Year 2018 – Default 2019

		Prediction	
		0	**1**
True class	**0**	4177	17
	1	0	3