





ARTICLE



<https://doi.org/10.1057/s41599-024-04047-5>

OPEN

Enhancing financial risk prediction with symbolic classifiers: addressing class imbalance and the accuracy–interpretability trade-off

Luis J. Mena¹, Vicente García ^{2✉}, Vanessa G. Félix^{1,3}, Rodolfo Ostos^{1,3}, Rafael Martínez-Peláez ^{1,4}, Alberto Ochoa-Brust⁵ & Pablo Velarde-Alvarado⁶

Machine learning for financial risk prediction has garnered substantial interest in recent decades. However, the class imbalance problem and the dilemma of accuracy gain by loss interpretability have yet to be widely studied. Symbolic classifiers have emerged as a promising solution for forecasting banking failures and estimating creditworthiness as it addresses class imbalance while maintaining both accuracy and interpretability. This paper aims to evaluate the effectiveness of *REMED*, a symbolic classifier, in the context of financial risk management, and focuses on its ability to handle class imbalance and provide interpretable decision rules. Through empirical analysis of a real-world imbalanced financial dataset from the Federal Deposit Insurance Corporation, we demonstrate that *REMED* effectively handles class imbalance, improving performance accuracy metrics while ensuring interpretability through a concise and easily understandable rule system. A comparative analysis is conducted against two well-known rule-generating approaches, *J48* and *JRip*. The findings suggest that, with further development and validation, *REMED* can be implemented as a competitive approach to improve predictive accuracy on imbalanced financial datasets without compromising model interpretability.

¹Academic Unit of Information Technology Engineering, Universidad Politecnica de Sinaloa, Mazatlan, Mexico. ²Department of Electrical and Computer Engineering, Universidad Autonoma de Ciudad Juarez, Ciudad Juarez, Mexico. ³Department of Economic and Administrative Sciences, Software Engineering Career, Universidad Autonoma de Occidente, Mazatlan, Mexico. ⁴Department of Systems and Computer Engineering, Universidad Catolica del Norte, Antofagasta, Chile. ⁵Faculty of Mechanical and Electrical Engineering, Universidad de Colima, Colima, Mexico. ⁶Academic Unit of Basic Sciences and Engineering, Universidad Autonoma de Nayarit, Tepic, Mexico. ✉email: vicente.jimenez@uacj.mx

Introduction

The subprime crisis of 2008 triggered the most profound post-war recession, catching most policymakers and the financial community by surprise and bringing commercial banks' financing and investment decisions into sharp focus (Cukierman 2019; Zubair et al. 2020). One of its primary causes was the flawed mortgage model that incentivised laxity that was used by financial entities in granting loans. These institutions, aware that they would not retain the mortgages, then sold them to other entities, which led to a lack of diligence in verifying applicant information (Hubbard and Navarro 2010). This resulted in the emergence of subprime mortgages, which were designed as high-risk loans that extended to borrowers who were unable to demonstrate the necessary credit scores or monthly income level corresponding to the requested credit amount (Jones and Sirmans 2019).

Consequently, financial institutions in the past two decades have prioritised implementing more effective decision-making methods for risk assessment with the goal of improving the accuracy of forecasting business failures and estimating creditworthiness. Thus, machine learning (ML) approaches have emerged as crucial tools for the financial sector (Leo et al. 2019; Quan and Sun 2024), such as by addressing the need to build automated bank fragility prediction (Shang et al. 2024) and credit scoring models from large datasets. These models aim to accurately classify cases into "good" or "bad" based on solvency ratios or estimated payment capacities (Bücker et al. 2022; Chen et al. 2016; Hussin-Adam-Khatir, Bee (2022); Nalić and Martinovic 2020).

Nevertheless, in practical applications of ML classification for financial risk management, an additional challenge called the class imbalance problem must be addressed (Niu et al. 2020; Shen et al. 2019) since most financial datasets exhibit a vastly greater number of solvent examples (majority class) than insolvent examples (minority class), resulting in financial-related datasets that are often strongly imbalanced (Hussin-Adam-Khatir, Bee (2022); Kennedy et al. 2010). Therefore, classification results tend to be skewed due to a bias towards the majority class (Shen et al. 2019), leading to poor performance of classifiers in identifying examples of the minority class (Niu et al. 2020; Wang et al. 2015). Notably, misclassifying an insolvent case as a solvent incurs a higher cost in risk management than missing out on an opportunity (Hussin-Adam-Khatir, Bee (2022); Shen et al. 2019).

On the other hand, ML approaches for financial assessment must maintain a balance between accuracy and interpretability (Florez-Lopez and Ramon-Jeronimo 2015; Hayashi 2016). Interpretability refers to a model's ability to provide information in a human-comprehensible form. This aspect holds significance because of both commercial and legal considerations, where financial managers need to understand the information received to combine it with their expert judgement for more accurate financial risk evaluation. Additionally, it is crucial for scenarios such as explaining to an applicant why their credit request was rejected (Bücker et al. 2022; Florez-Lopez and Ramon-Jeronimo 2015; Hayashi 2016). Similarly, in adherence to banking regulations and audit requirements in many countries, financial institutions are required to justify their decisions regarding accepting or denying finance requests (Bücker et al. 2022; Florez-Lopez and Ramon-Jeronimo 2015; Hayashi 2016; Tomczak and Zięba 2015).

In this sense, symbolic algorithms based on decision trees (DTs) and rule systems (Apté and Weiss 1997) in the final form of IF-THEN statements are the most commonly used methods for building expressive and human-readable representations of knowledge (Wu and Hsu 2012). Unlike neural networks (NNs) and support vector machines, which do not provide insight into how to generate their predictions (i.e., they are black box

methods) (Lantz 2013), rule solutions can be adequately incorporated for decision-making processes requiring the utmost clarity (Wu and Hsu 2012) and direct applicability in contexts such as financial risk management (Florez-Lopez and Ramon-Jeronimo 2015). However, the interpretability and conciseness of extracted rules pose a critical compromise (Hayashi 2016); a large set of rules or a higher average number of antecedents per rule results in more complex and less concise rules, diminishing interpretability and the ability to generalise from observed data to unseen data (a phenomenon known as *overfitting*) (Hayashi and Oishi 2018; Ying 2019). Hence, the simplification of extracted rules becomes crucial for enhancing interpretability in the decision-making process (Cano et al. 2011; Gacto et al. 2011; Lanzarini et al. 2017; Hayashi 2016), reducing the effort needed to understand their meaning (Gacto et al. 2011).

From this perspective, numerous research efforts have been implemented to assess financial risk using strategies to address the adverse effect of class imbalance on the predictive power of ML approaches; however, overall performance considering the trade-off between accuracy and interpretability has not been sufficiently addressed (Chen et al. 2024). Therefore, we present an overview of recent ML studies that have addressed the research gap in dealing with class imbalance problems in financial data and integrate rule solutions to improve the interpretability of the models.

Literature review

Many studies have developed ensemble methods by training multiple models and combining their predictions to improve performance in classifying financial risks from imbalanced datasets (He et al. 2018; Xia et al. 2020; Zhang et al. 2018). For example, Florez-Lopez and Ramon-Jeronimo (2015) proposed an ensemble model based on DTs as base learners, creating a correlated-adjusted decision forest (CADF) univariate to yield an accurate and comprehensible classification model for credit risk evaluation. The ensemble strategy involved merging four individual DT models from a single dataset. Feature and instance diversity were included via different wrapper-feature selection processes for each inductive model. A 10-fold *cross-validation* (where the dataset is randomly split into 10 mutually exclusive equal subsets for 10 training and testing sessions) was employed for DT construction. Additionally, *bootstrapping* (by sampling n instances uniformly from the data with replacement for training and using the remaining instances for testing) was implemented for out-of-sample validation. A penalty function was also introduced to generate adjusted-weighted votes using a mixed accuracy-correlation ranking scheme. CADF univariate was tested on the German credit risk dataset from the UCI repository. Comparative evaluations of predictive accuracy and interpretability against each DT classifier used to build the ensemble model, and other decision forest strategies revealed that CADF univariate outperforms any single classifier in terms of out-of-sample accuracy and emerged as the most interpretable among complex decision forest models.

Similarly, Hayashi et al. (2016) proposed another ensemble approach to increase the conciseness of extracted rules for automated financial risk models. They employed a recursive ML algorithm to extract classification rules (*Re-RX*) from a feedforward NN (Setiono et al. 2008), replacing the *C4.5* DT classifier (Quinlan 1993) with the unique variant *J48graft* from the *WEKA* workbench (Panigrahi and Borah 2018). Experiments were conducted in six two-class financial datasets, which considered discrete variables before analysing continuous data. *Re-RX* with *J48graft* resulted in a smaller and more accurate set of extracted

rules than the *Re-RX* algorithm for all the databases. Later, Hayashi extended his ensemble strategy by including the selection of continuous attributes (*Continuous Re-RX*) and sampling selection techniques (*Sampling Re-RX*) to achieve higher accuracy and interpretability (Hayashi 2016). The effectiveness and appropriateness of the *Re-RX* family algorithm were assessed in four real-life, two-class mixed (discrete and continuous attributes) financial datasets. The findings suggest that *Continuous Re-RX*, *Re-RX* with *J48graft*, and *Sampling Re-RX* comprised powerful management tools for creating accurate, concise, and interpretable decision support systems for financial risk analysis.

As an extension, Hayashi and Oishi (2018) proposed a straightforward two-stage sequential ensemble classifier to achieve a well-balanced rule extraction method that prioritises high accuracy while generating a concise number of rules. This approach employs a *backpropagation* NN alongside *Continuous Re-RX* with *J48graft* via recursive feedback. Experiments were performed on two mixed financial datasets, demonstrating that the proposed ensemble method represented the best trade-off solution, that offers both accuracy and interpretability.

Lanzarini et al. (2017) employed a hybrid ML approach based on a competitive learning vector quantisation (*LVQ*) NN (Kohonen et al. 2001) with particle swarm optimisation (*PSO*) (Wang et al. 2007) to establish classification rules. The method's performance was tested using a real consumer credit dataset against the *C4.5* and *PART* (Witten et al. 2011) classifier algorithms. Across ten independent runs, the *LVQ* network of 30 neurons + *PSO* showed a slightly lower accuracy than benchmark models but with a significantly lower average number of rules and antecedents.

Conversely, Wu and Hsu (2012) proposed an enhanced decision support model (*EDSM*) incorporating a relevance vector machine (*RVM*) (Tipping 2000) and *DT* for analysing financial rating domains. Their ensemble strategy employed a decompositional approach that deconstructed the *RVM* structure to obtain relevance vectors and predicted labels based on the best *cross-validation* result. These outputs were then fed into a rule-based classifier with explanatory capabilities, enabling it to understand and leverage the insights from the *RVM*. According to the results of the final rankings, the *EDSM* outperformed six other classifiers in forecasting solvent rating status using real data from the Taiwan financial system.

Similarly, another study used a Taiwanese customer dataset and proposed the *SPR-RIPPER* hybrid model (Xu et al. 2018). This model combines the *RELIEF* method (Kira and Rendell, 1992) to remove redundant features, enhancing model interpretability. Additionally, it utilises the synthetic minority class oversampling technique (*SMOTE*) (Chawla et al. 2002) to address the class imbalance by resampling minority classes with an increase of 100% in the training dataset and the *RIPPER* algorithm (Cohen 1995) for rule extraction. *SPR-RIPPER* exhibited higher precision, lower time complexity, and fewer rules extracted than base symbolic classifiers such as *C4.5* and *RIPPER*.

In the context of financial distress prediction, Kristóf and Virág (2022) evaluated ML approaches for reliably predicting the failure risk of the central banks of the 27 countries in the European Union. They applied *logit*, *C5.0 DT (Boosting C4.5)* (Quinlan 1996), and a deep learning NN method to 32,287 bank-year observations. The *C5.0* generated 100 different rule sets from 100 boosted *DTs* with depth values between 4 and 12. Notably, the ensemble *DT* method outperformed the other ML techniques in terms of predictive power.

Table 1 summarises the findings of the studies mentioned above regarding accuracy, interpretability, and dataset characteristics used for evaluating the performance of the proposed approaches. Additionally, we computed a *complexity* value (Eq. 1)

for a rule-based classifier introduced by Nauck (2002), who proposed an interpretability measure that associates the number of classes and antecedents per rule,

$$\text{complexity} = \frac{m}{\sum_{i=1}^r c_i} \quad (1)$$

where m is the number of classes, r is the number of rules, and c_i is the number of antecedents used in the i -th rule. Therefore, a higher value indicates a less complex and more comprehensible rule system since the classifier contains fewer rules and antecedents (Cano et al. 2011).

The experimental results of Chen et al. (2024) showed that class imbalance has a negative effect on the interpretation performance of ML approaches, which is consistent with our review, since for datasets with higher imbalance rates (EU-27 banks and Taiwan; see Table 1), the compressibility of models (*complexity* value) also decreased.

Therefore, we propose a novel approach to rule extraction for financial risk management that addresses the class imbalance problem with a high imbalance ratio (> 10) while considering the trade-off between accuracy and interpretability.

Methods

Financial dataset. In this study, we used the US database provided by the Federal Deposit Insurance Corporation that is publicly available online (Serrano-Cinca, Gutiérrez-Nieto (2013)). This strongly imbalanced dataset (Marqués et al. 2013) consists of financial accounting statements from 8292 banks segmented into 319 insolvent and 7973 solvent cases, resulting in an imbalance ratio of 24.99. The dataset considered 17 financial ratios as independent variables, aiming to cover the key indicators for diagnosing banking financial health. Table 2 presents each ratio along with its definition.

REMED. *REMED* is a symbolic *one-class* approach for binary classification introduced by Mena and Gonzalez (2009). The algorithm is strategically designed to address two key aspects: 1) addressing the class imbalance problem by constructing biased models aimed at recognising a target class by training both classes and 2) producing interpretable and concise rule-based systems that comprise only one rule with m antecedents for predicting the target class and another default rule without antecedents for predicting the default class. To achieve both aims, the *REMED* learning process unfolds in three main stages: 1) selection of antecedents, 2) selection of initial partitions, and 3) building the rule system.

Selection of antecedents. In the first stage (Algorithm 1), *REMED* estimates the probability p of an independent continuous variable X associated with the target class using simple *logistic regression* (Hosmer and Lemeshow 1989). Since the outcome is a probability, the dependent variable is bounded between 0 and 1. In *logistic regression*, a *logit* transformation (Armitage et al. 2008) is applied to the *odds ratio*, i.e., the probability of success divided by the probability of failure. The logistic function (Eq. 2) is represented by:

$$p = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X)}} \quad (2)$$

where coefficients β_0 and β_1 are estimated for each variable X using the *maximum likelihood function* (Aldrich 1997). Thus, an *odds ratio* of 1 indicates non-association, a *ratio* greater than 1 indicates a positive association (where an increase in X leads to an increase in p), and a *ratio* less than 1 indicates a negative association (where a decrease in X leads to an increase in p). X is considered a rule antecedent only if a statistically significant association is at a confidence level $> 99\%$. Depending on the

Table 1 Results and datasets of previous studies on ML for financial risk management.

Study		Accuracy		Interpretability			Financial datasets			
Authors, year	Approaches	Datasets	Performance metrics	Rules average	Antecedents average	Complexity	Name	Solvent cases	Insolvent cases	Imbalance ratio
Florez-Lopez and Ramon-Jeronimo (2015)	CADF univariate	German	AR = 0.758 AUC = 0.789 Type I error = 0.528 Type II error = 0.117	10.25	2.25	0.087	Australian	307	383	1.25
Hayashi et al. (2016)	Re-RX with J48graft	Bene1 Bene2 CARD1 CARD2 CARD3	AUC = 0.742 AUC = 0.755 AUC = 0.890 AUC = 0.876 AUC = 0.894	31.5 39.9 6.30 7.20 5.70	9.17 7.15 3.15 3.68 2.73	0.007 0.007 0.101 0.075 0.129	CARD1	307	383	1.25
Hayashi (2016)	Sampling Re-RX Re-RX with J48graft Continuous Re-RX Sampling Re-RX Re-RX with J48graft Continuous Re-RX Sampling Re-RX Re-RX with J48graft Continuous Re-RX Sampling Re-RX Re-RX with J48graft Continuous Re-RX	Australian Bene1 Bene2 German	AR = 0.805 AR = 0.865 AR = 0.864 AR = 0.860 AR = 0.862 AR = 0.869 AR = 0.869 AR = 0.721 AUC = 0.680 AR = 0.710 AUC = 0.669 AR = 0.725 AUC = 0.702 AR = 0.726 AUC = 0.600 AR = 0.707 AUC = 0.615 AR = 0.747 AUC = 0.648 AR = 0.732 AUC = 0.660 AR = 0.728 AUC = 0.650 AR = 0.752 AUC = 0.692	14.40 11.04 4.58 2.38 14.00 5.95 25.46 6.25 27.74 7.38 48.40 7.52 28.31 6.10 27.61 6.46 75.90 7.95 19.34 6.20 16.65 6.19 39.6 9.13	5.36 5.27 2.38 0.183 5.95 0.024 6.25 0.013 7.38 0.010 7.52 0.005 6.10 0.012 6.46 0.011 7.95 0.003 6.20 0.017 6.19 0.019 9.13 0.006	0.026 0.034 0.183 0.024 0.013 0.010 0.005 0.012 0.011 0.003 0.017 0.019 0.006	CARD2 CARD3 Bene1 Bene2	307 307 2083 1040 5033 2157	383 383 1040 1040 2157 2157	1.25 1.25 1.25 2.33
Hayashi et al. (2018)	Continuous Re-RX with J48graft	Australian German	AR = 0.884 AUC = 0.880 AR = 0.790 AUC = 0.757	15.4 5.66 44.9 7.68	5.66 0.023 7.68 0.006	0.023 0.006	German	700	300	2.33
Lanzarini et al. (2017)	LVQ + PSO	Ecuador	AR = 0.792 Type I error = 0.140 AR = 0.912 Sens = 0.886 Spec = 0.877 AUC = 0.707 F1-score = 0.534 AUC = 0.937	3.12	2.54	0.252	Ecuador	643	1604	2.49
Wu and Su (2012)	EDSM	Taiwan	AR = 0.912 Sens = 0.886 Spec = 0.877 AUC = 0.707 F1-score = 0.534 AUC = 0.937	8	3.25	0.077	Taiwan	23364	6636	3.52
Xu et al.(2018)	SPR-RIPPER	Taiwan	AUC = 0.707 F1-score = 0.534 AUC = 0.937	31	3	0.022				
Kristóf and Virág (2022)	CS.0	EU-27 banks	AUC = 0.937	100	8	0.003	EU-27 banks	32287	303	106.55

AR Accuracy rate, AUC Area under the receiver operation characteristic curve, Sens Sensitivity, Spec Specificity.

previously established association (positive or negative), it is possible to determine the relational operator (\geq or \leq) used to partition X within the feature space.

Algorithm 1. Selection of antecedents.

Selection of Antecedents (dataset, variables)

```

antecedents ← ∅
confidence_level ← 1-α // > 99%
ε ← 1/10k // convergence level
FOR x ∈ variables DO
    X [...] ← dataset[x] // instances for each continuous variable
    p, odds_ratio ← Logistic Regression (X [...], ε)
    IF p < (1 - confidence_level) THEN
        antecedents ← x, odds_ratio
    END-IF
END-FOR
    
```

Selection of initial partitions. Symbolic classifier learning is based on a DT scheme that divides the feature space from the top (root) to the bottom (leaf) until each instance is assigned to a unique class. Consequently, partitions are a set of exhaustive and excluding conditions for building a symbolic rule, which exhaustively classifies all instances by assigning them only one class (excluding). In this phase, REMED begins by sorting the values of an antecedent X in ascending order, computing its mean value \bar{X} , and then moving towards either the start (indicating negative association) or the end (indicating positive association) of $X[\dots]$, based on the odds ratio, to find the closest value to \bar{X} that belongs to the target class. Subsequently, REMED computes the average between the selected X value and its predecessor (in the case of positive association) or successor (in the case of negative association). This new estimation is performed only once for each antecedent because other displacements to calculate a new partition could include at least one instance of the target

Table 2 Financial ratios employed for banking distress analysis. Acronyms and definitions are taken from the Federal Deposit Insurance Corporation.

Variable	Definition
INTINCY	Yield on earning assets. Total interest income as a percent of average earning assets.
INTEXPY	Cost of funding earning assets. Annualised total interest expense on deposits and other borrowed money as a percent of average earning assets on a consolidated basis.
NIMY	Net interest margin. Total interest income less total interest expense as a percent of average earning assets.
NONIY	Noninterest income to earning assets. Income derived from bank services and sources other than interest bearing assets as a percent of average earning assets.
NONIXY	Noninterest expense to earning assets. Salaries and employee benefits, expenses of premises and fixed assets, and other noninterest expenses as a percent of average earning assets.
NOIY	Net operating income as a percent of average assets.
ROA	Return on assets (ROA). Net income after taxes and extraordinary items as a percent of average total assets.
ROAPTX	Pretax return on assets. Annualised pre-tax net income as a percent of average assets.
ROE	Return on Equity (ROE). Annualised net income as a percent of average equity on a consolidated basis.
ROEINJR	Retained earnings to average equity. Net income, less cash dividends declared, as a percent of average total equity capital.
EEFFR	Efficiency ratio. Noninterest expense, less the amortisation expense of intangible assets, as a percent of the sum of net interest income and noninterest income.
NPERFV	Noncurrent assets plus other real estate owned to assets. Noncurrent assets are defined as assets that are past due 90 days or more plus assets placed in nonaccrual status plus other real estate owned (excluding direct and indirect investments in real estate).
LNLSDEPR	Net loans and leases to deposits. Loans and lease financing receivables net of unearned income, allowances and reserves as a percent of total deposits.
EQV	Total equity capital as a percent of total assets.
RBC1AAJ	Core capital (leverage) ratio. Tier 1 (core) capital as a percent of average total assets minus ineligible intangibles.
RBC1RWAJ	Tier 1 risk-based capital ratio. Tier 1 (core) capital as a percent of risk-weighted assets as defined by the appropriate federal regulator for prompt corrective action during that time period.
RBCRWAJ	Total risk-based capital ratio. Total risk based capital as a percent of risk-weighted assets.

class instance on the opposite side of the classification threshold, thereby reducing the probability of belonging to the aim class. Algorithm 2 provides a detailed step-by-step description.

Algorithm 2. Selection of initial partitions.

Selection of Initial Partitions (*dataset, antecedents*)
partitions $\leftarrow \emptyset$
FOR *x*, *odds_ratio* \in *antecedents* **DO**
X [...] \leftarrow **Sort Ascending** (*dataset*[*x*]) // instances sorted ascending for each antecedent *x*
part \leftarrow **Average** (*X* [...]) // mean of the values of *X*
pointer \leftarrow **Position** (*X* [...], *part*) // searching *X* value closest to \bar{X}
k \leftarrow *pointer*
WHILE *X* [*k*].*class* \neq target class // searching the closest value to \bar{X} belonging to the target class
IF *odds_ratio* $>$ 1 **THEN**
k \leftarrow *k* + 1 // positive association
ELSE
k \leftarrow *k* - 1 // negative association
END-IF
END-WHILE
IF *pointer* \neq *k* **THEN**
IF *odds_ratio* $>$ 1 **then**
part \leftarrow (*X* [*k*].*value* + *X* [*k*-1].*value*) / 2 // positive association
ELSE
part \leftarrow (*X* [*k*].*value* + *X* [*k*+1].*value*) / 2 // negative association
END-IF
END-IF
partitions \leftarrow^{\cup} *part* // set of initial partitions
END-FOR

Building the rule system. After generating the initial partitions for each of the *m* antecedents, REMED builds a straightforward rule system with *m* conditions as follows:

IF *I* <relation> *part*₁ AND *i* <relation> *part*_{*i*} AND ... AND *m* <relation> *part*_{*m*} then \Rightarrow target class

ELSE \Rightarrow default class where <relation> is either \geq or \leq , depending on whether antecedent *i* is positively or negatively associated with the target class using *part*_{*i*} (the classification threshold).

In the final stage (Algorithm 3), REMED classifies instances using the initial system of rules. Subsequently, it aims to improve predictive performance by adjusting classification thresholds using the *bisection* method (Burden and Faires 2000). Initially, REMED defines the first searching interval for potential new partitions based on partition *i* and the maximum or minimum value (depending on association) for instances of antecedent *i*. The algorithm then builds a temporary rule system by modifying the current partition *i* with the new partition value, classifying the instances again, and retaining the new partition only if it decreases the number of incorrectly classified default class examples without decreasing the number of correctly classified target class examples. This step is repeated for each antecedent until the established convergence level is reached or when the current rule system no longer reduces the number of incorrectly classified default class instances. A detailed outline divided into instruction blocks (A-E) is as follows:

(A) REMED begins by constructing an initial rule system based on the current partitions, classifying instances, and then saving the results. REMED also saves the number of correctly classified (*k1*) and predicted (*k2*) target class instances.

(B) Later, REMED begins an iterative process (1... *m*) to improve the predictive value of each partition. It estimates a new partition for antecedent *i* by averaging its initial classification threshold with the maximum or minimum value of the examples for this antecedent (based on association). REMED saves a copy in the *copy_partitions* vector to evaluate the classification performance of the new partition.

(C) REMED creates a temporary rule system by changing the current partition of antecedent *i* with the new partition and

classifies examples again. *REMED* saves the number of correctly classified (k_3) and predicted (k_4) target class instances.

(D) *REMED* compares the results obtained with the new classifier. If the number of correctly classified target class examples decreases ($k_3 < k_1$), then *REMED* sets the current partition as the maximum interval value to estimate a new partition; otherwise, if the number of incorrectly classified default class examples decreases ($k_4 < k_2$), then *REMED* saves the number of predicted target class examples (k_5) and sets the current partition as the minimum interval value to estimate a new partition. *REMED* continues estimating new partitions for

antecedent i using the *bisection* method until the difference in absolute value between the maximum and minimum interval values does not overcome the established convergence level or until the current rule system no longer decreases the number of incorrectly classified default class instances.

(E) If the new partition for antecedent i improves the predictive value, it is included in the final set of partitions, and the number of predicted target class instances is updated ($k_2 = k_5$). This process is repeated for all m antecedents.

Algorithm 3. Building the rule system.

Algorithm 3. Building the rule system.

Building the Rule System (*dataset, antecedents, partitions*)

class[$\%i$] \leftarrow **Rule** (*dataset, partitions, odds_ratio*)

target_instances[$\%i$] \leftarrow **Calculate Target Instances** (*dataset.class, class* [$\%i$])

A

$k_1 \leftarrow$ **Add** (*target_class*[$\%i$]) // number of correctly classified target class instances

$k_2 \leftarrow$ **Add** (*class*[$\%i$]) // number of predicted target class instances

$\epsilon \leftarrow 1/10k$ // convergence level

FOR $i \leftarrow 1 \text{ } \% m$ **DO**

$min \leftarrow$ *partition*[i]

IF *odds_ratio*[i] > 1 **THEN**

$max \leftarrow$ **Maximum** (*dataset*[i]) // positive association

B

ELSE

$max \leftarrow$ **Minimum** (*dataset*[i]) // negative association

END-IF

$new_partition \leftarrow (min + max) / 2$

$copy_partitions[\%i] \leftarrow$ *partitions*[$\%i$]

WHILE **Abs** ($max - min$) $> \epsilon$ **DO**

$copy_partitions [i] \leftarrow new_partition$

$class [\%i] \leftarrow$ **Rule** (*dataset, copy_partitions, odds_ratio*)

$target_instances[\%i] \leftarrow$ **Calculate Target Instances** (*dataset.class, class* [$\%i$])

C

$k_3 \leftarrow$ **Add** (*target_class*[$\%i$]) // number of correctly classified target class instances

$k_4 \leftarrow$ **Add** (*class*[$\%i$]) // number of predicted target class instances

IF $k_3 < k_1$ **THEN** // least number of correctly classified target class instances

$max \leftarrow new_partition$

ELSE

IF $k_4 < k_2$ **THEN** // least number of incorrectly classified default class instances

$k_5 \leftarrow k_4$

$min \leftarrow new_partition$

D

ELSE

EXIT-WHILE

END-IF

END-IF

$new_partition \leftarrow (min + max) / 2$

END-WHILE

IF $min \neq$ *partitions*[i] **THEN**

$k_2 \leftarrow k_5$ // updating number of predicted target class instances

$partitions[\%i] \leftarrow min$ // updating set of partitions

E

END-IF

END-FOR

Therefore, the main goal of *REMED* is to maximise the classification performance of the target class. This begins with the selection of antecedents strongly associated with the aim class (*logistic regression* confidence level > 99%), which is the unique parameter needed; later, the search for partition thresholds upon encountering the first target class example (to prevent a decrease in the probability of belonging to the aim class) stops, and finally, the predictive performance of the rule system tries to make improvements without compromising the number of correctly classified target class examples.

J48. *J48* is a *Java*-implemented version of *C4.5* (Quinlan 1993), available through the *WEKA* workbench. *C4.5* is a widely recognised symbolic classifier used for financial risk management problems (Martens et al., (2007); Cano et al. 2011; Brown and Mues 2012; Florez-Lopez and Ramon-Jeronimo 2015; Tomczak and Zięba 2015; Hayashi 2016; Hayashi et al. 2016; Lanzarini et al. 2017; Hayashi and Oishi 2018; Xu et al. 2018). *C4.5* is a discrimination-based approach that can handle multi-class problems, generating a DT with class membership predictions for all the instances. The tree-building process employs a partition selection criterion known as information gain, an *entropy*-based metric that measures the purity between a partition and its sub-partitions. Employing a recursive procedure, *C4.5* selects antecedents that lead to purer child nodes (where a completely pure node includes examples belonging solely to one class) at each step. The *gain ratio* assists in identifying the best target antecedent. After the DT is built, *C4.5* applies a *pruning* strategy to prevent *overfitting* (Bramer (2002)).

JRip. *JRip* is a *Java*-based version of the *RIPPER* algorithm (Cohen 1995), also provided by the *WEKA* workbench. *RIPPER* is another symbolic classifier widely used in the early detection of financial risk that induces sets of classification rules (Cano et al. 2011; Sánchez-Garreta et al., (2012); Tomczak and Zięba 2015; Berka 2016; Obermann and Waack 2016; Xu et al. 2018; Otieno et al. 2020). Although *RIPPER* can handle multi-class problems, its learning process for binary classification tasks is particularly interesting. *RIPPER* employs a divide-and-conquer approach to iteratively build rules to cover previously uncovered training examples (generally target class examples) in growing and *pruning* sets. Rules are grown by adding conditions until each rule encompasses only a single example in the growing set, often from the default class. Thus, *RIPPER* usually generates rules starting from the target class and extending to the default class, offering an efficient method of learning rules specifically for the target class.

Performance assessment. To assess the accuracy performance of the rule-based systems generated by each symbolic classifier, we constructed a confusion matrix containing the predicted and actual values for binary classification (Table 3), where *TP* represents the number of true positives (instances correctly predicted as good), *FP* represents false positives (bad predicted as good), *FN* represents false negatives (good predicted as bad), and *TN* represents true negatives (bad predicted as bad).

Table 3 Confusion matrix for binary classification.

		Actual	
		Positive	Negative
Predicted	Positive	True Positive (TP)	False Positive (FP)
	Negative	False Negative (FN)	True Negative (TN)

Most studies dealing with imbalanced data typically denote the target class (minority) as positive and the default class (majority) as negative. However, in the financial context, labelling bad cases as positive and good cases as negative may seem unreasonable (Tomczak and Zięba 2015). Consequently, $y = 1$ usually denotes a good or solvent case, and $y = 0$ denotes a bad or insolvent case.

Another critical issue when dealing with class imbalance is how to satisfactorily assess classifier performance since using the typical accuracy rate (*AR*) (Eq. 3) can lead to misleading conclusions, as it measures only the overall percentage of correctly classified instances but does not evaluate the predictive performance for each class:

$$AR = \frac{TP + TN}{TP + FP + FN + TN} \tag{3}$$

Therefore, it is also necessary to determine the appropriate way to evaluate classifiers on datasets with uneven distributions; thus, we also used *Precision* measure (Eq. 4) to assess the predictive value for identifying solvent cases:

$$Precision = \frac{TP}{TP + FP} \tag{4}$$

Moreover, the *geometric mean* (*Gmean*) (Eq. 5) serves as a performance criterion to mitigate the impact of unbalanced data (Tomczak and Zięba 2015). It is considered a balancing measure between the correct classification of the positive and negative classes when evaluated separately (Tomczak and Zięba 2015):

$$Gmean = \sqrt{\frac{TP}{TP + FN} \times \frac{TN}{TN + FP}} \tag{5}$$

Another weakness of *AR* is that it ignores the misclassification cost (bad classified as good, or good classified as bad), which is an important issue in financial risk assessment. In this context, a binary classifier can incur two types of errors: *Type I error* (*FP*) when an insolvent case is classified as a solvent and *Type II error* (*FN*) when a solvent case is classified as an insolvent. Since the former can imply a loss of capital and the latter can imply only a missing business opportunity, the cost of *Type I error* is typically considered higher for financial institutions. Therefore, we also included the *Type I error* or the *FP rate* (Eq. 6) as another metric to assess the accuracy performance of the symbolic classifiers.

$$Type\ Error = \frac{FP}{FP + TN} \tag{6}$$

With respect to interpretability performance, we used criteria aimed at increasing the simplicity of the rule system. Following the principle of Occam’s razor, the best model constructs solutions with the smallest possible set of elements (Gacto et al. 2011). Thus, we computed *complexity* values (Eq. 1) as an interpretability measure related to the number of rules and antecedents per rule yielded by each symbolic classifier.

To prevent potential *overfitting*, we conducted a 10-fold *cross-validation* repeated five times with different seeds for all the classifiers. For the statistical comparison of the ML models, we used the nonparametric Friedman test and the Nemenyi post hoc test to determine significant differences among multiple classifiers and the Bonferroni–Dunn test to compare the performance of alternative approaches (*J48* and *JRip*) with that of *REMED* as a control classifier (Demšar 2006). We ranked the *k* symbolic classifiers according to the best-performing algorithm in terms of the *N* metrics used for assessing both accuracy and interpretability. A trade-off rank average was computed for each classifier, and a significance level of $p < 0.05$ was set for significant differences among the performances of all symbolic classifiers and pairwise comparisons with *REMED*.

Results and Discussion

The experimental results are summarised in Tables 4–7 and Figs. 1 and 2. Table 4 presents the confusion matrix for each classifier, allowing the calculation of their respective performance metrics. Additionally, Table 5 shows the rule systems yielded by *REMED* and *JRip* for estimating interpretability measures. Due to the large size of the DT generated by *J48*, only the number of antecedents per rule was indicated. *REMED*, as anticipated, only selected attributes significantly associated with the target class ($p < 0.0001$) and with the highest confidence level ($> 99.99\%$); consequently, this allowed the generation of a more concise set of rules utilising fewer antecedents (12 in total) for determining a

client’s insolvency and subsequently earned the best performance in terms of *complexity* indicator, unlike *JRip* and *J48*, which yielded multiple rules, particularly *J48* with an excessive number. Furthermore, when *REMED* addresses the issue of imbalanced classes, the rule system tends to identify most instances in the insolvent class.

Figures 1 and 2 compare the classifiers in terms of accuracy and interpretability, highlighting the superior performance of each approach concerning specific measures. A logarithmic scale was used to present the results more compactly for interpretability metrics. Notably, *REMED* is the best classifier in terms of *Gmean*, *Precision* and *Type I error*, which are three widely recognised metrics for assessing performance in class imbalance issues. An exception arises in *AR*, where *REMED* displays the lowest indicator among the ML models. Nonetheless, this behaviour is common, and the literature reports that solution proposals with superior performance in unbalanced classes may exhibit a decline in *AR* rates (Chao et al. 2022).

Table 6 shows the results of ranking the classifiers for each metric separately. Rank 1 represents the best-performing approach, rank 2 represents the second-best approach, and rank 3 represents the worst-performing approach. Higher values indicate better performance for accuracy metrics, except for *Type I error* (insolvent classified as solvent). In the case of interpretability measures, the best values are the lowest number of rules and average antecedents, along with the highest *complexity* indicator. The average rank was calculated by considering the trade-off between accuracy and interpretability.

Table 7 reports the results of the statistical tests used to detect significant differences among the average ranks of classifiers. The nonparametric Friedman test was applied to rank the $k = 3$ classifiers over the $N = 7$ metrics, yielding a statistic = 4.492 distributed according to the F distribution with degrees of freedom $k - 1 = 2$ and $(k - 1)(N - 1) = 12$. The critical value of $F(2,12)$ for a significance level of $p = 0.05$ was 3.885. Therefore, we rejected the null-hypothesis that all the classifiers were equivalent.

Later, we proceed with the Nemenyi post-hoc test to identify significant differences between the best-performing classifier (*REMED*) and the worst (*J48*). This was determined by examining their average ranks, which differed by at least a critical difference (*CD*), given by:

$$CD = q_{\alpha} \sqrt{\frac{k(k+1)}{6N}}$$

where the critical value of q_{α} , which is based on the studentised range statistics (Demšar 2006), at $\alpha = 0.05$ with $k = 3$ classifiers, was 2.343. Hence, the corresponding *CD* was computed as $2.343 \sqrt{\frac{3 \cdot 4}{6 \cdot 12}} = 1.252$; therefore, as the difference between the average ranks ($2.714 - 1.429 = 1.285$) was greater than the *CD* (1.252), we conclude that the post-hoc test had sufficient power to detect significant differences ($p < 0.05$) among the classifiers. Furthermore, upon comparing them, we identified two additional groups of algorithms: 1) *J48* shows no significant difference from *JRip*, and 2) *JRip* shows no significant difference from *REMED*.

Finally, to assess the significant performance of *REMED* as a control approach versus the other symbolic classifiers, we calculated the *CD* for the Bonferroni–Dunn test using the same equation as for the Nemenyi test but assigned a critical value for $\alpha/(k-1)$ (Demšar 2006). We find that for $q_{0.05} = 2.241$ and $k = 3$ classifiers, the corresponding *CD* was $2.241 \sqrt{\frac{3 \cdot 4}{6 \cdot 12}} = 1.198$, which indicated that *REMED* performed significantly better than *J48* ($2.714 - 0.429 = 1.285 > 1.198$). However, there were no significant differences with respect to *JRip* ($1.857 - 1.429 = 0.428 < 1.198$).

Table 4 Confusion matrix of the classifiers.			
<i>REMED</i>		Actual	
		Positive	Negative
Predicted	Positive	7652	131
	Negative	321	188
<i>J48</i>		Actual	
		Positive	Negative
Predicted	Positive	7878	188
	Negative	95	131
<i>JRip</i>		Actual	
		Positive	Negative
Predicted	Positive	7858	172
	Negative	115	147

Performance Metrics

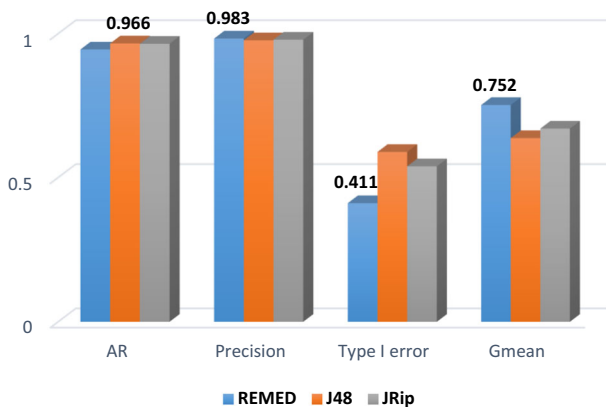


Fig. 1 Bar plot comparing the classification performance of the classifiers.

Interpretability Metrics

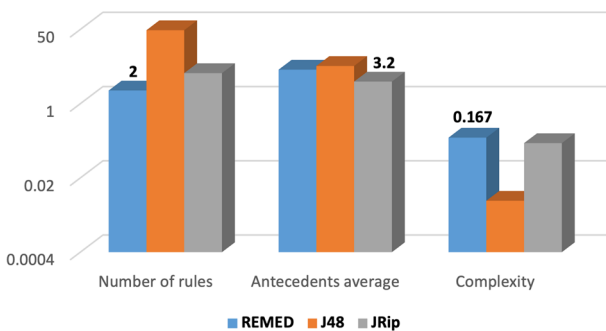


Fig. 2 Bar plot comparing the interpretability performance of the classifiers.

Table 5 The rule system of the symbolic classifiers.

REMED

IF (INTEXPY >= 2.5722) **AND** (NIMY <= 3.7568) **AND** (NOIJY <= 0.045512) **AND** (ROA <= -0.020157) **AND** (ROAPTX <= 0.33186) **AND** (ROE <= 2.4305) **AND** (ROEINJR <= -3.0748) **AND** (NPERFV >= 1.9392) **AND** (EQV <= 11.0556) **AND** (RBC1AAj <= 10.1801) **AND** (RBC1RWAJ <= 12.9377) **AND** (RBCRWAJ <= 14.2257) **THEN** insolvent
ELSE solvent

JRip

IF (RBC1RWAJ <= 9.365995) **AND** (NPERFV >= 8.743031) **THEN** insolvent
IF (ROEINJR <= -10.482257) **AND** (RBC1RWAJ <= 10.004857) **AND** (NPERFV >= 5.827501) **AND** (RBC1AAJ <= 6.549245) **THEN** insolvent
IF (ROE <= -6.834654) **AND** (RBCRWAJ <= 11.564752) **AND** (NPERFV >= 3.701495) **AND** (EQV <= 6.787471) **THEN** insolvent
IF (ROE <= -6.951554) **AND** (RBCRWAJ <= 12.718498) **AND** (NPERFV >= 5.754693) **AND** (LNLSDEPR <= 97.428213) **AND** (INTEXPY >= 3.280909) **AND** (NONIIY >= 0.03634) **THEN** insolvent
ELSE solvent

J48

5, 6, 6, 7, 7, 6, 11, 11, 11, 13, 13, 12, 9, 9, 9, 8, 10, 10, 9, 6, 3, 2, 4, 5, 7, 8, 8, 6, 3, 5, 5, 6, 7, 10, 10, 9, 8, 6, 6, 5, 7, 7, 6, 5, 7, 7, 6, 4 antecedents*

*DT with 48 leave nodes or rules

Table 6 Individual performance and rank of the classifiers for each assessment metric.

Metric	REMED		J48		JRip	
	Performance	Rank	Performance	Rank	Performance	Rank
AR	0.945	3	0.966	1	0.965	2
Precision	0.983	1	0.977	3	0.979	2
Type I error	0.411	1	0.589	3	0.539	2
Gmean	0.752	1	0.637	3	0.674	2
Number of rules	2	1	48	3	5	2
Antecedents average	6	2	7.29	3	3.2	1
Complexity	0.167	1	0.006	3	0.125	2
Average rank		1.429		2.714		1.857

Table 7 Statistical comparisons of classifiers.

Friedman test = 4.492 (p = 0.033)*

Control Classifier	Comparison Classifier	Nemenyi test q _{0.05} = 2.343 k = 3	Bonferroni-Dunn test q _{0.05} = 2.241 k = 3
REMED	J48	1.285 > 1.252*	1.285 > 1.198*
	JRip	---	0.428 < 1.198

*p < 0.05

Nevertheless, *REMED* was shown to be an unbiased symbolic classifier, suggesting its potential to identify insolvent cases, which have the highest erroneous classification cost. Moreover, non-parametric tests provide a more suitable statistical comparison among classifiers because these do not assume normal distributions or homogeneity of variance and can be applied to any evaluation measure (Demšar 2006). Therefore, including performance ranks for accuracy and interpretability metrics broadens the scope of statistical testing to find the best compromise solution.

On the other hand, cost-sensitive (Elkan 2001) and resampling strategies (Chawla et al. 2002) require constructing a cost matrix and determining under/oversampling rates (before running the algorithm) to address the class imbalance problem. In this sense, the fact that *REMED* requests a single parameter (confidence level for attribute selection) allows a more automated ML process.

Conclusions, Limitations and Future Work

ML for financial risk management has garnered substantial interest in the sector to improve the accuracy of forecasting banking failures and estimating creditworthiness.

However, the joint impact of the class imbalance problem and the dilemma of accuracy gain by loss of interpretability in ML approaches have not been widely studied, which constitutes a relevant gap in research for predicting bank failure and credit scoring models.

This study aimed to assess the performance of *REMED*, a symbolic classifier, in the context of financial risk prediction, using imbalanced datasets and considering a trade-off between accuracy and interpretability. A comparative analysis was conducted against two well-known rule-generating approaches, *J48* and *JRip*, using a dataset provided by the Federal Deposit Insurance Corporation.

The experimental results showed that *REMED* performed as a better and more direct ML method to address the problem of improving predictive accuracy from imbalanced financial data without affecting model interpretability. Furthermore, our study addresses a key research gap by examining how the class imbalance problem can affect interpretability performance, especially at extreme imbalance ratios.

Conversely, a notable limitation of this study is that the performance comparison was based on a single dataset of banking crisis analysis indicators. However, the selected sample collected a large amount of real data from 8292 banks represented by 17 independent financial attributes, which can reduce the effect of biased variance estimations due to dependencies between examples of a unique dataset.

A possible avenue for future research is to explore combining REMED with an oversampling rate to increase the representation of the target class or include cost ratio parameters to reduce misclassified examples instead of classification errors, as well as expanding the experimental framework to encompass a broader collection of datasets.

Data availability

The dataset generated and/or analysed during the current study is available in the repository of the Computer Security and Artificial Intelligence Research Laboratory at the Autonomous University of Nayarit: [https://securitylab.uan.mx/datasets/USbanks\(FDIC-2008\).csv](https://securitylab.uan.mx/datasets/USbanks(FDIC-2008).csv).

Received: 11 May 2024; Accepted: 18 October 2024;

Published online: 14 November 2024

References

- Aldrich J (1997) RA Fisher and the making of maximum likelihood 1912-1922. *Stat Sci* 12(3):162–176. <https://doi.org/10.1214/ss/1030037906>
- Apté C, Weiss S (1997) Data mining with decision trees and decision rules. *Future Gener Comput Syst* 13(2-3):197–210. [https://doi.org/10.1016/S0167-739X\(97\)00021-6](https://doi.org/10.1016/S0167-739X(97)00021-6)
- Armitage P, Berry G, Matthews JNS (2008) *Statistical methods in medical research*. Wiley-Blackwell, London
- Berka P (2016) Using the LISP-Miner system for credit risk assessment. *Neural Netw World* 26(5):497–518. <https://doi.org/10.14311/NNW.2016.26.029>
- Bramer M (2002) Using J-pruning to reduce overfitting in classification trees. *Knowl -Based Syst* 15(5-6):301–308. [https://doi.org/10.1016/S0950-7051\(01\)00163-0](https://doi.org/10.1016/S0950-7051(01)00163-0)
- Brown I, Mues C (2012) An experimental comparison of classification algorithms for imbalanced credit scoring data sets. *Expert Syst Appl* 39(3):3446–3453. <https://doi.org/10.1016/j.eswa.2011.09.033>
- Bücker M, Szepannek G, Gosiewska A, Biecek P (2022) Transparency, auditability, and explainability of machine learning models in credit scoring. *J Oper Res Soc* 73(1):70–90. <https://doi.org/10.1080/01605682.2021.1922098>
- Burden RL, Faires, JD (2000) *Numerical analysis*, Brooks Cole, Montgomery, Illinois
- Cano A, Zafra A, Ventura S (2011) An EP algorithm for learning highly interpretable classifiers. In: 2011 11th International Conference on Intelligent Systems Design and Applications, Cordoba, Spain, November 2011. Proceedings of the ISDA'11, IEEE, p 325-330. <https://doi.org/10.1109/isda.2011.6121676>
- Chao X, Kou G, Peng Y, Fernández A (2022) An efficiency curve for evaluating imbalanced classifiers considering intrinsic data characteristics: experimental analysis. *Inf Sci* 608:1131–1156. <https://doi.org/10.1016/j.ins.2022.06.045>
- Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP (2002) SMOTE: synthetic minority over-sampling technique. *J Artif Intell Res* 16:321–357. <https://doi.org/10.1613/jair.953>
- Chen N, Ribeiro B, Chen A (2016) Financial credit risk assessment: a recent review. *Artif Intell Rev* 45:1–23. <https://doi.org/10.1007/s10462-015-9434-x>
- Chen Y, Calabrese R, Martin-Barragan B (2024) Interpretable machine learning for imbalanced credit scoring datasets. *Eur J Oper Res* 312(1):357–372. <https://doi.org/10.1016/j.ejor.2023.06.036>
- Cohen W (1995) Fast effective rule induction. In: Twelfth International Conference on Machine Learning, Tahoe City, California, July 1995. Machine Learning Proceedings. Morgan Kaufmann, San Francisco, p 115-123. <https://doi.org/10.1016/b978-1-55860-377-6.50023-2>
- Cukierman A (2019) A retrospective on the subprime crisis and its aftermath ten years after Lehman's collapse. *Econ Syst* 43(3-4):100713. <https://doi.org/10.1016/j.ecosys.2019.100713>
- Demšar J (2006) Statistical comparisons of classifiers over multiple data sets. *J Mach Learn Res* 7(1):1–30
- Elkan C (2001) The foundations of cost-sensitive learning. In: The 17th International Joint Conference on Artificial Intelligence, Seattle, USA, August 2001. Proceedings of the IJCAI'01, vol 2. Morgan Kaufmann, San Francisco, p 973–978. <https://doi.org/10.5555/1642194.1642224>
- Florez-Lopez R, Ramon-Jeronimo JM (2015) Enhancing accuracy and interpretability of ensemble strategies in credit risk assessment. A correlated-adjusted decision forest proposal. *Expert Syst Appl* 42(13):5737–5753. <https://doi.org/10.1016/j.eswa.2015.02.042>
- Gacto MJ, Alcalá R, Herrera F (2011) Interpretability of linguistic fuzzy rule-based systems: an overview of interpretability measures. *Inf Sci* 181(20):4340–4360. <https://doi.org/10.1016/j.ins.2011.02.021>
- Hayashi Y (2016) Application of a rule extraction algorithm family based on the Re-RX algorithm to financial credit risk assessment from a Pareto optimal perspective. *Oper Res Perspect* 3:32–42. <https://doi.org/10.1016/j.orp.2016.08.001>
- Hayashi Y, Tanaka Y, Takagi T, Saito T, Iiduka H, Kikuchi H, Kikuchi H, Bologna G, Mitra S (2016) Recursive-rule extraction algorithm with J48graft and applications to generating credit scores. *J Artif Intell Soft Comput Res* 6(1):35–44. <https://doi.org/10.1515/jaiscr-2016-0004>
- Hayashi Y, Oishi T (2018) High accuracy-priority rule extraction for reconciling accuracy and interpretability in credit scoring. *N. Gener Comput* 36:393–418. <https://doi.org/10.1007/s00354-018-0043-5>
- He H, Zhang W, Zhang S (2018) A novel ensemble method for credit scoring: Adaptation of different imbalance ratios. *Expert Syst Appl* 98:105–117. <https://doi.org/10.1016/j.eswa.2018.01.012>
- Hosmer D, Lemeshow S (1989) *Applied logistic regression*. John Wiley and Sons. <https://doi.org/10.1002/0471722146>
- Hubbard RG, Navarro P (2010) Seeds of destruction: why the path to economic ruin runs through Washington, and how to reclaim American prosperity. FT Press
- Hussin-Adam-Khatir AA, Bee M (2022) Machine learning models and data-balancing techniques for credit scoring: what is the best combination? *Risks* 10(9):169. <https://doi.org/10.3390/risks10090169>
- Jones T, Sirmans GS (2019) Understanding subprime mortgage default. *J Real Estate Lit* 27(1):27–52. <https://doi.org/10.1080/10835547.2019.12090497>
- Kennedy K, Mac Namee B, Delany SJ (2010) Learning without default: a study of one-class classification and the low-default portfolio problem. In: Coyle L, Freyne J (eds) *Artificial Intelligence and Cognitive Science*. AICS 2009, Dublin, Ireland, August 2009. Lecture Notes in Computer Science, vol 6206. Springer, Berlin, Heidelberg, p 174–187. https://doi.org/10.1007/978-3-642-17080-5_20
- Kira K, Rendell LA (1992) A practical approach to feature selection. In: Kira K, Rendell, LA (eds) *The Ninth International Workshop on Machine Learning*. ML92, Aberdeen, Scotland, July 1992. Machine Learning Proceedings. Morgan Kaufmann, San Francisco, p 249–256. <https://doi.org/10.1016/B978-1-55860-247-2.50037-1>
- Kohonen T, Schroeder MR, Huang TS (2001) *Self-organizing maps*. Springer-Verlag. <https://doi.org/10.1007/978-3-642-56927-2>
- Kristóf T, Virág M (2022) EU-27 bank failure prediction with C5.0 decision trees and deep learning neural networks. *Res Int Bus Financ* 61:101644. <https://doi.org/10.1016/j.ribaf.2022.101644>
- Lantz B (2013) *Machine learning with R*. Packt Publishing Ltd, Birmingham
- Lanzarini LC, Villa-Monte A, Bariviera AF, Jimbo Santana P (2017) Simplifying credit scoring rules using LVQ+ PSO. *Kybernetes* 46(1):8–16. <https://doi.org/10.1108/k-06-2016-0158>
- Leo M, Sharma S, Maddulety K (2019) Machine learning in banking risk management: a literature review. *Risks* 7(1):29. <https://doi.org/10.3390/risks7010029>
- Marqués AI, García V, Sánchez JS (2013) On the suitability of resampling techniques for the class imbalance problem in credit scoring. *J Oper Res Soc* 64(7):1060–1070. <https://doi.org/10.1057/jors.2012.120>
- Martens D, Baesens B, Van-Gestel T, Vanthienen J (2007) Comprehensible credit scoring models using rule extraction from support vector machines. *Eur J Oper Res* 183(3):1466–1476. <https://doi.org/10.12139/ssrn.878283>
- Mena L, Gonzalez JA (2009) Symbolic one-class learning from imbalanced datasets: application in medical diagnosis. *Int J Artif Intell T* 18(2):273–309. <https://doi.org/10.1142/S0218213009000135>
- Nalić J, Martinovic G (2020) Building a credit scoring model based on data mining approaches. *Int J Softw Eng Know* 30(2):147–169. <https://doi.org/10.1142/s0218194020500072>
- Nauck D (2002) Measuring interpretability in rule-based classification systems. In: Proceedings of 12th IEEE International Conference on Fuzzy Systems, St. Louis, MO, May 2003. IEEE, p 196–201. <https://doi.org/10.1109/fuzz.2003.1209361>
- Niu K, Zhang Z, Liu Y, Li R (2020) Resampling ensemble model based on data distribution for imbalanced credit risk evaluation in P2P lending. *Inf Sci* 536:120–134. <https://doi.org/10.1016/j.ins.2020.05.040>
- Obermann L, Waack S (2016) Interpretable multiclass models for corporate credit rating capable of expressing doubt. *Front Appl Math Stat* 2:16. <https://doi.org/10.3389/fams.2016.00016>
- Otieno B, Wabwoba F, Musumba G (2020) Towards small-scale farmers fair credit scoring technique. In: 2020 IST- Africa Conference. 2020 IST-Africa, Kampala, Uganda, May 2020. IEEE, p 1–11
- Panigrahi R, Borah S (2018) Rank allocation to J48 group of decision tree classifiers using binary and multiclass intrusion detection datasets. *Procedia Comput Sci* 132:323–332. <https://doi.org/10.1016/j.procs.2018.05.186>

- Quan J, Sun X (2024) Credit risk assessment using the factorization machine model with feature interactions. *Humanit Soc Sci Commun* 11(234):1–10. <https://doi.org/10.1057/s41599-024-02700-7>
- Quinlan JR (1993) C4.5: Programs for machine learning. Morgan Kaufmann
- Quinlan JR (1996) Bagging, boosting, and C4.5. In: The Thirteenth National Conference on Artificial Intelligence, Portland, Oregon, August 1996. Proceedings of the Association for the Advancement of Artificial Intelligence, The AAAI Press, Menlo Park, California, p 725–730
- Sánchez-Garreta JS, García V, Marqués-Marzal AI (2012) Assessment of financial risk prediction models with multi-criteria decision making methods. Proceedings of 19th International Conference on Neural Information Processing, 60–67. https://doi.org/10.1007/978-3-642-34481-7_8
- Serrano-Cinca C, Gutiérrez-Nieto B (2013) Partial least square discriminant analysis for bankruptcy prediction. *Decis Support Syst* 54(3):1245–1255. <https://doi.org/10.1016/j.dss.2012.11.015>
- Setiono R, Baesens B, Mues C (2008) Recursive neural network rule extraction for data with mixed attributes. *IEEE T Neural Network* 19(2):299–307. <https://doi.org/10.1109/tnn.2007.908641>
- Shang L, Zhou B, Li J, Tang D, Boamah V, Pan Z (2024) Evaluating financial fragility: a case study of Chinese banking and finance systems. *Humanit Soc Sci Commun* 11(1):1–9. <https://doi.org/10.1057/s41599-024-02932-7>
- Shen F, Zhao X, Li Z, Li K, Meng Z (2019) A novel ensemble classification model based on neural networks and a classifier optimisation technique for imbalanced credit risk evaluation. *Phys A Stat Mech Appl* 526:121073. <https://doi.org/10.1016/j.physa.2019.121073>
- Tipping ME (2000) The relevance vector machine. In: SA Stolla, TK Leen, KR Mullar (Eds.) *Advances in neural information processing systems*, Denver, Colorado, November–December 1999. MIT Press, p 652–658
- Tomczak JM, Zięba M (2015) Classification restricted Boltzmann machine for comprehensible credit scoring model. *Expert Syst Appl* 42(4):1789–1796. <https://doi.org/10.1016/j.eswa.2014.10.016>
- Wang Z, Sun X, Zhang D (2007) A PSO-based classification rule mining algorithm. In: Huang, DS., Heutte, L., Loog, M. (eds) *Advanced Intelligent Computing Theories and Applications. With Aspects of Artificial Intelligence. ICIC 2007*, Qingdao, China, August 2007. Lecture Notes in Computer Science, vol 4682. Springer, Berlin, Heidelberg. p 377–384. https://doi.org/10.1007/978-3-540-74205-0_42
- Wang H, Xu Q, Zhou L (2015) Large unbalanced credit scoring using lasso-logistic regression ensemble. *PLoS One* 10(2):e0117844. <https://doi.org/10.1371/journal.pone.0117844>
- Witten IH, Frank E, Hall MA (2011) *Data mining practical machine learning tools and techniques*. Morgan Kaufmann Publishers. <https://doi.org/10.1016/c2009-0-19715-5>
- Wu TC, Hsu MF (2012) Credit risk assessment and decision making by a fusion approach. *Knowl -Based Syst* 35:102–110. <https://doi.org/10.1016/j.knosys.2012.04.025>
- Xia Y, Zhao J, He L, Li Y, Niu M (2020) A novel tree-based dynamic heterogeneous ensemble method for credit scoring. *Expert Syst Appl* 159:113615. <https://doi.org/10.1016/j.eswa.2020.113615>
- Xu P, Ding Z, Pan M (2018) A hybrid interpretable credit card users default prediction model based on RIPPER. *Concurr Comp Pr E* 30(23):e4445. <https://doi.org/10.1002/cpe.4445>
- Ying X (2019) An overview of overfitting and its solutions. *J Phys Conf Ser* 1168:022022. <https://doi.org/10.1088/1742-6596/1168/2/022022>
- Zhang H, He H, Zhang W (2018) Classifier selection and clustering with fuzzy assignment in ensemble model for credit scoring. *Neurocomputing* 316:210–221. <https://doi.org/10.1016/j.neucom.2018.07.070>
- Zubair S, Kabir R, Huang X (2020) Does the financial crisis change the effect of financing on investment? Evidence from private SMEs. *J Bus Res* 110:456–463. <https://doi.org/10.2139/ssrn.3514579>

Author contributions

Luis Mena: conceptualization, methodology, writing – original draft preparation, writing – reviewing and editing. Vicente García: conceptualization, writing – reviewing and editing. Vanessa G. Félix: visualization, investigation. Rodolfo Ostos: software. Rafael Martínez-Peláez: validation. Alberto Ochoa-Brust: formal analysis. Pablo Velarde-Alvarado: data curation.

Competing interests

The authors declare no competing interests.

Ethical approval

Ethical approval was not required as the study did not involve human participants.

Informed consent

This article does not contain any studies with human participants or animals performed by any of the authors. Therefore, informed consent was not applicable.

Additional information

Correspondence and requests for materials should be addressed to Vicente García.

Reprints and permission information is available at <http://www.nature.com/reprints>

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2024