# Early Warning System for Non-Performing Clients

Arnaud Germain[1] and Frédéric Vrins[1]

[1] *Louvain Institute of Data Analysis and Modeling in economics and statistics (LIDAM/LFIN), UCLouvain*

**Abstract**

In its "Guidance to banks on non-performing loans", ECB requires banks to implement an Early Warning System (EWS) to identify potential non-performing clients at a very early stage. Relying on a unique dataset provided by a systemic European bank including 5.5 million observations of anonymized data from 2018 to 2022, we aim to predict the corporate clients who will become non-performing in a given warning horizon. We propose two solutions to address time and client heterogeneity issues. Regarding the latter, we divide our dataset into several clusters using k-means, fit a prediction model on each cluster, and combine those models together. This boosts the out-of-sample performance compared to a case where we fit a single prediction model on the whole dataset and a case where we rely on domain knowledge to determine the clusters. Second, to address time heterogeneity, we forecast the unconditional probability to be positive using macroeconomic variables and then rescale the output of the prediction model using Bayes' theorem. This enhances the out-of-sample performance compared to a case where the macroeconomic variables are directly included as predictors of the prediction model. Both approaches are complementary in the sense that the best predictive performance is achieved by combining them together. Our findings help to increase the performance and the robustness of EWS but can also be useful in a wide range of pattern recognition problems.

## 1 Introduction

Credit risk is, by far, the largest risk ran by banks. It is therefore compulsory for them to properly monitor the latter. Although this is partly achieved via financial provisions and regulatory capital requirements, it is the primary interest of banks to limit defaults and to keep their loans performing. This becomes increasingly important in an era of changing macro factors such as climate change, inflation and geopolitical instability, which are known to impact credit risk of firms. In its "Guidance to banks on non-performing loans" (see https://www.bankingsupervision.europa.eu/ecb/pub/pdf/guidance_on_npl.en.pdf), European Central Bank (ECB) requires banks to implement an Early Warning System (EWS) "to identify and manage potential non-performing clients (NPC) at a very early stage, in order to monitor performing loans and prevent the deterioration of credit quality". ECB gives a relative freedom to banks in their EWS development. In this paper, we focus on corporate clients.

The 2008 financial crisis has highlighted a lack of accuracy from financial institutions in monitoring credit risk in their portfolios. EWS enable banks to better anticipate potential distresses. The main purpose of this model is to take appropriate actions to mitigate risk when a signal is raised,

as illustrated in Figure 1. An effective EWS can lower capital requirements and credit losses and therefore increase the return on equity of a bank's lending book. It has been increasingly observed that the clients with high risk behavioral nature tends to improve towards healthy direction when there is a contact made by the bank to indicate the possible risks the client may face within the near future, given the feasible circumstances the client possesses.

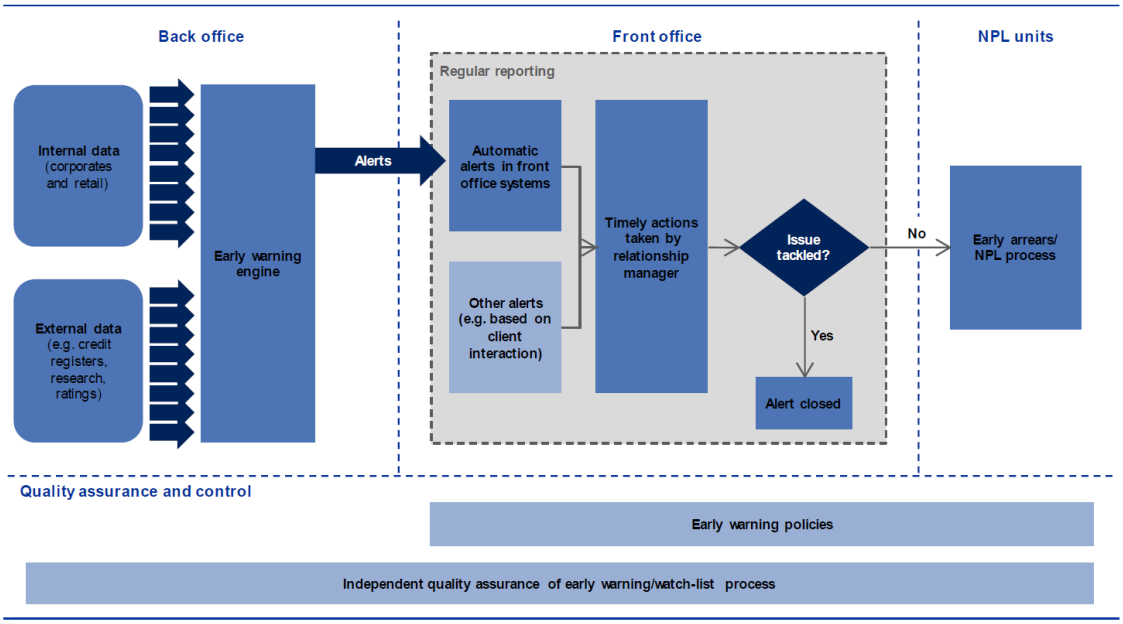


Figure 1: ECB example of an Early Warning System. Source: taken from "Guidance to banks on non-performing loans" (see https://www.bankingsupervision.europa.eu/ecb/pub/pdf/guidance_on_npl.en.pdf).

Following ECB's guidelines, EWS should be run on a monthly basis at least, and include both internal and external data. This could include macro variables, balance sheet information, transactional data (private data stored by the bank about a customer) or any other indicator that could be relevant. What should be the target variable is not precised by the ECB. In line with industry practice, we choose as target variable the corporate clients with products that will become non-performing (adopting the 90 days past-due, or DPD, definition) in a given warning horizon (e.g., within the next 6 months). A similar "warning horizon" is adopted in EWS designed in the context of financial crises detection.

Relying on a panel dataset to fit such a model might lead to (i) time heterogeneity issue and (ii) client heterogeneity issue. The dataset might displays label shift. More specifically, the monthly proportion of positive observations (also referred as positive rate) is likely to change across time. A model fitted on a training set might fail to capture those changes in an out-of-time testing set. We refer to this problem as time heterogeneity issue. Moreover, if one fits a model on the whole dataset, this might miss some heterogeneity in the clients' default dynamics. It might be claimed that clients' default dynamics are different in function of their sector or their size and that a global model might fail to capture that.

In collaboration with a major European bank that provides a unique database with a large sample of anonymized internal data, we build an EWS and propose two solutions to tackle time

and client heterogeneity issues. Firstly, to address client heterogeneity issue, we divide our dataset into several clusters using k-means, fit a prediction model on each cluster, and combine those models together. This boosts the out-of-sample performance compared to a case where we fit a single prediction model on the whole dataset and a case where we rely on domain knowledge to determine the clusters. Secondly, to address time heterogeneity issue, we forecast the unconditional probability to be positive using macroeconomic variables and then rescale the output of the prediction model using Bayes' theorem. This boosts the out-of-sample performance compared to a case where the macroeconomic variables are directly included as predictors of the prediction model. Both approaches are complementary in the sense that the best predictive performance is achieved by combining them together.

The paper is organized as follows. Section 2 reviews the literature about two related problems: default prediction and EWS for financial crisis. The model is presented in Section 3. It formalizes the target we want to predict and described the two different approaches to address respectively time and client heterogeneity issue. Section 4 presents the unique database before diving into the empirical analysis.

## 2 Literature review

Despite the practical relevance of this problem and the mandatory nature for banks to manage EWS, the literature dealing with the early detection of non-performing clients remains surprisingly scarce. This can be explained from the fact that this type of studies requires accessing to highly sensible private data. Moreover, publishing related results is very difficult due to obvious confidential and competition reasons. Nevertheless, we can identify two streams of related literature that can be leveraged in this context: default prediction/credit scoring and EWS for financial crisis. Default prediction literature predicts default either of loans or firms. EWS for financial crisis tries to predict at a very early stage when a country might be in financial crisis. In this section, we give an overview of those two literature and elaborate the differences and similarities with the problem of building an EWS for NPC.

There is a vast amount of research on EWS to detect financial crisis. Usually, the goal is to predict if a financial crisis will happen in a given warning horizon. Two classes of models coexist in this literature. First class is signaling model (also referred as KLR) proposed by Kaminsky et al. (1998). The idea is to transform each possible indicator into a binary signal using a critical threshold. The other class is limited dependent regression models (also referred as BP) proposed by Berg and Pattillo (1999). This can take the form of widely used models such as logit or probit. Kaminsky et al. (1998) claimed that the non-linear nature of BP models implies that it is difficult to compute the marginal contribution of each indicator. While this was maybe true at the time, interpretability of non-linear models has been extensively studied since. BP model can compute the likelihood of a crises (or a default in the credit risk context) considering all variables simultaneously. It can also extract new information from a variable that is independent from variables that are already included. Furthermore, information is lost in the KLR model by transforming each independent variable into binary variables compared to BP model. Berg and Pattillo (1999) showed that their model exhibits better in and out-of-sample results so that BP models became the standard practice. In this domain, most recent articles conclude in a supremacy of machine learning algortihms such

as SVM (Samitas et al., 2020) and XGBoost (Petropoulos et al., 2022) or of data mining algorithms such as Chi-Square Automatic Interaction Detector Decision Tree (Koyuncugil and Ozgulbas, 2012).

There is also a large literature on default prediction. In the same vein as for the EWS for financial crisis literature, we can divide default models into one-dimensional models (similar to KLR) based on the paper of Beaver (1966) and multi-dimensional models (similar to BP) based on the paper of Altman (1968). Furthermore, in a similar way to EWS for financial crisis literature, most recent articles conclude in a supremacy of machine learning algorithms, for instance Support Vector Machine (Zeng et al., 2020) or Artificial Neural Networks (Chen and Du, 2009).

EWS for NPC is in fact a particular case of default prediction model and has therefore some similarities with the seminal papers of Beaver (1966) and Altman (1968). The goal of both problems is to predict if a firm will be in default in the future using today's data. However, when a commercial bank builds an EWS to monitor credit risk, there are also some substantial differences with the traditional framework of the literature. Firstly, data is substantially different because academicians mainly use public data (e.g. financial statements) whereas the bank can leverage internal data. Secondly, the time horizon considered (e.g. 6 months) is shorter than in the traditional default prediction framework because we use monthly frequency data whereas public data is usually available in yearly frequencies (e.g. financial statements).

Sun and Li (2009) propose a model that take into account experts' experiential knowledge and non-financial information to build an EWS for commercial bank loans. Yang et al. (2001) use Artificial Neural Networks to build an EWS for commercial bank loan risk. Zhou et al. (2007) propose an EWS for credit risk assessment problems of commercial banks based on rare event simulation using a cross-entropy scheme. Tsai (2013) adopts multinomial logit models for an EWS of financial distress. All these papers are using only financial ratios as input data and are therefore working in yearly frequencies so that it does not comply with an EWS in the ECB sense. Kaluder and Klepac (2014) suggest a fuzzy expert system to build an EWS for credit risk management that complies with ECB guidelines. They found an increase in predictive power and robustness compared to purely statistical models. Their model also highlights the importance of interpretability in this domain to facilitate the discussion between stakeholders.

# 3 Model

In this section, we present first how the target is defined and then the methodologies we use to address time and client heterogeneity issues.

## 3.1 Target definition

In this section, we formalize the model describing the problem at hand. We aim to predict whether a default (90 DPD) will happen within a $h$-length window (e.g. 6 months) referred as warning horizon so that we can take actions to mitigate the consequences. Denoting by $d_{i,t}$ the binary variable indicating if client $i$ is in default at month $t$, our EWS target is the binary variable $y_{i,t,h}$ indicating if $d_{i,t}$ gets to one within the warning horizon $h$ :

$$y_{i,t,h} = \mathbb{1}_{\{\max_{s\in\{t+1,\dots,t+h\}} d_{i,s}=1\}} = \begin{cases} 1 \text{ if } d_{i,t+k} = 1 \text{ at any } k = 1, ..., h \\ 0 \text{ otherwise.} \end{cases} \tag{1}$$

Note that we are only interested to predict the default of clients that are not already in default. Hence, observations for which $d_{i,t} = 1$ will be removed from the dataset. Let $\mathbf{x}_{i,t} = (x_{i,1,t}, \ldots, x_{i,J,t})$ be a vector of $J$ features for client $i$ at time $t$. We want to solve the following classification problem:

$$\mathbb{P}[y_{i,t,h} = 1] = f(\mathbf{x}_{i,t}). \tag{2}$$

One could try to build up the most accurate system (in-sample) by having a different dynamic model for each client:

$$\mathbb{P}[y_{i,t,h} = 1] = f_{i,t}(\mathbf{x}_{i,t}). \tag{3}$$

On the one hand, the latter approach suffers from obvious practical issues. Which model should be used for a new client or for a new month? Furthermore, once the client is in default, it will unlikely return to non-default class after (which is different from the dynamics of the crisis since a country usually returns to a calm period after a crisis). On the other hand, even if building a model that is common to all clients and all months as in Equation 2 can solve those practical issues, this might fail to address accurately time and client heterogeneity issues. For the sake of conciseness, we drop the index $h$ since it remains constant throughout the paper.

### 3.2 Client heterogeneity

As explained before, if one fits a model on the whole dataset, this might miss some heterogeneity in the clients' default dynamics. It might be claimed that clients' default dynamics are different in function of their sector or their size and that a global model might fail to capture that. A first possible solution is to rely on some domain knowledge segmentation and fit a model on each segment. For instance, one can regroup customers by sector and fit a model on each sector or one can divide customers in buckets in function of their size and fit a model on each bucket. A second possible solution is to rely on a data-driven segmentation. This has been tested in the financial crisis EWS literature by van den Berg et al. (2008). They suggest a preliminary step forming optimal country clusters and build a model on each cluster. Data-driven segmentation (using Hausman's test) outperforms both global model and domain knowledge segmentation (geographical clusters).

We consider a data-driven segmentation based on k-means algorithm called *clagging* (cluster aggregating). The idea is to fit a prediction model on each k-means cluster of the training set, assign observations from the testing set to those clusters and use the corresponding models to obtain predictions. This method has been proposed by Germain and Vrins (2025). They argue that instead of choosing arbitrarily a number of cluster $k$, one can do this procedure for $k = 1, ..., K$. Hence, divide the dataset in one cluster (the original dataset) and fit one model, then in two clusters and fit two models on it and so on until $\bar{K} = 1 + 2 + ... + K = K(K-1)/2$ models are fitted. Then, they combine the $\bar{K}$ predictions together using a linear combination scheme. They propose the following function that takes into account the distance $\delta_{i,k}$ of a test point $i$ to the centroid of the cluster model $k$ was fitted on and also the number of explanatory variables $j$ in the dataset to compute the weight of each prediction:

$$w_{i,k} = \frac{\frac{1}{\delta_{i,k}}^{j-1}}{\sum_j \frac{1}{\delta_{i,k}}^{j-1}}. \tag{4}$$

Their empirical results obtained using several datasets and several classification algorithms underline that this method can outperform bagging and fitting a global model on the whole training set.

### 3.3 Time heterogeneity

As explained before, the dataset might displays concept drift. More specifically, the monthly proportion of positive observations (also referred as positive rate) is likely to change across time. A model fitted on a training set might fail to capture those changes in an out-of-time testing set. In order to address this time heterogeneity issue, one could include macro variables in the model. The first way to do that is to *directly* include a set of macro variables $\mathbf{z}_t$ as predictors so that Equation 2 becomes:

$$\mathbb{P}[y_{i,t} = 1] = f(\mathbf{x}_{i,t}, \mathbf{z}_t). \tag{5}$$

The value for a macro variable will be the same for every observations related to a specific month. This particular distribution for a feature might lead either to a very low feature importance[1] either to overfitting[2]. One could include macro variables *indirectly* so that:

$$\mathbb{P}[y_{i,t} = 1] = h(f(\mathbf{x}_{i,t}), g(\mathbf{z}_t)) \tag{6}$$

We propose to rescale each month the output of the model using macro variables (e.g. increase all the individual probabilities $\mathbb{P}[y_{i,t} = 1]$ for a specific month $t$ if we anticipate a high default rate). We rely on a standard rescaling function based on Bayes' theorem (see for instance Saerens et al. (2002)). Let $p_{train}(y_{i,t}|\boldsymbol{x})$ be the conditional probability to be positive (e.g. the output of the classification model) on training set and let $p_{train}(y_{i,t})$ and $p_{test}(y_{i,t})$ be the unconditional probability to be positive respectively on the training and testing sets. If one assume that within-class densities do not change from training to testing sets, then:

$$p_{test}(y_{i,t}|\boldsymbol{x}) = \frac{\frac{p_{test}(y_{i,t})}{p_{train}(y_{i,t})} p_{train}(y_{i,t}|\boldsymbol{x})}{\frac{p_{test}(y_{i,t})}{p_{train}(y_{i,t})} p_{train}(y_{i,t}|\boldsymbol{x}) + \frac{1-p_{test}(y_{i,t})}{1-p_{train}(y_{i,t})}(1 - p_{train}(y_{i,t}|\boldsymbol{x}))} \tag{7}$$

so that conditional survival and default probabilities still sum to one after rescaling. The unconditional probability $p_{train}(y_{i,t})$ can be estimated directly by observing the proportion of positive observations in the training set that we note $\bar{y}_{train}$. In our setting, we can estimate the unconditional probability of default on the testing set at each time $t$ using a set of macro variables so that the rescaling is tailored to each month. Let us note it $\bar{y}_t$ and build a model to estimate it:

$$\bar{y}_t = g(\boldsymbol{z}_t). \tag{8}$$

Note that, even if the time series $\bar{y}_t$ is likely to exhibit a lot of autocorrelation, we can not include past values of the target in our model because we need to wait the end of the warning horizon (e.g.

[1] For instance in a tree-based model, the lack of variance induced by the particular distribution of the macro variables might lead in high impurity when splitting a node.

[2] Again, assume a tree-based model, the lack of variance induced by the particular distribution of the macro variables might lead to bad (random) choices for the split points.

6 months) to observe the value of $\bar{y}_t$. Equation 7 becomes:

$$\mathbb{P}[y_{i,t} = 1] = \frac{\frac{g(\boldsymbol{z}_t)}{\bar{y}_{train}} f(\mathbf{x}_{i,t})}{\frac{g(\boldsymbol{z}_t)}{\bar{y}_{train}} f(\mathbf{x}_{i,t}) + \frac{1-g(\boldsymbol{z}_t)}{1-\bar{y}_{train}}(1 - f(\mathbf{x}_{i,t}))} \,. \tag{9}$$

# 4 Empirical analysis

## 4.1 Data

We consider a unique panel dataset from a systemic European bank containing 5.5 million observations from 2018 to 2022 in monthly frequencies. The scope of the model is every corporate clients (excluding single person companies). The bank has taken all standard measures to ensure that the data remains anonymous. The dataset contains clients from different countries, different sectors and different sizes and the total number of clients is 182,491. A vast majority of clients are Belgian. The dataset is obviously imbalanced with 3.27% of positive observations. Note that clients already in default at time $t$ have been excluded from the dataset. Observations from the first 54 months of the sample are assigned to a training set and observations from the last 6 months (from August 2022 to December 2022) are assigned to a testing set. If a client has a target equals to one at any time in the testing set, every observations related to this client are excluded from the training set. The total number of observations for the training and testing sets are respectively 4,869,060 and 707,994, which represents respectively 87.3% and 12.7% of all the observations. The dataset contains 692 features. For competitive reasons, we do not disclose the exact list of features. The dataset contains both external (e.g. demographics, financial information, balance sheet, ...) and internal (feature engineering using clients' accounts) data.

## 4.2 Results

In this section, we analyze the impact of the two heterogeneity mitigation modules discussed earlier using this unique dataset. We use the Area Under the ROC Curve (AUC) as a performance metric. Specifically, every time we report the AUC, it is computed on the testing set (i.e., "out-of-time") using a model fitted on the training set. We do not recalibrate the model every month. As a preprocessing step, we consider feature selection. Dimension reduction measures allow to face the curse of dimensionality and to avoid overfitting. We keep the top 50 features using feature importance[3]. As a robustness check, we compute the AUC of an XGBoost model fitted with all features and the AUC of an XGBoost model fitted with the top 50 features, giving respectively 86.26% and 86.19%. We only use those 50 features for the remaining of the section.

### 4.2.1 Client heterogeneity

We compare the performance of different classification algorithms in different configurations: data-driven segmentation using *clagging*, domain knowledge segmentation and no segmentation (i.e a model fitted on the whole training set). For the domain knowledge segmentation, we build 6 buckets of clients related to their sizes and we fit one model per bucket. Testing set observations are assigned to a bucket and only the corresponding bucket model is used for prediction. For

[3]Feature importance was computed using an XGBoost model.

clagging, we choose $K = 10$. We consider 3 classification algorithms: XGBoost, Random Forest and Ridge Logistic Regression. Results are displayed in Table 1.

| | Global model | Domain knowledge | Clagging |
|---|---|---|---|
| XGBoost | 86.19% | 86.25% | 86.48% |
| Random Forest | 85.99% | 86.14% | 86.18% |
| Ridge regression | 79.99% | 81.80% | 82.27% |

Table 1: AUC on the testing set for different algorithms considering 3 configurations: a global model (i.e. no segmentation), domain knowledge segmentation and clagging (data-driven segmentation).

Whatever the configuration, the best performing algorithm is XGBoost. Clagging outperforms the two other configurations independently of the considered algorithm. The increase of performance resulting from segmentation (being either domain knowledge or data-driven) is not negligible. In particular, the performance enhancement achieved by Clagging is prove to be significant according to Delong's test at 95% confidence level. It is worth stressing that XGBoost and Random Forest are state-of-the-art models in terms of predictive performance and are therefore very difficult to beat. Increasing their performance, even moderately so, is already quite a challenging task. Therefore, a statistically significant performance enhancement should be regarded as a notable achievement.

#### 4.2.2 Time heterogeneity

Having noticed that XGBoost performs best in every configuration, we restrict ourselves to discuss the results of this algorithm only. Assume a global model is fitted with XGBoost on the training set and then applied on the testing set, giving respectively in-sample and out-of-sample predictions. Time heterogeneity issue can be observed in Figure 2 that displays the monthly proportion of observed positive observations (black), of in-sample predicted positive observations (blue) and of out-of-sample predicted positive observations (red). The model fails to capture that the unconditional probability to be positive has changed between training and testing sets. The red line is rather flat where the corresponding black line is decreasing.

We use the 54 observations of $\bar{y}_t$ on the training set to estimate a forecasting model (see Equation 8) using the following set of macro variables: inflation, unemployment rate, short-term (3M) rate, 10Y rate, VIX, BEL20 and credit spread[4].

In Figure 3, 3M and 10Y rates are the only significant variables using an OLS linear regression. We therefore only keep the latter and we fit a linear regression with LASSO penalty to build our forecasting model. We plot in Figure 4 the out-of-time rolling-window forecasts for the 6 months of the testing set. The forecasting model is able to predict the decreasing trend of the monthly proportion of positive observations in the testing set based on the macro variables. Using those forecasts, we rescale the individual output probabilities given both by a XGBoost global model and by clagging using XGBoost using Equation 9. Figure 5 shows graphically the impact of the rescaling

[4]Inflation refers to the unadjusted growth of harmonized index of consumer prices in Belgium available on Eurostat. Unemployment rate refers to the unadjusted data of unemployment according to ILO definition in Belgium available on Eurostat. Short-term rate refers to the 3M interbank rate in Belgium available on Federal Reserve Bank of St. Louis. 10Y rate refers to the 10-Year government bond yields in Belgium available on Federal Reserve Bank of St. Louis. VIX refers to the monthly VIX Index converted from daily data available on CBOE. BEL20 refers to the rate of return of the BEL20 Index retrieved from daily prices available on Investing.com. Credit spread refers to the ICE BofA Euro High Yield Index Option-Adjusted Spread available on Federal Reserve Bank of St. Louis.

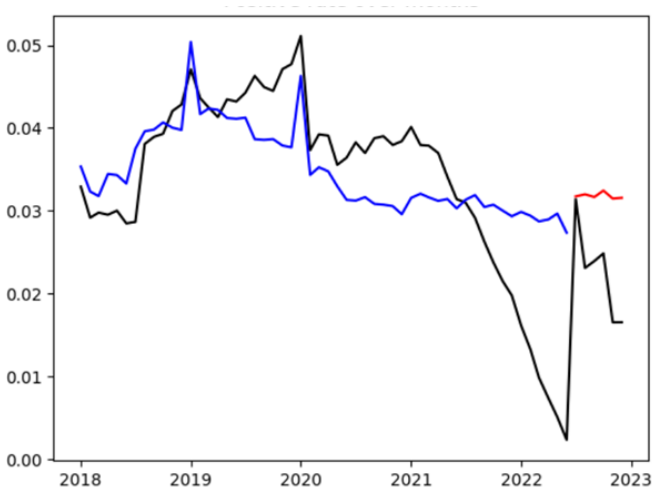


Figure 2: Monthly proportion of observed ($y$) positive observations (black), of in-sample predicted ($\hat{y}_{train}$) positive observations (blue) and of out-of-sample predicted ($\hat{y}_{test}$) positive observations (red).

OLS Regression Results

| | | | |
|---|---|---|---|
| Dep. Variable: | default_rate | R-squared: | 0.766 |
| Model: | OLS | Adj. R-squared: | 0.730 |
| Method: | Least Squares | F-statistic: | 21.04 |
| Date: | Wed, 25 Jun 2025 | Prob (F-statistic): | 2.96e-12 |
| Time: | 09:56:15 | Log-Likelihood: | 200.97 |
| No. Observations: | 53 | AIC: | -385.9 |
| Df Residuals: | 45 | BIC: | -370.2 |
| Df Model: | 7 | | |
| Covariance Type: | nonrobust | | |

| | coef | std err | t | P>\|t\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| const | 0.0706 | 0.014 | 5.088 | 0.000 | 0.043 | 0.099 |
| credit_spread | -0.0010 | 0.002 | -0.577 | 0.567 | -0.004 | 0.002 |
| BEL_20 | 0.0139 | 0.016 | 0.893 | 0.377 | -0.017 | 0.045 |
| unemployment | 0.0038 | 0.002 | 1.850 | 0.071 | -0.000 | 0.008 |
| ir_10Y | -0.0221 | 0.002 | -9.648 | 0.000 | -0.027 | -0.017 |
| inflation | -0.0006 | 0.001 | -0.577 | 0.566 | -0.003 | 0.002 |
| ir_3M | 0.1103 | 0.013 | 8.219 | 0.000 | 0.083 | 0.137 |
| VIX | -0.0001 | 0.000 | -0.635 | 0.529 | -0.000 | 0.000 |

| | | | |
|---|---|---|---|
| Omnibus: | 0.247 | Durbin-Watson: | 0.769 |
| Prob(Omnibus): | 0.884 | Jarque-Bera (JB): | 0.038 |
| Skew: | 0.065 | Prob(JB): | 0.981 |
| Kurtosis: | 3.022 | Cond. No. | 452. |

Figure 3: OLS regression of the unconditional probability to be positive with all macro variables.

by displaying the monthly proportion of observed positive observations (black) on the testing set, of unrescaled predicted positive observations (red) and of rescaled predicted positive observations (blue). The monthly proportion of rescaled predicted positive observations is able to better follow the decreasing trend than the the monthly proportion of unrescaled predicted positive observations.

As displayed in Table 2, the rescaling is able to increase the predictive performance in terms of AUC compared to the unrescaled output given by a XGBoost global model. We also tested to

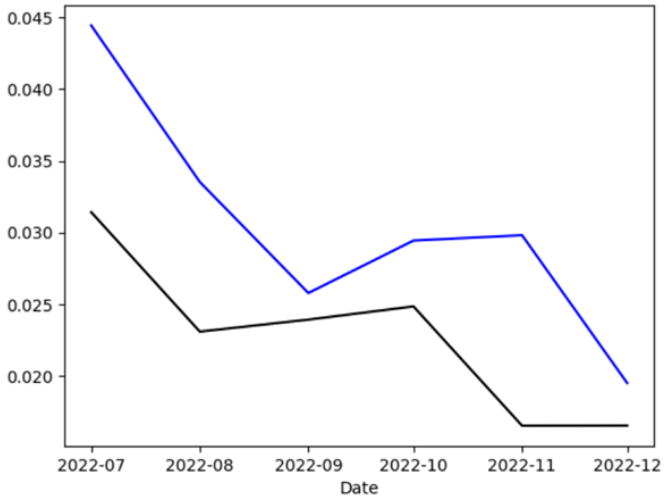


Figure 4: Monthly proportion of observed positive observations (black) and forecasted proportion (rolling-window) using a linear model with LASSO penalty including 3M and 10Y rates (blue) for the 6 months of the testing set.

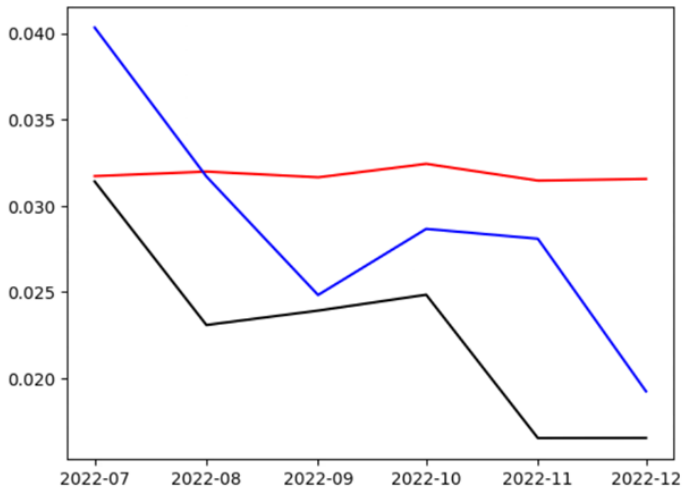


Figure 5: Monthly proportion of observed positive observations (black) on the testing set, of unrescaled predicted positive observations (red) and of rescaled predicted positive observations (blue).

include the set of macro variables directly as predictors of XGBoost. This deteriorates the AUC to 83.45%. Moreover, rescaling is also able to increase the predictive performance in terms of AUC compared to the unrescaled output given by clagging using XGBoost. Note that those increases are all significant using a Delong test with a 95% confidence level. This underlines the complementarity of our two approaches to address both time and client heterogeneity issues.

# 5 Conclusion

Monitoring properly credit risk is of primary importance for banks. An Early Warning System for Non-Performing Clients can mitigate this risk if appropriate actions are taken when a signal is

| | Global model | Clagging |
|---|---|---|
| Unrescaled | 86.19% | 86.48% |
| Rescaled | 86.46% | 86.76% |

Table 2: AUC on the testing set for XGBoost algorithm considering a global model and data-driven segmentation with clagging, both with rescaled and unrescaled outputs.

raised.

A key step in an Early Warning System is the estimation of the probability to default in a given warning horizon. This a challenging problem because fitting a model on the whole panel dataset might lead to time and client heterogeneity issue. In this paper, we propose to address client heterogeneity issue by dividing the dataset into several clusters using k-means, fitting a prediction model on each cluster, and combining those models together using a weighting function based on the distance between a test point and each cluster. This boosts the out-of-sample performance compared to a case where we fit a single prediction model on the whole dataset and a case where we rely on domain knowledge to determine the clusters. To address time heterogeneity issue, we forecast the unconditional probability to be positive using macroeconomic variables and then rescale the output of the prediction model using Bayes' theorem. Both approaches individually boost the out-of-sample performance and this performance can further be increased by combining the two approaches together, underlining their complementary nature. Our empirical analysis is based on a unique database provided by a systemic European bank.

Our contribution is useful for any bank that should build an Early Warning System following ECB guidelines. Moreover, since the approaches are independent from the target and the set of features, they could be used in a wide range of pattern recognition problems.

## Acknowledgment

The work of Arnaud Germain is funded by the ING chair. Frédéric Vrins benefits from the financial support of the Fonds de la Recherche Scientifique F.S.R.-FNRS (grant J.0225.24) as well as of the Belgian Federal Science Policy Office (grant ARC 18-23/089).

## References

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

Beaver, W. H. (1966). Financial ratios as predictors of failure. *Journal of accounting research*, pages 71–111.

Berg, A. and Pattillo, C. (1999). Predicting currency crises:: The indicators approach and an alternative. *Journal of International Money and Finance*, 18(4):561–586.

Chen, W.-S. and Du, Y.-K. (2009). Using neural networks and data mining techniques for the financial distress prediction model. *Expert systems with applications*, 36(2):4075–4086.

Germain, A. and Vrins, F. (2025). Clagging: Cluster aggregating as an efficient alternative to bootstrap aggregating.

Kaluder, I. and Klepac, G. (2014). Credit risk early warning system using fuzzy expert systems. In *Central European Conference on Information and Intelligent Systems*, page 250. Faculty of Organization and Informatics Varazdin.

Kaminsky, G., Lizondo, S., and Reinhart, C. (1998). Leading indicators of currency crises. *International Monetary Fund*, 45.

Koyuncugil, A. S. and Ozgulbas, N. (2012). Financial early warning system model and data mining application for risk detection. *Expert systems with Applications*, 39(6):6238–6253.

Petropoulos, A., Siakoulis, V., and Stavroulakis, E. (2022). Towards an early warning system for sovereign defaults leveraging on machine learning methodologies. *Intelligent Systems in Accounting, Finance and Management*, 29(2):118–129.

Saerens, M., Latinne, P., and Decaestecker, C. (2002). Adjusting the outputs of a classifier to new a priori probabilities: a simple procedure. *Neural computation*, 14(1):21–41.

Samitas, A., Kampouris, E., and Kenourgios, D. (2020). Machine learning as an early warning system to predict financial crisis. *International Review of Financial Analysis*, 71:101507.

Sun, J. and Li, H. (2009). Financial distress early warning based on group decision making. *Computers & Operations Research*, 36(3):885–906.

Tsai, B.-H. (2013). An early warning system of financial distress using multinomial logit models and a bootstrapping approach. *Emerging Markets Finance and Trade*, 49(sup2):43–69.

van den Berg, J., Candelon, B., and Urbain, J.-P. (2008). A cautious note on the use of panel models to predict financial crises. *Economics Letters*, 101(1):80–83.

Yang, B., Li, L. X., Ji, H., and Xu, J. (2001). An early warning system for loan risk assessment using artificial neural networks. *Knowledge-Based Systems*, 14(5-6):303–306.

Zeng, S., Li, Y., Yang, W., and Li, Y. (2020). A financial distress prediction model based on sparse algorithm and support vector machine. *Mathematical Problems in Engineering*, 2020.

Zhou, H., Qiu, Y., and Wu, Y. (2007). An early warning system for loan risk assessment based on rare event simulation. In *Asian Simulation Conference*, pages 85–94. Springer.