



Achieving Performance Trade-offs in Imbalanced Credit Default Prediction

Guanzhi Li

Department of Economics, Shenzhen MSU-BIT University, Shenzhen, China
1120240810@smbu.edu.cn

Abstract. This study explores the challenge of achieving a balance in performance when predicting credit defaults on imbalanced datasets. In this paper, a probability-weighted ensemble model is proposed to integrate the advantages of the two basic models, i.e., logistic regression and random forests. It studied in depth the performance of several models on imbalanced datasets, analysing data pre-processing, features, and multi-dimensional metrics. Based on this, some optimisation approaches such as dynamic gating hybridisation and residual-enhanced hybridisation were proposed and verified the balancing performance by combining them with the balancing strategies such as SMOTE and random oversampling. Finally, the robustness of the results was validated by five-fold hierarchical cross-validation. The results show that, as far as recall is concerned, the best model is logistic regression, while as far as precision is concerned, the best model is random forests. The hybrid model did not offer a substantial improvement in any one metric but showed stable and good performance in the F1 score, a composite metric, indicating a superior capacity to manage the trade-off between precision and recall. Moreover, for the different optimisation approaches, complicated models and optimisation strategies did not lead to significantly improved balancing performance. In summary, the hybrid model's strength has never been in optimising a single metric, but in the synergistic optimisation of many metrics; blindly increasing model complexity cannot significantly boost the balancing capacity for unbalanced classification.

Keywords: Credit Default Prediction; Imbalanced Classification; Logistic Regression; Random Forest; Ensemble Learning

1 Introduction

Accurate credit default prediction is of great significance for risk management in the financial sector. Accurate identification of high-risk clients can not only reduce the default losses for financial institutions but also directly affect the efficiency of loan approval processes and the distribution of resources [1, 2]. In datasets, default

samples are usually a small part of the whole, resulting in severe class imbalance. This leads to models being biased towards the majority class, which makes it harder for them to detect clients who are not paying their bills [3, 4].

In this context, the primary issue in credit risk modelling is to improve default detection skills while managing the false positive rate. The major problem in credit risk modelling is to improve the identification of defaults and reduce false positives. Due to biases in model performance, one model is not able to optimise several measures. Hand presented a comprehensive review methodology instead of traditional metrics, highlighting that a thorough assessment must consider more than surface-level indicators to factor in the misclassification costs [4]. Thus, research has moved beyond enhancing single models to investigate more complex strategies. Mena et al. have investigated the accuracy-interpretability trade-off by utilising symbolic classifiers, showing that they may enhance prediction performance even in the presence of significant class imbalance [5]. The in-depth work of Ayari et al. points out the need to choose appropriate machine learning frameworks and evaluation methods for complex and imbalanced datasets in present-day financial settings [6].

Although Logistic Regression and Random Forests have their respective advantages in interpretability and accuracy, a single model is typically not able to satisfy several performance metrics simultaneously [7, 8]. This paper proposes a probability-weighted hybrid model based on the comparison of logistic regression and random forests and tests its performance using numerous criteria. The objective of this research is to address the trade-offs between several performance indicators and to validate the hybrid model's capability to achieve balanced performance in predicting credit defaults in imbalanced datasets.

2 Methodology

2.1 Data Source and Processing

All data utilised in this work is derived from the dataset on Kaggle. The data files were acquired via KaggleHub and processed by systematic data cleaning and preparation. The data file comprises clients' financial attributes and default status.

The default indicator is the dependent variable, and the rest are independent variables for modelling. Primary variables: Default (binary: 1=default, 0=no default); Balance (customer debt amount); Income (repayment capacity); Student (categorical identity variable).

The dataset has 10,000 observations with no missing values. The mean of default is 0.0333. Data pre-processing includes missing value check, data segmentation, and standardisation. No missing values were found, so no imputation was needed. Stratified sampling was used to split datasets, which can keep the default rates of the

training set and the test set to remain consistent and thereby improve the stability of prediction. Continuous variables 'income' and 'balance' were standardised (mean=0, variance=1) for the LR model.

2.2 Models

This study employs baseline models like logistic regression and random forest and constructs a weighted ensemble model combining them to maximize the performance trade-off. Previous comparative studies have also shown that more complex models, such as deep learning approaches, do not always outperform classical machine learning methods in structured financial data scenarios [9].

Logistic Regression. Logistic regression is a conventional linear classification model that has benefits like a straightforward structure, great computing efficiency, and strong interpretability, making it prevalent in credit scoring applications. The model's predictive logic is rooted in its ability to map a linear combination of features to a probability space. Specifically, given the feature vectors, $x \in \mathbb{R}^d$, the LR model is expressed as:

$$P(Y = 1|x) = \frac{1}{1 + e^{-(\beta_0 + \beta^T x)}} \quad (1)$$

In this context, $Y \in \{0,1\}$ denotes the classification outcome (where 1 represents default and 0 represents non-default), β_0 is the intercept term, and β is the vector of model parameters.

Random Forests. Random Forests are an ensemble learning method that uses decision trees. It trains multiple sub-models and combines them with a majority voting scheme in terms of prediction [8]. It shows strong non-linear modelling power and good generalisation performance. This model greatly reduces the chance of false positives in credit default prediction, thereby showing improved accuracy. For classification problems, the final prediction of an RF is typically obtained through a majority voting mechanism, expressed as:

$$\hat{y} = \text{mode}\{h_1(x), h_2(x), \dots, h_T(x)\} \quad (2)$$

Where, $h_t(x)$ denotes the prediction of the t-th decision tree, T represents the total number of trees, and mode indicates the majority voting operation.

Hybrid Model. In this paper, a weighted hybrid model is proposed to consider these performance differences. This model combines the advantages of both model types by

linearly weighting and combining the prediction probabilities from logistic regression and random forests, which are defined as:

$$P_{\text{hybrid}}(x) = \alpha P_{\text{LR}}(x) + (1 - \alpha) P_{\text{RF}}(x) \quad (3)$$

Where, $P_{\text{LR}}(x)$ represents the LR model's predicted default probability for sample x , $P_{\text{RF}}(x)$ represents the RF model's predicted probability, and $\alpha \in [0,1]$ controls the balance between the two models in the final prediction. Bitetto's Studies have shown that model fusion techniques can improve the overall performance for unbalanced classification problems to some degree [2]. The weighting factors can be adjusted to obtain a more desirable compromise among the various performance measures by balancing recall and precision.

Evaluation framework. In predicting credit default, the typical class imbalance in the data renders conventional accuracy metrics inadequate for a comprehensive assessment of model performance [5]. The study utilises a multi-faceted evaluation methodology to assess the model thoroughly in terms of accuracy, recall, and the F1-score. Precision pertains to the management of false positives, while recall focuses on the identification of actual default cases. In imbalanced data domains, a distinct trade-off typically exists between these two metrics, where enhancing recall may lead to an increase in false positives, while pursuing higher overall accuracy may elevate the likelihood of missed detections [4].

3 Result

3.1 Performance Analysis of Baseline Models

Fig. 1 illustrates that the ROC curve for the logistic regression model is predominantly situated at the top-left corner, with an AUC value of 0.95, signifying that the model exhibits robust classification proficiency and can adeptly differentiate between default and non-default samples. Fig. 1(a) illustrates that the ROC curve of the Random Forest model has commendable performance, with an AUC value of 0.89, signifying robust classification skills, but marginally inferior to those of the logistic regression model. At the same time, this trade-off is also seen in the PR curve in Fig. 1(b), where precision decreases dramatically as recall improves, leading to an average precision (AP) score of 0.52.

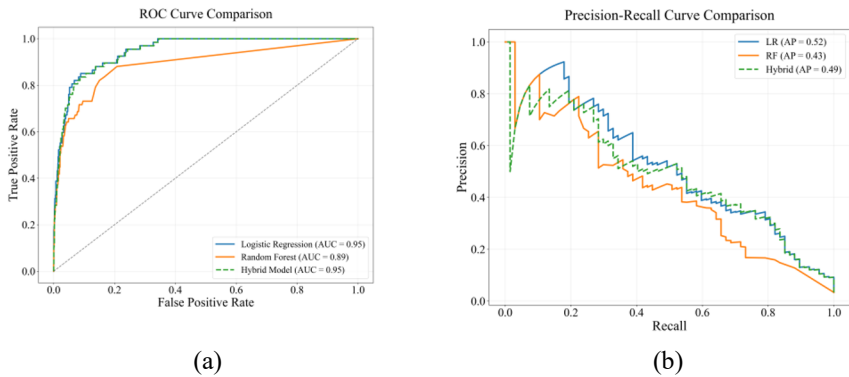


Fig. 1. Comparison of Model Performance: (a) ROC Curves and (b) Precision-Recall Curves for LR, RF, and Hybrid Models (Photo/Picture credit: Original).

The individual performance metrics are displayed in Table 1. Subsequent examination reveals that the model has a recall of 0.8806 for the default class (Class 1), signifying effective detection of default samples; however, its precision is merely 0.1730, suggesting a substantial false positive rate in default predictions. Random forest models, however, exhibit superior control over false positives. The model’s precision for the default class (Class 1) is 0.5429, significantly higher than that of the logistic regression model. It demonstrates the effectiveness of reducing false positives in predicting default samples. However, the recall is merely 0.2836, indicating that a significant number of default samples remain undetected.

Table 1. Performance Summary of Baseline Models for Class 1

Model	Precision	Recall	F1-score	AUC	AP
LR	0.1730	0.8806	0.2892	0.9500	0.5200
RF	0.5429	0.2836	0.3725	0.8900	0.3400

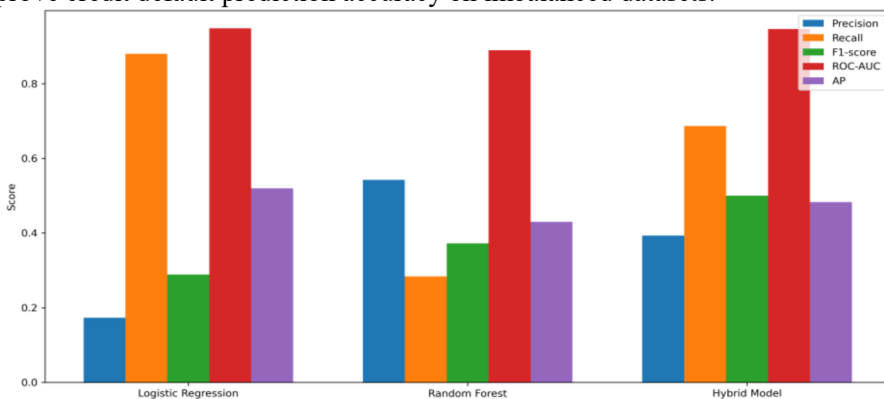
3.2 Hybrid Model Analysis

Since the performance of the logistic regression and random forest models is significantly different, in this study, a probability-weighted hybrid model is proposed, and its performance is thoroughly examined with different weighting combinations. As Table 2 shows, changing the weight settings has a big impact on the performance measurements of the model.

Table 2. Performance of the hybrid model under different weight settings

LR Weight	RF Weight	Precision	Recall	F1-score	AUC	AP
0.2	0.8	0.5106	0.3582	0.4211	0.9411	0.4616
0.4	0.6	0.4471	0.5672	0.5000	0.9454	0.4772
0.5	0.5	0.3932	0.6866	0.5000	0.9466	0.4834
0.6	0.4	0.3046	0.7910	0.4398	0.9475	0.4886
0.8	0.2	0.2135	0.8507	0.3413	0.9484	0.4940

The performance differences between the logistic regression and the random forest models may be seen. To study these discrepancies further, a probability-weighted hybrid model was built and tested with different weight combinations. As shown in Fig. 2, the hybrid model obtained the best F1-score, which represents a more balanced connection between precision and recall. The weighting experiments also show that the model works best when the weights of logistic regression and random forest are relatively balanced such as 0.4/0.6 or 0.5/0.5. That is, model ensembling is not simply averaging predictions, but involves balancing weights across multiple evaluation metrics. The benefit of the hybrid model is not in the optimisation of a single statistic but that it provides a more steady trade-off between numerous metrics. The overall results show that the ensemble strategy can blend the strengths of many models and improve credit default prediction accuracy on imbalanced datasets.

**Fig. 2.** Three-Model Comparison Chart (Photo/Picture credit: Original).

3.3 Model Comparison and Robustness Analysis

To explore the analysis further, additional models are introduced. The Gating-based Dynamic Hybrid (GDH) model adaptively modifies the weights using the gating method, and the Error-correction Residual Hybrid (ERH) model provides residual compensation. These strategies are evaluated within a single framework, allowing a

clear comparison of their effectiveness. The goal is to reach a performance equilibrium on unbalanced datasets.

The new proposed methods are the Dynamic Hybrid model and the Residual Boost Hybrid model (Table 3). The experimental results show that both methods can achieve some improvement on some specific metrics, but they are not better than the fixed-weight hybrid model in terms of comprehensive metrics, which indicates that the complex optimization strategy does not necessarily lead to a better global balance. These results show that more complicated model architectures or optimisation tactics do not always increase performance in imbalanced classification tasks. This also supports this paper's fundamental claim that, in practice, maintaining a stable trade-off is more valuable than optimizing a single statistic.

Table 3. Model comparison across different methods

Category	Method	Precision	Recall	F1	AUC	AP
Baseline	Fixed Hybrid (0.5/0.5)	0.393	0.687	0.500	0.947	0.483
Baseline	Random Forest	0.543	0.284	0.373	0.890	0.429
Baseline	Logistic Regression	0.173	0.881	0.289	0.949	0.524
Optimization	Dynamic Hybrid (Gating)	0.243	0.836	0.377	0.948	0.508
Optimization	Residual Boost Hybrid	0.522	0.358	0.425	0.945	0.475
Sampling	Random OverSampler + Hybrid	0.338	0.672	0.450	0.944	0.475
Sampling	SMOTE+ Hybrid	0.215	0.836	0.341	0.939	0.453

In addition to model optimisation, 5-fold stratified cross-validation (Table 4) is used to optimize the model and assess the robustness of the results by calculating the mean and standard deviation of each statistic. The F1-score is high over the data partitions, and the standard deviation of each metric is lower than that of other models, thereby showing remarkable output stability. This indicates that the hybrid model's performance gain is due to its capacity to balance several measures in unbalanced data settings. Hence, the multi-model comparison and cross-validation demonstrate that the hybrid model's strength is in a more stable and predictable trade-off between accuracy and recall. This property is very useful for imbalanced classification problems such as credit default prediction.

Table 4. Robustness analysis via 5-fold cross-validation (Mean $\pm$ Std)

Model	Precision	Recall	F1-score	AUC	AP
Hybrid	0.375 $\pm$	0.591 $\pm$	0.459 $\pm$	0.945 $\pm$	0.459 $\pm$
(0.5/0.5)	0.04	0.024	0.037	0.014	0.047
Logistic	0.187 $\pm$	0.895 $\pm$	0.309 $\pm$	0.948 $\pm$	0.535 $\pm$
Regression	0.018	0.065	0.028	0.015	0.062
Random	0.563 $\pm$	0.309 $\pm$	0.398 $\pm$	0.885 $\pm$	0.389 $\pm$
Forest	0.047	0.017	0.016	0.02	0.042

4 Discussion

The results of the experiment show significant differences in performance between the fundamental models: logistic regression is more aggressive, focusing on recall (0.8806), whereas random forests are more cautious, focusing on accuracy (0.5429). This discrepancy reinforces the tradeoff between accuracy and interpretability described by Mena et al., and also illustrates the challenges of relying on a single metric such as AUC in credit prediction as discussed by Hand [4, 5]. The hybrid model in this experiment used weight distributions of 0.4/0.6 and 0.5/0.5, which resulted in an F1-score of 0.5 by effectively balancing default detection and false positive management.

Results also showed that increasing the model complexity, e.g. in the form of residual augmentation or dynamic gating, did not significantly improve the trade-off, suggesting that arbitrary increases in model structure do not lead to gains. Thus, the optimisation should move from increasing model depth to improve the accuracy of weight allocation with more robust algorithms for keeping a dynamic balance between precision and recall.

While the study has made some progress, several limitations need to be acknowledged. While the Kaggle dataset is fairly representative, the static structure of the dataset does not fully portray the volatility of today's credit markets. Second, logistic regression and random forest are very basic models, and the room for hybrid models to improve is limited. In addition, the degree of weight usage in hybrid approaches is rather low.

Hence, in subsequent research, the combination of ensemble models and weighted comparison models might be explored to improve the prediction capability of the models on equilibrium [10, 11]. In addition, the complexity of weighting processes should be considered, and in addition to static weighting approaches, the exploration of local or global dynamic weighting is required to deal with sophisticated real-world conditions. Adaptive adjustments under the existing frameworks may be the future direction of study. For example, Snoek successfully increased the stability of hybrid classifiers in the processing of multidimensional financial data by the application of Bayesian optimization [12]. Finally, as Ayari mentioned, adding social media behaviour and credit ratings to the dataset might reduce the delays associated with traditional indicators significantly [6].

5 Conclusion

No single model can simultaneously optimize all evaluation metrics in imbalanced credit default prediction. The biggest problem is to find a stable trade-off between precision and recall, instead of simply increasing these two numbers. The results show that some models perform better on some measures, but the overall performance of them is biased. The hybrid model does not substantially exceed baseline models in any one statistic. Nonetheless, it attains a more stable and consistent equilibrium across several criteria, resulting in enhanced overall performance. So it is more critical to maintain consistent trade-offs across measures than to increase model complexity or to optimise particular indicators. This is an important balanced performance, especially in practical applications such as credit risk prediction, where the identification of risks and the control of false positives are equally important.

References

1. Alrasheedi, M.A.: Enhancing fraud detection in credit card transactions: A comparative study of machine learning models. *Computational Economics* (2025)
2. Bitetto, A., Cerchiello, P., Filomeni, S., et al.: Can we trust machine learning to predict the credit risk of small businesses? *Review of Quantitative Finance and Accounting* 63, 925–954 (2024)
3. Han, S., Jung, H., Yoo, P.D., et al.: Note: Non-parametric oversampling technique for explainable credit scoring. *Scientific Reports* 14, 26070 (2024)
4. Hand, D.J.: Measuring classifier performance: A coherent alternative to the area under the ROC curve. *Machine Learning* 77(1), 103–123 (2009)
5. Mena, L.J., García, V., Félix, V.G., et al.: Enhancing financial risk prediction with symbolic classifiers: Addressing class imbalance and the accuracy–interpretability trade-off. *Humanities and Social Sciences Communications* 11, 1540 (2024)
6. Ayari, H., Guetari, P.R., Kraïem, P.N.: Machine learning powered financial credit scoring: A systematic literature review. *Artificial Intelligence Review* 59, 13 (2026)
7. Ariza-Garzón, M.J., Arroyo, J., Caparrini, A., Segovia-Vargas, M.J.: Explainability of a machine learning granting scoring model in peer-to-peer lending. *IEEE Access* 8, 64873–64890 (2020)
8. Teles, G., Rodrigues, J.J., Rabêlo, R.A., Kozlov, S.A.: Comparative study of support vector machines and RFs machine learning algorithms on credit operation. *Software: Practice and Experience* 51(12), 2492–2500 (2021)
9. Guetari, R., Ayari, H., Sakly, H.: Computer-aided diagnosis systems: A comparative study of classical machine learning versus deep learning-based approaches. *Knowledge and Information Systems* 65(10), 3881–3921 (2023)
10. Parvin, A.S., Saleena, B.: An ensemble classifier model to predict credit scoring–comparative analysis. In: *IEEE International Symposium on Smart Electronic Systems (iSES)*, pp. 27–30 (2020)

11. Zhang, et al.: An ensemble credit scoring model based on LR with heterogeneous balancing and weighting effects. *Expert Systems with Applications* 212, 118732 (2023)
12. Snoek, J., Larochelle, H., Adams, R.P.: Practical Bayesian optimization of machine learning algorithms. *arXiv preprint arXiv:1206.2944* (2012)

Open Access This chapter is licensed under the terms of the Creative Commons Attribution-NonCommercial 4.0 International License (<http://creativecommons.org/licenses/by-nc/4.0/>), which permits any noncommercial use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

