Journal of the Brazilian Computer Society, 2026, 32:1, doi: 10.5753/jbcs.2026.6415

On the Limits of Genetically-Optimized Homogeneous Ensembles for Credit Risk Classification

Ricardo Franceli da Silva [**University of Sao Paulo** | ***ricardo.franceli.silva@alumni.usp.br***]
Leandro dos Santos Maciel [**University of Sao Paulo** | ***leandromaciel@usp.br***]

Department of Business Administration, School of Economics, Business, Accounting and Actuary, University of Sao Paulo, Av. Prof. Luciano Gualberto, 908, Butanta, Sao Paulo, SP, 05508-010, Brazil.

Received: 22 August 2025 • **Accepted:** 03 May 2026 • **Published:** 25 June 2026

Abstract. Credit risk assessment is a challenging task with significant economic and financial impacts. It requires capturing complex nonlinear patterns and interactions between variables to accurately predict creditworthiness and minimize the risk of default. This study investigates the performance of a genetically-optimized homogeneous ensemble composed of five Multilayer Perceptron (MLP) models applied to a large-scale peer-to-peer lending dataset. Individual models achieved competitive precision scores (up to 80.57%); however, the optimized ensemble failed to surpass the best-performing individual model under economically viable conditions. Ensemble scenarios operating under a stricter classification threshold achieved higher precision gains but yielded negative Expected Profit per loan, rendering them impractical for real-world credit granting. This finding was confirmed by a robustness check, where the experiment was repeated after removing the top model. A pairwise error correlation analysis revealed consistently high correlations among base learners (0.762-0.918), with co-occurring error rates between 79.43% and 93.93%, providing empirical evidence that the base classifiers lack the predictive diversity necessary for synergistic ensemble gains. The results reveal a critical boundary condition for ensemble methods: when base classifiers share the same underlying learning algorithm, thereby lacking conceptual diversity, synergistic gains are unattainable; instead, the optimization process converges on weighting the strongest component. This study concludes that classifier diversity is a fundamental principle for an ensemble to deliver superior performance, regardless of the strength of its individual learners, challenging the assumption that optimized ensembles universally outperform their strongest individual components in machine learning.

Keywords: Credit Risk, Machine Learning, Ensemble, Genetic Algorithm, Optimization

1 Introduction

In the financial industry, accurate credit risk assessment is crucial for managing exposure to potential defaults. Credit risk models aim to classify customers based on their likelihood of meeting financial obligations. A key metric in this context is precision, especially the ability to correctly identify customers who are highly likely to fulfill their obligations—those with a very low risk of default. High precision here minimizes the risk of misclassifying high-risk customers as low-risk, thus preventing potential losses from individuals likely to default. This is important even if it means missing some business opportunities by misclassifying low-risk customers.

Multilayer Perceptron (MLP) models are widely used in credit risk assessment due to their ability to capture complex patterns in data. However, individual MLP models, despite their effectiveness, may not fully optimize precision score. Ensemble methods, which combine the predictions of multiple models, offer a promising approach to enhance model performance. This paper explores the use of a genetic algorithm to optimize an ensemble of five MLP models, maximizing precision score in the context of credit risk assessment. The goal is to refine the prediction of low-risk customers, ensuring that those predicted to have a very low risk of default are accurately identified.

In the financial industry, the provision of credit serves as a fundamental pillar, fueling economic growth, enabling investment, and facilitating consumption across various sectors. However, this vital function inherently carries significant risks, primarily the potential for default, which can severely impact the stability and profitability of financial institutions. Effective credit risk assessment is therefore not merely an operational necessity but a strategic imperative for managing exposure to these potential defaults.

Credit risk models are designed to classify customers based on their likelihood of fulfilling financial obligations. The stakes in this classification are high, as misjudgments can lead directly to substantial financial losses. Specifically, when a high-risk customer is erroneously classified as low-risk and granted a loan, the institution faces the potential loss of both the principal amount extended and the anticipated interest earnings. This direct financial impact underscores the importance of achieving highly accurate customer classification, particularly in correctly identifying those who pose a genuine default risk. While missing business opportunities by misclassifying low-risk customers can occur, preventing these direct losses from defaults is often the primary concern.

MLP models are widely utilized in credit risk assessment due to their inherent ability to capture complex, non-linear patterns within intricate datasets. Despite their individual effectiveness, single MLP models may not always achieve optimal performance across all critical metrics. This limitation

often stems from the inherent trade-offs in model training and the complexity of real-world financial data. Consequently, ensemble methods, which strategically combine the predictions of multiple individual models, offer a promising avenue to enhance overall model robustness and predictive power.

This paper explores the application of a genetic algorithm to optimize such an ensemble, specifically comprising five distinct MLP models. The core objective of this optimization is to maximize the precision score in the context of credit risk assessment. Genetic algorithms are particularly well-suited for this task due to their capacity to efficiently search vast solution spaces, identifying optimal or near-optimal combinations of weights that might be elusive through conventional methods [Goldberg, 1989]. This evolutionary approach allows for a refined tuning of the ensemble, aiming to ensure that customers predicted to have a very low risk of default are identified with the highest possible accuracy, thereby directly contributing to the mitigation of financial losses.

This study investigates whether genetic algorithm optimization can enhance the predictive performance of a homogeneous ensemble of MLP models beyond that of the best-performing individual model in a large-scale credit risk classification task. The research objective is to evaluate the practical boundaries of ensemble optimization when base learners share the same underlying learning algorithm, thereby lacking conceptual diversity. To this end, the following hypotheses are formally stated: the null hypothesis (H_0) posits that the precision score achieved by the genetically-optimized homogeneous ensemble is greater than or equal to that of the best-performing individual MLP model; the alternative hypothesis (H_1) posits that the ensemble's precision score does not exceed that of the best individual model, suggesting that homogeneity among base learners constitutes a binding constraint on ensemble performance gains.

The presented results indicate that the effectiveness of ensembling with optimization may not always enhance the standalone performance of artificial neural networks in credit risk assessment. While individual MLP models demonstrated robust precision scores, ranging from 78.57% to 80.57%, the genetically optimized ensemble did not achieve a higher precision score than the best-performing individual model. This suggests that in certain contexts, particularly when base models already exhibit strong performance, the optimization process may heavily weight the strongest existing component, rather than discovering novel synergistic combinations that yield further gains.

The study provides strong empirical evidence that a genetically-optimized ensemble of high-performing, yet homogeneous, MLP models does not improve precision over the best-performing individual model in a large-scale credit risk classification task. The findings also demonstrate a practical boundary condition for the effectiveness of ensemble optimization, showing that benefits are marginal or non-existent when individual models already exhibit exceptionally high performance and, critically, lack diversity due to sharing the same underlying learning algorithm. Furthermore, the results offer insight into the optimizer's behavior; in the absence of synergistic gain, it logically concentrated the weights on the single best-performing model.

The remainder of this paper is organized as follows. Section 2 reviews the relevant literature on credit risk modeling, the application of neural networks, and ensemble techniques. Section 3 details the methodology, describing the dataset, the architecture of the individual MLP models, the optimization metric, and the genetic algorithm framework used to create the ensemble. In Section 4, we present the empirical results for both the individual and ensemble models, including the findings from a robustness check. Finally, Section 5 discusses the implications of these results, addresses the study's limitations, and concludes the paper with directions for future research.

2 Related Research

Credit risk, intrinsically linked to the possibility of financial losses arising from a borrower's failure to meet their contractual obligations, constitutes a central and constant concern for financial institutions and investors globally [Saunders and Cornett, 2018; Duffie and Singleton, 2003; Caouette *et al.*, 1998]. The inability to accurately assess and manage this risk can lead to substantial financial losses, impacting not only individual institutions through the loss of principal and interest but also potentially destabilizing broader financial markets. Effective management of this risk is therefore not merely an operational necessity but a strategic imperative, safeguarding the stability of lending entities and fostering efficiency and confidence in the financial system as a whole. Within the scope of predictive modeling for such evaluations, ensemble learning approaches have emerged as robust paradigms that seek to enhance predictive performance through the combination of multiple learning models [Rokach, 2010; Zhou, 2012]. However, the mere aggregation of classifiers does not always guarantee optimal performance; thus, optimization in ensembles emerges as a crucial step. This process involves the systematic adjustment of parameters, the selection of committee members, or the definition of weighting and combination schemes for individual predictions, with the aim of maximizing a specific performance metric and, consequently, the generalization capability of the consolidated model in complex tasks [Kuncheva and Rodríguez, 2007; Gandomi and Haider, 2015].

Credit scoring models are designed to enable the distinction between good credit and bad credit groups, thereby informing lending decisions [Chen and Huang, 2003]. Models that accurately predict the good credit group offer significant business advantages, such as reducing the operational costs associated with credit analysis, facilitating faster decision-making processes, ensuring effective credit collections, and minimizing potential default risks [Tsai and Wu, 2007; West, 2000]. Conversely, the precise identification of the 'bad credit' group is equally critical, as it directly mitigates the risk of loan defaults, thereby safeguarding capital and ensuring the sustainability of lending operations.

Since the seminal research of Altman [1968], approaches to modeling inherent default risk in credit have continuously evolved. Early methodologies proposed various statistical methods, including probit analysis [Abdou *et al.*, 2008] and logistic regression [Abdou *et al.*, 2008; Crook *et al.*, 2007], with the primary objective of predicting credit default risk. More

recent technological advancements have ushered in complex techniques from the field of machine learning (ML), such as decision trees, bagging and boosting ensembles like Random Forest (RF), Support Vector Machines (SVM), Artificial Neural Networks (ANN), fuzzy models, genetic algorithms, and deep learning neural networks [Louzada and Ara, 2016]. This evolution reflects a shift from models relying on more restrictive statistical assumptions to data-driven approaches capable of uncovering intricate, non-linear relationships within vast and complex financial datasets.

The predictive power of deep learning models is well documented in the literature [Sadhwani *et al*., 2020], with the Multilayer Perceptron (MLP) standing out for its ability to model intricate structures in high-dimensional data. An MLP is a type of feed-forward Artificial Neural Network (ANN) consisting of an input layer, one or multiple hidden layers of nodes, and an output layer [Liu *et al*., 2024]. This architecture allows MLPs to approximate non-linear functions, providing improved prediction accuracy in complex problems such as credit risk assessment [Liu *et al*., 2024]. Consequently, MLPs have been widely adopted as classifiers in various financial decision-making contexts, including bankruptcy prediction and credit risk evaluation [Atiya, 2001; Huang *et al*., 2004; Lee *et al*., 2006; Angelini *et al*., 2008; Battiston *et al*., 2012; Haykin, 1999; Zhao *et al*., 2015; Chi Guotai and Moula, 2017; Liu *et al*., 2024]. However, findings regarding their consistent superiority can vary. For instance, a comparative study by West [2000] evaluating various neural network architectures and traditional classifiers for credit scoring concluded that while the MLP is widely recognized, it was not the topology that consistently demonstrated superior outcomes. Conversely, Zhao *et al*. [2015], in a comparative review of methodologies between 2011 and 2015, identified the MLP as a top-performing model. More recently, Louzada and Ara [2016] compared numerous algorithms across different datasets and suggested that neural networks are not consistently among the best predictors of credit risk, especially in imbalanced datasets, despite their prevalence in research. Such limitations, particularly concerning data imbalance, underscore the need for advanced techniques, such as ensemble methods, to bolster predictive robustness [Louzada and Ara, 2016].

To improve the performance of individual classifiers, the integration of multiple classifiers into a combined system has been extensively investigated, with positive effects [Bhuria *et al*., 2025; Li *et al*., 2018; Al-Maari *et al*., 2025; Nanni and Lumini, 2009; Ghatge and Halkarnikar, 2013; Abellán and Castellano, 2017; Chen *et al*., 2020; Li *et al*., 2018; Lessmann *et al*., 2015; Louzada and Ara, 2016; Tsai and Wu, 2007]. These ensemble systems merge the outputs of several classifiers, aiming to achieve better overall performance compared to any single classifier, by voting or weighting the results [Nanni and Lumini, 2009]. A fundamental principle for improving the accuracy of these ensembles is the promotion of diversity among their base classifiers [Breiman, 1996]. Indeed, a broader spectrum of characteristics among these classifiers generally correlates with increased robustness and precision in the final ensemble capacity [Abellán and Mantas, 2014]. The need for diversity and an optimal combination often necessitates sophisticated optimization techniques to fine-tune the ensemble's structure and weighting mechanisms.

The use of heterogeneous ensembles can enhance the performance of individual classifiers, leveraging their diversity to yield significantly positive results [Bhuria *et al*., 2025]. However, some studies show that such ensembles may provide only marginal performance gains over the best-performing individual models [Nanni and Lumini, 2009; Al-Maari *et al*., 2025].

Some authors have explored the use of homogeneous ensembles. For example, [Nanni and Lumini, 2009] evaluated the application of stand-alone models and homogeneous ensembles of Levenberg-Marquardt neural networks, MLPs, 5-nearest neighbor, and Radial Basis Function Support Vector Machines. Nanni and Lumini [2009] found that while ensembles generally improved the performance of individual models, MLP-based ensembles exhibited the least improvement over their stand-alone counterparts when measured by AUC (Area Under the Curve). A similar finding was reported by Tsai and Wu [2007], who identified that well-structured individual MLP models can have better performance than ensembles in terms of accuracy; however, Type I and Type II errors do not always follow the same trend. Their work concluded that for optimizing specific outcomes, such as minimizing Type I or Type II errors, the results may vary.

3 Methodology

This section details the methodology and is organized into four parts: Dataset and Preprocessing, which describes the data and the transformations applied; Base Learners Specification, which specifies the individual model architectures; Performance Metrics, which defines the criteria for assessing performance; and Ensemble Optimization, which details the genetic algorithm used to combine the models.

3.1 Dataset and Preprocessing

The dataset utilized in this research was the All Lending Club Loan Data[1], with loans accepted from Q1 2007 to Q4 2018. The dataset was preprocessed, highly correlated features were reduced (a threshold of 80% was considered), and categorical features were transformed into binary (dummy) variables. All variables that could not be confirmed as pre-origination, that is, those not explicitly documented as available at the time of the credit decision, were removed from the dataset. The resulting dataset comprises 68 features and 1,341,026 observations[2].

Table 1 summarizes the descriptive statistics for the numeric variables utilized in this study, after preprocessing. For detailed variable definitions, readers are referred to the data

[1] `https://www.kaggle.com/datasets/wordsforthewise/lending-club`

[2] The retained variables are: *loan_amnt, term, int_rate, annual_inc, dti, fico_range_high, pub_rec, collections_12_mths_ex_med, charge-off_within_12_mths, pub_rec_bankruptcies, and tax_liens (numeric); and the following categorical variables encoded as binary dummies: A2-G5 (subgrade), Source Verified, Verified, Joint App (verification status), w (initial list status), credit_card, debt_consolidation, educational, home_improvement, house, major_purchase, medical, moving, other, renewable_energy, small_business, vacation, wedding (loan purpose), MORTGAGE, OTHER, OWN, RENT (home ownership), and DirectPay (payment plan)*

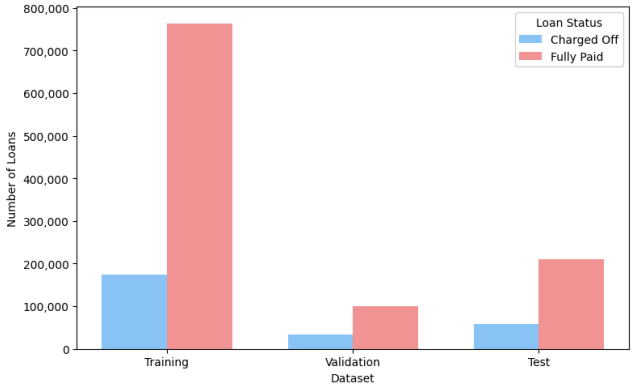

Figure 1. Distribution of loan status across datasets.

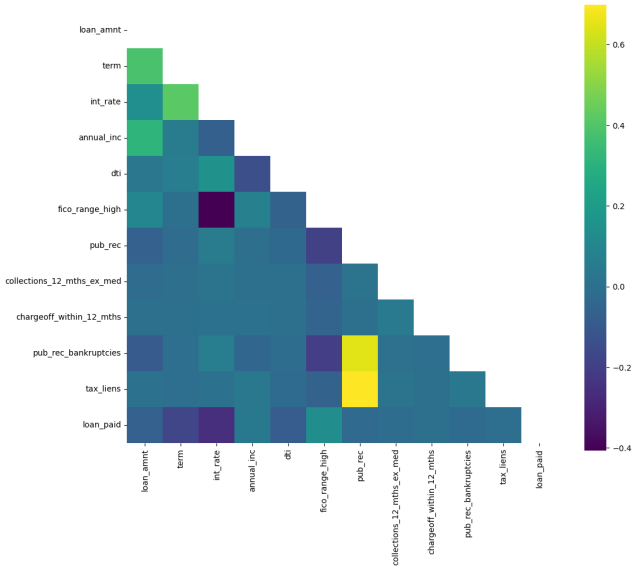

Figure 2. Correlation matrix of numeric variables retained after preprocessing.

dictionary provided with the Lending Club dataset[3]. The focus of this research, however, is on the aggregate predictive power of the ensembles and individual learners rather than on an individual critical evaluation of each feature.

The loan status variable was transformed into a binary outcome: 1 for "fully paid" (indicating a performing loan) and 0 for "not paid" (indicating a defaulted loan, see **Figure 1**). The dataset was split chronologically by origination date into training (70%, 938,718), validation (10%, 134,105, used as the basis for ensemble optimization), and test (20%, 268,206) sets, reflecting real-world deployment conditions where models are trained on historical data and evaluated on subsequent observations. The proportion of defaulted loans across the three sets was 18.6%, 25.2%, and 21.4% for training, validation, and test, respectively, reflecting the natural temporal drift in default rates over the sample period. All models were trained with the same data.

3.2 Base Learners Specification

To evaluate the ensemble's performance, a pool of base classifiers with distinct architectural configurations was required. Accordingly, MLP models were constructed, each with a unique architecture as specified in **Table 2**. The primary focus was on generating architectural diversity in conceptually homogenous models to assess the ensemble's ability to improve upon individual model performance, rather than on the intrinsic performance of each individual learner based on prior studies. Consistent with standard practices for binary classification, all models utilized the ReLU activation function **(Eq. 1)** for the hidden layers and the sigmoid function **(Eq. 2)** for the output layer.

$$f(z) = \max(0, z) \tag{1}$$

where z is the input for the activation function.

The sigmoid function constrains the model's output to a value between 0 and 1, which can be interpreted as a probability, and is defined as:

$$\sigma(z) = \frac{1}{1 + e^{-z}} \tag{2}$$

where z is the input to the activation function.

The models were compiled with the binary cross-entropy loss function and optimized using the Adam optimizer. The Adam algorithm is a method for stochastic optimization that relies on adaptive estimates of the first and second moments of the gradients [Kingma and Ba, 2015]. The weight update for a weight w at time step t is given, in general, by:

$$w_{t+1} = w_t - \frac{\eta}{\sqrt{\hat{v}_t} + \epsilon}\hat{m}_t \tag{3}$$

where η is the learning rate, $\hat{m}_t$ is the bias-corrected first moment estimate of the gradient, $\hat{v}_t$ is the bias-corrected second moment estimate of the gradient, and ϵ is a small value to prevent division by zero.

The binary cross-entropy loss function is suitable for binary classification problems and is formulated as:

$$L(\hat{y}, y) = -[y \log(\hat{y}) + (1 - y) \log(1 - \hat{y})] \tag{4}$$

where y is the true label (0 or 1) and $\hat{y}$ is the predicted output of the model. Each model was trained individually, and cross-validated with the validation dataset.

3.3 Performance Metrics

The precision score was chosen as the primary optimization metric to prioritize the accurate identification of customers likely to pay their loans (Class 1). This approach aims to minimize the misclassification of defaulters as payers (i.e., reducing False Positives for Class 1), thereby mitigating financial losses by ensuring credit is predominantly granted to low-risk individuals [Caouette *et al.*, 1998].

Precision score is defined as:

$$Precision = \frac{TruePositives}{TruePositives + FalsePositives} \tag{5}$$

Here, True Positives are customers correctly identified as payers, and False Positives are customers incorrectly identified as payers (who are actually defaulters).

[3] `https://resources.lendingclub.com/LCDataDictionary.xlsx`

Table 1. Descriptive Statistics of Numeric Variables

Variable	mean	std	min	25%	50%	75%	max	kurtosis	skewness
loan_amnt	14,422	8,712	500	8,000	12,000	20,000	40,000	-0.08	0.78
term	42	10	36	36	36	36	60	-0.54	1.21
int_rate	13	5	5	10	13	16	31	0.50	0.71
annual_inc	76,262	69,867	16	45,996	65,000	90,000	10,999,200	4840.92	46.50
dti	18	11	-1	12	18	24	999	2120.50	27.18
fico_range_high	700	32	664	674	694	714	850	1.67	1.29
pub_rec	0	1	0	0	0	0	86	743.81	11.56
collections_12_mths_ex_med	0	0	0	0	0	0	20	695.23	14.92
pub_rec_bankruptcies	0	0	0	0	0	0	12	20.02	3.44
chargeoff_within_12_mths	0	0	0	0	0	0	10	595.48	18.10
tax_liens	0	0	0	0	0	0	85	3671.43	32.75
loan_paid	1	0	0	1	1	1	1	0.29	-1.51

Note: Table presents the descriptive statistics of the numeric variables retained after preprocessing.

Table 2. Characteristics of the five MLP models.

Model	Hidden Layers	Epochs	Batch Size
Model 1	[78, 39, 19, 8, 4]	40	512
Model 2	[64, 32, 8, 4]	40	512
Model 3	[8]	40	512
Model 4	[128, 64, 32, 8, 4]	40	128
Model 5	[128, 64, 32, 8, 4]	40	32

To provide a comprehensive evaluation of model performance, five complementary metrics are reported, based on Lessmann *et al*. [2015]. Categorical prediction accuracy is assessed through Recall, Accuracy, and F1-score, which capture the model's ability to correctly classify both performing and defaulting loans under a fixed decision threshold. Discriminatory capacity is measured by the Area Under the ROC Curve (AUC), which evaluates the model's ability to rank borrowers by default risk across all possible thresholds, independently of any specific cutoff. Probabilistic prediction accuracy is assessed through the Brier Score, which measures the mean squared error between predicted probabilities and actual outcomes, penalizing both overconfident and poorly calibrated predictions. Finally, a cost-sensitive metric, the Expected Profit per Loan (EP), is included to align model evaluation with the financial objectives of the lending institution, capturing the asymmetric economic consequences of misclassification that are obscured by standard statistical metrics.

The Expected Profit per Loan was computed following the profit-based framework proposed by Verbraken *et al*. [2014], which explicitly incorporates the asymmetric costs and revenues associated with credit granting decisions. For each approved loan correctly identified as performing (True Positive), the model generates an interest revenue equal to $r = W \times i \times t$, where W is the loan amount, i is the annual interest rate, and t is the loan term in years. For each defaulting loan incorrectly approved (False Positive), the institution incurs a principal loss equal to W. For each performing loan incorrectly rejected (False Negative), the institution foregoes the interest revenue. The net profit is then normalized by the total number of approved loans to yield the Expected Profit per Loan, enabling comparison across models with different approval volumes — a normalization that is particularly relevant in the presence of degenerate classifiers that approve all applicants[4].

Initially, a probability threshold of 0.5 (50%) was used to classify customers as payers (1). To evaluate the model's performance under a stricter criterion for identifying good payers, an additional analysis was conducted using a higher probability threshold of 0.8 (80%).

3.4 Ensemble Optimization

To further enhance precision score, an ensemble approach was adopted [Abellán and Mantas, 2014]. The predictions from the five MLP models were combined using a weighted sum to form the ensemble prediction ($\hat{y}_{ensemble}$):

$$\hat{y}_{ensemble} = w_1 \cdot \hat{y}_1 + w_2 \cdot \hat{y}_2 + w_3 \cdot \hat{y}_3 + w_4 \cdot \hat{y}_4 + w_5 \cdot \hat{y}_5 \quad (6)$$

where $\hat{y}_i$ is the prediction of individual model i, and w_i is the weight assigned to model i by the genetic algorithm.

The weights assigned to each model were freely optimized using a genetic algorithm, which is well-suited for finding optimal solutions in complex search spaces, such as large, nonlinear and noisy ones [Goldberg, 1989].

The genetic algorithm was configured to maximize the precision score, specifically targeting the accurate identification of customers likely to pay their loans (Class 1). This optimization was achieved by evolving a population of potential weight combinations across multiple generations. Each individual within the population represented a unique set of weights applied to the predictions of the individual MLP models. The fitness of each individual was evaluated based on the precision score it yielded, thereby guiding the optimization process towards the combination of weights that maximized the desired outcome.

[4] In this study, the financial parameters were derived from the descriptive statistics of the dataset: an average loan amount of $14,422, an average annual interest rate of 13%, and an average loan term of 3.5 years (42 months), yielding an interest revenue of $6,562 per performing loan and a principal loss of $14,422 per defaulting loan approved, assuming a Loss Given Default (LGD) of 100%. A sensitivity analysis with Exposure at Default (EAD) of $7,211 and LGD of 50% presented similar results.

To investigate the impact of population size and generational progression on performance, two distinct optimization scenarios were developed:

- Scenario 1: Employed a population of 200 individuals and ran for 200 generations, with a crossover probability (‘cxpb‘) of 0.6 and a mutation probability (‘mutpb‘) of 0.05. Classification Threshold: 0.5.
- Scenario 2: Utilized a larger population of 1000 individuals and ran for 1000 generations, maintaining the same ‘cxpb‘ of 0.6 and ‘mutpb‘ of 0.05. Classification Threshold: 0.5.

To provide a comprehensive analysis, two additional optimization scenarios were developed, varying the classification threshold:

- Scenario 3: Employed a population of 200 individuals and ran for 200 generations, with a crossover probability (‘cxpb‘) of 0.6 and a mutation probability (‘mutpb‘) of 0.05. Classification Threshold: 0.8.
- Scenario 4: Utilized a larger population of 1000 individuals and ran for 1000 generations, maintaining the same ‘cxpb‘ of 0.6 and ‘mutpb‘ of 0.05. Classification Threshold: 0.8.

These scenarios aimed to determine whether smaller population sizes and fewer generations are sufficient to improve credit risk classification, or if larger populations and more extensive generational evolution are necessary to significantly enhance the ensemble’s performance beyond that of the individual models.

Credit risk prediction by individual models may leave room for improvement in precision score, which can be mitigated with ensemble models optimized by genetic algorithms.

3.5 Ensemble Ablation Analysis

To formally investigate the conditions under which ensemble weighting yields synergistic gains versus collapsing onto the best individual model, two complementary analyses were conducted.

The first analysis computed the pairwise error correlation between base learners as an empirical measure of predictive diversity within the ensemble, following the framework established by Kuncheva and Whitaker [2003]. For each pair of models, the Pearson correlation coefficient between their binary error vectors was computed, where an error vector assigns 1 to instances misclassified by a given model and 0 to correctly classified instances. Complementarily, the co-occurring error rate was calculated as the proportion of instances on which a given model erred that were also misclassified by each other model — formally, for models i and j:

$$CoError(i,j) = \frac{\sum_{k=1}^{n} \mathbb{1}[e_i^k = 1 \wedge e_j^k = 1]}{\sum_{k=1}^{n} \mathbb{1}[e_i^k = 1]} \tag{7}$$

where e_i^k denotes whether model i erred on instance k. Under this formulation, a value of 100% on the diagonal is expected by definition, and off-diagonal values approaching 100% indicate that the models commit errors on virtually the same instances, leaving no complementary predictive signal for the ensemble to exploit.

The second analysis examined the weight distribution produced by the genetic algorithm across all four optimization scenarios. The optimized weights were inspected to determine whether the optimizer converged on a uniform distribution, indicative of synergistic combination, or concentrated the weights on a single component, indicative of collapse onto the best individual model. This analysis was conducted across all combinations of classification threshold (0.5 and 0.8) and population size (200 and 1,000 individuals) to assess the robustness of the convergence behavior to variations in the optimization configuration.

To formally test the stated hypotheses, the McNemar test [McNemar, 1947; Dietterich, 1998] was applied to assess whether the difference in predictive performance between the best-performing individual model and each optimized ensemble scenario was statistically significant. The McNemar test is a non-parametric paired test particularly suited for comparing two classifiers evaluated on the same test set, focusing on discordant cases, instances where one model correctly classifies an observation while the other does not [Dietterich, 1998]. It is especially recommended when models are expensive or impractical to retrain multiple times, as is the case with large neural networks trained on large-scale datasets [Dietterich, 1998]. The test statistic is defined as:

$$\chi^2 = \frac{(|n_{01} - n_{10}| - 1)^2}{n_{01} + n_{10}} \tag{8}$$

where n_{01} denotes the number of instances correctly classified by the best individual model but misclassified by the ensemble, and n_{10} denotes the reverse. Under the null hypothesis, the statistic follows a χ^2 distribution with one degree of freedom. The exact binomial variant, recommended when the number of discordant cases is small ($n_{01} + n_{10} < 25$) [Fagerland *et al*., 2013], is defined as:

$$p = 2 \sum_{i=n_{10}}^{n_{01}+n_{10}} \binom{n_{01}+n_{10}}{i} \left(\frac{1}{2}\right)^{n_{01}+n_{10}} \tag{9}$$

The exact binomial variant was adopted for the economically viable scenarios given the small number of discordant cases observed, while the asymptotic variant was used for Scenarios 3 and 4, where discordant cases were abundant.

4 Results

The results of each model with the training and validation datasets are presented in **Table 3**. The table demonstrates a good performance for most individual models, as they correctly classify 81% to 83% the good payers (precision) in the validation dataset. These are the individuals who would have their loans granted based on the models’ analysis. While these results are already commendable, the remaining difference between 100% precision and the achieved precision for each model represents potential financial losses due to misclassified defaulters (False Positives), which could result in bad

debts. It is precisely these losses that the ensemble aims to mitigate through its optimized performance.

Table 3. Performance metrics for MLP models in the training and validation datasets.

	Training		Validation	
Modelo	**Precision**	**Accuracy**	**Precision**	**Accuracy**
Model 1	82.77%	81.86%	76.73%	75.27%
Model 2	82.42%	81.82%	76.41%	75.42%
Model 3	82.44%	81.72%	76.57%	75.37%
Model 4	82.64%	81.96%	76.44%	75.34%
Model 5	81.39%	81.39%	74.84%	74.84%

The outputs of each optimization scenario, evaluated on the validation dataset, are presented in **Table 4**. It is crucial to emphasize that the development of the individual models and the subsequent ensemble optimization process were conducted without any prior exposure to the final test dataset, ensuring an unbiased evaluation of the ensemble's performance.

Table 4. Performance metrics for MLP models ensembles in the validation dataset.

	Validation		Test	
	Prec.	**Accur.**	**Prec.**	**Accur.**
Scenario 1	76.74%	75.33%	80.57%	77.64%
Scenario 2	76.74%	75.33%	80.57%	77.64%
Scenario 3	85.07%	60.69%	88.81%	56.01%
Scenario 4	85.07%	60.69%	88.81%	56.01%

Table 5 presents the final results of both the individual models and the optimized ensembles on the unseen test dataset. Model 5 exhibited degenerate behavior, predicting all instances as performing loans and thereby achieving no discriminative power (AUC = 0.500), as discussed in Section 5. A key observation concerns the impact of the classification threshold on the economic performance of the ensemble. Scenarios 3 and 4, operating under a threshold of 0.8, achieved higher precision but at the cost of substantially lower accuracy, as a large volume of creditworthy borrowers was incorrectly rejected. This overly conservative approval policy resulted in a loss-generating portfolio, as the foregone interest revenue from rejected good payers outweighed the savings from avoided defaults, yielding a negative Expected Profit per loan.

Among the profitable scenarios, those operating under a threshold of 0.5, the precision achieved by the optimized ensembles is equivalent to that of the best-performing individual model. This indicates that the ensemble optimization process did not yield a direct improvement in precision over the strongest standalone model. However, the optimization led to marginal gains in accuracy and equal Expected Profit per loan, suggesting that while the ensemble did not surpass the best individual model in its primary optimization metric, it produced a slightly more balanced outcome across the full set of evaluation criteria.

Further analysis of the optimized weights reveals that Scenarios 1 and 2 exhibited a strong concentration of weight on Model 1, which aligns with Model 1's superior individual performance, suggesting that the genetic algorithm converged on a solution that primarily leveraged the strengths of the already high-performing base learner rather than discovering synergistic combinations across the ensemble. Scenarios 3 and 4, in contrast, concentrated weights more strongly on Model 3, the second-best individual model in terms of precision, a combination that produced a significant gain in precision but imposed an excessively restrictive approval policy, reducing the volume of approved loans, impairing accuracy, and ultimately destroying economic value.

The adjustment of the classification threshold led to enhanced precision but did not translate into economic improvement. On the contrary, Scenarios 3 and 4 demonstrate that threshold-driven precision maximization can destroy portfolio value when the resulting approval rate becomes too restrictive, underscoring the importance of aligning the optimization metric with the financial objectives of the lending institution.

The genetic algorithm optimized weights, including the concentration observed in scenarios 1 and 2, and the limited variation between scenarios 3 and 4, are presented in **Table 9**.

Figure 6 presents the Precision-Recall curves for both the individual MLP models and the optimized ensemble scenarios, evaluated on the test dataset. Unlike the ROC curve, the Precision-Recall curve is particularly informative in the context of imbalanced datasets, as it focuses on the performance of the model with respect to the minority class, in this case, defaulting borrowers, without being influenced by the large number of true negatives.

Among the individual models, Models 1, 2, and 3 exhibit nearly identical curves, with Average Precision (AP) scores ranging from 0.8872 to 0.8879, providing further visual evidence of the low predictive diversity among the base learners. Model 4, in contrast, shows a notably inferior curve (AP = 0.7857), consistent with its weaker individual performance reported in Table 5 and with the negligible weight assigned to it by the genetic algorithm across all scenarios (**Table 9**).

The ensemble panel reveals that all four optimization scenarios collapsed onto a single, indistinguishable Precision-Recall curve (AP = 0.8872), regardless of the classification threshold or genetic algorithm configuration employed. Crucially, this curve is not superior to the best individual model, it is virtually identical to the curve of Model 3, the strongest base learner. This finding provides direct visual confirmation that the genetic algorithm converged on a solution that replicates the best individual component rather than discovering synergistic combinations across the ensemble. Taken together with the high pairwise error correlations (0.762–0.918) and co-occurring error rates (79.43%–93.93%) reported in **Figure 5**, these results establish a consistent and mutually reinforcing body of evidence: in the absence of conceptual diversity among base learners, ensemble optimization collapses onto the strongest individual model, yielding no discriminative gain regardless of the optimization effort invested.

Figure 3 presents the calibration curves for both individual models and ensemble scenarios, evaluated on the test dataset. Calibration and discrimination are distinct and complementary properties of predictive models, while discrimination, measured by AUC, quantifies the model's ability to rank

Table 5. Confusion Matrices and Performance metrics for MLP models and ensembles in the test dataset.

Model		Confusion Matrix		Precision	Recall	Accuracy	F1	AUC	Brier	EP
Optim.		0	1							
Model 1	0	9,582	47,891	80.57%	94.24%	77.62%	86.87%	0.6897	0.1594	2,162
	1	12,142	198,591							
Model 2	0	7,821	49,652	80.21%	95.53%	77.97%	87.20%	0.6909	0.1572	2,164
	1	9,430	201,303							
Model 3	0	9,527	47,946	80.53%	94.11%	77.49%	86.79%	0.6898	0.1593	2,146
	1	12,414	198,319							
Model 4	0	8,162	49,311	80.28%	95.24%	77.88%	87.12%	0.6861	0.1596	2,160
	1	10,029	200,704							
Model 5	0	0	57,473	78.57%	100.00%	78.57%	88.00%	0.5000	0.1692	2,065
	1	0	210,733							
Scenario 1	0	9,579	47,894	80.57%	94.27%	77.64%	86.89%	0.6902	0.1592	2,164
	1	12,074	198,659							
Scenario 2	0	9,578	47,895	80.57%	94.27%	77.64%	86.89%	0.6902	0.1592	2,164
	1	12,074	198,659							
Scenario 3	0	44,101	13,372	88.81%	50.35%	56.01%	64.27%	0.6899	0.1593	-1,533
	1	104,625	106,108							
Scenario 4	0	44,101	13,372	88.81%	50.35%	56.01%	64.27%	0.6899	0.1592	-1,532
	1	104,623	106,110							

Note: F1: F1-score; AUC: Area Under the ROC Curve; Brier: Brier Score; EP: Expected Profit per loan.

borrowers by default risk, calibration assesses whether the predicted probabilities accurately reflect the empirical frequencies of the positive class [Guo *et al*., 2017; Van Calster *et al*., 2019]. A model can exhibit strong discriminatory capacity while remaining poorly calibrated, and vice versa, as these properties are orthogonal by nature. In the low-probability region (0.0-0.3), all individual models systematically underestimate the true probability of performing loans, the calibration curves lie above the diagonal, indicating that the observed fraction of positive outcomes substantially exceeds the predicted probabilities.

This underestimation is consistent with the temporal drift in default rates across the dataset splits: the model was trained on a period with lower default rates (18.6%), leading it to assign conservative repayment probabilities to borderline cases. When applied to the test set — covering the most recent period (2017–2018), which exhibits a higher default rate (21.4%) — the observed fraction of performing loans in the low-probability region substantially exceeds the model's predictions, manifesting as underconfidence [Guo *et al*., 2017].

In the high-probability region (0.6-1.0), all models converge toward the diagonal, indicating well-calibrated predictions for borrowers with strong repayment signals. Notably, Model 3, despite exhibiting the greatest deviation from perfect calibration in the low-probability region, is the strongest discriminator — a result consistent with the orthogonality between calibration and discrimination. The ensemble scenarios under threshold 0.5 collapse onto a single calibration curve, further corroborating the finding that the optimizer converged on a single component, while Scenarios 3 and 4, operating under threshold 0.8, exhibit greater deviation in the low-probability region, consistent with the drift.

The Brier Scores, reported in **Table 5**, are consistent with the calibration curves: since the underconfidence is concentrated in the low-probability region — which represents a small fraction of observations given the dominance of performing loans (80.2%) — its impact on the overall Brier Score is limited, explaining the similar scores observed across models (0.159). Model 5, with a Brier Score of 0.169, is the sole exception, as its degenerate behavior of predicting all instances as performing loans introduces systematic errors across the entire probability space.

To formally investigate the conditions under which ensemble weighting yields synergistic gains versus collapsing onto the best individual model, two complementary analyses were conducted. First, the pairwise error correlation between base learners was computed as an empirical measure of predictive diversity within the ensemble, following the framework established by Kuncheva and Whitaker [2003]. The results, presented in Figure 5, reveal consistently high correlations across all model pairs, ranging from 0.762 to 0.918. The co-occurring error analysis further corroborates this finding, showing that between 79.43% and 93.93% of errors are shared across model pairs — that is, when one model misclassifies an instance, the remaining models are highly likely to commit the same error. This indicates that the base learners, despite their architectural differences in depth, width, and batch size, fail on nearly identical instances and therefore offer no complementary predictive signal that the ensemble could exploit. Second, the weight concentration behavior of the genetic algorithm was examined across all four optimization scenarios, as reported in **Table** 9. In every scenario — regardless of

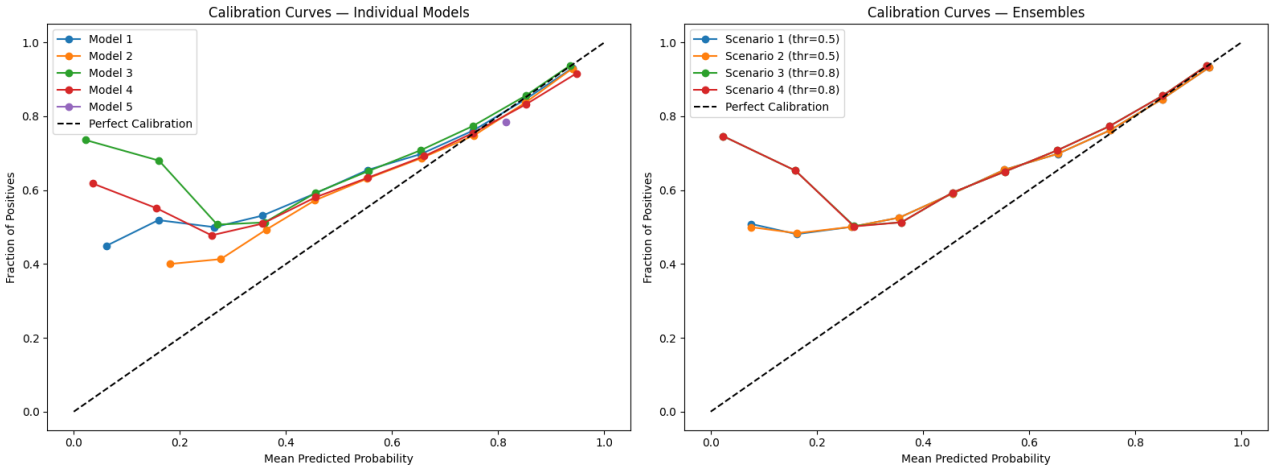

Figure 3. Panel A: Calibration curves - individual models.
Panel B: Calibration curves - ensembles.

classification threshold (0.5 or 0.8) or population size (200 or 1,000 individuals) — the optimizer converged on allocating approximately 100% of the ensemble weight to Models 1 and 3, the best-performing individual learners. This convergence behavior is not a failure of the optimization process; on the contrary, it reflects the algorithm operating correctly in the absence of combinatorial gain. When base learners share the same underlying learning algorithm and exhibit highly correlated errors, no weighting scheme — regardless of its sophistication or the computational effort invested — can produce synergistic improvement. The optimizer therefore logically concentrates the weights on the strongest available component. Taken together, these two analyses establish a consistent and mutually reinforcing body of empirical evidence: high error correlation among base learners is both a diagnostic indicator of insufficient diversity and a sufficient condition for weight concentration in optimized ensembles. This finding provides a formal empirical basis for the boundary condition identified in this study, contributing to a deeper understanding of when ensemble optimization adds value in credit risk classification tasks.

The findings show that, despite individual MLP models exhibiting competitive precision scores, ranging from 78.57% to 80.57%, the ensemble approach, even when augmented with genetic optimization, did not yield a precision score superior to that of the best-performing individual model among the economically viable scenarios, i.e., those operating under a classification threshold of 0.5. Scenarios operating under a threshold of 0.8, while achieving higher precision, resulted in negative expected profit per loan and are therefore not considered practically deployable in a real-world credit granting context.

To validate these findings, the models were re-trained and another test was performed ensembling only the three individual architectures with the worst performance. The results are presented in the next subsection.

4.1 Robustness Check

The models used to check the robustness of the findings were the models 2, 4 and 5, presented in **Table 2**. The results of re-trained models with the training and validation datasets are presented in **Table 6**. The results demonstrate the same good performance for individual models, correctly classifying 76% the good payers (precision) in the validation dataset.

Table 6. Robustness check - performance metrics for MLP models in the training and validation datasets.

	Training		Validation	
Modelo	**Precision**	**Accuracy**	**Precision**	**Accuracy**
Model 2	82.40%	81.81%	76.31%	75.43%
Model 4	82.51%	81.93%	76.32%	75.38%
Model 5	82.22%	81.79%	75.90%	75.33%

The same four distinct optimization scenarios were developed. The outputs of each optimization scenario in this robustness check are presented in **Table 7**.

Table 7. Robustness check - performance metrics for MLP models ensembles in the validation dataset.

	Validation		Test	
	Prec.	**Accur.**	**Prec.**	**Accur.**
Scenario 1	76.33%	75.42%	80.15%	77.99%
Scenario 2	76.33%	75.43%	80.15%	77.99%
Scenario 3	84.90%	60.79%	88.32%	57.77%
Scenario 4	84.90%	60.79%	88.32%	57.78%

Table 8 presents the final results for individual models and the optimized ensembles on the unseen test dataset, for both the original analysis and the robustness check round. The maximum precision achieved by the optimized ensembles in both the original test and the robustness check was equivalent to that of the best-performing individual model in the economically viable scenarios — those operating under a classification threshold of 0.5 — where a strong concentration of weight on the most precise base learner was observed. This indicates that the ensemble optimization process, using only MLPs as base learners, did not yield any precision improvement over

the strongest standalone model under practical deployment conditions. Ensembles operating under a threshold of 0.8 achieved higher precision; however, this came at the cost of substantially lower accuracy and negative Expected Profit per loan, rendering these scenarios economically unviable. In the robustness check, this precision gain was attained by distributing weights across the two best-performing individual models rather than concentrating entirely on a single component. This result underscores that precision maximization in isolation does not constitute a sufficient optimization criterion for credit risk applications, as overly restrictive approval policies can destroy portfolio value by rejecting a large volume of creditworthy borrowers

The same results were verified about the weights concentration and the benefit of a different threshold in the original test and in the robustness check round (see **Table 9**).

The robustness check confirms that individual MLP models exhibit competitive precision scores, however, the ensemble approach, when applied in individual MLP learners with genetic optimization, did not improve the precision score. Interestingly, while the optimization process was designed to maximize precision, it led to a slight improvement in the ensemble's overall accuracy. This effect was observed in both the main test scenarios and the subsequent robustness check (**Fig. 4**).

The McNemar test results are presented in **Table 10**. For Scenarios 1 and 2, operating under a classification threshold of 0.5, the test identified statistically significant differences between Model 1 and the ensemble ($p = 0.020$ and $p = 0.021$, respectively). However, the direction of the discordance reveals that the ensemble marginally outperformed the best individual model, with $n_{10} > n_{01}$ (approximately 411 vs. 346 and 407 vs. 343 discordant cases, respectively). Despite statistical significance, attributable to the large sample size of 268,206 test instances, the practical magnitude of this difference is negligible, corresponding to fewer than 65 additional correct classifications out of 268,206 observations (0.024%), with no measurable impact on precision, AUC, or Expected Profit per loan. For Scenarios 3 and 4, operating under a threshold of 0.8, the test confirmed that the ensembles were significantly inferior to the best individual model ($n_{01} \gg n_{10}$: approximately 92,538 vs. 34,535 and 92,494 vs. 34,524), consistent with the substantially lower accuracy and negative Expected Profit per loan reported in **Table 5**.

5 Discussion

This research aimed to assess whether genetic algorithm optimization could enhance the standalone performance of artificial neural networks in credit risk assessment. The empirical results revealed a nuanced outcome: despite individual MLP models demonstrating competitive precision scores (ranging from 78.57% to 80.57%), the genetically optimized ensemble did not achieve a higher precision score than the best-performing individual model under economically viable conditions. This suggests that in contexts where base models already exhibit good performance and are exclusively MLPs, the optimization process tends to converge by heavily weighting the strongest existing component, rather than discovering novel synergistic combinations. The results were confirmed by a robustness check with three MLP models.

Although the MLP architectures varied in terms of layers, neurons, and batch size, the underlying learning algorithm, backpropagation with the Adam optimizer, remained the same. This configuration likely resulted in conceptually similar models, leading to low diversity within the ensemble. The pairwise error correlation analysis corroborated this interpretation, revealing consistently high correlations across all model pairs (0.762–0.918), with co-occurring error rates between 79.43% and 93.93%, providing direct empirical evidence that the base learners fail on nearly identical instances and therefore offer no complementary predictive signal for the ensemble to exploit. Consequently, an absence of synergistic diversity was observed. This should not be viewed as a failure of the optimizer; on the contrary, its behavior was logical, in the absence of any combinatorial gain, the algorithm correctly concentrated the weights on the best-performing individual model.

The McNemar test results provide formal statistical grounding for these empirical observations. For the economically viable scenarios (threshold = 0.5), the statistically significant result must be interpreted in the context of practical significance: with a test set of 268,206 instances, even trivial differences in classification behavior can yield significant p-values [Dietterich, 1998]. The marginal advantage of the ensemble over the best individual model, fewer than 65 discordant instances, does not translate into any measurable gain in precision, AUC, Brier Score, or Expected Profit per loan, reinforcing the conclusion that the optimization process converged on the strongest individual component rather than discovering synergistic combinations. For the scenarios operating under a threshold of 0.8, the test formally confirms what the economic analysis already suggested: the ensemble's aggressive approval restriction, driven by weight concentration on Model 3, resulted in a substantially higher number of misclassifications relative to the best individual model, ultimately destroying portfolio value. Taken together, these results formally support the alternative hypothesis H_1: under economically viable conditions, the genetically-optimized homogeneous ensemble does not yield a practically meaningful improvement in predictive performance over the best-performing individual MLP model.

It is worth noting that precision, while effective at minimizing false positives, does not incorporate the economic consequences of false negatives, creditworthy borrowers incorrectly rejected. As demonstrated by the cost-sensitive analysis, scenarios that maximized precision through a stricter classification threshold yielded negative Expected Profit per loan, highlighting a fundamental tension between statistical optimization metrics and financial objectives. This suggests that profit-based metrics, such as the Expected Profit per loan [Verbraken *et al.*, 2014], may constitute more appropriate optimization targets in credit risk applications where the asymmetric costs of misclassification are economically significant.

Our results align with existing literature. They are consistent with Nanni and Lumini [2009], who found that MLP-based ensembles exhibited the least improvement over their stand-alone counterparts. Similarly, Tsai and Wu [2007]

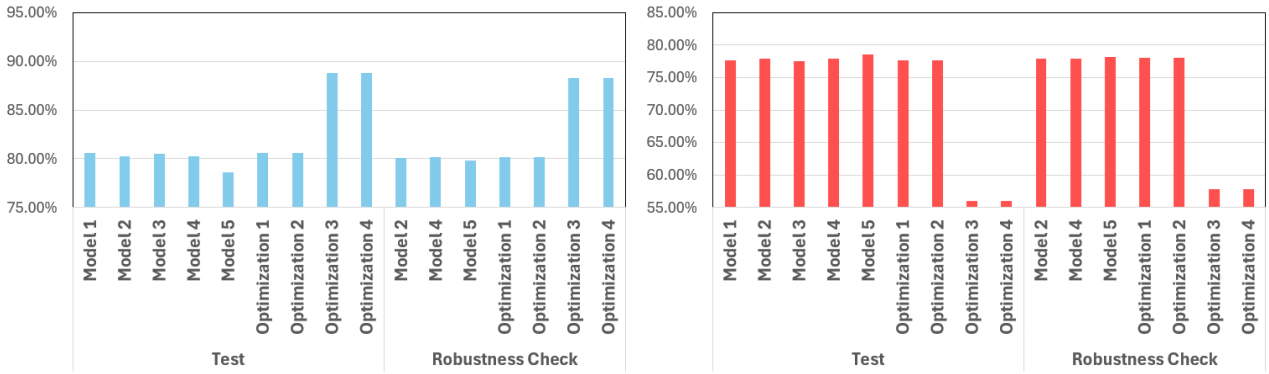

Figure 4. Panel A: Precision score of individual classifiers and ensembles, both in test and robustness check. Panel B: Accuracy score of individual classifiers and ensembles, both in test and robustness check.

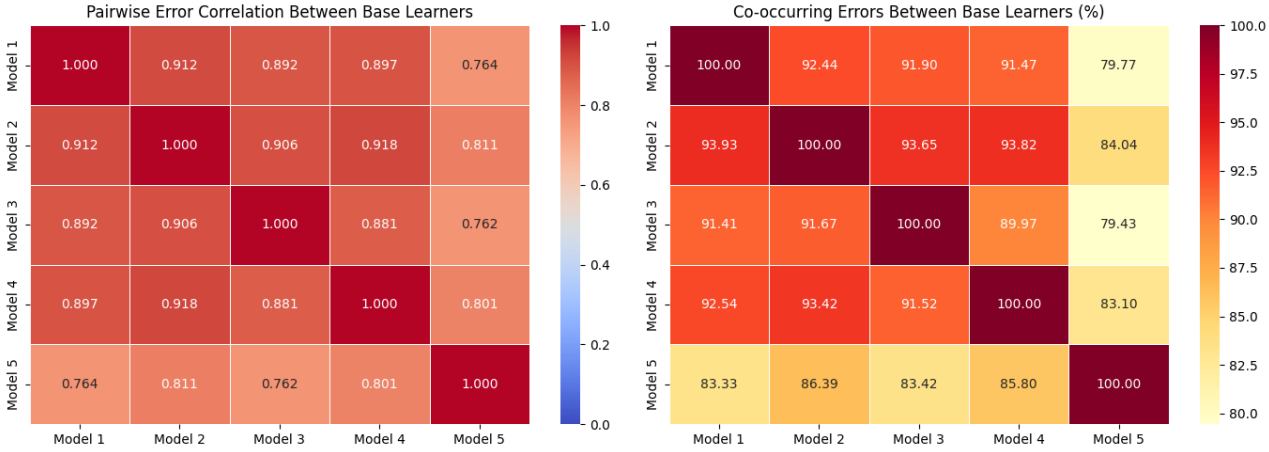

Figure 5. Panel A: Pairwise error correlation between baselearners. Panel B: Co-occurring errors between base learners (%).

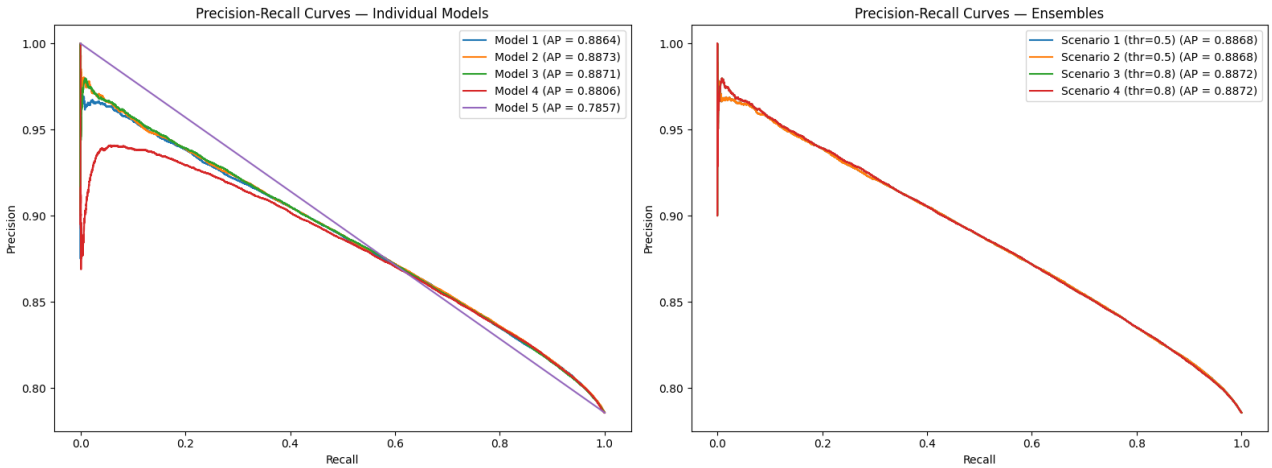

Figure 6. Panel A: Precision-Recall curves - individual models. Panel B: Precision-Recall curves - ensembles.

Table 8. Robustness check - Confusion Matrices and Performance metrics for MLP models and ensembles in the test dataset.

Model		Confusion Matrix		Precision	Accuracy	Original	Original	EP
Optim.		0	1			Precision	Accuracy	
Model 2	0	7,432	50,041	80.12%	77.96%	80.21%	77.97%	2,154
	1	9,072	201,661					
Model 4	0	7,569	49,904	80.15%	77.97%	80.28%	77.88%	2,158
	1	9,189	201,544					
Model 5	0	5,880	51,593	79.81%	78.23%	78.57%	78.57%	2,151
	1	6,785	203,948					
Scenario 1	0	7,527	49,946	80.15%	77.99%	80.57%	77.64%	2,159
	1	9,088	201,645					
Scenario 2	0	7,521	49,952	80.15%	77.99%	80.57%	77.64%	2,160
	1	9,070	201,663					
Scenario 3	0	42,616	14,857	88.32%	57.77%	88.81%	56.01%	-965
	1	98,395	112,338					
Scenario 4	0	42,611	14,862	88.32%	57.78%	88.81%	56.01%	-965
	1	98,384	112,349					

Note: Original Precision and Accuracy refer to the results obtained in the full model set (Table 5), reported for comparison purposes. EP: Expected Profit per loan.

Table 9. Optimized Weights for Scenarios 1 to 4 - original test and robustness check

Optimization	Model 1	Model 2	Model 3	Model 4	Model 5
Original test					
Scenario 1	91.20%	0.02%	8.53%	0.25%	0.00%
Scenario 2	91.25%	0.16%	8.38%	0.22%	0.00%
Scenario 3	1.13%	0.02%	98.62%	0.21%	0.01%
Scenario 4	1.13%	0.07%	98.61%	0.18%	0.00%
Robustness check					
Scenario 1	-	9.77%	-	90.22%	0.00%
Scenario 2	-	8.25%	-	90.74%	1.01%
Scenario 3	-	45.00%	-	54.88%	0.12%
Scenario 4	-	45.48%	-	54.29%	0.13%

Table 10. McNemar test results: Model 1 vs. optimized ensemble scenarios.

Comparison	n_{01}	n_{10}	Statistic	p-value
Scenario 1	346	411	346.0	0.0200
Scenario 2	343	407	343.0	0.0214
Scenario 3	92,538	34,535	34,535.0	0.0000
Scenario 4	92,494	34,524	34,524.0	0.0000

Note: n_{01}: instances correctly classified by Model 1 but misclassified by the ensemble; n_{10}: instances correctly classified by the ensemble but misclassified by Model 1. Exact binomial test used for Scenarios 1/2; asymptotic test used for Scenarios 3/4. Significance with *p-value* < *0.05*.

reported that well-structured individual MLP models can achieve better accuracy than ensembles. A key point of divergence from Tsai and Wu [2007], however, lies with Type I errors: our ensemble also showed no improvement in this metric compared to the best individual model. Furthermore, the findings underscore the inherent potential of ensemble methods when augmented by evolutionary optimization techniques, particularly in critical financial applications where highly accurate risk prediction is paramount, while simultaneously revealing the boundary conditions under which such gains are unattainable.

A critical consideration for the adopted optimized ensemble model pertains to its interpretability and explainability. Given that the rationale for the chosen weights is primarily derived from overall model performance, without deeper insight into their individual contributions, the model's decisions may be susceptible to bias and prove challenging to articulate to customers whose credit applications are denied. Furthermore, the inherent class imbalance (defaulters versus non-defaulters) represents a pervasive characteristic of credit risk datasets, posing a continuous challenge for robust modeling. Additionally, the dynamic nature of economic conditions can significantly alter default patterns, necessitating continuous model adaptation. A limitation of this study concerns the absence of confidence intervals or measures of variability around the reported performance metrics. Given the stochastic nature of the genetic algorithm, whose results may vary across runs due to random initialization and evolutionary operators such as crossover and mutation, multiple

independent executions would be recommended to assess the stability of the optimized weights and the robustness of the reported precision scores. Future work should consider reporting results as means and standard deviations across multiple runs to strengthen the statistical validity of the findings.

Despite these observations, this investigation provides insights into the dynamics of evolutionary optimization when applied to complex machine learning models within domains characterized by specific challenges, such as credit risk assessment. In future studies, the same dataset may be evaluated with heterogeneous ensembles, empirically demonstrating that structural diversity alone, as explored in this work, is insufficient to yield performance gains, and that conceptual heterogeneity among base classifiers is a necessary condition for synergistic ensemble improvement. Additionally, experimentation with other datasets can establish new benchmarks for performance and robustness in this field, ultimately contributing to improved capital allocation and a substantial reduction in default losses for financial institutions.

Declarations

Authors' Contributions

RS contributed to the conception of this study, data curation, formal analysis, investigation and writing the original draft. LM performed the methodological proposition, supervision, validation and review this manuscript. All authors read and approved the final manuscript.

Competing interests

The authors declare that they have no competing interests.

Availability of data and materials

The datasets (and/or softwares) generated and/or analysed during the current study will be made upon request.

Acknowledgements

The authors gratefully acknowledge the reviewer for their insightful comments and valuable contributions to this research.

Funding

The authors also wish to thank CERC SA for their generous financial support, without which this work would not have been possible.

References

Abdou, H., Pointon, J., and El-Masry, A. (2008). Neural nets versus conventional techniques in credit scoring in egyptian banking. *Expert Systems with Applications*, 35(3):1275–1292. DOI: 10.1016/j.eswa.2007.08.030.

Abellán, J. and Mantas, C. J. (2014). Improving experimental studies about ensembles of classifiers for bankruptcy prediction and credit scoring. *Expert Systems with Applications*, 41(8):3825–3830. DOI: 10.1016/j.eswa.2013.12.003.

Abellán, J. and Castellano, J. G. (2017). A comparative study on base classifiers in ensemble methods for credit scoring. *Expert Systems with Applications*, 73:1–10. DOI: 10.1016/j.eswa.2016.12.020.

Al-Maari, A.-A., Abdulnabi, M., Nathan, Y., Ali, A., Ali, U., and Khan, M. (2025). Optimized credit card fraud detection leveraging ensemble machine learning methods. *Engineering, Technology & Applied Science Research*, 15(3):22287–22294. DOI: 10.48084/etasr.10287.

Altman, E. I. (1968). Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *The Journal of Finance*, 23(4):589–609. DOI: 10.1111/j.1540-6261.1968.tb00843.x.

Angelini, E., Di Tollo, G., and Roli, A. (2008). A neural network approach for credit risk evaluation. *The quarterly review of economics and finance*, 48(4):733–755. DOI: 10.1016/j.qref.2007.04.001.

Atiya, A. F. (2001). Bankruptcy prediction for credit risk using neural networks: A survey and new results. *IEEE Transactions on neural networks*, 12(4):929–935. DOI: 10.1109/72.935101.

Battiston, S., Puliga, M., Kaushik, R., Tasca, P., and Caldarelli, G. (2012). Debtrank: Too central to fail? financial networks, the fed and systemic risk. *Scientific reports*, 2:541. DOI: 10.1038/srep00541.

Bhuria, R., Gupta, S., Kaur, U., Bharany, S., Ur Rehman, A., Hussen, S., Tejani, G. G., and Jangir, P. (2025). Ensemble-based customer churn prediction in banking: a voting classifier approach for improved client retention using demographic and behavioral data. *Discover Sustainability*, 6(28). DOI: 10.1007/s43621-025-00807-8.

Breiman, L. (1996). Bagging predictors. *Machine learning*, 24:123–140. DOI: 10.1007/BF00058655.

Caouette, J. B., Altman, E. I., and Narayanan, P. (1998). *Managing credit risk: the next great financial challenge*. John Wiley & Sons. Book.

Chen, M.-C. and Huang, S.-H. (2003). Credit scoring and rejected instances reassigning through evolutionary computation techniques. *Expert Systems with Applications*, 24(4):433–441. DOI: 10.1016/S0957-4174(02)00191-4.

Chen, Z., Chen, W., and Shi, Y. (2020). Ensemble learning with label proportions for bankruptcy prediction. *Expert Systems with Applications*, 146:113155. DOI: 10.1016/j.eswa.2019.113155.

Chi Guotai, M. Z. A. and Moula, F. (2017). Modeling credit approval data with neural networks: an experimental investigation and optimization*. *Journal of Business Economics and Management*, 18(2):224–240. DOI: 10.3846/16111699.2017.1280844.

Crook, J., Edelman, D., and Thomas, L. (2007). Recent developments in consumer credit risk assessment. *European Journal of Operational Research*, 183(3):1447–1465. DOI: 10.1016/j.ejor.2006.09.100.

Dietterich, T. G. (1998). Approximate statistical tests for comparing supervised classification learning algorithms. *Neural Computation*, 10(7):1895–1923. DOI: 10.1162/089976698300017197.

Duffie, D. and Singleton, K. J. (2003). *Credit Risk: Pricing, Measurement, and Management*. Princeton University

Press. DOI: 10.1515/9781400829170.

Fagerland, M. W., Lydersen, S., and Laake, P. (2013). The mcnemar test for binary matched-pairs data: mid-p and asymptotic are better than exact conditional. *BMC Medical Research Methodology*, 13:91. DOI: 10.1186/1471-2288-13-91.

Gandomi, A. H. and Haider, M. (2015). Beyond the hype: Big data concepts, methods, and analytics. *International Journal of Information Management*, 35(2):137–144. DOI: 10.1016/j.ijinfomgt.2014.10.007.

Ghatge, A. and Halkarnikar, P. (2013). Ensemble neural network strategy for predicting credit default evaluation. *International Journal of Engineering and Innovative Technology (IJEIT)*, 2(7):223–225. Avaialble at:`https://scispace.com/pdf/ensemble-neural-network-strategy-for-predicting-credit-552ojfm1db.pdf`.

Goldberg, D. E. (1989). *Genetic Algorithms in Search, Optimization and Machine Learning*. Addison-Wesley, Reading, MA, USA. DOI: 10.5860/choice.27-0936.

Guo, C., Pleiss, G., Sun, Y., and Weinberger, K. Q. (2017). On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70, pages 1321–1330. DOI: 10.48550/arxiv.1706.04599.

Haykin, S. (1999). *Neural Networks: A Comprehensive Foundation*. Prentice Hall. Book.

Huang, M.-C., Chen, M.-H., Hsu, C.-H., Chen, P.-C., and Wu, D.-M. (2004). Credit rating analysis with support vector machines and neural networks: a market comparative study. *Decision Support Systems*, 37(4):543–558. DOI: 10.1016/S0167-9236(03)00086-1.

Kingma, D. P. and Ba, J. (2015). Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*. DOI: 10.48550/arXiv.1412.6980.

Kuncheva, L. I. and Rodríguez, J. J. (2007). A weighted voting framework for ensembles of classifiers. *Journal of Artificial Intelligence Research*, 30:691–717. DOI: 10.1007/s10115-012-0586-6.

Kuncheva, L. I. and Whitaker, C. J. (2003). Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy. *Machine learning*, 51(2):181–207. DOI: 10.1023/a:1022859003006.

Lee, T.-H., Chiu, C.-S., Chou, P.-H., and Lu, C.-C. (2006). Credit scoring using the hybrid neural discriminant technique. *Expert Systems with Applications*, 30(4):773–782. DOI: 10.1016/S0957-4174(02)00044-1.

Lessmann, S., Baesens, B., Seow, H.-V., and Thomas, L. C. (2015). Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research. *European Journal of Operational Research*, 247(1):124–136. DOI: 10.1016/j.ejor.2015.05.030.

Li, W., Ding, S., Chen, Y., and Yang, S. (2018). Heterogeneous ensemble for default prediction of peer-to-peer lending in china. *IEEE Access*, 6:54396–54406. DOI: 10.1109/ACCESS.2018.2810864.

Liu, Y., Baals, L. J., Osterrieder, J., and Hadji-Misheva, B. (2024). Leveraging network topology for credit risk assessment in p2p lending: A comparative study under the lens of machine learning. *Expert Systems with Applications*, 252:124100. DOI: 10.1016/j.eswa.2024.124100.

Louzada, F. and Ara, A. (2016). Classification methods applied to credit scoring: Systematic review and overall comparison. *Expert Systems with Applications*. DOI: 10.1016/j.sorms.2016.10.001.

McNemar, Q. (1947). Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2):153–157. DOI: 10.1007/BF02295996.

Nanni, L. and Lumini, A. (2009). An experimental comparison of ensemble of classifiers for bankruptcy prediction and credit scoring. *Expert Systems with Applications*, 36(2, Part 2):3028–3033. DOI: 10.1016/j.eswa.2008.01.018.

Rokach, L. (2010). Ensemble-based classifiers. *Artificial Intelligence Review*, 33(1-2):1–39. DOI: 10.1007/s10462-009-9124-7.

Sadhwani, A., Giesecke, K., and Sirignano, J. (2020). Deep Learning for Mortgage Risk*. *Journal of Financial Econometrics*, 19(2):313–368. DOI: 10.1093/jjfinec/nbaa025.

Saunders, A. and Cornett, M. M. (2018). *Financial Institutions Management: A Risk Management Approach*. McGraw-Hill Education, 9th edition. Book.

Tsai, C.-F. and Wu, J.-C. (2007). Using neural network ensembles for bankruptcy prediction and credit scoring. *Expert Systems with Applications*, 34(4):2639–2649. DOI: 10.1016/j.eswa.2007.05.019.

Van Calster, B., McLernon, D. J., van Smeden, M., Wynants, L., and Steyerberg, E. W. (2019). Calibration: the achilles heel of predictive analytics. *BMC Medicine*, 17:230. DOI: 10.1186/s12916-019-1466-7.

Verbraken, T., Bravo, C., Weber, R., and Baesens, B. (2014). Development and application of consumer credit scoring models using profit-based classification measures. *European Journal of Operational Research*, 238(2):505–513. DOI: 10.1016/j.ejor.2014.04.001.

West, D. (2000). Neural network credit scoring models. *Computers & Operations Research*, 27(11-12):1131–1152. DOI: 10.1016/S0305-0548(99)00149-5.

Zhao, Z., Xu, S., Kang, B. H., Kabir, M. M. J., Liu, Y., and Wasinger, R. (2015). Investigation and improvement of multi-layer perceptron neural networks for credit scoring. *Expert Systems with Applications*, 42(7):3508–3516. DOI: 10.1016/j.eswa.2014.12.006.

Zhou, Z.-H. (2012). *Ensemble Methods: Foundations and Algorithms*. CRC Press. DOI: 10.1201/b12207.