# Return on Investment in Artificial Intelligence : The Case of Bank Capital Requirement

Henri Fraisse and Matthias Laporte, Banque de France

Research Seminar of the SCOR Foundation for Science

Paris, le 26 janvier 2023

# Introduction

- Boom in AI and their potential application in the financial services industry
- 2004 Basel II accord implemented in 2008 in France :
  - Banks can use Internal models upon regulatory approval
  - Internal models : quantitative tools to compute credit risk parameters.
  - These credit risk parameters are used as input for computing capital requirement
- French banks : Internal models rely on “traditional” modelling approach (e.g. logistic regression)
- This paper : what if they were willing to use AI models ?

# Related literature

- Capital requirement computation and effects :
  - Recent literature on the link between capital requirement and corporate outcomes or credit distribution using granular data (Behn et al. 2016, Fraisse et al. 2020)
  - Capital requirement and manipulation (Behn et al. 2016, Plosser and Santos 2018)
- AI and the financial industry
  - Forecasting credit default in the retail sector using AI techniques (Khandani et al. 2010, Albanesi et al., 2019)
  - Forecasting credit default in the corporate sector using AI techniques (Barboza et al. 2017, Moscatelli et al. 2019)

# Policy context

- EBA repair program
  - Guidelines on the definition of the credit risk parameters
  - Guidelines on the estimation of internal models
- Enforcement : ECB on site campaign : Targeted Review of Internal Model (“TRIM”)
  - High default portfolio (2016-2018) : retail, SME
  - Low default portfolio (2018-2019) : financial institution, large corporate
  - Production of TRIM guides
- AI and regtechs :
  - EC Artificial Intelligence Act
  - ECB & BIS networks
  - EBA consultation paper

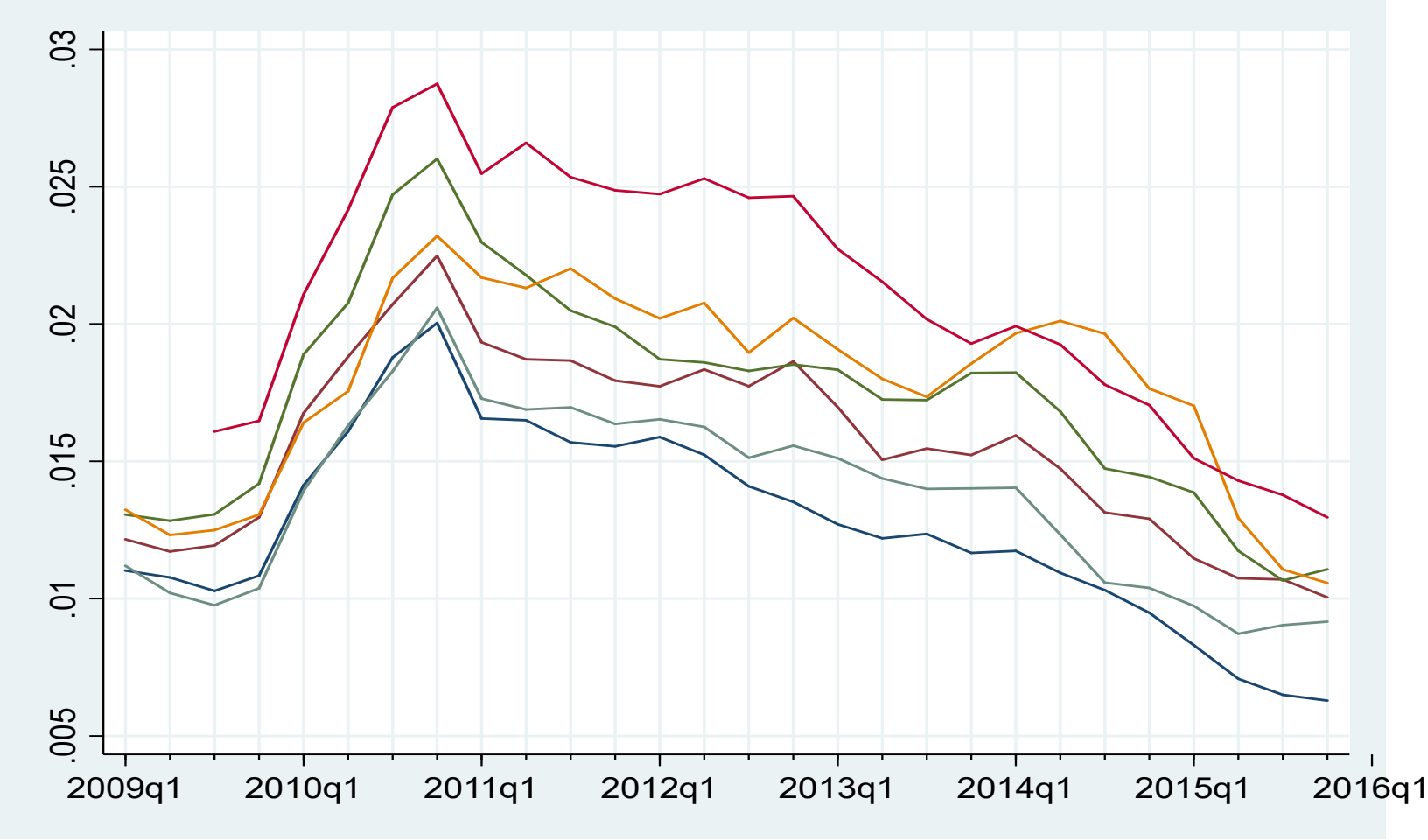
# Contribution

- Using a granular data set, we construct pseudo internal models for each bank and each popular AI
- Using as benchmark the current methodology used by banks for their internal models we proceed to a comparative analysis of the AI models in their ability to :
  - get the regulatory approval
  - the reduction in capital charge they might lead to

# Data

- We use the data coming from the French credit national register :
  - loan-level information from all banks on individual borrowers with total bank debt higher than € 25 000
  - we know the type of credit, the location of the firm, its industry, its size, its Banque de France rating
  - 2009-2016 period : calibration 2009-2015 / backtesting 2016
- We complement these data with firm-level accounting data (“FIBEN’)
- Six largest banking groups
- Measure of default :
  - Banque de France has been recognized as an external credit assessment institution
  - Defaut : legal proceedings (restructuring or liquidation) or payments incidents reported by one or more credit institutions within the next year

# Data


.005
.01
.015
.02
.025
.03
2009q1
2010q1
2011q1
2012q1
2013q1
2014q1
2015q1
2016q1

# From the set up of internal models to…

1ST block : a continuous **risk score** expressed as a function of risk drivers (usually a logistic regression)

**The distinction between AI models and the benckmark model operates only through the first block of the internal model.**

2nd block : A discrete **risk scale ("rating system")** expressed as a function of the continuous risk score : firms are grouped into *R* rating

3rd block : A **probability of default** $PD_R$ is associated to each rating

# …the computation of capital requirement

- The Basel formula allows to compute the capital requirement corresponding to the $r(j)$ facility belonging to the rating class *r* as a function of the probability of default of this rating class:

$$CR_{r,r(j)} = f\left(PD_r, Exposure_{r(j)}\right)$$

- The total capital requirement is then obtained by summing $CR_{r,r(j)}$ over the *R* rating grades. that we normalize by the total of exposures :

$$CR = \sum_{r=1}^{R} \sum_{j=1}^{J_r} CR_{r,r(j)}$$

# …the computation of capital requirement

- The Loss-Given-Default risk parameter is set at 45% in line with the foundation approach
- Conforming with article 501 of the CRR we apply the supporting factor to exposures granted to SME (e.g. a 25% discount of the capital requirement).

# Risk Score

- Benchmark model : mixture of a quantitative approach and a qualitative approach
  - Qualitative : a pre-selection of key risk drivers following discussions with the business lines : here the Banque de France staff in charge of rating corporate
  - Quantitative : logistic regression (AUC)

- AI models : starting from the pre-selection of risk drivers above
  - random forests
  - gradient boosting
  - ridge regressions
  - neural networks (Multilayers Perceptrion with one, two or three hidden layers : MLP1, MLP2 or MLP3)

# Risk Score : the traditional approach –qualitative step

**TABLE 1- LIST OF VARIABLES USED FOR PREDICTING CORPORATE DEFAULTS**

| Variables | Expert Judgement |
|---|---|
| Customer Accounts and Discounts / Turnover | Activity |
| Trade Payable / External Purchases and Expenses | Activity |
| Turnover / Value Added | Financial Autonomy |
| Financial Debt/Turnover | Financial Autonomy |
| Finance Costs/Gross Operating Surplus | Financial Autonomy |
| Interest and Finance Charges / Gross Operating Surplus | Financial Autonomy |
| Personnel Expenses / Turnover | Financial Structure |
| Own Funds / Social Capital | Financial Structure |
| Own Funds/Total Assets | Financial Structure |
| Net Own Funds / Total Asset | Financial Structure |
| Provision and Depreciation / Tangible Capital Asset | Financial Structure |
| Short term assets / long term assets | Liquidity |
| Cash flows / financial debt | Liquidity |
| Net Cash Flow / Turnover | Liquidity |
| Personnel Expenses / Full Time Employees | Productivity |
| Value Added / Tangible Capital Asset | Productivity |
| Personnel Expenses / Value Added | Profit Sharing |
| Finance Costs / Value Added | Profit Sharing |
| Financing Capacity / Value Added | Profit Sharing |
| Gross Operating Surplus/Turnover | Profitability |
| Gross Operating Surplus/Output | Profitability |
| Operating Income before taxes / Turnover | Profitability |
| Return on Asset | Profitability |
| Gross Operating Surplus / Tangible Capital Asset | Rentability |
| Default | Risk |
| Industries ( 8 buckets) | Structural Characteristics |
| Turnover (13 buckets) | Structural Characteristics |
| Legal (4 buckets) | Structural Characteristics |
| Age of firm | Structural Characteristics |

*Notes: list of variables that have been tested in the econometric analysis. Those variables have been pre-selected by credit experts from the Banque de France. They are used in their qualitative analysis. They are produced when a bank put a request in the national credit register for a given counterparpy.*

# Risk Score : the traditional approach –quantitative step

- We discard indicators with more than 20% of missing values
- When two indicators are highly correlated (above 0.7), we keep the ones with fewer missing values
- We discretize each indicator into 5 quintiles and a class of missing value
- For each discretized indicator, we run a logistic regression including industry, judicial status and size fixed effects on the default indicator. We select the indicator for which the highest AUC (standing for Area Under (ROC) Curve) is reached.
- We then add indicators sequentially in the logistic regression while the AUC is increasing. We stop when each additional variable tested lead to the same AUC within a 0.2% range.

# Risk Score : AI approach

- Qualitative step : similar to the traditional approach
- Quantitative step :
  - Imputation for missing values
  - No discretization of continuous variables
  - Each of the classifiers are fitted using standard techniques including cross-validation in order to tune hyper-parameters to prevent overfitting.
  - Use the *sklearn* open source package for Python
  - 6-fold cross-validation based on a one year (4 quarters) vs. all chronological split

# Risk Score

- We come up with a risk score :
  - for each estimating technique (logistic regression, gradient boosting, neural networks, random trees, ridge regressions)
  - for each bank

# Internal rating scale : supervisory expectations

- The 2019 ECB TRIM guide used for on-site missions during the ECB 2019 campaign of on site missions reviewing internal models
- Main expectations :
  - Ability to accurately predict defaults within a risk grade
  - Risk differentiation across risk grades / risk homogeneity within a risk grade
- Supervisory expectations translated into quantitative tests :
  - Change in AUC over the recent period
  - Use of Z-test testing the difference in default rate between risk grades
  - Stability : no risk inversion. e.g. the default rate observed for a better grade should not become higher than the default rate of the adjacent worse rating grade

# Algorithm to set up a rating scale given supervisory expectations

For each quarter, each risk score, each bank :

- we sort the firms by ascending risk score
- Search for an optimal threshold score splitting the scale in two segments with minimal intra-class variance on each side
- Z-test at a 10% p-value level for comparing the default rate on each segment
- If significant difference : split again each of those two resulting classes independently
- the recursion stops as none of the sub-classes can be split anymore

# Algorithm to set up a rating scale given supervisory expectations

- For a given bank and a given model, we have as many rating scale as quarters of observation
- For each type of model and each bank we select the rating scale leading to :
  - Maximizing the share of Z-tests successfully passed between adjacent grades over the training period
  - Stop when there are only 7 grades left (regulatory constraint)

# Calibration of risk parameters

- For each grade of the final rating scale, we compute a long run average of the default rates

$$\overline{PD_g} = \sum_{t=1}^{T} \frac{1}{N_{g,t}} \cdot \sum_{i=1}^{N_{g,t}} D_{i,t}$$

- to which we apply a margin of conservatism : the 95% percentile of the upper bounds of the 95% confidence intervals of the empirical default rate (normality assumption) over all the training quarters

# Indicators for comparative analysis

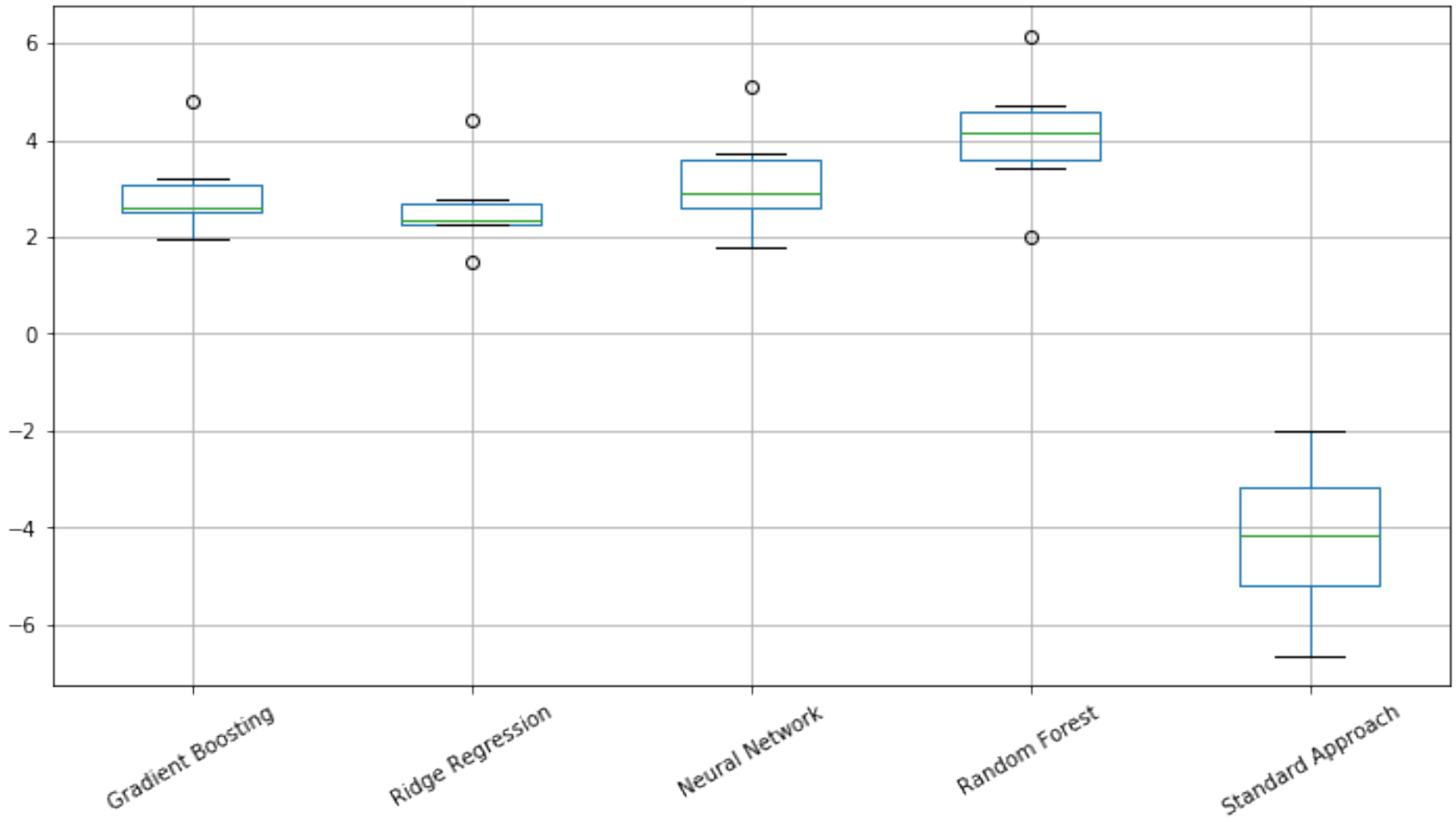
**First component : likelihood to get the regulatory approval**

- **Backtesting** : Performance in predicting corporate default using ROC analysis and AUC indicators in-sample and out-of-sample
- **Risk differentiation :** the share of Z-tests successfully passed between adjacent grades averaged over the training period
- **Stability :** the number of times that two adjacent grades change in the ranking over the period. We standardize this indicator by dividing it by the total number of times a change in order of adjacent grades can be potentially observed.

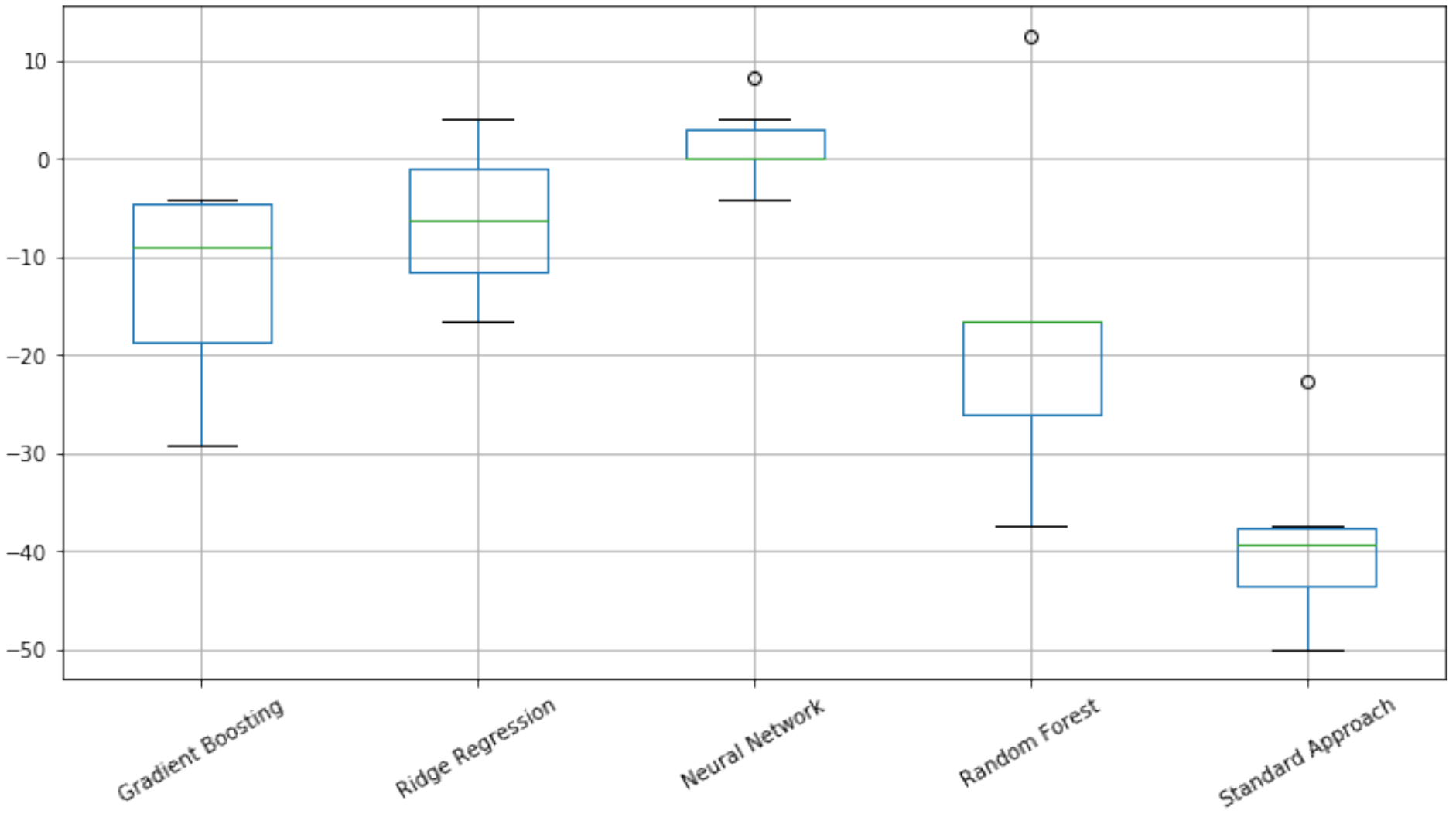
**Second component : banks incentive**

- Reduction in capital requirement : RWA density

# Results : Predictive Accuracy



# Results : Risk Differentiation

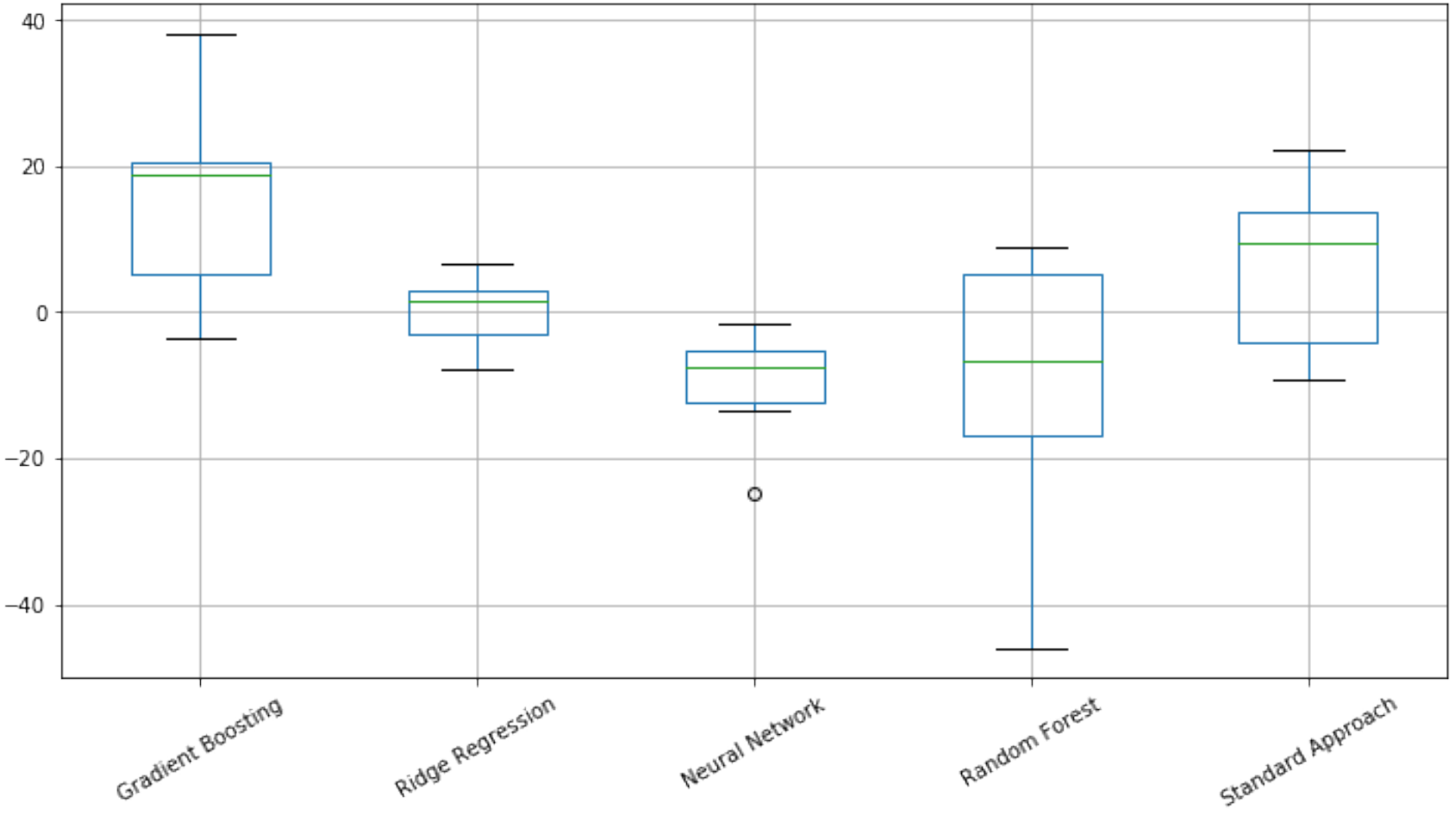


# Results : Density RWA

TABLE 8- RWA DENSITY

| Models | Criteria | Sample | Bank 1 | Bank 2 | Bank 3 | Bank 4 | Bank 5 | Bank 6 | Average |
|---|---|---|---|---|---|---|---|---|---|
| GB | QQ | Train | 38 | 1 | 20 | -4 | 21 | 17 | 16 |
| | | Test | 41 | 2 | 24 | -2 | 25 | 20 | 18 |
| | MM | Train | 44 | 4 | 23 | -3 | 40 | 21 | 22 |
| | | Test | 47 | 5 | 26 | -1 | 46 | 23 | 24 |
| LogReg | QQ | Train | 3 | 1 | 7 | -8 | 3 | -4 | 0 |
| | | Test | 3 | 1 | 8 | -9 | 4 | -5 | 0 |
| | MM | Train | 5 | 1 | 7 | -9 | 12 | -9 | 1 |
| | | Test | 5 | 2 | 9 | -11 | 14 | -10 | 2 |
| Logit | QQ | Train | | | | | | | |
| | | Test | | | | | | | |
| | MM | Train | | | | | | | |
| | | Test | | | | | | | |
| MLP1 | QQ | Train | 15 | 0 | 4 | -9 | 4 | -3 | 2 |
| | | Test | 16 | 0 | 4 | -10 | 4 | -4 | 2 |
| | MM | Train | 19 | 1 | 3 | -9 | 11 | -6 | 3 |
| | | Test | 20 | 1 | 3 | -11 | 12 | -6 | 3 |
| MLP2 | QQ | Train | 12 | -3 | -5 | -14 | 6 | -9 | -2 |
| | | Test | 14 | -3 | -6 | -14 | 8 | -10 | -2 |
| | MM | Train | 18 | -3 | -2 | -13 | 9 | -12 | 0 |
| | | Test | 19 | -3 | -3 | -13 | 12 | -13 | 0 |
| MLP3 | QQ | Train | -2 | -25 | 14 | 1 | -6 | 9 | -1 |
| | | Test | -2 | -27 | 17 | 4 | -5 | 12 | 0 |
| | MM | Train | 0 | -26 | 14 | 2 | 5 | 9 | 1 |
| | | Test | 0 | -28 | 17 | 5 | 7 | 12 | 2 |
| RF | QQ | Train | 9 | 0 | 7 | -18 | -46 | -14 | -10 |
| | | Test | 12 | 3 | 13 | -13 | -34 | -10 | -5 |
| | MM | Train | 16 | 6 | 10 | -15 | -27 | -12 | -4 |
| | | Test | 19 | 8 | 16 | -10 | -14 | -8 | 2 |
| Standard | QQ | Train | 14 | -8 | 22 | 11 | -9 | 8 | 6 |
| | | Test | 15 | -5 | 27 | 16 | -4 | 9 | 10 |
| | MM | Train | 32 | 3 | 37 | 25 | 24 | 20 | 23 |
| | | Test | 32 | 7 | 41 | 29 | 30 | 22 | 27 |

# Results : RWA impact



# Robustness checks

- Sample size
- Data Processing

# Conclusion

- The RF prone to an overfitting of the data in the training sample
- Except RF, no strong difference from a model to another in term of predictive accuracy
- Neural networks and the traditional model lead to the more robust rating system
- Neural networks lead to the strongest decrease in capital requirement in some cases.
- Bottom line : The traditionnal approach shows good performances but some incentives for banks to adopt Neural networks
- Byproduct : modeling technique might be per se a source of RWA variability

# Conclusion

- Other criterias for selecting internal models : P&L rather than RWA
- IA on other credit risk parameters : LGD and CCF
- Regulatory obstacles : switching costs, governance, transparency,  shortage of skills

# Basel II risk-weight formula

**Empirical risk-weight formula:**

$$RW(\text{LGD,PD,M},\rho) = 1.06{\cdot}12.5{\cdot}\text{LGD}{\cdot}\left[\Phi\left(\frac{\Phi^{-1}(\text{PD}) + \sqrt{\rho}\Phi^{-1}(0.999)}{\sqrt{1-\rho}}\right) - \text{PD}\right]{\cdot}f(\text{PD,M})$$

**Basel II risk-weight formula:**

$$RW(\text{LGD,PD,M}) = 1.06{\cdot}12.5{\cdot}\text{LGD}\left[\Phi\left(\frac{\Phi^{-1}(\text{PD}) + \sqrt{\rho(\text{PD,S})}\Phi^{-1}(0.999)}{\sqrt{1-\rho(\text{PD,S})}}\right) - \text{PD}\right] f(\text{PD,M})$$

where

$$\rho(\text{PD,S}) = \frac{1-e^{-50\text{PD}}}{1-e^{-50}}{\cdot}0.12 + \left(1 - \frac{1-e^{-50\text{PD}}}{1-e^{-50}}\right){\cdot}0.24 - 0.04\left(1 - \frac{\min\{50, \max\{S, 5\}\} - 5}{45}\right)$$

Other retail: turnover $<$ 2.5 m €; S:= turnover; M:= maturity

# Results

**TABLE 7- PREDICTIVE ACCURACY : A SUMMARY**

| Model | AUC average | F-score average |
|---|---|---|
| GB | 87 | 17 |
| Logreg | 85 | 13 |
| Logit | 83 | 11 |
| MLP1 | 85 | 13 |
| MLP2 | 86 | 15 |
| MLP3 | 87 | 15 |
| RF | 92 | 41 |
| Standard | 77 | 7 |

*Note: This table shows the AUC and the F-score averaged across banks and samples (e.g. the training sample and the testing sample).*

# Results

**TABLE 4- ROBUSTNESS OF THE RATING SYSTEM : A SUMMARY**

| Model | Risk differentiation Average | Default rate inversion Average |
|---|---|---|
| GB | 84 | 9 |
| Logreg | 84 | 5 |
| Logit | 87 | 3 |
| MLP1 | 83 | 6 |
| MLP2 | 88 | 6 |
| MLP3 | 89 | 3 |
| RF | 81 | 7 |
| Standard | 45 | 13 |

*Note: This table shows the risk differentiation indicator and the rating inversion indicator averaged across banks and samples (e.g. the training sample and the testing sample).*

# Robustness check (1) : sample size

- Focus on the largest bank
- Consider four portfolios consisting in 20, 40, 60 and 80 % of the exposures randomly drawn +the entire portfolio (100%)
- Rerun the analysis considering these five portfolio as five different banks
- Predictive accuracy is not strongly dependent on the sample size. Whathever the model we consider, the AUC and the F-score do not change significantly when the sample size increases.
- As for the rating system, once again MLP offers the best balance between a low inversion rate and a strong risk differentiation across all the sample size

# Robustness check (2) : data processing stage

- We compare the AI logreg model to the benchmark model.
- those two models are very closed in term of predictive accuracy and robustness of rating system.
- discretization leads to a slight outperformance of the benchmark model on the testing sample.