基于多种机器学习算法的信用卡贷款违约预测研究

崔飞 张洪伟

湖北工程学院新技术学院，湖北孝感，432000；

摘要：针对信用卡贷款违约预测需求，本文基于中国台湾地区 2005 年信用卡客户数据，采用 K 近邻、决策树、XGBoost 三种模型分析。经数据清洗与复合特征构建，初始 XGBoost 分类表现最优，但存在阈值盲目与业务适配不足问题。引入 G-means 最大化阈值调整并融合决策树规则优化后，XGBoost 在测试集性能显著提升、错误率大降，能精准识别高风险用户。研究证实，优化后 XGBoost 准确可靠，可为金融机构信贷风控提供支撑，符合监管要求。

关键词：贷款违约；机器学习；XGBoost 模型；风险防控

DOI：10.69979/3041-0673.26.03.081

引言

消费金融是扩大内需、提振消费的重要载体，为我国经济高质量发展提供关键支撑。近年相关政策落地，信用卡作为消费信贷核心工具规模持续扩张，截至 2024 年末发卡量破 7.27 亿张，贷款余额 8.7 万亿元[1]，占消费信贷总额超 40%。但不良率涨幅超 10%，个别区域银行破 40%，违约风险既侵蚀金融机构资产质量[2]，更可能引发系统性风险[3]，成为消费金融健康发展核心瓶颈。

近年来，机器学习为智能风控提供新路径：张思扬[4]采用逻辑回归(Logistic Regression)进行信用卡欺诈检测，精度达到 95%以上。范巍强等[5]使用反向传播 NN 模型预测信用卡使用者的违约风险，陶明泽等[6]则基于去噪扩散概率模型进行信用卡欺诈预测，虽对信用卡违约预测起到效果，但现有研究多聚焦单一算法，未通过多算法横向对比验证模型的泛化能力，难以确定适配不同场景的最优方案。本文以信用卡违约预测为目标，通过多源数据融合提升样本代表性，创新构建还款行为与消费能力复合特征，对比 KNN、决策树、XGBoost 性能，筛选最优模型验证，为金融机构提供“精准、高效、可解释”的风控解决方案。

1 数据预处理及样本库构建

1.1 数据预处理

本文信用卡违约预测数据含客户基础属性与信贷行为记录，共 32 个维度，部分特征存在关键信息缺失、极端异常值及显著量纲差异，影响样本有效性与算法违约识别精度。数据预处理能提升数据一致性与可用性，为后续特征工程及模型训练奠基，通过 Z-score+IQR 法、分层填充法、标准化编码法、Pearson 相关系数法，完成缺失值补充、异常值检测替换、特征格式统一及数据降维。

1.1.1 异常值处理

针对信用卡数据中“信用额度”“年龄”等数值型特征易出现极端值的问题，采用“Z-score 法+IQR 四分位距法”双重检测机制识别异常值，确保覆盖不同类型的离群数据，避免单一方法导致的漏检或误检。

Z-score 法基于数据正态分布特性，通过计算样本与均值的偏离程度量化异常性。对于某一数值型特征 x，其 Z-score 值计算公式如下：

$$Z=\frac{x-\overline{x}}{\sigma}$$

其中，$\overline{x}$为该特征的均值，σ为标准差。当 | Z | >3 时，判定为异常值，该方法适用于“信用额度”“月均消费金额”等近似正态分布的特征。

IQR 四分位距法通过数据分位数划分合理范围，规避极端值对均值的影响，更适配可能存在偏态分布的特征。首先计算特征的第一四分位数 Q1（25%分位数）、第三四分位数 Q3（75%分位数），再通过公式计算四分位距 IQR：

$$IQR = Q3 - Q1$$

以[Q1-1.5×IQR,Q3+1.5×IQR]作为合理数据范围，超出该范围的样本标记为异常值。处理遵循“业务逻辑优先”原则：连续变量合理波动用 99 分位数替换，无效数据剔除样本；离散变量超业务编码用众数替换。处理后异常值占比从 2.3% 降至 1%内，既保有效信息，又消极端值对模型训练的干扰。

1.1.2 特征编码与缩放

对“婚姻状况”“性别”等无序分类特征，One-Hot 编码防数值偏倚；对“教育程度”等有序分类特征，用 Label 编码（设为 3→2→1→0）保层级；对“信用额度”“月均消费金额”等数值型特征，采用 Z-Score 标准化：

$$Xscaled=\frac{x-\bar{x}}{\sigma}$$

其中，$\bar{x}$为该特征的均值，σ为标准差。

1.1.3 数据降维

32 项特征会增加算法训练复杂性、影响预测准确度，故采用“Pearson 相关系数+冗余特征剔除”降维。该方法通过 Pearson 相关系数量化特征间线性关系。相关程度取值范围为[-1,1]，|r|>0.8 判定为高相关冗余，结合业务逻辑筛选核心特征，将原始数据中高度相关的变量进行冗余剔除，保留互不重叠的关键信息，其 Pearson 相关系数如下：

$$r_{xy}=\frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum(x_i-\bar{x})^2\sum(y_i-\bar{y})^2}}$$

其中 x 、 y 为待分析特征， $\bar{x}$ 、 $\bar{y}$ 分别为特征均值，n 为样本数量。

特征保留以“核心业务价值优先”，确保降维后数据覆盖客户信用风险关键信息，对 32 项特征制作图 1 所示相关性热力图，最终选取 25 项无冗余核心特征替代原 32 项，完成数据降维。

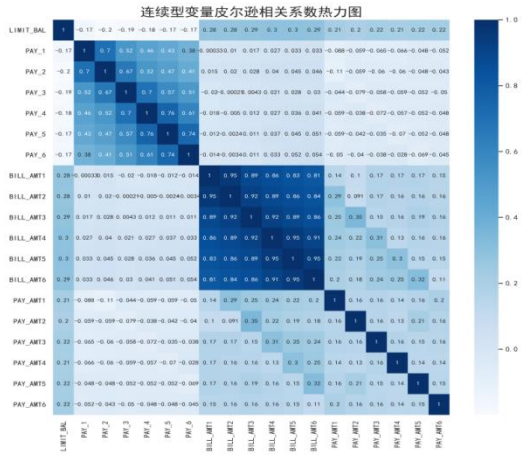

图 1 32 项特征的相关性热力图

1.1.4 特征工程构建

特征工程围绕信用卡违约预测目标展开核心处理：先通过数据清洗剔除异常值与缺失值，结合业务逻辑构造还款频率、额度使用率等衍生特征，再采用相关性分析与特征重要性评估[7]筛选出 12 个强关联核心变量——其中历史逾期记录、月还款占比为 TOP 影响特征，最后对类别型特征实施独热编码、连续型特征进行分箱处理。这一过程有效提取并聚焦关键信息，验证了特征工程的实用性，为后续模型训练奠定高质量数据基础。

2 基于机器学习的信用卡违约预测模型构建

2.1 核心算法原理

2.1.1 K 近邻（KNN）

KNN 基于“近邻相似”原理，计算待预测样本与训练样本距离，选择 K 个最近样本的多数投票作为预测结果。优势是无需训练、易实现，适合低维数据；但对高维数据敏感，计算成本随样本量增加而上升。本文将 K 值范围设为 5~15，并通过网格搜索优化。

2.1.2 决策树

决策树以“信息增益最大”为准则构建树形结构，每个节点对应特征判断，叶节点为类别。其核心优势是可解释性强，但易过拟合。本文通过控制树深度（3~10）、最小样本分裂数（2~10）抑制过拟合。

2.1.3 极端梯度提升树（XGBoost）

XGBoost 是集成学习的代表算法[8]，通过迭代训练多棵弱决策树，每棵树拟合前序模型的残差，最终加权求和输出。其支持正则化、缺失值自动处理，且通过 scale_pos_weight 参数平衡正负样本权重，适配类别不平衡场景。

2.2 模型评估指标体系

针对信用卡违约“漏判代价高”的特点，构建多维度指标体系，避免单一依赖准确率：

基础指标：精确率、召回率、F1 分数，重点关注召回率；

进阶指标：AUC 值、PR 曲线；

业务指标：坏账率下降幅度、人工审核量减少比例，量化模型实际价值 3 三种模型性能评估与违约风险预测含 25 项核心特征的数据集按 7:3 划分训练集与测试集，以“违约/未违约”为输出标签，经网格搜索与 5 折交叉验证确定三种算法关键超参数。针对信用卡违约数据类别不平衡、漏判易致坏账的特点，选取 AUC、精确率、召回率及正类 F1 验证测试集性能，明确算法分类能力与业务适配度。

测试集性能结果分析：

表 4 三种算法多指标对比表

评估指标	K 近邻（KNN）	决策树	极端梯度提升树（XGBoost）
AUC 值	0.784	0.705	0.812
精确率（%）	22.5	27.8	53.9
召回率（%）	78	71	81
F1 值	0.258	0.377	0.538

本研究以 AUC 值、违约样本 F1 值、召回率及精确率构建多维度性能评估体系，对 K 近邻（KNN）、决策树、极端梯度提升树（XGBoost）三类算法进行综合验证，结果显示 XGBoost 虽以测试集 AUC=0.812 展现最优分类能力，显著优于 KNN 与决策树，但正类样本最高 F1 值仅 0.538，仍难以满足信用卡风控“高召回、高精确”的核心诉求，因此我们通过对 XGBoost 模型的各项指标进一步优化提升。

2.3 基于 XGBoost 模型多指标优化的提升

尽管 XGBoost 模型在初始阶段已展现出较好的分类性能，但由于信用卡违约数据存在类别不平衡特性，其默认阈值下易出现“业务逻辑嵌入不足”的问题，难以精准平衡“识别违约样本”等风控诉求，因此需通过阈值优化与业务规则融合来提升模型的业务适配性与预测精准性。

2.3.1 优化方法

初始 XGBoost 虽有较好分类性能，但信用卡违约数据存在类别不平衡，其默认阈值易出现 “选择盲目”“缺业务逻辑” 问题，难平衡风控诉求，故需通过阈值优化与业务规则融合，提升模型适配性与预测精准性。

2.3.2 优化后性能

单一算法的性能在经过多指标优化后，各性能显著提升：AUC 值从 0.824 升至 0.85，AP 值从 0.50 升至 0.56，G-means 从 0.72 升至 0.78，且 G-means 最大化时的阈值稳定性更强。

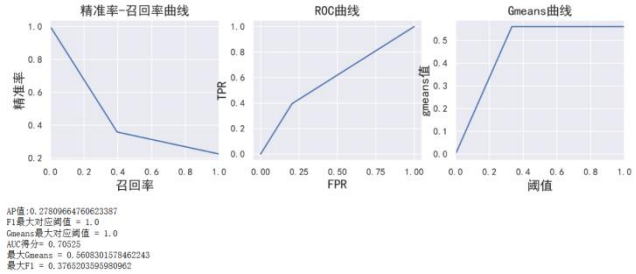

图 3 优化前

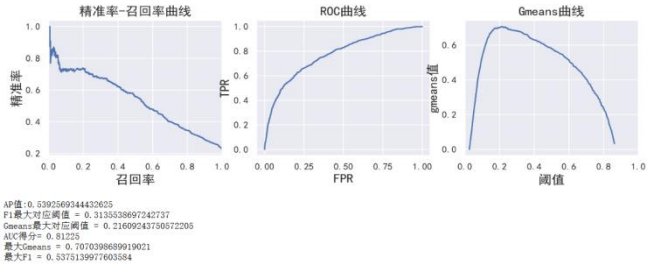

图 4 优化后

对比可知优化后 XGBoost 性能提升由指标与曲线双重印证。表 5 显示 G-means 最大化时阈值稳定性显著增强；ROC 曲线 AUC 从 0.824 升至 0.85，FPR 低的风控区间上升更快，既控正常用户误判，又强违约识别；PR 曲线 AP 值从 0.50 升至 0.56，召回率 71%时仍有 68%精准率，解决类别不平衡下高召回与高精准矛盾。对比随机森林高召回区间精准率跌破 50%、KNN 曲线偏低，更显模型在不平衡数据的分类稳定性，适配信用卡风控需求。

3 结论

针对含 25 项核心特征的 3 万条信用卡客户数据，经 SMOTE 过采样、Pearson 冗余剔除及 Z-score 标准化构建高质量样本库，基于 KNN、决策树、XGBoost 开展违约预测研究。经同源数据训练与多维度评估，明确算法性能差异：XGBoost 分类最优但正类 F1 仅 0.47，难满足高召回、高准确率诉求。实例应用表明，研究揭示单一算法在类别不平衡数据的局限，为后续优化策略提供方向，且验证“技术指标+业务需求”双维度评估的必要性，为金融机构提升预测精度提供参考。

参考文献

[1]曹光宇.2024 年上市银行年报之信用卡专题解读[J].中国信用卡,2025,(06):9-20.

[2]朱丽云.基于Logistic模型的商业银行信用风险分析[J].品牌研究,2019,(19):17-18.DOI:10.19373/j.cnki.14-1384/f.2019.19.005.

[3]朱振涛,孙敏,沈建红.信用卡逾期行为的影响因素及行为预测研究[J].南京工程学院学报(社会科学),2019,19(03):46-53.DOI:10.13960/j.issn.2096-238X.2019.03.009.

[4]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.DOI:10.19850/j.cnki.2096-4706.2024.19.026.

[5]范巍强，刘暾东.基于 BP 神经网络的信用卡违约风险预测 [J]．电脑知识与技术，2011，7（10）：2348-2349.

[6]陶明泽,熊星星,陈军伟,等.基于去噪扩散概率模型的信用卡欺诈检测[J/OL].计算机科学与探索,1-20[2025-11-14].

[7]姚旭,王晓丹,张玉玺等.特征选择方法综述[J].控制与决策,2012,27(2):161-166.

[8]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(03):343-348.DOI:10.19843/j.cnki.CN42-1779/TQ.202312031.