Computer Science and Application 计算机科学与应用, 2025, 15(12), 199-208
Published Online December 2025 in Hans. https://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2025.1512335

Hans汉斯

融合机器学习算法的银行客户信用风险评估研究

安英博[1]，许美玲[2]，李　奕[1]，杨冉冉[1]

[1]河北金融学院金融科技学院，河北　保定
[2]河北金融学院信息与人工智能学院，河北　保定

收稿日期：2025年11月4日；录用日期：2025年11月26日；发布日期：2025年12月11日

摘　要

在金融数字化转型加速的背景下，银行客户信用风险决策面临“样本不均衡”与“误判成本高昂”的双重挑战，亟需兼顾整体精度与少数类召回的稳健模型。本文基于阿里云天池22,500名银行客户信贷数据，系统对比逻辑回归、随机森林与XGBoost三种机器学习模型，发现单一模型在召回率与F1指标上的不足；进一步提出融合模型框架，结合软投票加权平均与代价敏感学习，在不改变数据分布的前提下放大正类梯度权重，有效提升召回率至80.17%，较最优单一模型提高4.07%，AUC达0.8913，准确率保持在83.13%，为银行评估高风险客户提供了可解释、可落地的技术路径。

关键词

机器学习，银行客户信用风险，XGBoost，融合模型

Research on Credit Risk Assessment of Bank Customers Based on Integrating Machine Learning Algorithm

Yingbo An[1], Meiling Xu[2], Li Yi[1], Ranran Yang[1]

[1]School of Financial Technology, Hebei Finance University, Baoding Hebei
[2]School of Information and Artificial Intelligence, Hebei Finance University, Baoding Hebei

Received: November 4, 2025; accepted: November 26, 2025; published: December 11, 2025

Abstract

In the context of the accelerating digital transformation in finance, bank customer credit risk

文章引用: 安英博, 许美玲, 李奕, 杨冉冉. 融合机器学习算法的银行客户信用风险评估研究[J]. 计算机科学与应用, 2025, 15(12): 199-208. DOI: 10.12677/csa.2025.1512335

decision-making faces the dual challenges of "sample imbalance" and "high misjudgment costs". This situation calls for a robust model that balances overall precision and minority class recall. This study systematically compares three machine learning models—Logistic Regression, Random Forest, and XGBoost—based on the credit data of 22,500 bank customers from Alibaba Cloud Tianchi, finding that a single model falls short in recall rate and F1 metrics. It further proposes an ensemble model framework that integrates soft voting weighted averaging and cost-sensitive learning, effectively increasing the positive class gradient weight without altering the data distribution, thereby boosting the recall rate to 80.17%, an improvement of 4.07% over the optimal single model, with an AUC of 0.8913 and an accuracy maintained at 83.13%, providing an interpretable and actionable technical pathway for banks to assess high-risk customers.

Keywords

Machine Learning, Bank Customer Credit Risk, XGBoost, Ensemble Model

1. 引言

近年来，金融科技的迅猛发展促使大数据与人工智能技术广泛应用于金融领域。信用风险评估作为金融核心议题，直接影响银行的信贷决策与市场稳定性。传统评估方法因依赖主观判断而存在局限，无法动态处理复杂多变的金融环境。机器学习作为一种先进的数据处理技术，以其精准的预测能力为改善银行信用风险管理提供了新的解决方案。顾洲一等运用非平衡样本处理算法，构建 Logistic 模型计算违约概率[1]。周永圣等引入 XGBoost 算法处理数据样本，依据得出的重要性得分筛选个人信用风险评估指标，构建改进的随机森林模型的预测效果良好[2]。王培培等提出一种基于麻雀搜索算法的随机森林模型，利用 SSA 优化 RF 模型中决策树和最小节点数，具有较高的准确率[3]。

XGBoost、随机森林等集成学习方法通过增强分类器的性能，显示出推动风险评估精确度的可能性[4] [5]。然而，由于数据不均衡及特征维度高等问题，评价模型的优化仍面临挑战[6]。因此，本研究在现有成果基础上，针对数据处理与模型构建展开深入探究。

2. 数据描述与数据处理

2.1. 数据描述

本文使用的数据集来源于阿里云天池数据库提供的银行个人信用数据，训练集由 22,500 名客户的 22 个特征信息构成，测试集包含 7500 名客户的 21 个特征信息。数据中的特征包括客户个人基本信息、贷款记录、违约记录等，其中，特征 Subscribe 为目标属性，代表是否为高风险客户，即 Subscribe = yes 的客户信用风险较高，在信贷审批阶段，银行会依照 Subscribe 的预测结果来调整审批流程和放款利率。该特征中，no 占比约 86.88% (19,548/22,500)，yes 占比约 13.12% (2952/22,500)。训练集数据的基本特征如表 1 所示：

Table 1. Features of training dataset
表 1. 训练集数据特征

特征	字段描述	数据特征	实例
id	客户编号，用于唯一标识每个客户	整数型	22,500

续表

age	客户年龄	整数型	35
job	客户职业	分类型	technician、admin、blue-collar 等
marital	客户婚姻状况	分类型	single、married、divorced 等
education	客户教育程度	分类型	professional.course、high.school、basic.9y 等
default	客户是否有违约记录	分类型	no、yes、unknown
housing	客户是否有住房贷款	分类型	yes、no
loan	客户是否有其他贷款	分类型	yes、no
contact	与客户的联系渠道	分类型	cellular、telephone
month	联系客户的月份	分类型	aug、may、apr 等
day_of_week	联系客户的星期几	分类型	mon、thu、wed 等
duration	与客户通话的时长(秒)	整数型	3295
campaign	本次营销活动中联系该客户的次数	整数型	1
pdays	距离上次联系该客户的天数	整数型	476
previous	在本次营销活动之前联系该客户的次数	整数型	0
poutcome	上次营销活动的结果	分类型	nonexistent、failure
emp_var_rate	就业变化率	实数型	1.4
cons_price_index	消费者物价指数	实数型	95.37
cons_conf_index	消费者信心指数	实数型	−33.04
lending_rate3m	3 个月贷款利率	实数型	3.63

数据来源：阿里云天池。

2.2. 数据预处理

(1) 缺失值处理

通过数据探查可知，数值型特征缺失值数量为 17，分类型特征缺失值数量为 18，使用 Pipeline 构建预处理流程填补缺失数据：针对数值型特征，使用 SimpleImputer 以均值填补缺失数据，规避因缺失值而致使的数据分布偏移情况，并用 StandardScaler 工具对数据集实施标准化处理；针对类别型特征，使用众数进行缺失值填补，采用 OneHotEncoder 独热编码，把每一类转化成二进制向量，便于后续计算和处理。

(2) 特征值相关性分析

为了进一步洞察不同特征之间、特征与客户风险之间的相关性，为后续的特征选择、模型构建提供依据，使用 corr_matrix 计算特征属性与 subscribe 的相关系数、对数值型特征采用热力图分析、对分类型特征采用卡方检验。

由图 1 可以看出 Subscribe 与特征属性的相关度情况：就业变化率(emp_var_rate, −0.2701)、3 个月贷款利率(lending_rate3m, −0.1809)呈现显著负相关；营销活动次数(campaign, 0.1640)、3 月(month_mar, 0.1537)及有违约记录(default_yes, 0.1527)等特征与 Subscribe 呈正相关，其中营销活动次数的增加直接推动转化，而违约客户的高订阅率需警惕其潜在信用风险。

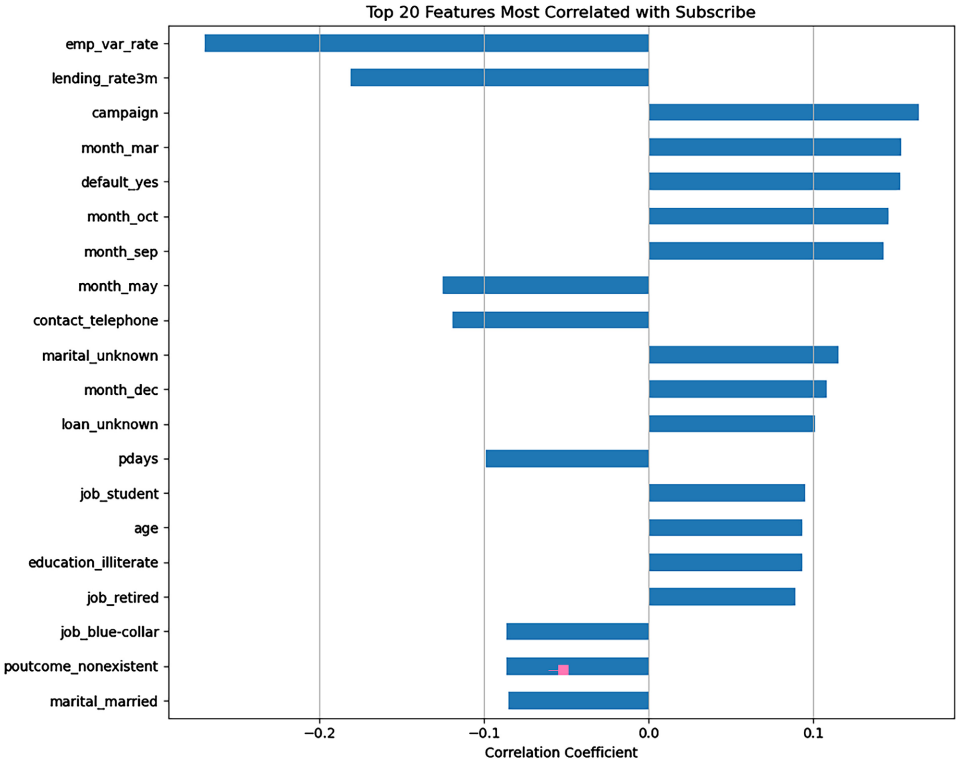

Figure 1. Correlation between subscribe and feature values
图 1. Subscribe 与特征值相关性

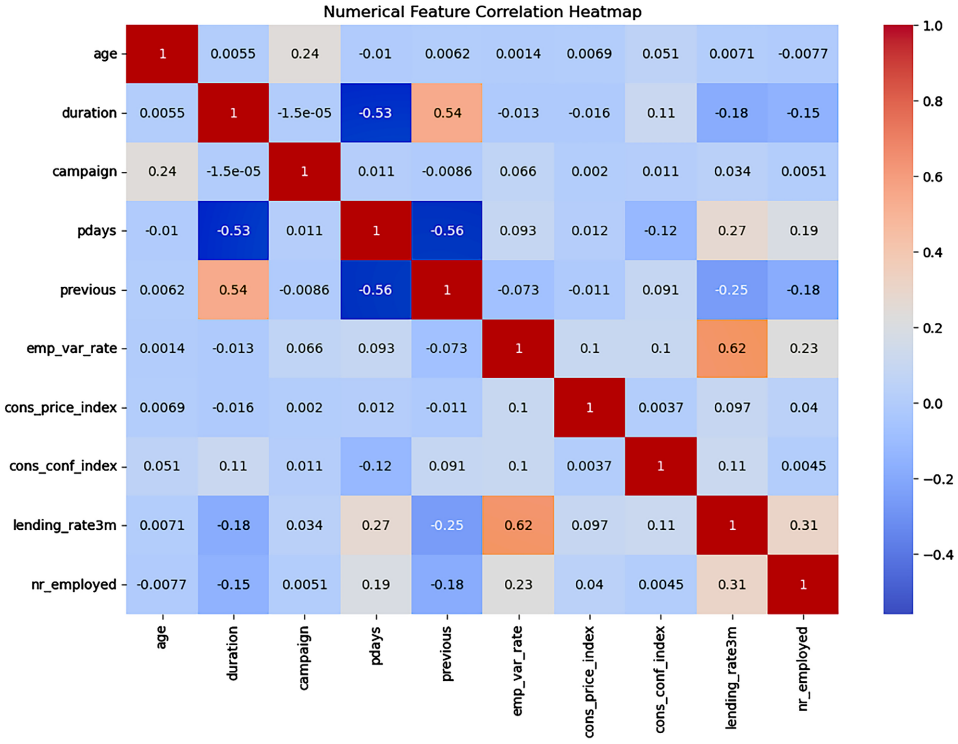

Figure 2. Correlation analysis heatmap of numerical features
图 2. 数值型特征相关性分析热力图

分析图 2 可以看出，emp_var_rate 和 lending_rate3m 相关系数为 0.62，表明就业变化率和短期贷款利率正相关；duration 与 previous 相关系数为 0.54，表明与客户通话的时长和营销活动前联系次数正相关；duration 与 pdays 相关系数为−0.53，表明与客户通话的时长与距上次联系天数负相关。

Feature	Chi2 Score	P-value
job	605.0475	0.0000
marital	425.9471	0.0000
education	286.8155	0.0000
default	617.9258	0.0000
housing	122.2295	0.0000
loan	240.4844	0.0000
contact	319.5970	0.0000
month	2067.8158	0.0000
day_of_week	21.6471	0.0002
poutcome	209.1941	0.0000

Figure 3. Correlation analysis of categorical features
图 3. 分类型特征相关性分析

由图 3 卡方结果可知，job、marital、education 等类别特征与目标变量 Subscribe 的 p 值几乎为 0，均存在显著关联。因此，通过以上分析，特征属性删除 id，保留 20 个特征。

3. 构建模型

本研究对训练集数据进行切片，80%的数据训练模型，20%的数据进行验证。先分别选用逻辑回归[7]、随机森林[8]、XGBoost 算法构建模型，通过交叉验证对预测效果进行比较；为了进一步提高模型的性能指标，采用模型融合技术，尝试对三种模型赋予不同的权重，通过网格搜索法确定召回率最高的权重；最后通过融合模型，将三种算法的预测结果进行加权平均，对客户进行风险评估。

为了有效评估模型的预测效果，本文选取二元分类模型评估的五个常用指标：准确率(*Accuracy*)、精确率(*Precision*)、召回率(*Recall*)、精确率和召回率的调和平均值(*F*1)、*AUC* (ROC 曲线下的面积)，计算方法见表 2 和公式(1)~(5)。

Table 2. Confusion matrix
表 2. 混淆矩阵

	预测值真	预测值负
真实值真	*TP*	*FN*
真实值负	*FP*	*TN*

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \tag{1}$$

$$precision = \frac{TP}{TP + FP} \tag{2}$$

$$recall = \frac{TP}{TP + FN} \tag{3}$$

$$F1 = 2 * \frac{\mathrm{Pr}ecision * \mathrm{Re}call}{\mathrm{Pr}ecision + \mathrm{Re}call} \tag{4}$$

$$AUC = \int_0^1 TPR\ d\left(FPR\right) \tag{5}$$

3.1. 逻辑回归模型

逻辑回归使用 sigmoid 函数，将线性组合转化为概率估计。本文将处理好的数据集带入 Logistic Regression 逻辑回归模型训练，采用网格搜索，以召回率为目标进行超参数调优。针对不同求解器设置参数组合，其中 C 尝试取值[0.01, 0.1, 1, 10]以探索合适的正则化强度，penalty 考虑 l1、l2、elasticnet 等正则化类型，solver 涵盖 liblinear、saga、newton-cg 等多种求解方式。经五折交叉验证，确定最佳参数组合 C = 0.01，l1_ratio = 0.1，penalty = elasticnet，solver = saga，将参数代入逻辑回归模型进行训练和验证，获取的评估指标数据如表 3 所示。

Table 3. Evaluation-metric results of logistic-regression model
表 3. 逻辑回归模型评估指标结果

Model	*Accuracy*	*Precision*	*Recall*	*F*1	*AUC*
Logistic Regression	0.7944	0.3586	0.7203	0.4789	0.8089

3.2. 随机森林模型

随机森林属于基于决策树的集成学习方法，构建由多棵决策树组成的森林，每棵树从样本中随机抽取，在节点分裂时仅使用部分特征，使得模型在增强预测能力的同时具备更好的泛化性能。本文采用网格搜索，以召回率为目标对 RandomForestClassifier 算法进行超参数调优。为了尝试不同决策树的数量来平衡模型性能与计算成本，设定树的数量 n_estimators = [100, 200, 300]；设定最大深度 max_depth = [3, 4, 5, 6, 7, 8]，来更好地把控模型，防止模型过拟合；考虑控制内部节点的分裂样本情况，防止树过度增长，设定分裂最小样本数 min_samples_split = [3, 4, 5, 6, 7, 8, 9, 10]。经过交叉验证比对，确定 n_estimators = 300，max_depth = 5，min_samples_split = 10 为最佳参数，将参数代入随机森林模型进行训练和验证，获取的评估指标数据如表 4 所示。

Table 4. Evaluation-metric results of random-forest model
表 4. 随机森林模型评估指标结果

Model	*Accuracy*	*Precision*	*Recall*	*F*1	*AUC*
Random Forest	0.7878	0.3555	0.7610	0.4846	0.8626

3.3. XGBoost 模型

XGBoost 是一种增强型的梯度提升决策树算法，通过逐步添加弱分类器，来优化整体误差并提升精度，其损失函数考虑了误差项和正则化项，加速收敛。训练 XGBoost 模型时，使用网格搜索和交叉验证，以召回率为评估目标进行参数调优。由于较多数量的树能够让模型具有更佳的拟合能力，设定弱评估器的数量 n_estimators = [100, 200, 300]；为了平衡复杂度与泛化能力，设定树的最大深度 max_depth = [3, 4, 5, 6, 7, 8]；设定学习率 learning_rate = [0.01, 0.1, 0.2]，从样本中采样的比例 subsample 取值为[0.6, 0.8, 1.0]，构造每棵树时随机抽样的特征占比 colsample_bytree 取值为[0.6, 0.8, 1.0]。交叉验证结果表明，n_estimators = 300，max_depth = 7，learning_rate = 0.2，subsample 和 colsample_bytree 均为 1 时，模型效果最佳，表明使用全部样本和特征有助于模型充分学习信息。将该参数组合导入 XGBoost 模型进行训练和验证，各

项评估指标数据如表 5 所示。

Table 5. Evaluation-metric results of XGBoost model
表 5. XGBoost 模型评估指标结果

Model	*Accuracy*	*Precision*	*Recall*	*F*1	*AUC*
XGBoost	0.8753	0.8372	0.0610	0.1137	0.8874

三种模型的 *Accuracy*、*Precision*、*Recall*、*F*1、*AUC* 性能指标结果如图 4 所示。从整体判别能力来看，XGBoost 在 *Accuracy* (0.8753)、*AUC* (0.8874)、*Precision* (0.8732)三项指标上均列第一，显著优于其他模型，说明其阈值分类综合性能最佳；Random Forest 次之，其 *AUC* 为 0.8626，但 *Accuracy* 为 0.7878，*Precision* 为 0.3555，相比 XGBoost 分别降低了 8.75、51.77 个百分点；Logistic Regression 模型的 *AUC* 最低为 0.8089，具备可接受的排序能力，但 *Accuracy*、*Precision* 结果和随机森林差不多，相比 XGBoost 模型的整体判别效果较差。

从正类查全能力来看，Random Forest 的召回率 *Recall* 最高(0.7610)，Logistic Regression 与之接近，XGBoost 仅为 0.0610，不足前两者的一成，表明 XGBoost 对正样本的敏感性严重不足，可能因阈值优化偏向高 *Precision* 而牺牲了 *Recall*。

综合对比可以看出，三种模型在负类样本数据上表现较为突出：逻辑回归模型各项指标的表现处于中等水平，在精确率等方面仍然还存在提升空间，因此，在对模型性能要求不是特别高或者更看重模型可解释性的情况下，可以考虑使用该模型；随机森林模型各项指标的表现比较均衡，整体性能较为稳定，*Precision* 最低意味着后续需投入较高人工复核成本；XGBoost 模型区分正负类的整体能力较强，准确率达到 87.53%，精确率为 83.72%，具有明显优势，但其召回率仅为 6.1%，这使其在实际应用中容易遗漏重要的正类样本，而此类样本是风险较高的用户，银行需要重点关注。因此，三种模型不宜单独使用，若要应用，需进一步改进优化。

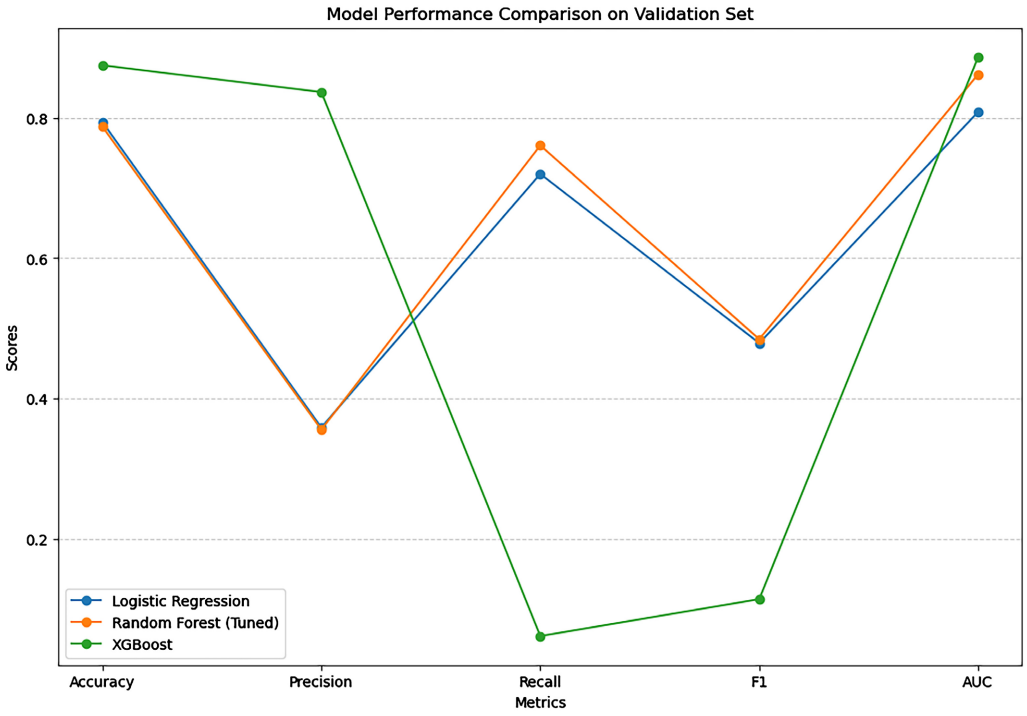

Figure 4. Performance-metric comparison of three models
图 4. 三种模型的性能指标对比

3.4. 融合模型

上述三种模型的实验表明，单一模型受到自身算法以及数据集复杂性的限制，不能完全获取数据蕴含的所有信息，因此要提升模型的泛化性能和预测精准度，本文采用融合模型策略，该策略旨在结合三种算法的优势，减少单一模型的不稳定性与局限性，从而提升风险评估的整体性能，即采用集成学习方法中的软投票加权平均策略来实现这一目的。

(1) 软投票加权平均

通过计算每个模型预测结果的加权平均值来提升模型性能，从而全面发挥每个模型的长处[9]，具体做法为逻辑回归、随机森林和 XGBoost 三个基模型赋予不同的权重，利用 VotingClassifier 形成投票分类器，对三种模型的概率输出进行加权计算，最终得到综合预测结果。

(2) 类别不平衡处理

上述三种单一模型在正类样本的预测性能一般，有可能受到数据集类别不平衡的影响(yes 占比约 13%)，模型训练容易被多数类梯度主导，导致决策边界过度保守、对少数类召回不足。因此本文尝试两种方法进行调整优化：第一，采用 SMOTE 方法对训练集进行过采样，改变数据分布；第二，采用代价敏感方法[10]，在 XGBoost 里设置参数 scale_pos_weight = *N*negative/*N*negative，把梯度偏向正类，在不改变数据分布的前提下引导树更多地向少数类区域分裂。

(3) 权重调整

为三种模型设置不同的权重，如逻辑回归权重 = 1，随机森林和 XGBoost 权重 = 2，使得整体预测结果更加均衡，权重的选择基于每个模型的单独性能指标，对于银行客户信用风险而言，银行除了关注准确率指标，更关注是否会遗漏高风险客户的风险，即偏向关注 *Recall* 召回率。因此，通过网格搜索法，尝试不同权重组合[1, 1, 1]、[1, 2, 2]、[2, 1, 1]、[2, 2, 1]、[1, 1, 2]、[2, 2, 2]、[3, 1, 1]、[1, 3, 1]、[1, 1, 3]进行参数调优，针对不同权重的融合模型生成的 Recall 指标进行对比，交叉验证来确定最优组合。

(4) 实验结果

表 6 给出了两种类别不平衡处理策略下融合模型的最优性能。可以看出，采用 SMOTE 过采样，虽然 *Recall* 达到 0.9508，但 *Precision* 骤降至 0.1629，*F*1 仅为 0.2781，*Accuracy* 跌至 0.3531，表明过采样在提升查全的同时引入过多假正例，整体判别性能显著下降；相比之下，融合模型通过设置 scale_pos_weight 进行代价敏感训练，在未改变原始数据分布的前提下，将正类误差梯度加权，采用网格搜索的最佳权重[1, 1, 3]，即逻辑回归权重和随机森林权重为 1，XGBoost 权重设为 3，实现了各评估指标的更好权衡：*Recall* 提升到 0.8017，*Precision* 为 0.4242，*F*1 提高至 0.5548，*Accuracy* 与 *AUC* 分别达到 0.8313 和 0.8913。

Table 6. Evaluation-metric results of ensemble model
表 6. 融合模型评估指标结果

数据优化方法	最佳权重	*Accuracy*	*Precision*	*Recall*	*F*1	*AUC*
SMOTE	[1, 3, 1]	0.3531	0.1629	0.9508	0.2781	0.8335
设置 scale_pos	[1, 1, 3]	0.8313	0.4242	0.8017	0.5548	0.8913

如图 5 所示，逻辑回归、随机森林、XGBoost 及融合模型的性能指标结果证实，融合模型在 *AUC* (0.8913)和 *Recall* (0.8017)指标上显著优于单一模型，同时 *Accuracy* (0.8313)保持在较高水平，表明使用代

价敏感方法的融合模型能够在保持整体分类性能稳定的同时，有效提升了整体判别能力。准确率提高方面，融合模型较随机森林和逻辑回归分别提升了 4.35 和 3.69 个百分点，虽略低于 XGBoost 的 0.8753，但结合更高的 *Recall* 和 *F*1 值，整体分类性能更均衡，显著降低了高风险客户的误判风险，同时假阳性率得到合理约束，更适合银行风险管理的实际需求。

```
================================== 所有模型性能对比 ==================================

| Model                | Accuracy | Precision | Recall   | F1       | AUC      |
|----------------------|----------|-----------|----------|----------|----------|
| Logistic Regression  | 0.7944   | 0.3586    | 0.7203   | 0.4789   | 0.8089   |
| Random Forest (Tuned) | 0.7878   | 0.3555    | 0.7610   | 0.4846   | 0.8626   |
| XGBoost              | 0.8753   | 0.8372    | 0.0610   | 0.1137   | 0.8874   |
| Fusion Model         | 0.8313   | 0.4242    | 0.8017   | 0.5548   | 0.8913   |
```

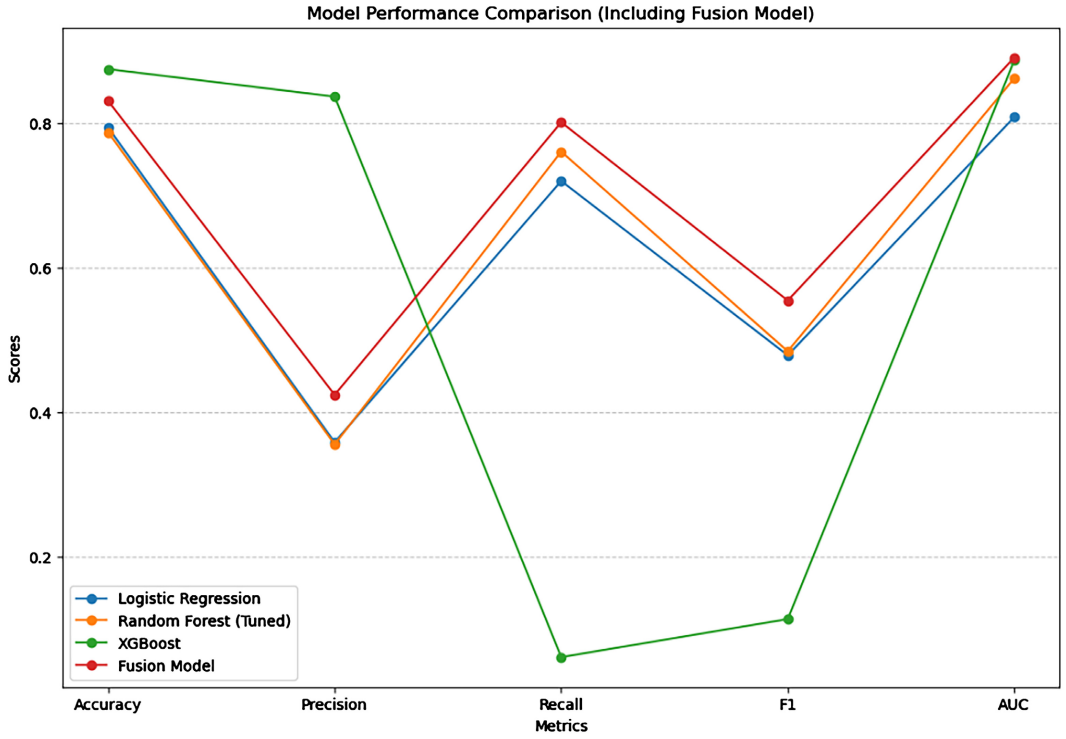

Figure 5. Performance-metric comparison across all models
图 5. 所有模型的性能指标结果对比

4. 结论

本文针对银行客户信用数据，应用机器学习算法进行了深入研究，通过系统比较逻辑回归、随机森林、XGBoost 三种模型在银行客户信用风险评估中的性能表现，发现单一模型存在显著局限性，于是提出通过软投票加权平均结合代价敏感学习的融合模型，实验表明，该模型策略在 *Recall* 与 *AUC* 上显著超越单一模型，同时能够维持较高的 *Accuracy* 和 *F*1 值，有效平衡了精确性与查全率的矛盾，实现了性能突破，更契合银行风险控制中“精准识别”与“全面覆盖”的双重目标，同时能够为同类场景提供优化路径的参考。

未来研究可进一步探索动态权重调整机制，结合实时业务反馈优化模型组合，引入深度学习模型增强特征表征能力，以应对复杂多变的金融风险环境。

基金项目

2025 年度河北省金融科技应用重点实验室课题(课题编号：2025003)。

参考文献

[1] 周永圣, 崔佳丽, 周琳云, 等. 基于改进的随机森林模型的个人信用风险评估研究[J]. 征信, 2020, 38(1): 28-32.

[2] 顾洲一, 胡丽娟. 机器学习视角下商业银行客户信用风险评估研究[J]. 金融发展研究, 2022(1): 79-84.

[3] 王培培, 周小平, 陈佳佳, 等. 基于麻雀搜索算法与随机森林融合模型的个人信用评估[J]. 上海师范大学学报(自然科学版中英文), 2024, 53(2): 241-246.

[4] 张淼, 顾海燕. 基于优化决策树参数的随机森林模型预测全国 GDP[J]. 中国林业经济, 2025(4): 39-50.

[5] 赵阳, 张杰萌, 严国义. 基于 SMOTE-XGBoost 算法的信用卡违约预测模型研究[J]. 武汉工程大学学报, 2025, 47(3): 343-348.

[6] Chen, T.Q. and Guestrin, C. (2016) XGBoost: A Scalable Tree Boosting System. *CoRR*, abs/1603.02754.

[7] 张思扬. 基于逻辑回归模型的信用卡逾期风险预测及优化[J]. 现代信息科技, 2024, 8(19): 141-145, 151.

[8] 邱泽国, 贺百艳. 机器学习算法下信用风险评估体系构建研究——基于中国银联数据的个人信用风险评价分析[J]. 价格理论与实践, 2021(10): 89-92, 194.

[9] 曹伟萍, 张劲松. 基于不平衡数据处理与加权软投票异质集成的农户贷款违约风险预测[J]. 计算机应用与软件, 2025, 42(8): 71-79.

[10] 王丹, 吴腾, 于振华, 等. 基于联邦学习的代价敏感卷积神经网络分类方法[J]. 西安科技大学学报, 2025, 45(3): 591-606.