JIMEA | Jurnal Ilmiah MEA (Manajemen, Ekonomi, dan Akuntansi)
Vol. 9 No.2, 2025

XGBOOST MODEL FOR DEFAULT PREDICTION IN CREDIT SCORING OF CONVENTIONAL BANK

Hilmi Suftandar[1]; Meditya Wasesa[2]; Utomo Sarjono Putro[3]
School of Business and Management, Institut Teknologi Bandung, Bandung[1,2,3]
Email : hilmi_suftandar@sbm-itb.ac.id[1]; meditya.wasesa@sbm-itb.ac.id[2];
utomo@sbm-itb.ac.id[3]

ABSTRACT

We develop an XGBoost-based classification model for predicting loan default in the context of credit scoring for a conventional commercial bank in Indonesia. The model aims to improve predictive performance in identifying high-risk borrowers using historical loan data. For this purpose, we use two years of consumer loan records, consisting of over forty thousand observations, including borrower demographic, credit score, and loan characteristics. To address the severe class imbalance within the dataset, we employ the Random OverSampling Examples (ROSE) technique on the training subset. The model is trained and evaluated using standard classification performance metrics, including precision, recall, F1 score, and area under both ROC and Precision-Recall curves. Our results show that the XGBoost model performs well in detecting non-defaults with high sensitivity and precision, particularly in the training set. However, performance on the test set indicates a significant drop in recall for the default class, suggesting model overfitting under imbalanced conditions. These findings highlight the potential and limitations of using ensemble learning methods such as XGBoost in real-world credit risk evaluation, especially when data imbalance remains a major concern.

Keywords : XGBoost; Credit Default; Credit Scoring; Imbalance Data; ROSE

ABSTRAK

Penelitian ini bertujuan mengembangkan model klasifikasi berbasis XGBoost untuk memprediksi risiko gagal bayar pada penilaian kredit di bank konvensional. Model dirancang untuk meningkatkan akurasi identifikasi debitur berisiko tinggi dengan menggunakan data historis pinjaman konsumer selama dua tahun, mencakup lebih dari 40.000 observasi. Data terdiri atas karakteristik demografis debitur, skor kredit, serta jenis produk pinjaman. Ketidakseimbangan kelas pada data ditangani dengan pendekatan Random OverSampling Examples (ROSE) pada data pelatihan. Evaluasi model dilakukan dengan metrik klasifikasi seperti precision, recall, F1 score, serta area di bawah kurva ROC dan Precision-Recall. Hasil menunjukkan bahwa model XGBoost mampu mengenali nasabah non-gagal bayar secara akurat pada data pelatihan, namun mengalami penurunan performa dalam mengenali gagal bayar pada data pengujian. Hal ini menunjukkan potensi overfitting akibat ketimpangan data yang belum sepenuhnya teratasi. Temuan ini menyoroti potensi dan keterbatasan penggunaan metode ensemble seperti XGBoost dalam evaluasi risiko kredit, serta perlunya strategi penyeimbangan data yang lebih adaptif untuk implementasi di dunia nyata.

Kata Kunci : Xgboost; Gagal Bayar; Penilaian Kredit; Data Tidak Seimbang; ROSE

INTRODUCTION

The growing adoption of digital technology in the banking sector has accelerated the need for more advanced and data-driven approaches to credit assessment.

05/05/2025 |Accepted : 04/06/2025 |Published : 05/08/2025
P-ISSN; 2541-5255 E-ISSN: 2621-5306 |

Traditional credit scoring systems, which often rely on fixed criteria and manual evaluation, are increasingly considered inadequate in capturing the dynamic risk profiles of borrowers. In the context of consumer lending, especially within conventional banks, predicting loan default with greater precision has become essential to maintaining portfolio quality and managing non-performing loans. Machine learning models, with their ability to uncover complex patterns in large datasets, offer promising alternatives to enhance predictive accuracy in credit risk evaluation.

This study proposes the application of the Extreme Gradient Boosting (XGBoost) algorithm to predict default risk in consumer loans using historical credit data from a conventional bank in Indonesia. To address the significant class imbalance commonly found in real-world loan datasets, the study applies the Random OverSampling Examples (ROSE) technique during training. The model's performance is evaluated using key classification metrics, including precision, recall, F1 score, and the area under ROC and Precision-Recall curves. Based on this context, the study aims to answer the key research questions:

To what extent can the XGBoost model accurately predict default risk in an imbalanced credit dataset of a conventional bank?

LITERATURE REVIEW AND HYPHOTHESIS DEVELOPMENT

The increasing complexity of borrower behavior and financial products has led to the growing application of machine learning (ML) techniques in credit scoring. Unlike traditional statistical models, ML algorithms are capable of modeling non-linear relationships and uncovering hidden patterns in large-scale financial datasets. Studies have demonstrated the effectiveness of ML methods such as Logistic Regression, Decision Trees, Random Forests, and Gradient Boosting in enhancing the predictive performance of credit risk models (Lessmann et al., 2015). Among these, Extreme Gradient Boosting (XGBoost) has gained prominence due to its scalability, regularization capabilities, and robustness in handling noisy and heterogeneous data (Chen & Guestrin, 2016). XGBoost has been widely applied in various domains of finance, including credit default prediction, portfolio optimization, and fraud detection.

One of the major challenges in credit scoring is the class imbalance problem, where default cases are significantly outnumbered by non-default cases. This imbalance can bias standard classification models toward the majority class, leading to poor recall

for high-risk borrowers. Several techniques have been proposed to address this, including data-level methods such as SMOTE and ROSE. The ROSE method generates synthetic examples for the minority class using a smoothed bootstrap approach and has been found effective in classification tasks with severe imbalance (Menardi & Torelli, 2014). While studies on ML-based credit scoring have flourished in global literature, there is still limited research focusing on their application in conventional banking institutions in Indonesia. Therefore, this study aims to fill this gap by applying XGBoost and the ROSE balancing technique to a real-world dataset from an Indonesian bank. Based on this review, we propose the following research hypotheses:

H1: *The XGBoost model achieves satisfactory predictive performance in default classification on an imbalanced loan dataset.*

RESEARCH METHODOLOGY

Method is a method of work that can be used to obtain something. While the research method can be interpreted as a work procedure in the research process, both in searching for data or disclosing existing phenomena (Zulkarnaen, W., et al., 2020:229). This study employs a quantitative approach by applying the Extreme Gradient Boosting (XGBoost) classification algorithm to historical consumer loan data sourced from a conventional Indonesian bank. The dataset comprises 40,726 observations over a two-year period, containing borrower demographic information, credit scores, loan types, and default status. The target variable is binary, with "1" indicating a default and "0" indicating a non-default. Initial data preprocessing involves checking for missing or invalid values and transforming categorical variables into factors (See Table 1).

Due to the class imbalance in the default variable, the Random OverSampling Examples (ROSE) technique is used to generate a balanced training dataset. The data is then split into training (80%) and testing (20%) sets. XGBoost is trained using the balanced dataset and evaluated on the original imbalanced test set to assess generalizability. Key performance metrics include precision, recall, F1 score, and area under the ROC and Precision-Recall curves (AUC-PR). Feature importance is also extracted to identify the most influential predictors. All modeling and analysis are conducted using R programming language, with packages such as xgboost, caret, ROSE, pROC, and PRROC.

RESULT AND DISCUSSION

The XGBoost model was trained using a balanced dataset generated via the ROSE method and then tested on the original imbalanced test dataset. The results show a clear difference in performance between the training and testing phases, revealing important insights about the model's behavior under real-world class distribution.

On the training dataset, the confusion matrix heatmap (see Figure 1) shows strong performance with high accuracy and recall across both classes. The model achieved a balanced accuracy of 83.90%, with a recall of 87.69% for the default class and 80.11% for the non-default class. This indicates that the model learned to distinguish between defaulters and non-defaulters effectively when trained on a synthetically balanced dataset. The ROC curve on the training set (see Figure 2) demonstrates strong separability with an AUC of 0.9246, while the Precision-Recall (PR) curve (see Figure 3) shows high sensitivity to positive class prediction with an AUC-PR of 0.9213. These results highlight XGBoost's ability to capture nonlinear patterns in borrower profiles when provided with balanced data.

However, evaluation on the test dataset reveals performance degradation, particularly in detecting defaulters. As shown in the confusion matrix heatmap (see Figure 4), the model's recall for the default class drops to 47.84%, while recall for the non-default class increases to 87.64%. The resulting balanced accuracy falls to 67.74%, and the Kappa statistic drops from 0.6773 in the training set to 0.3242 in the test set, suggesting mild overfitting. The ROC curve for the test dataset (see Figure 5) yields an AUC of 0.8208, while the PR curve (see Figure 6) presents a more modest AUC-PR of 0.4209, indicating lower precision in identifying defaults under natural class imbalance.

Overall, these results demonstrate the strength of XGBoost in modeling complex borrower behavior, particularly when class balance is addressed. However, the decline in test performance underlines the challenge of imbalanced classification in real-world settings. As suggested by Menardi & Torelli (2014), data-level balancing techniques like ROSE can enhance sensitivity but may require combination with threshold tuning or cost-sensitive learning to ensure robust generalization.

CONCLUSION

This study explores the application of the Extreme Gradient Boosting (XGBoost) algorithm in predicting loan default within a conventional bank's credit scoring system.

By utilizing a two-year dataset consisting of over 40,000 consumer loan records, the model was trained on a balanced sample using the ROSE technique and evaluated against an imbalanced real-world test set. The results confirm that XGBoost delivers high predictive performance on training data, particularly in identifying both defaulters and non-defaulters with balanced accuracy and strong AUC metrics.

However, the model's performance on the test dataset reveals notable sensitivity to data imbalance, especially in identifying default cases. The drop in recall and AUC-PR on the test set highlights a limitation in model generalization when deployed in naturally skewed environments (See Table 2). In addition, the feature importance analysis suggests that borrower attributes such as job type, number of dependents, and year of disbursement are among the most influential predictors of default risk. These findings support The Hypothesis, confirming the predictive strength of XGBoost under balanced training conditions, and the study successfully addresses the Research Questions:

Research Question – *To what extent can the XGBoost model accurately predict default risk in an imbalanced credit dataset of a conventional bank?* – the findings show that while XGBoost performs well on synthetically balanced data, its accuracy and sensitivity to defaults significantly decline under real-world class imbalance. Although the model maintains high precision for non-default predictions, its lower recall for default cases suggests a risk of under-identifying high-risk borrowers. This partially supports Hypothesis , highlighting the algorithm's potential under certain conditions but also its vulnerability to class distribution.

Despite these contributions, the study is subject to several limitations. First, the model relies solely on historical loan data from one institution, which may limit the generalizability of findings across different banking contexts. Second, the ROSE balancing technique, while effective, may not fully capture the intricacies of real minority-class behavior, potentially affecting out-of-sample recall. Future research is encouraged to explore hybrid balancing methods, multi-institutional datasets, and algorithm-level improvements such as threshold tuning or cost-sensitive learning to enhance robustness and practical applicability.

REFERENCES

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785–794. https://doi.org/10.1145/2939672.2939785

Lessmann, S., Baesens, B., Seow, H.-V., & 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. https://doi.org/10.1016/j.ejor.2015.05.030

Menardi, G., & Torelli, N. (2014). Training and assessing classification rules with imbalanced data. Data Mining and Knowledge Discovery, 28(1), 92–122. https://doi.org/10.1007/s10618-012-0295-5

Zulkarnaen, W., Fitriani, I., & Yuningsih, N. (2020). Pengembangan Supply Chain Management Dalam Pengelolaan Distribusi Logistik Pemilu Yang Lebih Tepat Jenis, Tepat Jumlah Dan Tepat Waktu Berbasis Human Resources Competency Development Di KPU Jawa Barat. *Jurnal Ilmiah MEA (Manajemen, Ekonomi, & Akuntansi)*, *4*(2), 222-243. https://doi.org/10.31955/mea.vol4.iss2.pp222-243.

FIGURE AND TABLE

Table 1. The Dataset (first 5 rows)

NAMA KANTOR	DealReff	Disburse_Year	CIF	Loan type	Cust_Job	Marital_Status	Dependants	Credit_Score	Kota_Alamat	Purpose	Default_Status
CABANG RAWAMANGUN	1230102000001	2023	17Q5BD	G6I	Pegawai Negeri	Single	0	0	Jakarta	Pendidikan	1
CABANG PANDEGLANG	1230102000007	2023	759028	G2B	Pensiunan	Married	1	2	Pandeglang	Pendidikan	0
CAB. KEBAYORAN BARU	1230102000008	2023	13W7N6	G6C	Pegawai Negeri	Single	0	0	Jakarta	Pendidikan	0
CABANG RAWAMANGUN	1230102000011	2023	17RKMC	G6C	Pegawai Negeri	Single	0	2	Jakarta	Pengobatan	0
CABANG BOGOR	1230102000014	2023	560309	G2B	Pensiunan	Married	0	2	Bogor	Elektronik	0

Table 2. Performance metrics of the XGBoost model evaluated on the training dataset (balanced using ROSE) and the original imbalanced test dataset. The model demonstrates strong predictive power during training, with high recall and F1 score, but experiences a drop in recall and F1 on the test set, indicating reduced sensitivity to defaults under real-world class imbalance.

Metric	Train Dataset (Balanced via ROSE)	Test Dataset (Original Imbalanced)
Accuracy	83.85%	81.99%
Precision (Non-def)	86.98%	91.04%
Recall (Default)	87.69%	47.84%
F1 Score	87.33%	62.72%
Balanced Accuracy	83.90%	67.74%
ROC AUC	0.9246	0.8208
PR AUC	0.9213	0.4209
Kappa	0.6773	0.3242

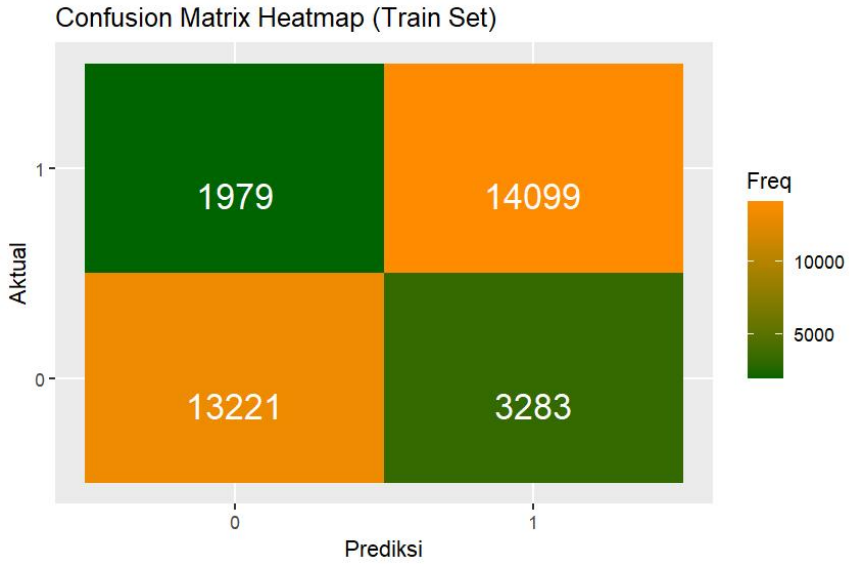

Figure 1. Confusion matrix heatmap of the XGBoost model evaluated on the balanced training dataset. The model demonstrates strong classification performance, with high recall for both default and non-default classes. These results indicate effective learning under the synthetically balanced data generated by the ROSE technique.

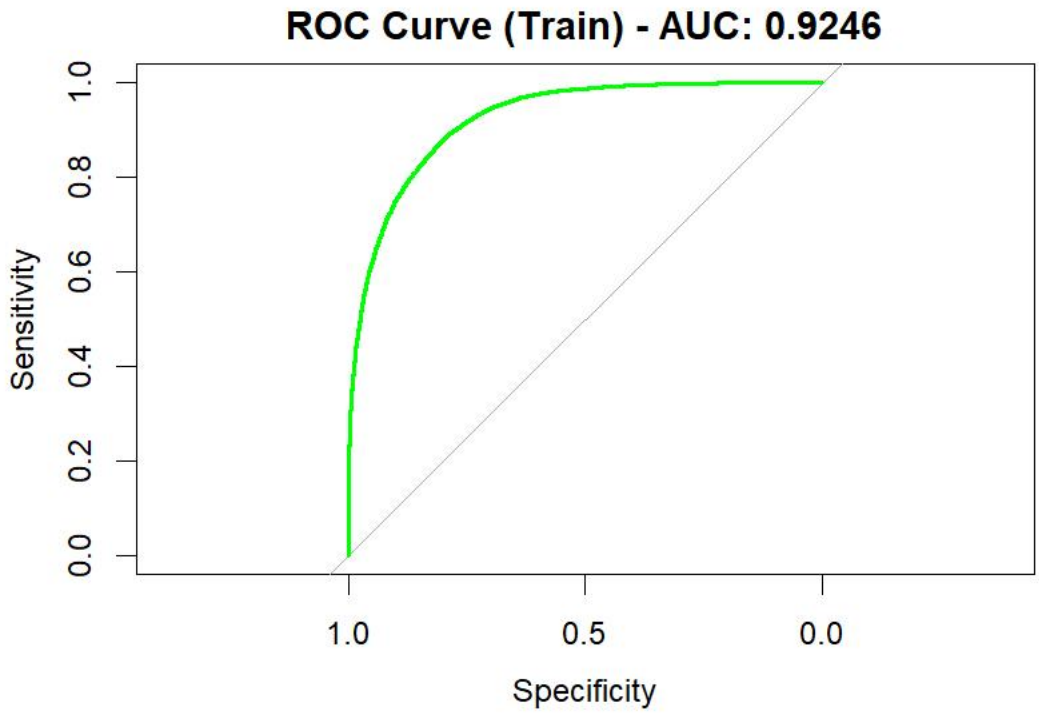

Figure 2. ROC curve of the XGBoost model on the training dataset, showing excellent discriminative ability with an AUC of 0.9246. The curve indicates high separability between defaulters and non-defaulters under balanced conditions.

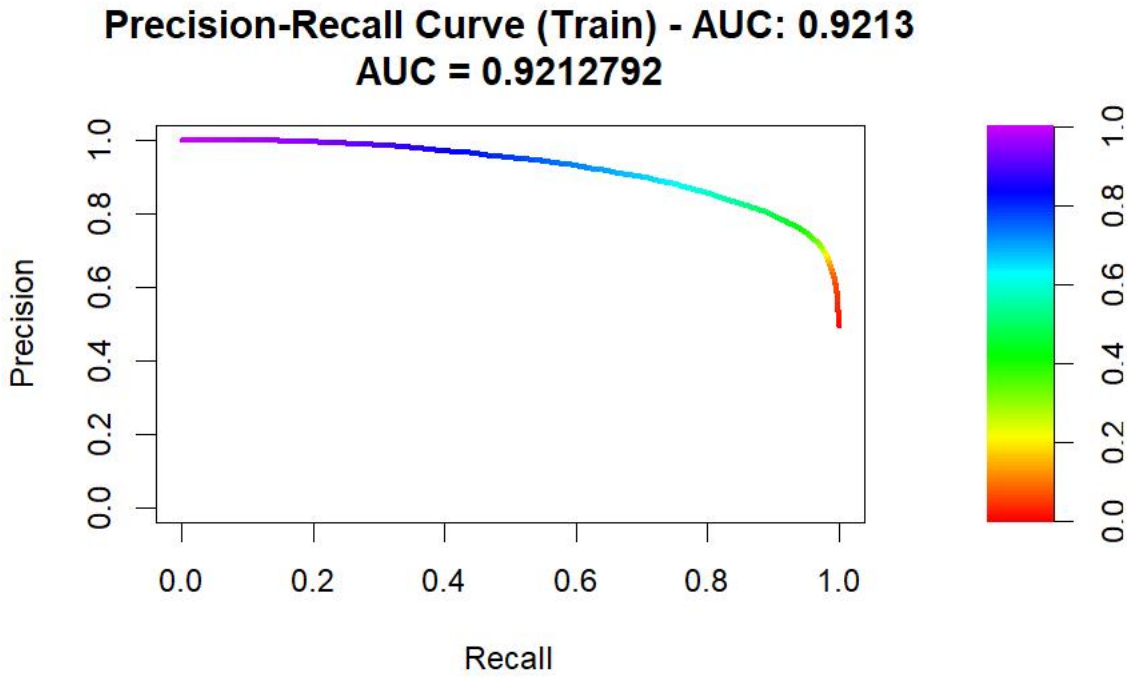

Figure 3. Precision-Recall (PR) curve of the XGBoost model on the training dataset. The curve yields a high AUC-PR of 0.9213, reflecting strong sensitivity to default cases in the presence of balanced data.

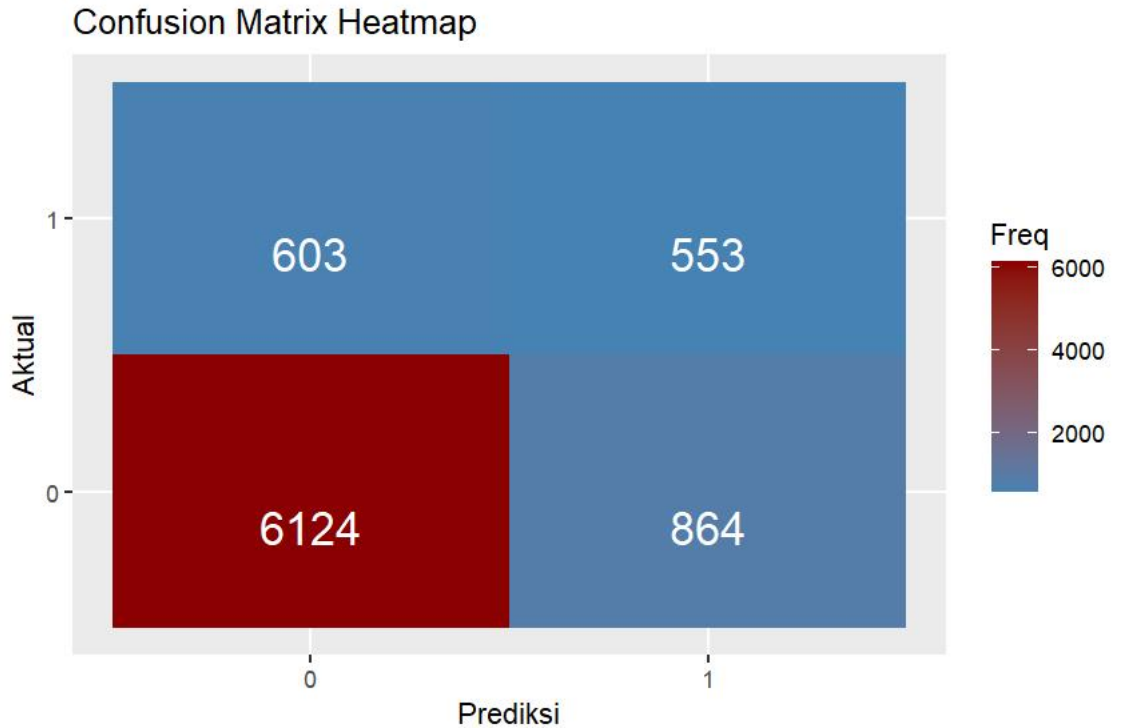

Figure 4. Confusion matrix heatmap of the XGBoost model on the imbalanced test dataset. The model achieves high accuracy for non-defaults but exhibits reduced recall for default cases, highlighting the impact of class imbalance on model performance.

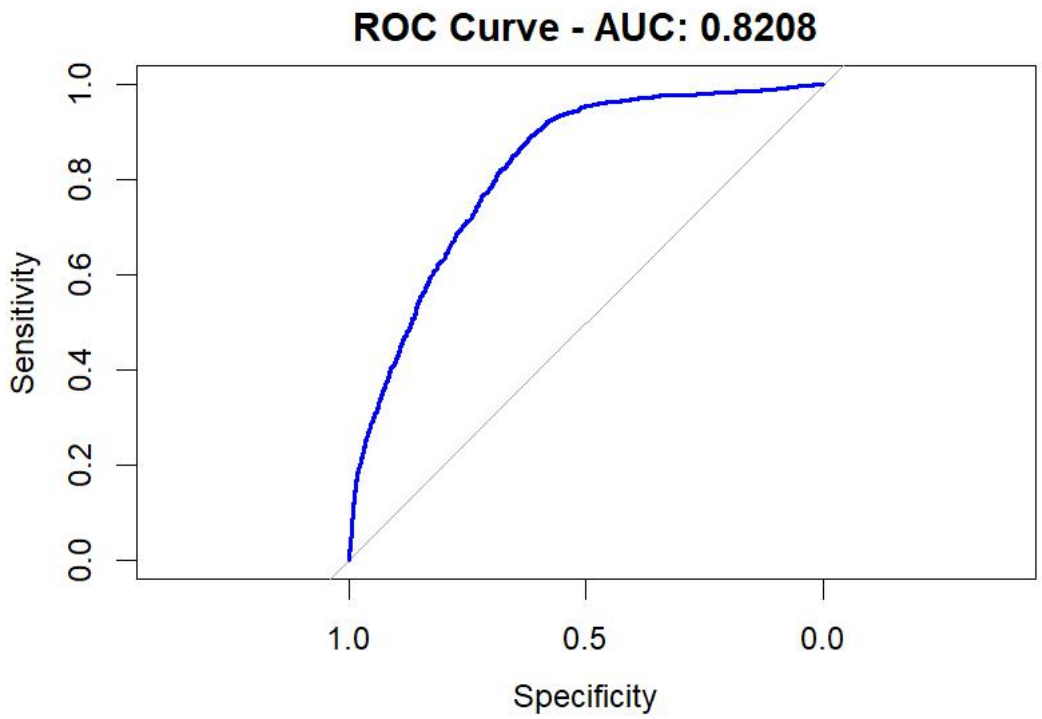

Figure 5. ROC curve of the XGBoost model evaluated on the test dataset. While the AUC of 0.8208 remains strong, it is notably lower than the training result, suggesting a decline in generalizability under imbalanced real-world conditions.

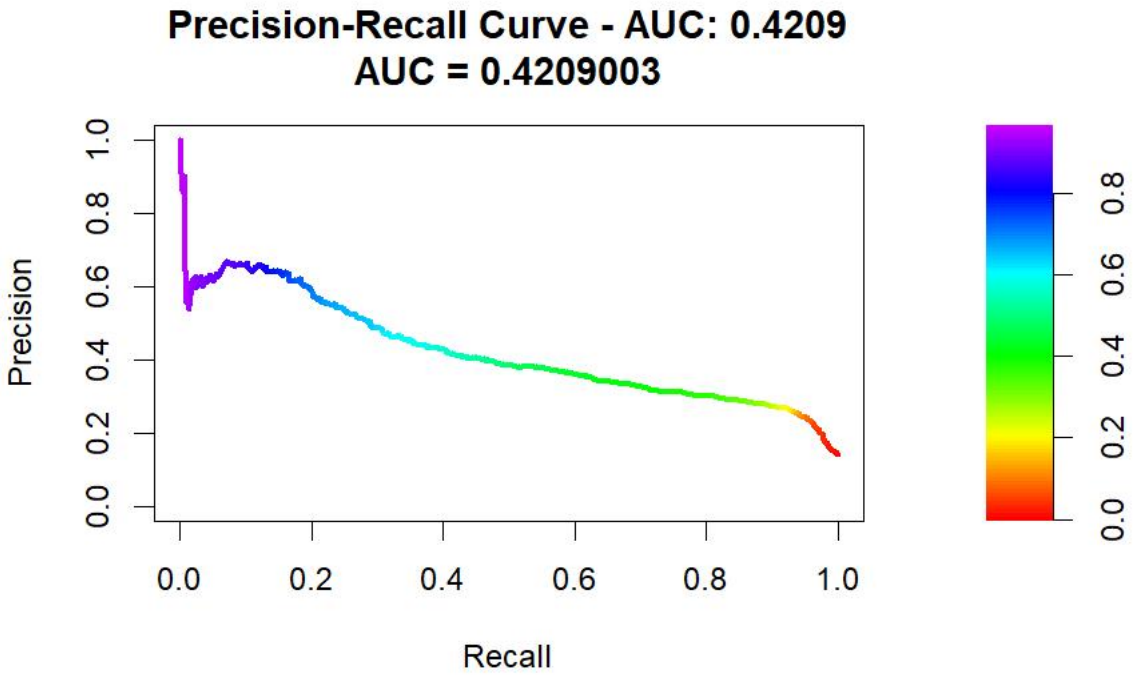

Figure 6. Precision-Recall (PR) curve of the XGBoost model on the test dataset, with an AUC-PR of 0.4209. The reduced area under the curve emphasizes the model's limited ability to identify default cases accurately in naturally skewed data.