RISK MANAGEMENT MAGAZINE

Vol. 19, Issue 2
May – August 2024

EXCERPT

https://www.aifirm.it/rivista/progetto-editoriale/

A Novel Supervised-Unsupervised Approach for Past-Due Prediction

Giampaolo Gabbi, Daniele Tonini, Michele Russo

A Novel Supervised-Unsupervised Approach for Past-Due Prediction

Giampaolo Gabbi (SDA Bocconi School of Management), Daniele Tonini (SDA Bocconi School of Management), Michele Russo (SDA Bocconi School of Management)

Corresponding author: Giampaolo Gabbi (gabbi@sdabocconi.it)

Article submitted to double-blind peer review, received on 9th February 2024 and accepted on 5th April 2024

Abstract

In the current landscape of banking and financial services, a primary concern for industry practitioners revolves around predicting the probability of default (PD) and categorizing raw data into risk classes. This study addresses the challenge of predicting payment past-due for customers of Residential Mortgage-Based Securities (RMBS) and Small and Medium Enterprises (SMEs) within the Italian banking sector, employing an innovative approach that integrates a classification model (Random Forest) with an anomalies detection technique (Isolation Forest). The models are trained on a substantial dataset comprising performing loans from the 2020-2022 period. Notably, this research stands out not only for its novel modeling approach but also for its focus on the arrear status of RMBS and SME customers as the target variable. By concentrating on past-due rather than the broader concept of probability of default, this approach enhances understanding of customers' financial stress levels, enabling proactive monitoring and intervention by decision-makers. The ultimate aim of this experimentation is to develop a robust and effective algorithm applicable in real-world scenarios for predicting the likelihood of past-due among individual customers and companies, thereby supporting management decision-making processes. Empirical results demonstrate that the proposed framework surpasses conventional statistical and machine learning algorithms in credit risk modeling, exhibiting robust performance on new data (validated against 2023 data) and thus proving its operational suitability.

JEL Classification: G21, G24, G32

Aknowledgments

The authors would like to express their deep gratitude to Roberta Fasano, Giovanni Fischetto, Massimiliano Gianfreda, Leonardo Mangia, Palmalisa Marra, Costantino Mele, Francesco Mello, Francesco Sannicola, Simone Fabio Tarantino and Carla Totaro for their invaluable support throughout this research. Their insightful suggestions on methodological approaches and financial choices were instrumental in guiding the direction and quality of this research. The research activity took place in the context of the "Invisible Business" industrial research project, financed by Links Management and Technology S.p.A. and the Regione Puglia through the tender "Aiuti ai programmi integrati promossi da MEDIE IMPRESE ai sensi dell'articolo 26 del Regolamento".

1. Foreword

Credit risk modeling is a cornerstone of financial research and risk management, especially in the aftermath of financial crises. Accurate and comprehensive tools to assess credit risk are essential for mitigating potential losses and ensuring the stability of financial institutions. This section aims to provide a thorough review of the key methodological approaches used in the literature for modeling the probability of default (PD). It includes insights from both empirical applications and academic research, identifies existing literature on credit risk, and explores new empirical methodologies to underscore the novelty of the proposed model.

Traditional models often employ binary classifications to determine credit default, focusing on whether a borrower is over 90 days in past-dues. However, this approach can result in the loss of valuable information by reducing the continuous measure of days past due to a simple binary variable. The financial crisis of 2008-2009 particularly heightened interest in understanding the factors affecting credit access for small and medium-sized enterprises (SMEs), which are heavily dependent on direct lenders and were significantly impacted by reduced credit availability following banking shocks.

Historically, discrete choice methods have been used to model credit default. These models typically define a binary dependent variable based on a standardized definition of default, reducing continuous measures like days past due to binary outcomes. While this method is straightforward, it potentially overlooks valuable information that could enhance model accuracy and risk prediction.

Credit default indicators exhibit persistence over time, suggesting that using lagged days past due could improve default prediction by leveraging temporal information. This analysis focuses on two borrower categories: SMEs and household borrowers, both of which play crucial roles in the economy. SMEs, in particular, are vital for employment, income generation, and fostering innovation and growth.

For residential mortgages, credit risk assessment primarily focuses on the borrower's equity in the property as a key default determinant. Risk management in financial modeling has led to extensive experimentation with various algorithms to achieve optimal classification performance. This review covers traditional statistical models, machine learning techniques, and hybrid approaches, evaluating their effectiveness in predicting default probabilities.

The article follows a structured approach that begins with a comprehensive literature review, examining key methodological approaches in credit risk modeling. This review explores both empirical applications and academic perspectives, providing a foundation for understanding current practices and identifying gaps in existing literature. Subsequently, the article delves into detailed case studies, examining specific datasets and scenarios pertinent to credit risk assessment, particularly focusing on residential mortgage-backed securities (RMBS) and small and medium-sized enterprise (SME) loans. Following this empirical foundation, the article presents a robust methodological framework that integrates supervised and unsupervised learning techniques, aiming to enhance predictive accuracy in default probability modeling. Finally, the article concludes with insightful remarks, discussing the implications of the proposed model and suggesting avenues for future research and application in the field of credit risk management.

2. Literature Review

The objective of this section is to review the main methodological approaches available in the literature to model the probability of default, both in empirical applications and from an academic perspective. Furthermore, we aim to identify the existing literature on credit risk and explore new empirical methodologies. In doing so, we aim to highlight the novelty of the proposed model.

Credit risk modelling is a critical area of research in finance, particularly relevant in light of the financial crises, which have highlighted the need for more accurate and comprehensive risk assessment tools. Traditional models have typically used binary classifications to determine credit default, focusing mainly on whether a borrower is more than 90 days. However, these models can lose valuable information by simplifying days past due into a dichotomous variable. The financial crisis of 2008-2009 increased the interest of economists and regulators in understanding the factors affecting access to credit for small and medium-sized enterprises (SMEs). SMEs, which are highly dependent on direct lenders, are particularly affected by reductions in credit availability following banking shocks (Berger and Udell, 2002; Wehinger, 2014).

Credit default has historically been modelled using discrete choice methods, first proposed by Altman (1968) and later developed by others such as Löffler and Maurer (2011), Bonfim (2009) and Carling *et al*. (2007). These models typically define a binary dependent variable based on the Basel Committee on Banking Supervision's (BCBS, 2006) definition of default, which considers a borrower to be in default if he or she is more than 90 days in past-dues. Although effective, this approach reduces a continuous measure (days past due) to a binary outcome, thereby losing potentially useful information that could improve model accuracy and risk prediction.
Credit default indicators are known to be persistent over time. Once a borrower has defaulted, the likelihood of a quick return to compliance is low. Similarly, once the number of days in default becomes positive, it tends to remain so, showing positive serial correlation. This persistence suggests that the use of the number of lagged days could improve the prediction of future defaults by exploiting this temporal information, an advantage that standard default prediction models typically do not exploit.
This analysis is conducted for two categories of borrowers: SMEs and household borrowers.
SMEs play a crucial role in the economy, generating employment and income and fostering innovation and growth. In the euro area, SMEs account for around 99% of all enterprises, employ around two-thirds of the labour force and contribute around 60% of value added (Gagliardi-Main *et al.*, 2013). The economic importance of SMEs is particularly pronounced in southern European countries such as Italy, Spain and Portugal. During the financial crisis, SMEs experienced a double shock: a significant reduction in demand for goods and services combined with tighter credit conditions, which severely affected their cash flows.
The sovereign debt crisis of 2011 further exacerbated these challenges, particularly for Italian banks (Bofondi, Carpinelli and Sette, 2013). SMEs generally face higher credit risk than large firms due to greater information asymmetries. Banks often have limited access to detailed financial information on SMEs, making it difficult to accurately assess their creditworthiness (Berger and Udell, 1995; Degryse and Van Cayseele, 2000). This information gap leads to higher perceived risk and may result in tighter credit conditions for SMEs (Ivashina, 2009). SMEs generally have less stringent accounting requirements and fewer incentives to invest in detailed disclosure practices (Baas and Schrooten, 2006), contributing to banks' reluctance to lend.
Credit risk assessment for residential mortgages focuses mainly on the borrower's equity in the property as a key factor in the default decision. If the market value of the house exceeds the value of the mortgage, the borrower has a financial incentive to sell the property rather than default. Option-based theories view mortgage default as a put option, where the borrower can transfer the property to the lender to pay off the debt. Borrowers exercise this option when the market value of the house falls significantly below the value of the mortgage, although high transaction costs and reputational damage reduce the likelihood of a 'merciless' default. Equity-related factors influencing default rates include the initial loan-to-value ratio, house price appreciation rates, mortgage seniority, mortgage term and current interest rates. A mortgage interest rate below current market levels discourages default, as a new mortgage would have a higher interest rate.

Risk management has always been a primary concern in financial modeling, prompting extensive experimentation with various algorithms and techniques to achieve optimal classification performance. In this section, we will provide detailed evidence of the different methodologies employed historically and contemporarily in credit risk modeling. This review will cover traditional statistical models, machine learning techniques, and hybrid approaches, evaluating their effectiveness in predicting default probabilities.

At a broad level, the probability of default (PD) problem can be framed as the development of an algorithm or methodology to predict a target variable (Y), typically encoded as a binary variable (0/1), where a value of 1 indicates the occurrence of a default event and 0 otherwise. A diverse range of variables can be utilized to predict the probability of default, encompassing both intrinsic characteristics of the borrower, such as demographics in the context of business-to-consumer (B2C) lending or industry and firm size in business-to-business (B2B) applications, and financial indicators and key performance indicators (KPIs) related to the financial behavior of the subjects under study.

Variables commonly employed as independent predictors in credit risk measurement can be categorized into quantitative variables (based on financial ratios), behavioral variables, and qualitative or "soft" factors (Gabbi, Matthias and Giammarino, 2019). Among the most frequently used models in credit risk management, quantitative variables derived from historical balance sheet data and trends are predominant for both loans and bonds (Gabbi & Sironi, 2005). Due to the historical nature of many of these data, they can often induce procyclicality effects (Gabbi & Vozzella, 2013). Regulatory authorities have acknowledged that the Basel II framework contributed to undesirable effects on system stability during financial crises, resulting in credit crunch phenomena that particularly affected small and medium-sized enterprises (SMEs) whose access to credit may be influenced by regulation (Gabbi & Vozzella, 2020).

There is compelling research highlighting the efficacy of qualitative variables in approximating future business dynamics, management plans, and company perspectives (Brunner *et al*., 2000; Morales *et al*., 2000; Grunert *et al*., 2005). Several studies (Lehmann, 2003; Grunert, Norden, & Weber, 2005; Godbillon-Camus & Godlewski, 2005) have identified the opacity of information

processed by banks as a significant challenge in assessing the credit risk of loans to SMEs. The utilization of forward-looking information enables SMEs to mitigate information asymmetries relative to larger companies and reduces the risk of credit crunch (Grunert & Norden, 2012; Howorth & Moro, 2012). While regulation for internal models does not mandate specific variables, it encourages banks to diversify their input sources to adequately capture the complexity of credit risk (Basel Committee on Banking Supervision).

Several systematic literature review publications on banking probability default methodologies are available, Dastile *et al* (2020) and Alaka *et al* (2016), Brown and Mues (2012), all pointing out in the direction of two main class of techniques developed and applied to default prediction, namely statistical techniques and Machine Learning and Artificial Intelligence based techniques.

With reference to classical statistical technique, most of the relevant publications implementes logistic regression modeling like in Steenackers and Goovaert (1989), Arminger *et al* (1997) and West (2000) or alternatively linear and quadratic discrminant analysis as in Desai *et al* (1996), West (2000) and Baesens *et al* (2003). Regarding the main results, these techniques proved to be quite good at predicting the investigated phenomenon, providing - above all - interpretable results on the variables that most influence the outcome. However, in cases of dataset where the relationship between predictors and the target variable follows non linearities, interaction and complex effects these methods are not well-suited, unless the functional form of the relationships is known or discovered *ex-ante*.

Alternatively, Machine Learning based techniques has been experimented for the same purpose with a very good level of performance. More specifically, tree-based methods and Artifical Neural Networks have been found wide applications in this domain. With respect to tree-based methodologies, Classification Trees has been tested like in Arminger *et al* (1997), Yobas *et al* (2000) and more recently in Feldman and Gross (2005) for mortgage default prediction. In addition, ensemble methods such as Random Forest algorithm (Brieman, 2001) as well as Gradient Boosting Methods (Friedman, 2001 and Friedman, 2002) have been implemented proving to obtain relevant results in this domain of application like in Zhu *et al* (2019) and Ma *et al* (2019). In addition, neural networks architectures have been also widely applied for loan default predictions both as experimental methodologies like in Angelini *et al* (2008) and Khashman (2010) as well as in comparative algorithm performance studies like in Petropoulos *et al* (2019). However, despite being very performative in practice, the implementation of these algorithms comes with limited or none interpretability of the results, making extremely challenging to understand which are the financial ratios, KPIs and demographics that could potentially most influence the probability of default. To address this problem, not only circumscribed to this kind of applications, several tools of explainable AI have been developed in recent years, among which the most used are Variable Importance (Fisher *et al*, 2019), Partial Dependence Plot (Friedman, 2001) and SHAP (Shapley value) plot as described in Song *et al* (2016) and Frye *et al* (2020). Several examples of application of explainable AI tools are available in this regard: Brake *et al* (2019) showed how explainable machine learning could be used in the finance sector, whereas Bussmann *et al* (2021) provide evidence on how these techniques could be potentially applied to credit risk management, focusing on SHAP value and variable importance. Besides this supervised approach, it is worth noting that some applications are trying to leverage on unsupervised learning methodologies as well, like implementing Isolation forests (Liu, 2008) for credit card transactions has been addressed by Ounacer *et al* (2018).

It is relevant to note that the previous literature review is not exhaustive of the vast domain of application under investigation, but this section of the work has been organized bringing in the most relevant academic references for the followed methodological approach.

3. Case

This study focuses on developing an algorithm to predict loan arrears within two segments of the portfolio: Residential Mortgage-Backed Securities (RMBS) and Small and Medium Enterprises (SMEs). Unlike traditional credit risk models that predominantly emphasize borrower default, this research innovatively centers on forecasting loan arrears, which serves as an early indicator of potential defaults. Specifically, this section aims to achieve two primary objectives:

- Providing an overview of the dataset utilized in the algorithm's application and implementation;
- Detailing the dataset restructuring process undertaken for analysis and outlining the classification of various Key Performance Indicators (KPIs) computed for this purpose.

3.1 Data

The data used in the current analysis are coming from a wide database of loans of different banks. The banks that provided data can be considered medium to small in the context in which they operate. From a geographical point of view, the banks in the sample are spread all over the Italian territory. The data were provided in anonymised form by a private company that manages certain information on behalf of these banks. As described above, RMBS and SME data has been analyzed: in particular, roughly 4 million of cases for the former, while over 600.000 cases for the latter has been included in the dataset aggregating data from different bank sources. Each row represents a monthly snapshot of a loan, tracked over time to predict payment delays. Key columns include *Loan Identifier* for unique loan identification, *Originator* for the associated bank, and *Pool Cut-off date* for data registration timing of each observation. Other variables pertain to borrowers or loan characteristics, detailed in the following report section.

As already discussed, the focus of this work shifts from defaults to payment delays, specifically measuring the number of months in past-dues. The new definition of default and the line drawn between 90 days past due and non-performing are consistent with the choice made in this study. In particular, we have verified that our target variable was a client when it simultaneously exceeds, for more than 90 consecutive days, the absolute threshold: 100 euros for retail exposures; 500 euros for other non-retail exposures, and the relative threshold: 1% of the total amount of all exposures arising from the relationships that the customer has with the bank. The threshold for past-dues is set at four months for RMBS loans and three months for SME loans. This decision enhances the ability to

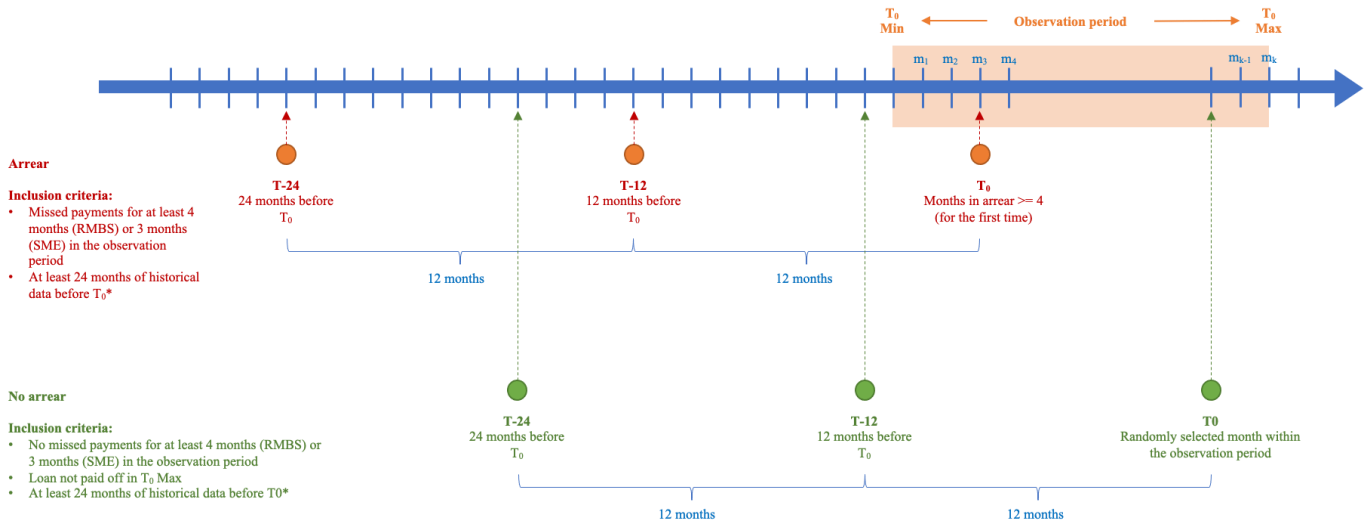
identify critical situations before defaults occur and optimizes intervention strategies to manage payment delays and prevent potential defaults. Given the peculiar features of the two cases, data restructuring has been carried out differently for RMBS and SME cases. Further details are provided here below.

With reference to the RMBS subsample, the observation period for the analyzed phenomenon was set to 2022, with the previous two years used to predict payment delays of four months or more. The main steps included creating the target dummy (0/1) variable for past-dues based on evidence of past-due presence in 2022 identifying the first past-due date, and reconstructing predictive variable values accordingly, registering the data 12 and 24 months before the first evidence of past-due. It is important to note that only cases with at least 24 months of historical data previous to the first past-due identification have been taken into account in the analysis. Similarly, to what has been discussed for the RMBS, data for SMEs has been restructured identifying the first past-due in 2022 (assuming 3 months of delay in payments) and then all the other variables have been dynamically restructured.

More specifically, the inclusion criteria for the past-due case are the missing payment for more than 4 months (RMBS) or 3 months (SME) for the first time in 2022 and the availability of at least 24 months of historical data, given the data of first past-due. Simmetrically, non past-due observations have been identified if not payments have been missed in the observation period and having 24 months of available historical data. Furthermore, for non past-due cases random sampling has been applied to rebalance the dataset: more specific details will be given below.

A diagram representing the above-described process is available in Figure 1.

Figure 1: Workflow of the KPIs creation and sampling process



3.2 Feature Engineering

After having restructured the data, an appropriate phase of feature engineering has been carried out in order to enhance the quality and depth of the available data for the following modeling step. More specifically, the variables in the dataset could be classified into two broad groups:

- Static Variables - These variables have a single value for each loan (numeric or categorical) throughout the observation period. They are usually related to the borrower's demographic information or specific loan characteristics. Static variables are useful in the model construction phase to differentiate using structural characteristics that may indicate a higher propensity for payment delays. Among the static variables there are - for example - the type of the borrower (RMBS), the nationality (RMBS), the credit quality (RMBS), the geographic area (RMBS and SME), NACE code industry (SME) and purpose of the loan (RMBS and SME).

- Dynamic Variables - These variables change over time and capture variations in the loan flow elements or the credit situation of the loan holder(s). The reference value for dynamic variables might be the value 12/24 months before the past-due or an index calculated during the observation periods. Some of the dynamic variables are the loan to value (RMBS), number of months in past-due (RMBS and SME), maximum number in past-due (RMBS and SME), borrower deposit amount (SME) and the ratio between the average past-due value and the average installment. More specifically, some of the included dynamic variables are coming directly from the dataset, while most of them have been computed as KPIs or ratio mainly using original variables like the installment value, the number of months in past-due, the past-due amount: starting from these values several metrics has been calculated (ratio of means, measures of variability, maxima and minima).

3.3 Data Rebalancing

Before moving to the actual description of the applied methodology, it is worth underlying that the restructured dataset shows a very strong imbalance in the classes of the target variable (past-due). More specifically, the proportion of positive cases, those facing past-due in 2022, is less than 0.05% for the RMBS subsample and 2.22% for the SME case. This evidence could potentially bias the testing of the new algorithm because it is extremely likely - in presence of usage of unbalanced dataset for a classification problem - to overtrain the ability of detecting the majority class, while learning much worse the specific features for the minority class.

For the aforementioned reasons, a specific rebalancing strategy has been implemented to define the final dataset for the model testing phase. In particular, a random undersampling technique has been applied to the majority class, achieving a 1/20 ratio between class in the end: evidence of the effectiveness of a similar approach has been discussed by Hasanin and Khoshgoftaar (2018) in a simulated experiment on class imbalance. Despite still having a quite unbalanced dataset, this intervention on the original sources is aimed at obtaining a more balanced dataset and to consequently let the algorithms being more effective in learning better, while training, the relationships that link the features with the minority class of the target variable.

4 Methodological Framework

In this section of the work, the methodological approach to the modeling problem will be described. As pointed out in the literature review the two main approaches to model a credit risk problem are the supervised one (classical as well as Machine learning based) and unsupervised. The main idea of this application is to merge the two solutions in order to improve the performance of both methodologies.

4.1 Description of the Algorithm

More specifically, the algorithm wants to integrate two tree-based models, namely a Random Forest (supervised block of the model) with an Isolation Forest (unsupervised part of the same): the former will serve the purpose of modeling the classical classification objective, while the latter will be used as anomalies detection tool.

The key steps and rationale behind this integrated model are detailed below:

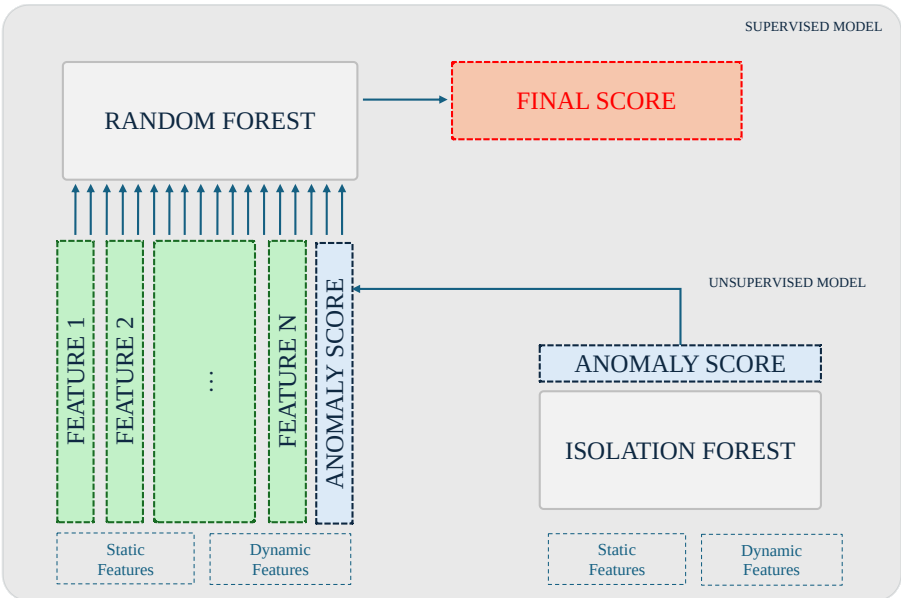
- Creation of the Unsupervised Isolation Forest Model for anomalies detection - All previously mentioned variables were used as inputs for the Isolation Forest model to estimate an anomaly score for each observation. The target variable (past-due information) was not included, focusing solely on identifying anomalous cases regardless of their connection to payment delays. The anomaly score has been included as extra predictor variable in the Random Forest Classifier.
- Creation of the Supervised Random Forest Model for the classification - As discussed above, a Random Forest model was selected due to its effectiveness in handling missing values and its proven performance in similar classification applications, as highlighted in the literature review.

At a broad level, the algorithm of Random Forest (Brieman, 2001) is a tree ensemble learning method, based on the idea of growing in parallel multiple trees (either classification or regression trees) on bootstrapped sample and using a random selection of the original features set. The predictions of the different trees are aggregated using a majority voting scheme (in case of a classification problem) or averaging (in the case of a regression problem). This method proved to be very effective in a lot of data science application, mainly for the extremely good ability in limiting overfitting and handing missing data.

Similarly, Isolation Forest (Liu, 2008) is an algorithm based on the detection of anomaly points using binary classification trees: in particular, the method is based on the applications of recursive splits of the dataset using features of the data at random and generating an anomaly score to quantify how a certain element is different from the rest of the data points.

From a theoretical standpoint, the integrated approach aims to improve prediction accuracy by including an additional variable that captures relevant anomaly information regarding each client's credit behavior before the onset of payment delays. The proposed approach could be considered theorically sound, given similar implementation in related domain as in Zakrzewska (2007), Bijak and Thomas (2012) and Bao *et al.* (2019), despite the different types of algorithms implemented. The experimental application of this approach yielded excellent prediction results for both RMBS and SME loans, achieving a high balance in performance. More specific details regarding the performance of the proposed framework will be described in the following section. For the sake of clarity, a diagram representing the modeling approach is reported Figure 2.

Figure 2: Conceptualization of the proposed algorithm



In the following paragraph the results of the testing and benchmarking of the algorithm will be presented to assess its effectiveness in terms of performance, in this section. More specifically, the novel model has been benchmarked with different algorithms, both classical as well as Machine Learning based to gain a complete and multifaceted assessment of its performance. The performance of the different algorithms has been assessed using hold-out approach (75% of the observation has been used for the training of the algorithm, while the remaining 25% for testing on fresh sample). A complete list, along with a brief description of the algorithm, is presented in Table 1.

Table 1: Descriptions of tested algorithms

Model	Description
Logistic Regression	A statistical model that uses a logistic function to model the probability of a binary phenomenon (0/1)
Logistic Regression with Regularization	An extension of the Logistic Regression model that includes types of penalization (L1, L2, or ElasticNet) on coefficients to prevent overfitting and improve the generalizability of a classification model. In the case under analysis, ElasticNet has been implemented
Random Forest	An ensemble learning model based on the parallel construction of multiple decision trees with the aim of reducing overfitting problems
XGBoost	A gradient boosting (ensemble) algorithm based on the sequential construction of decision trees
H2O AutoML Model	An automated machine learning framework that explores various models and data pre-processing techniques to find the best possible model such as GLM (Generalized Linear Models, DRF (Random Forest & Extremely Randomized Trees), XGBoost, GBM (Gradient Bosteed Methods), Deep learning (Fully connected multilayer ANNs) and StackEnsemble. This solution will be tested to (i) validate the results obtained from the Random Forest and XGBoost algorithms and to (ii) include a performance benchmark coming from an automatic yet robust and performative modeling framework

4.2 Hyperparameters tuning

When building and assessing the performance of a Machine Learning model, it is extremely important to perform the tuning of hyperparameters: this is because the final effectiveness of the algorithm massively depends on the combination of the different tunable parameters of the different models.

For each of the included models, different hyperparameters' configurations have been tested and results have been validated using a 5-fold Cross Validation. The validation of the hyperparameters has been conducted through the implementation of Random Discrete search, uniformly sampling from a grid that encloses all the possible combination of hyperparameters.

All the models have been trained and tested using the H2O framework's for excluding any possible external bias related to the developer of the library or package.

The selection of the optimal hyperparameter combination for each algorithm was based on maximizing the Area Under the ROC Curve (AUC) metric.

This metric, indeed, is particularly useful in comparing models with different hyperparameters' configurations and it is independent of the threshold value set for classifying positive and negative cases, unlike other metrics such as sensitivity, specificity, accuracy, and F-measure.

For Logistic Regression with regularization, after initially employing a grid search with commonly adopted penalization degrees, manual testing of specific regularization values was conducted to gain greater sensitivity to the final output. However, it was observed that the final output exhibited minimal changes in performance even to significant variations in the penalization degree.

Given the experimental nature of this work, more specific information on the tuning of hyperparameters of the Supervised-Unsupervised model will be provided here below.

Regarding the Random Forest model, optimization was performed through a grid search of the following hyperparameters, limited to these values:

- Max Depth: 3, 5, 10, 20, 30

- Mtries (sampling column): 5, 10, 20
- Sample rate (sampling row proportion): 0.5, 0.632, 0.75
- Ntrees: 100, 200, 500

With reference to the Isolation Forest, it is extremely important to highlight that the default setting of the hyperparameters has been used given the unsupervised nature of the algorithm. This approach has been followed both for the RMBS subsample as well as for the SME.

Once selected for each of the tested models, the best configuration of the hyperparameters assessment on the test set has been applied. More details will be given in the following paragraph.

4.3 Validation Strategy of the tested algorithms

The current section of the work will present the approach implemented to validate the different algorithms. More specifically, following the usual procedures to validate a classification model, the algorithms have been compared according to several metrics that are summarized here below:

- Number of False Positives
- Number of False Negatives
- Precision
- Sensitivity
- Specificity
- AUC (Area under the curve of the Receiver Characteristic Operating curve)

It is important to note that the accuracy metric computed as proportion of the cases correctly classified into their respective classes, despite being widely used in classification problems, is extremely sensitive to the set threshold to classify the cases into the positive or negative class.

For this reason, a more exhaustive and less sensitive measure, like the AUC, will be used to select the most performing model.

The AUC, area under the ROC curve, is indeed computed by varying all the possible values of the classification threshold and then computing the values of specificity and sensitivity before plotting them, providing in the end a more holistic validation of the algorithm[1].

It is worth noting that for all the models the threshold for the different metrics obtained from the Confusion Matrix is reported: as an overall approach, the threshold has been selected to balance the sensitivity and specificity of the prediction through the maximization of the Youden's index[2].

In the next pages detailed results on all the previously mentioned metrics will be provided both for RMBS as well as for SME.

In this section the testing results for the two different subsamples will be presented along with the rationale behind the choice of the best model. Table 2 and Figure 3 report all the detailed performance metrics for the RMBS sample.

Table 2: RMBS - Performance metrics for the validated models (computed on 9580 cases of which 46 are past-dues)

Model	Threshold	False Pos.	False Neg.	Precision	Sensitivity	Specificity	AUC
Logistic Regression	0.006	1066	12	0.031	0.739	0.888	0.831
H2O Auto ML (GLM)	0.004	1508	7	0.025	0.848	0.842	0.863
Random Forest + Isolation Forest	0.006	1519	8	0.024	0.826	0.841	0.877
Pen. Logistic Regression	0.003	2457	5	0.016	0.891	0.742	0.851
XGBoost	0.001	1512	10	0.023	0.783	0.841	0.845

[1] the AUC metric ranges from 0 to 1. A model with an AUC of 0.5 is extremely poor (random guess), while an AUC of 1 represents the perfect model. In practice, values of AUC greater than 0.75 characterize good classification models

[2] the Jouden's index is computed as $J = Sensitivity + Specificity - 1$

Figure 3: RMBS - Performance comparison for the validated models (Specificity is reported in green, Sensitivity is reported in red, AUC is reported in blue)

As it is possible to note from the results, the proposed model seems to be very performative with respect to most of the metrics included in the analysis. Looking at the AUC, the Random Forest + Isolation Forest (RF+IF from now on) algorithm outperforms all the other tested ones, scoring a 0.877 of AUC compared to the second-best model that shows a value of 0.863 on the same metric. This result shows that the model proves quite good at detecting both the positive as well as negative class.

It is true that - by focusing the attention on the number of False Negative (past-due cases which are predicted as not in past-due) - the algorithm with the best performance is the Penalized Logistic Regression (only 5 cases are false negative); however, this evidence is counterbalanced by a very high number of False Positive cases (more than 2400 false positive).

Since the scope of the algorithm is to have a relatively good balanced in predicting both the classes under analysis, it was considered not advisable to select as best model one with such a high number of false positive cases because it could trigger in practice a too harsh contract revision policy from the institute.

For the sake of completeness, it is important to report that the same performance metrics have been computed setting the classification threshold through the maximization of the F1 score[3], a performance metric widely used in case of unbalanced dataset. However, when setting the threshold in this fashion the number of False Negative increases to 34 in the case of the most performative algorithm - RF + IF (according to the AUC) - making this choice not suitable at all from a practical standpoint.

Given all the evidence previously detailed, the RF+IF seems the best model in terms of balance between different metrics, electing it as most suitable for a real case application scenario. Detailed values are reported in the Appendix (Table 7).

Similarly, to what has been discussed for the RMBS sample, Table 3 and Figure 4 report the performance results for the SME data points.

In general, all the metrics are slightly better for the SME case compared to the RMBS but there are a lot of similarities between the two sub samples. Assuming the same approach followed for RMBS, the metrics reported in the table are those obtained setting the threshold when maximizing the Youden's index.

More specifically, the most performative model in terms of AUC is still the RF +IF (0.957) followed by the H2O Auto ML (0.95) and the XGBoost (0.93).

Regarding the number of False Positive, the RF + IF algorithm is still the best one in the group (only 2 cases are misclassified as false negative); while the lowest value of false positive could be found when implementing the Penalized Logistic Regression (31 misclassified cases).

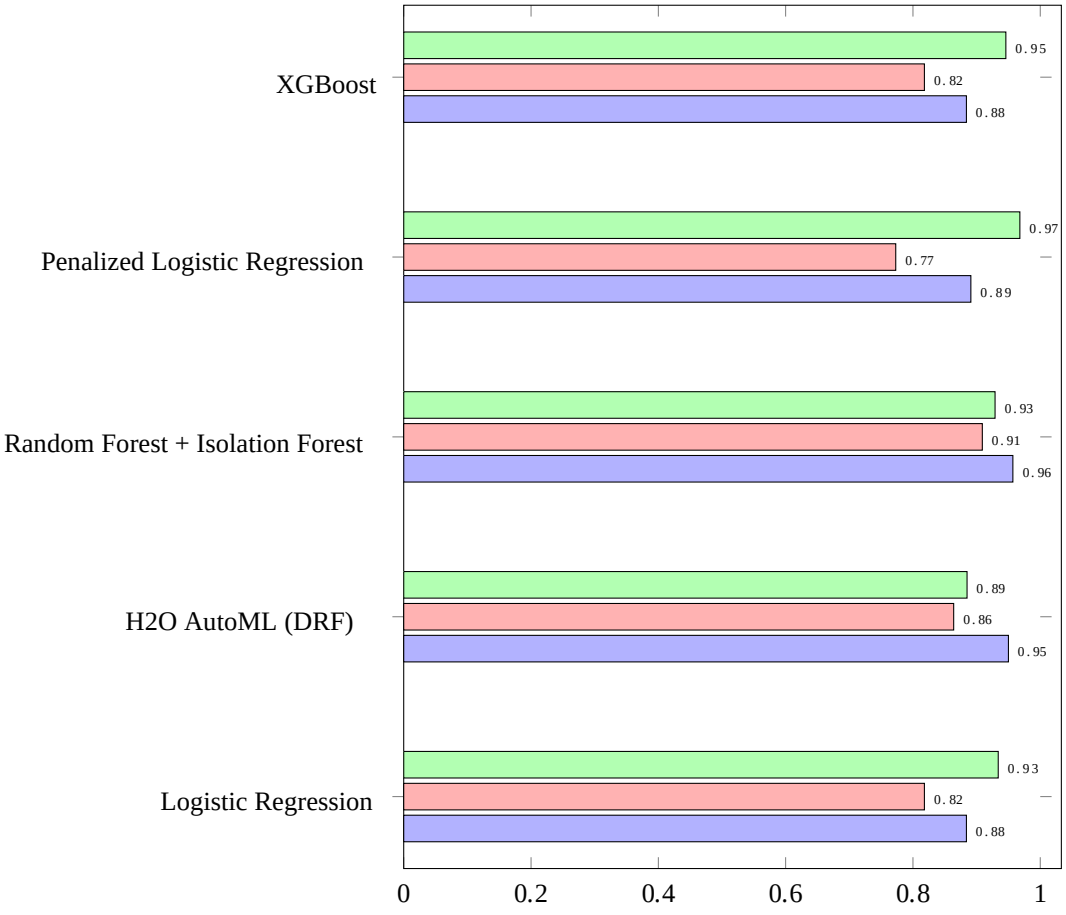
[3] the F1 score is computed as $\frac{2tp}{2tp+fp+fn}$, where tp are the true positive, fp are the false positive and fn are the false negative

Also in this case, the performance has been double checked computing the same performance metrics but setting the threshold through the maximization of the F1 score: however, similarly to what has been discussed for the RMBS case, the absolute number of false negative increased for all the algorithms suggesting to discharge this approach. Detailed results are provided in Table 8

Table 3: SME - Performance metrics for the validated models (computed on 993 cases of which 22 are past-dues)

Model	Threshold	False Pos.	False Neg.	Precision	Sensitivity	Specificity	AUC
Logistic Regression	0.024	64	4	0.22	0.818	0.934	0.884
H2O Auto ML (DRF)	0.026	112	3	0.145	0.864	0.885	0.95
Random Forest + Isolation Forest	0.036	69	2	0.225	0.909	0.929	0.957
Pen. Logistic Regression	0.041	31	5	0.354	0.773	0.968	0.891
XGBoost	0.032	52	4	0.257	0.818	0.946	0.93

Figure 4: SME - Performance comparison for the validated models (Specificity is reported in green, Sensitivity is reported in red, AUC is reported in blue)



4.4 Random Forest + Isolation Forest, Variable Importance and Partial Dependence Plot

To complement the performance analysis just exposed in the previous paragraphs, the variable importance has been computed in order to understand which are the variables that most impact on the past-due both for RMBS as well as for the SME cases. As mentioned in the literature review, machine learning based methodologies are usually, like in the case under analysis, better in terms of performance compared to classical models but one of the main drawbacks of these algorithms is the lack of interpretability of results. More specifically, when dealing with regression models it is easy to assess the effect of one feature on the target variable, both in terms of sign as well as magnitude, by interpreting the coefficient; this is not possible with most of the machine learning methods: for this reason, several methodologies have been developed to indirectly estimate these effects.

Here below, the variable importance in predicting the past-due for RMBS and SME is reported (the ten most important variables are shown in Table 4). As it is possible to see, the variables that are most important in predicting the past-due for RMBS loans are geographic area, the age of the debtor, some ratios and KPIs (std. dev. of the ratio past-due/installment; max number of months in

past-due etc.), the current interest rate and the isolation forest anomaly score. On the other hand, for the SME cases, the most important variables turn out to be the industry of the company, the number of months in past-due, the past-due balance and the geographic area. In this case, the anomaly score of the isolation forest is relevant but it is not included in the ten most important features. In order to understand how each value or level of these variables could potentially impact the past-due, partial dependence plot (PDP) are shown in Figure 5, Figure 6 and Figure 7. For example, it is possible to note (Figure 5) that the Isolation Forest Anomaly Score has a quite weak positive effect on the probability of past-due; similarly, for SME higher number of months with positive past-due value (last year) increases the likelihood of missing the monthly payment.

Table 4: Variable Importance (RMBS and SME)

Variable Importance (RMBS)	Percentage
ST_geographic_region	11.0%
24_LY_std_past-due_over_installment	2.8%
ST_age	2.8%
12_LY_max_num_months_past-due	2.5%
24_current_interest_rate	2.2%
24_LY_max_num_months_past-due	2.2%
IF_anomaly_score	2.1%
12_number_months_past-dues	2.1%
24_current_interest_rate_margin	2.1%
12_LY_n_months_positive_past-due	2.0%

Variable Importance (SME)	Percentage
ST_ industry_code	18.4%
12_LY_n_months_positive_past-due	6.3%
12_LY_n_months_positive_past-due_balance	5.4%
24_mean_total_past-due_balance	4.6%
24_ LY_n_months_positive_past-due	4.4%
24_LY_n_mnoth_positive_past-due	4.0%
12_ mean_total_past-due_balance	3.2%
ST_geographic_region	2.9%
24_st_dev_tot_past-due_balance	2.3%
24_max_tot_past-due_balance	2.1%

Figure 5: RMBS – Partial Dependence plot (Isolation Forest Anomaly score)

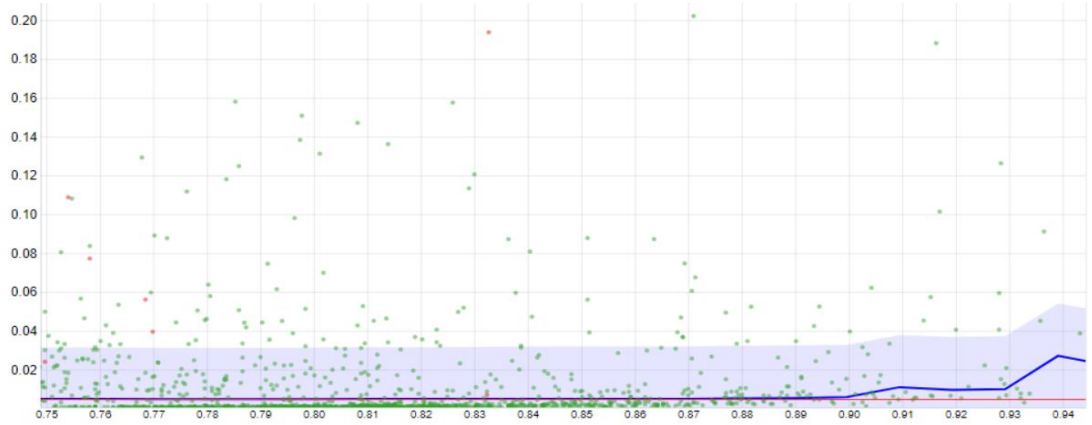


Figure 6: RMBS – Partial Dependence plot (Ratio between Standard Deviation of Past-dues value and the mean installment amount)

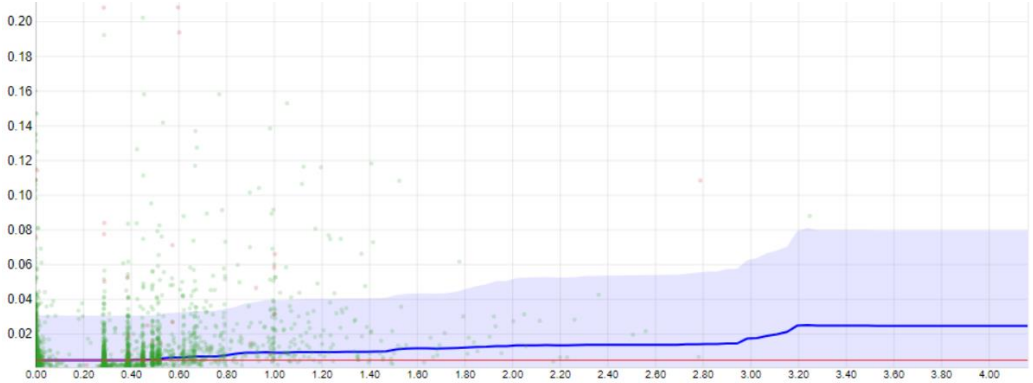
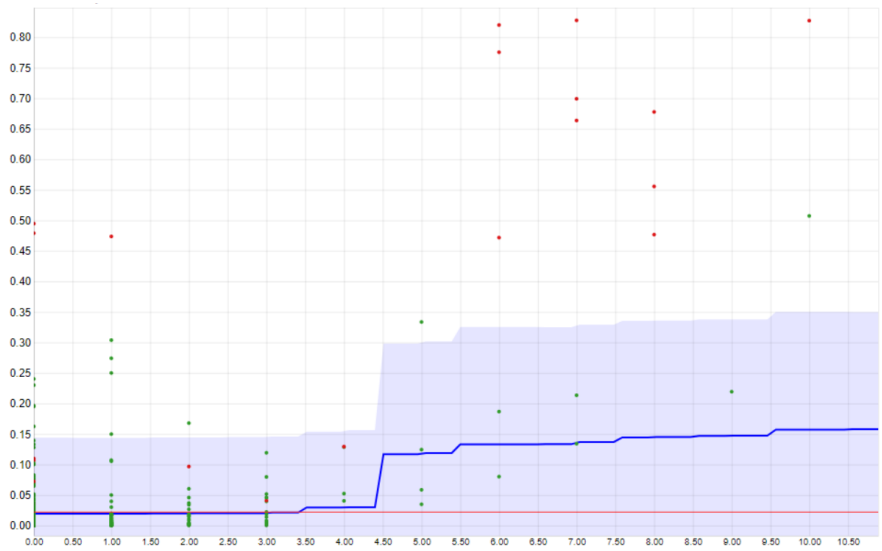


Figure 7: SME – Partial Dependence plot (Number of months with positive past-due value (last year))



4.5 Algorithm Validation on 2023 Data

With the purpose of further validation, the selected algorithm has been tested with specific reference to its robustness and effectiveness on new observed cases, coming from 2023 data. This activity is mainly aimed at further testing the selected model but – at the same time – applying it into a realistic scenario, very similar to the one to which it will be exposed once deployed in practice. Table 5 shows the performance metrics computed on 2023 data both for RMBS and for SME.

Table 5: RMBS and SME performance metrics for 2023 data

Model	Threshold	False Pos.	False Neg.	Precision	Sensitivity	Specificity	F-Measure	AUC
RMBS – Random Forest + Isolation Forest (computed on 30583 cases of which 207 are past-dues)	0.024	64	4	0.22	0.818	0.934	0.346	0.884
SME – Random Forest + Isolation Forest (computed on 2835cases of which 75 are past-dues)	0.036	52	4	0.257	0.818	0.946	0.391	0.93

As it is possible to see from the results reported in the table, the algorithm seems to be quite good at predicting the past-due of the customers: by looking at the number of False Negative in the case of RMBS only 4 have been misclassified and 4 also for SME. Focusing the attention on other performance metrics, it is worth noting that the value of the AUC is very good (0.884 in case of RMBS and 0.93in case of SME), aligning with the performance results in the testing set. Likewise, the values of Sensitivy and Specificity are satisfactory being respectively 0818 and 0.934 for RMBS and 0.818 and 0.946 for SME.

5 Final Remarks

It is possible to assert, given all the previously reported results, that the developed model seems to be a good and well-suited altervative to more diffused methodologies in the credit risk estimation domain. From a technical standpoint, the model has achieved very good performance under all the considered criteria, outperforming in the most relevant ones all the challenging methodologies both for RMBS as well as SME. In addition, by testing the model on fresh data (2023) the level of its effectiveness has been validated, confirming the robustness of the approach that strengthens the flexibility of the supervised classification model (Random Forest) with the anomalies detection properties of the unsupervised one (Isolation Forest).

From an operational standpoint, this new model could be implemented in real application to help the management in monitoring the current loan portfolio and consequently taking informed decisions on specific case. It is, indeed, important to remark that this work is innovative not only in terms of implemented methodology but also in terms of the targeted phenomenon (past-due prediction). The past-due prediction could be then used to foresee the eventual default, allowing the decision makers to put in practice specific actions just for a circumscribed subgroup of customers that are more likely – according to the model – to not repay entirely the loan.

As a final remark, an application of the developed algorithm will be briefly discussed hereafter. It is relevant to remember that the output of the model – as it is typical for a binary classification algorithm – is a numeric score ranging from 0 to 1: the higher this value, the more likely the past-due is. As described above, it is possible to convert this score into two classes (past-due and no past-

due) depending on the set threshold. Alternatively, it is possible to use this score to define risk classes that could be more helpful in practice to have a better level of detailed of the positions that need to be monitored carefully.

Table 6 reports the data of risk classes for RMBS loan. From a technical standpoint, the classes are not overlapping (each observation will be in one class only) and strictly increasing in terms of associated past-due probability, with class 7 being the one with the highest risk. The classes have been built using statistical criteria starting from the entire distribution of predicted scores. For instance, considering the first three riskiest classes (risk class 5 up to risk class 7) it is possible to correctly detect roughly 67% of the total past-due cases, proving this approach to be very handy for a practical sperimentation to new cases.

Table 6: RMBS risk classes

Risk class	Tot. Loan	N. no past-due	N. past-due	% past-due in each class	% detected past-due (n past-due class/ tot. past-due)	Cumulative % detected past-due	Rel. dimension of the class	Cumulative % of loan
7	172	155	17	9.88%	37.0%	37.0%	1.8%	1.8%
6	288	279	9	3.13%	19.6%	56.5%	3.0%	4.8%
5	288	283	5	1.74%	10.9%	67.4%	3.0%	7.8%
4	480	475	5	1.04%	10.9%	78.3%	5.0%	12.8%
3	960	956	4	0.42%	8.7%	87.0%	10.0%	22.8%
2	2784	2781	3	0.11%	6.5%	93.5%	29.1%	51.9%
1	4608	4605	3	0.07%	6.5%	100.0%	48.1%	100.0%
Total	**9580**	**9534**	**46**	**0.48%**	100.0%		100.0%	

In the field of banking and financial services, a critical focus for industry stakeholders is the accurate prediction of probability of default (PD) and the effective classification of raw data into risk classes. This study addresses the challenge of predicting PD for Residential Mortgage-Based Securities (RMBS) and Small and Medium Enterprises (SMEs) within the Italian banking sector. It presents an innovative methodology that combines a Random Forest classification model with an Isolation Forest anomaly detection technique, trained on a comprehensive dataset covering the period 2020-2022.

What sets this research apart is its unique emphasis on the delinquency status of RMBS and SME clients as the primary target variable. By focusing on arrears rather than the broader concept of PD, this approach provides deeper insights into customer financial stress, facilitating proactive monitoring and intervention strategies for decision-makers.

The ultimate goal of this study is to develop a robust, practical algorithm capable of accurately predicting both individual customer and corporate delinquencies, thereby improving management decision making. Empirical results highlight the superiority of the proposed framework over traditional statistical and machine learning algorithms in credit risk modelling, demonstrating robust performance validated with 2023 data and confirming its operational readiness.

However, when selecting and deploying a machine learning model such as the one proposed in this article, there are a number of critical aspects that need to be considered. Practitioners must consider that validity of the model is closely linked to the quality and representativeness of the data set used for training (2020-2022). If historical data does not accurately reflect future economic conditions or changes in customer behavior, predictions may be inaccurate. Although the model performed well on test data and was also validated on 2023 data, there is always a risk of overfitting, especially with complex machine learning models that are based on many features. In order to avoid performance degradation on new, previously unseen data, it's always recommended to retrain the model, at least on annual basis.

Random forest and isolation forest models are known to be less interpretable than simpler models. This lack of transparency can make it difficult for decision makers to understand and trust the model's predictions, even if we might use XAI tools (such as partial dependence plots) to improve interpretability, as shown in the article.

The division into risk classes and the definition of thresholds for classification (past due and not past due) can introduce bias. If the thresholds are not properly calibrated, classification errors can occur, leading to incorrect management decisions. Isolation Forest is designed to detect anomalies, but may have difficulty detecting anomalies in contexts with high variability or complex data patterns. This can affect the accuracy of predicting failure. Models may not be able to adapt quickly to sudden changes in market conditions, such as financial crises or regulatory changes, limiting their effectiveness in situations of economic stress.

Implementing and maintaining these complex models can be costly for financial institutions, both in terms of computational resources and the expertise required to run and update the models. Although the model has been validated with fresh data from 2023, its performance may not be fully generalisable to other geographical contexts or sectors beyond Italian banks.

In conclusion, financial institutions are encouraged to adopt advanced credit risk models that combine Random Forest classification with Isolation Forest anomaly detection. This recommendation is based on the superior performance of the hybrid model in predicting delinquencies, suggesting that it could improve the accuracy of credit risk assessments. Implementation of the developed model can significantly improve the monitoring of loan portfolios. By accurately identifying loans at higher risk of delinquency, banks

can proactively mitigate potential losses. The model's ability to segment customers into risk classes enables more targeted and effective management strategies.

By focusing on predicting delinquencies rather than defaults, the model provides a nuanced understanding of borrowers' financial stress. This enables financial institutions to design and implement early intervention strategies, such as restructuring loans or offering financial counselling to at-risk borrowers, potentially preventing defaults.

Regulators could consider updating guidelines to require the use of sophisticated credit risk models. The effectiveness and robustness of the model in predicting delinquencies could help financial institutions meet regulatory requirements more efficiently and accurately.

The classification of loans into risk classes allows banks to optimise the allocation of resources. Higher-risk loans can be monitored more closely, while lower-risk loans require less oversight, resulting in more efficient use of human and technology resources. Detailed risk classifications allow financial institutions to refine their risk-based pricing strategies. By aligning loan pricing with the predicted risk of default, banks can better balance risk exposure and profitability.

Understanding the likelihood of default enables more effective customer engagement. Banks can offer personalised communication and support to high-risk customers, improving satisfaction and potentially reducing churn.

To maintain the accuracy and effectiveness of the model, financial institutions should establish policies for continuous data collection, updating and analysis. The model's reliance on comprehensive and recent data (e.g. 2020-2022) underscores the importance of a data-driven approach.

Investment in staff training is critical for banks to effectively use advanced credit risk models. Appropriate training ensures that the insights provided by the model are correctly interpreted and applied in the decision-making process. In addition, the success of the model encourages further collaboration between academic researchers, financial institutions and technology providers. Continuous innovation and validation of such models is essential to keep pace with evolving market conditions and emerging risks.

By adopting these policy implications, financial institutions can use the developed model to improve their credit risk management practices. This adoption could lead to more stable and resilient financial systems and improve overall efficiency, compliance and customer relations in the banking sector.

Appendix

Table 7: RMBS - Performance metrics for the validated models, threshold set maximizing F1 score (computed on 9580 cases)

Model	Threshold	False Pos.	False Neg.	Precision	Sensitivity	Specificity	F-Measure	AUC
Logistic Regression	0.076	78	36	0.114	0.217	0.992	0.149	0.831
H2O Auto ML (GLM)	0.086	42	39	0.143	0.152	0.996	0.147	0.863
Random Forest + Isolation Forest	0.067	67	34	0.152	0.261	0.993	0.192	0.877
Pen. Logistic Regression	0.088	21	39	0.25	0.152	0.998	0.189	0.851
XGBoost	0.026	62	35	0.151	0.239	0.993	0.185	0.845

Table 8: SME - Performance metrics for the validated models, threshold set maximizing F1 score (computed on 993 cases)

Model	Threshold	False Pos.	False Neg.	Precision	Sensitivity	Specificity	F-Measure	AUC
Logistic Regression	0.152	8	9	0.619	0.591	0.992	0.605	0.884
H2O Auto ML (DRF)	0.299	2	9	0.867	0.591	0.998	0.703	0.95
Random Forest + Isolation Forest	0.342	1	9	0.929	0.591	0.999	0.722	0.957
Pen. Logistic Regression	0.131	11	8	0.56	0.636	0.989	0.596	0.891
XGBoost	0.146	5	8	0.737	0.636	0.995	0.683	0.93

Table 9: RMBS – List of Static Features considered in the models

Feature name	Description
ST_Borrower Type	Debtor type
ST_Number of Debtors	Number of debtors
ST_Borrower's Employment Status	Debtor's employment status
ST_First-time Buyer	First-time Buyer
ST_Class of Borrower	Class of debtor
ST_Primary Income	Primary debtor's annual income
ST_Secondary Income	Secondary debtor's annual income
ST_Resident	Residence
ST_Origination Channel / Arranging Bank or Division	Sales channel, arranging bank or division
ST_Purpose	Purpose of financing
ST_Amount Guaranteed	Guaranteed amount
ST_Loan Currency Denomination	Currency
ST_Original Balance	Initial amount
ST_Fractioned / Subrogated Loans	Fractioned loan
ST_Repayment Method	Repayment method
ST_Payment Frequency	Installment frequency
ST_Type of Guarantee Provider	Type of guarantor
ST_Guarantee Provider	Name of guarantor
ST_Pre-payment Amount	Amount of prepayments or early reductions
ST_Interest Rate Type	Interest rate type
ST_Geographic Region List	Province code
ST_Property Type	Property type
ST_Original Loan to Value	Loan to value
ST_Valuation Amount	Original appraisal amount
ST_Additional Collateral Provider	Provider of additional real guarantees
ST_Income Verification for Primary Income	Primary debtor income certification
ST_Income Verification for Secondary Income	Secondary debtor income certification
ST_Valuation Date	Original appraisal date
ST_Shared Ownership	Shared ownership
ST_Restructuring Arrangement	Restructured loan indicator
ST_Property Rating	Property rating
ST_Lien	Mortgage grade
ST_Length of Payment Holiday	Duration of suspensions
ST_Interest Cap Rate	Interest rate cap
ST_Loan Term	Original loan duration
ST_Mortgage Inscription	Mortgage registration amount
ST_Mortgage Mandate	Mortgage registration mandate
ST_New Property	New property
ST_Prior Repossessions	Previous mortgage possession
ST_Principal Grace Period	Number of months of grace period
ST_Payment Type	Payment type
ST_Prepayment_ratio	Prepayment amount/Original balance ratio
ST_Tot_Income	Sum of Primary + Secondary income
ST_Age	Age of the borrower at t0

Table 10: RMBS – List of Dinamic Features considered in the models (each feature is measured at t-12 and t-24)

Feature name	Description
Current Balance	Outstanding debt balance
Payment Due	Contractual amount of the installment

Debt to Income	Installment to income ratio
Cumulative Pre-payments	Total prepayments or early reductions
Current Interest Rate Index	Reference rate
Current Interest Rate	Applied rate
Current Interest Rate Margin	Spread
Interest Rate Reset Interval	Rate review
Current Loan to Value	Current Loan to Value
Current Valuation Amount	Updated appraisal amount
Current Valuation Type	Type of updated appraisal
Current Valuation Date	Date of updated appraisal
Date Last in Arrears	Date since the debtor is in arrears
Arrears Balance	Balance of arrear amounts
Number Months in Arrears	Number of months in arrears
Arrears 1 Month Ago	Balance of arrear amounts recorded the previous month
Arrears 2 Months Ago	Balance of arrear amounts recorded two months earlier
Months in Arrears Prior	Number of months in arrears at the end of the month preceding the repayment date
LY_max_num_month_arrear	Maximum number of months in arrear (Last Year)
LY_N_month_pos_arrear	Number of months in which the arrears balance is positive (Last Year)
LY_max_balance_arrear	Maximum value of arrears balance (Last Year)
LY_N_balance_pos_arrear	Number of times the arrears balance is positive (Last Year)
LY_avg_arrear_over_payment	Average Arrears/average installment ratio (Last Year)
LY_std_ arrear_over_payment	Standard Deviation Arrears/average installment ratio (Last Year)
LY_Payment_Income_Ratio	Installment/Income ratio (Last Year)

Table 11: SME – List of Static Features considered in the models

Feature name	Description
ST_Geographic Region	Geographic province
ST_Obligor Legal Form / Business Type	Debtor type
ST_Borrower Basel III Segment	Segment to which the bank's client (debtor) belongs according to Basel III regulations
ST_Syndicated	Syndicated loan
ST_ Industry Code	Debtor's sector
ST_Original Loan Balance	Initial loan amount
ST_Securitised Loan Amount	Securitised loan amount, i.e., the outstanding debt at the securitisation date
ST_Purpose	Purpose
ST_Principal Payment Frequency	Frequency of principal payment
ST_Interest Payment Frequency	Frequency of interest payment
ST_Weighted Average Life	Weighted average life (considering the type of amortisation and the maturity date) at the pool cut-off date
ST_Prepayment Penalty	Prepayment penalties
ST_Interest Floor Rate	Interest rate floor (lower limit)

ST_Final Margin	Final spread
ST_Interest Reset Period	Reference index review interval
ST_Turnover of Obligor	Debtor's turnover
ST_Short Term Financial Debt	Short-term financial debts
ST_Earnings Before Interest, Taxes, Depreciation and Amortisation (EBITDA)	EBITDA
ST_Number of Employees	Number of employees
ST_EBITDA/Turnover	EBITDA/Turnover

Table 12: SME – List of Dinamic Features considered in the models (each feature is measured at t-12 and t-24)

Feature name	Description
Total credit limit granted to the loan	Credit limit granted to the loan
Total Credit Limit Used	Credit used
Borrower deposit amount	Borrower's deposit amount (current account balance)
Borrower deposit currency	Borrower's deposit currency
Loan Hedged	Loan protection to offset currency risk losses (underlying risk)
Current Balance	Outstanding debt
Maximum Balance	Maximum outstanding debt
Amortization Type	Type of amortization
Regular Principal Instalment	Principal installment
Regular Interest Instalment	Interest installment
Balloon Amount	A loan with a large final installment
Payment type	Payment method
Prepayment Penalty	Prepayment penalties
Current Interest Rate	Applied rate
Interest Cap Rate	Cap (upper limit of the rate)
Interest Floor Rate	Floor (lower limit of the rate)
Interest Rate Type	Type of interest rate
Current Interest Rate Index	Reference rate
Current Interest Rate Margin	Spread
Revised Interest Rate Index	Revised interest rate index (post option exercise)
Final Margin	Final spread
Interest Reset Period	Reference index review interval
Currency of Financials	Financial statement currency
Number of Days in Interest Arrears	Number of days in interest arrears
Number of Days in Principal Arrears	Number of days in principal arrears
Days in Arrears Prior	Number of days in arrears in the month preceding repayment
Sum_arrear_balance	Total arrear balance
Regular_instalment	Total installment
LY_N_month_pos_arrear	Number of months in which the arrear balance is positive (Last Year)
LY_N_balance_pos_arrear	Number of times in which the arrear balance is positive (Last Year)
LY_mean_arrear_balance	Average arrear balance (Last Year)
LY_std_arrear_balance	Standard deviation of arrear balance (Last Year)
LY_max_arrear_balance	Maximum arrear balance (Last Year)
LY_tot_interest	Total interest (Last Year)

References

- Alaka, Hafiz A., Lukumon O. Oyedele, Hakeem A. Owolabi, Vikas Kumar, Saheed O. Ajayi, Olugbenga O. Akinade, and Muhammad Bilal. "Systematic Review of Bankruptcy Prediction Models: Towards a Framework for Tool Selection." Expert Systems with Applications 94 (March 15, 2018): 164–84. https://doi.org/10.1016/j.eswa.2017.10.040
- Altman E., Financial ratios, discriminant analysis and the prediction of corporate bankruptcy, J. Finance, 23 (4) (1968), pp. 589-609
- Angelini, Eliana, Giacomo di Tollo, and Andrea Roli. "A Neural Network Approach for Credit Risk Evaluation." The Quarterly Review of Economics and Finance 48, no. 4 (November 1, 2008): 733–55. https://doi.org/10.1016/j.qref.2007.04.001
- Arminger, Gerhard, Daniel Enache, and Thorsten Bonne. "Analyzing Credit Risk Data: A Comparison of Logistic Discrimination, Classification Tree Analysis, and Feedforward Networks." SSRN Scholarly Paper. Rochester, NY, April 8, 1997. https://papers.ssrn.com/abstract=4801
- Baas, T., and M. Schrooten. 2006. "Relationship Banking and SMEs: A Theoretical Analysis." Small Business Economics 27: 127–137. Bank of Italy. 2017. Annual report, Year 2016.
- Baesens, B., T. Van Gestel, S. Viaene, M. Stepanova, J. Suykens, and J. Vanthienen. "Benchmarking State-of-the-Art Classification Algorithms for Credit Scoring." Journal of the Operational Research Society 54, no. 6 (June 1, 2003): 627–35. https://doi.org/10.1057/palgrave.jors.2601545
- Bao, Wang, Ning Lianju, and Kong Yue. "Integration of Unsupervised and Supervised Machine Learning Algorithms for Credit Risk Assessment." Expert Systems with Applications 128 (August 15, 2019): 301–15. https://doi.org/10.1016/j.eswa.2019.02.033
- BCBS, 2006. International Convergence of Capital Measurements and Capital Standards: A Revised Framework Comprehensive version.
- Berger, A. N., and G. F. Udell. 1995. "Relationship Lending and Lines of Credit in Small Firm Finance." The Journal of Business 68:351–381.
- Berger, A. N., G. F. Udell 1994. Did risk-based capital allocate bank credit and cause a "credit crunch" in the United States? Journal of Money, Credit and Banking 26 (3): 585–628
- Bijak, Katarzyna, and Lyn C. Thomas. "Does Segmentation Always Improve Model Performance in Credit Scoring?" Expert Systems with Applications 39, no. 3 (February 15, 2012): 2433–42. https://doi.org/10.1016/j.eswa.2011.08.09
- Bofondi, M., L. Carpinelli, and E. Sette. 2013. "Credit Supply during a Sovereign Debt Crisis." Bank of Italy Temi di Discussione, (Working Paper) No, 909
- Bonfim D., Credit risk drivers: evaluating the contribution of firm level information and of macroeconomic dynamics, J. Bank. Finance, 33 (2009), pp. 281-299
- Bracke, Philippe, Anupam Datta, Carsten Jung, and Shayak Sen. "Machine Learning Explainability in Finance: An Application to Default Risk Analysis." SSRN Electronic Journal, January 1, 2019. https://doi.org/10.2139/ssrn.34351
- Breiman, Leo. "Random Forests." Machine Learning 45, no. 1 (October 1, 2001): 5–32. https://doi.org/10.1023/A:1010933
- Brown, Iain, and Christophe Mues. "An Experimental Comparison of Classification Algorithms for Imbalanced Credit Scoring Data Sets." Expert Systems with Applications 39, no. 3 (February 15, 2012): 3446–53. https://doi.org/10.1016/j.eswa.2011.09.033
- Brunner, Antje, Jan Pieter, and Martin Weber. 2000. Information production in credit relationship: On the role of internal ratings in commercial banking. CFS Working Paper 10.
- Bussmann, Niklas, Paolo Giudici, Dimitri Marinelli, and Jochen Papenbrock. "Explainable Machine Learning in Credit Risk Management." Computational Economics 57, no. 1 (January 1, 2021): 203–16. https://doi.org/10.1007/s10614-020-10042-0
- Carling K, T. Jacobson, J. Linde, K. Roszbach, Corporate credit risk modeling and the macroeconomy, J. Bank. Finance, 31 (2007), pp. 845-868
- Chi, Bo-Wen, and Chiun-Chieh Hsu. "A Hybrid Approach to Integrate Genetic Algorithm into Dual Scoring Model in Enhancing the Performance of Credit Scoring Model." Expert Systems with Applications 39, no. 3 (February 15, 2012): 2650–61. https://doi.org/10.1016/j.eswa.2011.08.120
- Dastile, Xolani, Turgay Celik, and Moshe Potsane. "Statistical and Machine Learning Models in Credit Scoring: A Systematic Literature Survey." Applied Soft Computing 91 (June 1, 2020): 106263. https://doi.org/10.1016/j.asoc.2
- Degryse, H., and P.Van Cayseele. 2000. "Relationship Lendingwithin a Bank-Based System: Evidence from European Small Business Data." Journal of Financial Intermediation 9 (1): 90–109
- Desai, Vijay S., Jonathan N. Crook, and George A. Overstreet. "A Comparison of Neural Networks and Linear Scoring Models in the Credit Union Environment." European Journal of Operational Research 95, no. 1 (November 22, 1996): 24–37. https://doi.org/10.1016/0377-2217(95)00246-4
- Feldman, David, and Shulamith Gross. "Mortgage Default: Classification Trees Analysis." The Journal of Real Estate Finance and Economics 30, no. 4 (June 1, 2005): 369–96. https://doi.org/10.1007/s11146-0057013-7
- Fisher, Aaron, Cynthia Rudin, and Francesca Dominici. "All Models Are Wrong, but Many Are Useful: Learning a Variable's Importance by Studying an Entire Class of Prediction Models Simultaneously." Journal of Machine Learning Research: JMLR 20 (2019): 177
- Friedman, Jerome H. "Greedy Function Approximation: A Gradient Boosting Machine." The Annals of Statistics 29, no. 5 (2001): 1189–1232.
- Friedman, Jerome. "Stochastic Gradient Boosting." Computational Statistics Data Analysis 38 (February 1, 2002): 367–78. https://doi.org/10.1016/S0167-9473(01)00065-2
- Gabbi, Giampaolo and Andrea Sironi. 2005. Which factors affect corporate bonds pricing? Empirical evidence from eurobonds primary market spreads. The European Journal of Finance 11: 59–74.
- Gabbi, Giampaolo and Pietro Vozzella. 2013. Asset Correlation and Bank Capital Adequacy. European Journal of Finance 19: 55–74.
- Gabbi, Giampaolo, and Pietro Vozzella. 2020. What is good and bad with the regulation supporting the SME's credit access. Journal of Financial Regulation & Compliance Emerald Group Publishing Limited, vol. 28(4), pages 569-586.
- Gabbi, Giampaolo, Massimo Matthias and Michele Giammarino. 2019. Modelling Hard and Soft Facts for SMEs. Some International Evidence. Journal of International Financial Management and Accounting 30: 203–22.
- Gagliardi-Main, D.,P.Muller, E. Glossop, C. Caliandro,M. Fritsch, G. Brtkova, andR.Ramlogan. 2013. Annual Report onEuropean SMEs 2012/2013: A recovery on the Horizon? SME Performance Review.

- Godbillon-Camus, Brigitte, and Christophe J. Godlewski. 2005. Credit risk management in banks: Hard information, soft information and manipulation. Working Paper, University of Strasbourg.
- Grunert, Jens, and Lars Norden. 2012. Bargaining power and information in SME lending. Small Business Economics 39.2: 401-417.
- Grunert, Jens, Lars Norden, and Martin Weber 2005. The role of non-financial factors in internal credit ratings. Journal of Banking & Finance 29.2: 509-531.
- Hasanin, Tawfiq, Taghi M. Khoshgoftaar, Joffrey L. Leevy, and Richard A. Bauder. "Investigating Class Rarity in Big Data." Journal of Big Data 7, no. 1 (December 2020): 23. https://doi.org/10.1186/s40537020-00301-0
- Howorth, Carole, and Andrea Moro. 2012. Trustworthiness and interest rates: an empirical study of Italian SMEs. Small Business Economics 39.1: 161-177.
- Ivashina, V. 2009. "Asymmetric Information Effects on Loan Spreads." Journal of Financial Economics 92 (2): 300–319.
- Khashman, Adnan. "Neural Networks for Credit Risk Evaluation: Investigation of Different Neural Models and Learning Schemes." Expert Systems with Applications 37, no. 9 (September 1, 2010): 6233–39. https://doi.org/10.1016/j.eswa.2010.02.101
- Lehmann, Bina. 2003. Is it worth the while? The relevance of qualitative information in credit rating. The Relevance of Qualitative Information in Credit Rating. Working Paper presented at the EFMA 2003, Helsinki, pp. 1-25
- Liu, Fei Tony, Kai Ming Ting, and Zhi-Hua Zhou. "Isolation Forest." In 2008 Eighth IEEE International Conference on Data Mining, 413–22, 2008. https://doi.org/10.1109/ICDM.2008.17
- Löffler G., A. Maurer, Incorporating the dynamics of leverage into default prediction, J. Bank. Finance, 35 (2011), pp. 3351-3361
- Ma, Xiaojun, Jinglan Sha, Dehua Wang, Yuanbo Yu, Qian Yang, and Xueqi Niu. "Study on a Prediction of P2P Network Loan Default Based on the Machine Learning LightGBM and XGboost Algorithms According to Different High Dimensional Data Cleaning." Electronic Commerce Research and Applications 31 (September 1, 2018): 24–39. https://doi.org/10.1016/j.elerap.2018.08.002
- Morales, Ann, Rene Sacasas, and Paul Munter. 2000. Safe harbor' under the Private Securities Litigation Reform Act of 1995. The CPA Journal 70.8: 66.
- Ounacer, Soumaya, Hicham Ait el Bour, Younes Oubrahim, M. Ghoumari, and Mohamed Azzouazi. "Using Isolation Forest in Anomaly Detection: The Case of Credit Card Transactions." Periodicals of Engineering and Natural Sciences (PEN) 6 (November 24, 2018): 394. https://doi.org/10.21533/pen.v6i2.533
- Petropoulos, Anastasios, Vasilis Siakoulis, Evaggelos Stavroulakis, and A. Klamargias. "A Robust Machine Learning Approach for Credit Risk Analysis of Large Loan Level Datasets Using Deep Learning and Extreme Gradient Boosting." IFC Bulletins Chapters, 2019. https://www.semanticscholar.org/paper/A-robust-machinelearning-approach-for-credit-risk-Petropoulos-Siakoulis/cbae059d97bf674e02d391f939297b31319032ec
- Song, Eunhye, Barry L. Nelson, and Jeremy Staum. "Shapley Effects for Global Sensitivity Analysis: Theory and Computation." SIAM/ASA Journal on Uncertainty Quantification 4, no. 1 (January 2016): 1060–83. https://doi.org/10.1137/15M1048070
- Steenackers, A., and M. J. Goovaerts. "A Credit Scoring Model for Personal Loans." Insurance: Mathematics and Economics 8, no. 1 (March 1, 1989): 31–34. https://doi.org/10.1016/0167-6687(89)90044-9
- West, David. "Neural Network Credit Scoring Models." Computers Operations Research 27, no. 11 (September 1, 2000): 1131–52. https://doi.org/10.1016/S0305-0548(99)00149-5
- Yobas, Mumine B.; Crook, Jonathan N.; Ross, Peter. "Credit Scoring Using Neural and Evolutionary Techniques." IMA Journal of Management Mathematics 11, no. 2 (March 1, 2000): 111–25. https://doi.org/10.1093/imaman/11.2.111
- Yu, Lean, Shouyang Wang, and Kin Keung Lai. "Credit Risk Assessment with a Multistage Neural Network Ensemble Learning Approach." Expert Systems with Applications 34, no. 2 (February 1, 2008): 1434–44. https://doi.org/10.1016/j.eswa.2007.01.009
- Zakrzewska, Danuta. "On Integrating Unsupervised and Supervised Classification for Credit Risk Evaluation." Information Technology and Control 36 (January 1, 2007)
- Zhu, Lin, Dafeng Qiu, Daji Ergu, Cai Ying, and Kuiyi Liu. "A Study on Predicting Loan Default Based on the Random Forest Algorithm." Procedia Computer Science, 7th International Conference on Information Technology and Quantitative Management (ITQM 2019): Information technology and quantitative management based on Artificial Intelligence, 162 (January 1, 2019): 503–13. https://doi.org/10.1016/j.procs.2019.12.017