Advancing credit scoring models: integrating explainable AI for fair and transparent financial decision-making

Dingyuan Liu[a], Feng Feng[*]

[a]liudingyuan5908@163.com, [*]feng_f@nxu.edu.cn

Ningxia University, 489 Helan Mountain West Road,Yinchuan City, Ningxia

Abstract. In the context of the big data era, internet finance is flourishing. However, it is also confronted with an increasing number of credit risk challenges, posing significant threats to economic security. The global nature of the internet means the impact and scope of individual and corporate credit defaults have widened, making it crucial to mitigate credit risk for stable economic development. Today, with the relative maturity of machine learning technology, several popular credit scoring technological methods have emerged, which this article summarizes. However, given the heightened demands, especially in terms of interpretability, placed by many financial institutions and organizations in recent years, previous methods or models cannot be straightforwardly applied and need to be explained in conjunction with Explainable AI (XAI). This paper also proposes a fusion model that simultaneously considers both the accuracy of credit default prediction and interpretability, which can adapt to the current economic environment.

Keywords: internet finance, credit risk, economic security, machine learning, XAI

1 Introduction

In the evolving landscape of financial services, credit scoring remains a critical component, influencing decisions on loan approvals. The advent of machine learning (ML) and artificial intelligence (AI) has ushered in a new era of predictive analytics, offering enhanced accuracy in credit risk assessment. However, the opaque nature of these "black box" models raises significant concerns regarding interpretability and fairness, prompting a shift towards Explainable AI (XAI). This paper explores the intersection of ML and XAI within credit scoring, examining the efficacy of popular algorithms and the imperative of transparency. Additionally, we propose optimizations for the PermuteAttack framework, aiming to improve the interpretability and applicability of ML models in credit scoring. This investigation not only highlights the current state of credit scoring methodologies but also charts a course for future research, emphasizing the need for balance between predictive power and ethical considerations in financial modeling.

ECIT 2024, March 15-17, Changsha, People's Republic of China

DOI 10.4108/eai.15-3-2024.2346182

2 Popular methods in credit scoring

This chapter summarizes and synthesizes the popular prediction and evaluation methods previously used in the field of credit scoring.

2.1 Logistic regression

Logistic regression, as a common method in the field of credit scoring, is the most widely used technique for constructing scorecards[1]. Logistic regression is a multivariate technique that estimates the probability of an event occurrence or non-occurrence by predicting a binary dependent variable from a set of independent variables. The formula for logistic regression is shown in Equation (1).

$$y = \frac{1}{1+e^{-z(x)}} = \frac{1}{1+e^{-(w^T x+b)}} \tag{1}$$

y represents the predicted default probability of a customer, *x* represents the features of the customer across various dimensions, and b is the bias term.

Despite its wide application, there are some clear limitations when dealing with complex credit scoring issues. Firstly, the logistic regression model is based on the assumption that there is a linear relationship between the independent variables and the log odds. However, in the real-world financial context, the relationship between financial indicators and the probability of default is often nonlinear. This simplified linear assumption limits the model's ability to capture nonlinear relationships between variables, which may negatively affect the accuracy of predictions. [2] proposed the PLTR, which can capture nonlinear relationships that emerge and is also competitive in terms of performance. This provides new insights for the continued application of logistic regression in the field of credit scoring.

2.2 Random forest

The effectiveness of Random Forests in the field of credit scoring has been proven by numerous studies [3]. A Random Forest is an ensemble model composed of many decision trees, each constructed independently, forming a tree-based decision framework by setting class-conditional probabilities at the ends of the branches. Starting from the root node, each decision tree develops subtrees with internal nodes connected by branches, culminating in what are known as leaf nodes. Each internal node represents a test on a feature (for example, determining whether a borrower owns property), while branches represent binary splits of the features. Two main challenges arise in the construction of decision trees: (1) how to select the optimal splitting feature at each internal node, and (2) determining the depth of the tree, that is, deciding when to stop further splitting. When constructing a Random Forest model for credit scoring, we use a dataset containing customer features along with their credit scores or default records as a training base. Through the training process, the model reveals the correlations between features and credit scoring, thereby predicting the credit status or default risk of new clients. Despite its value in the field of credit scoring, the integration of multiple decision trees in a Random Forest leads to weak interpretability, and the complexity of explaining individual decision processes reduces the overall interpretability of the model.

Although Random Forests themselves have weak interpretability, this can be improved by employing specific interpretative tools and methods (such as feature importance scores, Local Interpretable Model-agnostic Explanations (LIME), or SHAP values) to enhance the model's transparency and interpretability.

2.3 Support vector machine

In [4], the authors conducted a validation of various models, with the results indicating that Support Vector Machines (SVM) are promising and outperform other methods. SVM aims to identify a hyperplane that can divide borrowers into two distinct categories, "good credit" and "bad credit," while ensuring that the distance (i.e., the margin) between data points closest to this decision boundary is maximized, the classification formula for SVM is given by Equation (2).

$$y = \begin{cases} +1, b + \alpha^T x \geq +1 \\ -1, b + \alpha^T x \leq -1 \end{cases} \tag{2}$$

y represents the predicted default probability of a customer, *x* represents the features of the customer across various dimensions, and b is the bias term.

This method allows SVM to precisely identify and differentiate borrowers of varying credit levels, thereby significantly enhancing the accuracy and reliability of credit scoring. By applying the concept of hyperplanes in a high-dimensional feature space to separate categories, SVM demonstrates its strength in handling complex classification issues. Additionally, SVM can utilize kernel functions to perform nonlinear transformations of data, mapping it into a higher-dimensional space where it can be more easily separated. In credit scoring, this means SVM can handle more complex data patterns and better recognize the credit ratings of borrowers. While SVM is effective in areas such as credit scoring, its black-box nature is a major drawback that discourages financial practitioners from using it [5]. [6] proposes a Hybrid Credit Scoring Model (HCSM), which addresses the black-box nature of SVM to a certain extent.

2.4 Boosting

Boosting, a strategy based on ensemble learning, is considered one of the most important achievements in the field of machine learning and has shown strong performance in the domain of credit scoring [7]. It constructs a superior composite classifier by aggregating multiple weak classifiers. During this process, it employs an iterative approach where the weights of the samples misclassified in the previous round are specifically increased, ensuring that these samples receive more attention in subsequent training iterations. As a result, with each iteration, the Boosting method gradually improves its ability to identify borrowers' samples that are difficult to differentiate, significantly enhancing the model's predictive accuracy in the complex scenarios of credit scoring and default prediction.The core formula of the Boosting as Equation (3).

$$F_m(x) = F_{m-1}(x) + \gamma h_m(x) \tag{3}$$

$F_{m\text{-}1}(x)$ represents the model after $m-1$ iterations, $h_m(x)$ is the new base learner added in the *m*-th iteration (typically a decision tree), γ is the learning rate, controlling the step size of each iteration.

The computation of $h_m(x)$ depends on the loss function. In regression problems, the mean squared error (MSE) is commonly used. Thus, in the Boosting algorithm, $h_m(x)$ is obtained by fitting the residuals of the current model $F_{m-1}(x)$. Specifically, $h_m(x)$ can be represented as Equation (4).

$$h_m(x) = \arg\min_h \sum_{i=1}^{n} L(y_i, F_{m-1}(x_i) + h(x_i)) \tag{4}$$

Here, $L(y_i, F_{m-1}(x_i) + h(x_i))$ is the loss function, usually $(y_i - F_{m-1}(x_i) - h(x_i))^2$.

Currently, AdaBoost, XGBoost, and LightGBM are among the most frequently used Boosting algorithms. AdaBoost is known for its simplicity and ease of use, while XGBoost and LightGBM are more efficient when processing large-scale data. However, they all share one critical drawback — compared to simple linear models, the Boosting models composed of many decision trees are more difficult to interpret. [8] notes that "the right to explanation" must be maintained for automated decisions. The ability to explain the outcomes of predictions is crucial for risk decisions, such as when the financial sector denies loan applications or the criminal justice system refuses or grants parole.

3 Explainable artificial intelligence (XAI) in cretid scoring

In financial applications such as credit scoring and default prediction, although efficient machine learning models can provide highly accurate forecasts, the "black box" nature of these models often results in a lack of transparency in the decision-making process. Therefore, the application of Explainable AI (XAI) technologies becomes key to addressing this deficiency, as they can offer understandable explanations for model decisions, increasing user trust in model predictions, while also meeting regulatory requirements. Current explainability techniques fall into two main categories: local and global methods. Local Interpretable Model-agnostic Explanations (LIME) and SHAP (Shapley Additive exPlanations) are commonly used XAI techniques in this context, which can employ feature importance for both local and global interpretability [9]. LIME approximates complex model predictions with locally interpretable models, helping to reveal why a particular borrower is predicted to be high risk or low risk, and which features most influence the model's decision. SHAP uses Shapley values from game theory to quantify each feature's contribution to the model's prediction, unveiling which customer attributes are most critical in assessing credit risk.

In addition to the two methods mentioned above, Counterfactual explanations have emerged as a new method for validating and explaining machine learning (ML) models in financial retail credit scoring in recent years. In past work, counterfactual reasoning has been used to provide local explanations, but it also has the capability to identify model decision boundaries within specific data neighborhoods [9]. Counterfactual explanations offer an intuitive way to understand model decisions by constructing scenarios of "what if not this, but that." In credit scoring, they can clearly point out which changes (such as increasing income or reducing debt) could improve a borrower's credit rating, offering concrete suggestions for credit improvement.

Indeed, counterfactual explanations hold great potential for enhancing model transparency and interpretability.

4 Proposed hybrid model for the era of intelligent risk control

The paper proposes a hybrid model framework based on the PermuteAttack algorithm proposed by [10]. This method combines both good predictive performance and a more comprehensive explanatory mechanism. The operational process is illustrated in Figure 1.

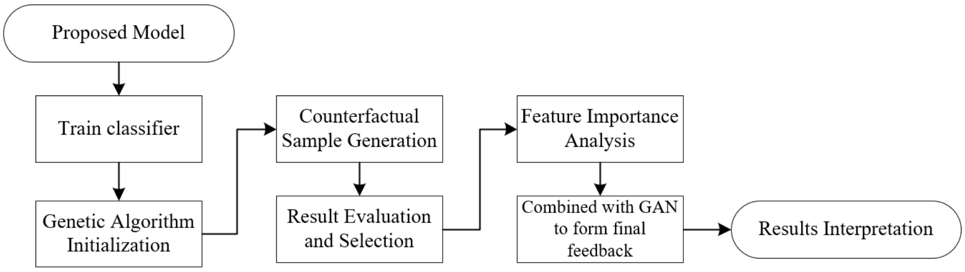
Proposed Model
Train classifier
Genetic Algorithm Initialization
Counterfactual Sample Generation
Result Evaluation and Selection
Feature Importance Analysis
Combined with GAN to form final feedback
Results Interpretation

Fig. 1. Proposed hybrid model flow chart.

1. Train classifier: Using CatBoost as the classifier for the predictive model and training it on the dataset German Credit, the specific operations will be mentioned later.

2. Genetic Algorithm Initialization: Define the genetic algorithm's parameters, {population size=30, number of mating parents=10, mutation probability, number of iterations=100,number of mutations=1.}.

3. Counterfactual Sample Generation: Use the PermuteAttack to attack a selected test sample, aiming to change the model's prediction by modifying feature values; in each iteration, using the genetic algorithm generates a new sample population through selection, crossover, and mutation;

4. Result Evaluation and Selection: Choose individuals with the highest fitness from multiple generations, which are considered successful counterfactual samples. Analyze these samples to determine which feature changes have the most significant impact on model predictions.

5. Feature Importance Analysis: Use SHAP values to assess the importance of each feature in the model; This step provides a preliminary metric to measure the contribution size of features to model predictions, that is, the greater the sum of absolute SHAP values for a feature, the greater its impact on the model's prediction outcome. Further processing is conducted for features that have been one-hot encoded. For each group of one-hot encoded features, the sum of SHAP values for all individual features within that group is calculated and observed.

6. Combined with GAN to form final feedback: Train a GAN using the feature data of "good" borrowers and "bad" borrowers respectively, to generate exemplary labels for "good" results and "bad" results as auxiliary explanatory illustrations.

In step 1, [10] selected random forest as the classifier; however, in fact, the classification performance of the random forest in this scenario is not optimal. This paper compares the classification performance of random forest, AdaBoost, XGBoost, LightGBM, and CatBoost models in this scenario.

Using the German dataset, the classifier aims to categorize loan applicants based on their information, where applicants classified as "good" are eligible for loans, while those classified as "bad" are denied loans. The dataset consists of a total of 1000 records with 20 features in total, including 13 categorical features (one-hot encoded). The dataset is split into training and testing sets with a 60-40 ratio. The ROC curve and AUC value of CatBoost are shown in Figure 2.

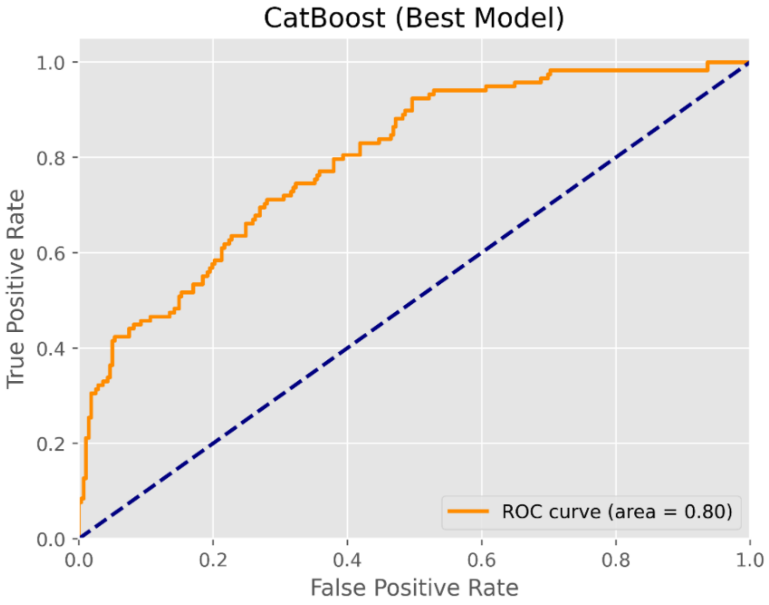

Fig. 2. ROC curve of CatBoost

The performance of different models is shown in Table 1, as evident from Figure 3, the data is highly imbalanced. Therefore, we utilize AUC as the primary evaluation metric. Furthermore, as indicated in the table, CatBoost outperforms Random Forest. Therefore, for the PermuteAttack framework, opting for the better-performing CatBoost to replace the original Random Forest is advisable, the corresponding optimal parameter is {depth:15, iterations:1000}.

Table 1. Performance of Different Models

Models	AUC
Random Forest	0.77
Decision Trees	0.59
Logistic Regression	0.75
LightGBM	0.77
AdaBoost	0.72
XGBoost	0.77
CatBoost	**0.80**

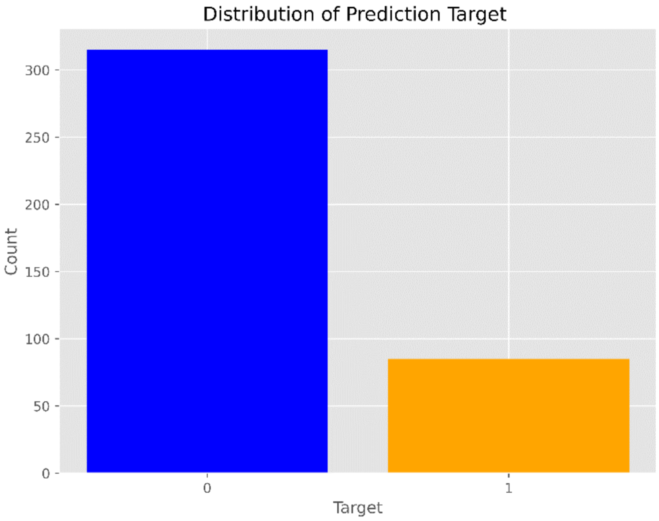

Fig. 3. The target distribution of the dataset

Secondly, only 49% of the samples generated in this framework are sufficiently similar to real samples, which is still not considered excellent in terms of success rate. We split the original dataset into two sets, "good" and "bad," and train GANs separately on each set. Then, we use the trained models to generate exemplary labels for "good" and "bad" instances. The aim is to combine these generated labels with the adversarial samples produced by previous algorithms to further enhance the interpretability and transparency of the hybrid model.

5 Conclusion

This paper has explored various dimensions of credit scoring, including the application of popular prediction methods, the integration of Explainable Artificial Intelligence (XAI).Furthermore, a hybrid model framework is proposed, which has better performance and interpretability, outperforming the method framework used in [10] .The framework of this paper employs CatBoost, and as shown in Table 2, when compared with the Random Forest used in [10], CatBoost exhibits superior performance on the German Credit dataset. It can also be inferred that CatBoost should outperform Random Forest in other similar, more balanced datasets. A comparison of the classifiers used by the two frameworks is shown in Table 2.

Table 2. Comparison of two different classifiers

Performance	CatBoost	Random Forest
ACC	0.7825	0.7750
F-score	0.5700	0.5161
Precision	0.7123	0.6761
Recall	0.4407	0.4174
Specificity	0.9255	0.9193
FPR	0.0745	0.0807

As can be seen, CatBoost's performance comprehensively outperforms Random Forest in this scenario. Besides, we also use GANs as an auxiliary explanation, which is better than the previous method in terms of the diversity and stability of the framework structure.

In conclusion, while significant progress has been made in the field of credit scoring through the adoption of advanced machine learning techniques and the integration of explainability frameworks, ongoing efforts are essential to address the challenges of interpretability and fairness. Future research should aim at enhancing the synergy between predictive accuracy and transparency, ensuring that credit scoring models are both effective and equitable.

Acknowledgments. Special thanks are extended to the Ningxia Key Research and Development Plan for the key project "Research and Application of Key Technologies for Dynamic Monitoring, Early Warning, and Emergency Rescue of Major Disasters in Coal Mines Based on Multi-source Sensing" (Project Number: 2022BEG02016), which significantly facilitated the research presented in this paper.

References

[1] Bolton, C.: Logistic Regression and Its Application in Credit Scoring. University of Pretoria, South Africa (2009)

[2] Elena, D.: Machine Learning for credit scoring: Improving logistic regression with non-linear decision-tree effects. Vol. 3,pp. 1178-1192. European Journal of Operational Research (2022)

[3] Germanno, T.: Comparative study of support vector machines and random forests machine learning algorithms on credit operation. Vol.12, pp. 2492-2500. Software: Practice and Experience(2021)

[4] Young Chan, L.: Application of support vector machines to corporate credit rating prediction. Vol. 33, pp. 67-74. Expert Syst. Appl(2007)

[5] Ri Young, G: Credit scoring: a review on support vector machines and metaheuristic approaches. Adv. Operat. Res. pp. 1-30(2019)

[6] Defu, Z.: A Hybrid Credit Scoring Model Based on Genetic Programming and Support Vector Machines. Vol. 7, pp. 8-12. 4th International Conference on Natural Computation (2008)

[7] Marceau, L.: A comparison of deep learning performances with other machine learning algorithms on credit scoring unbalanced data. arXiv preprint arXiv:1907.12363

[8] Paul, V.: The EU General Data Protection Regulation (GDPR). Cham:Springer International Publishing. pp. 70-85(2017)

[9] Andreas, C.: Machine learning interpretability for a stress scenario generation in credit scoring based on counterfactuals. Expert Systems with Applications. P.117271(2022)

[10] Masoud, H.: PermuteAttack: Counterfactual Explanation of Machine Learning Credit Scorecards. arXiv:2008.10138(2020)