Advances in Applied Mathematics 应用数学进展, 2025, 14(12), 201-214
Published Online December 2025 in Hans. https://www.hanspub.org/journal/aam
https://doi.org/10.12677/aam.2025.1412500

Hans汉斯

基于GBDT模型对于金融防欺诈的应用研究

朱宏伟[1,2*]，刘佳兴[3]，李兴博[2]，焦海涵[2]

[1]河北省科技金融与协同创新中心，河北 保定
[2]河北金融学院金融科技学院，河北 保定
[3]河北金融学院管理学院，河北 保定

收稿日期：2025年11月11日；录用日期：2025年12月5日；发布日期：2025年12月10日

摘 要

近年来，我国互联网消费金融飞速发展，促进消费需求，为低收入群体开辟借贷新途径，推动经济增长。但其快速扩张也导致借贷违约风险增加，风险特点异于传统金融，行业亟需创新违约风险管理、更新治理策略。本文用机器学习集成模型预测借款人违约风险：先研究互联网消费金融发展现状、借款人特征、违约情况及成因等；再分析消费金融欺诈分类、特点与防范，概述反欺诈模型；接着以Kaggle网站数据为基础，预处理后用XGBoost、LightGBM和CatBoost模型训练，获最优参数，建立综合评价模型实证分析；最后基于研究提出加强违约风险预测管理的建议。

关键词

互联网消费金融，机器学习，GBDT，防欺诈

Research on the Application of GBDT Model in Financial Fraud Prevention

Hongwei Zhu[1,2*], Jiaxing Liu[3], Xingbo Li[2], Haihan Jiao[2]

[1]Hebei Center for Technology Finance and Collaborative Innovation, Baoding Hebei
[2]College of Financial Technology, Hebei Finance University, Baoding Hebei
[3]School of Management, Hebei Finance University, Baoding Hebei

Received: November 11, 2025; accepted: December 5, 2025; published: December 10, 2025

Abstract

In recent years, China's internet consumer finance has developed rapidly, boosting consumer demand,

*通讯作者。

文章引用: 朱宏伟, 刘佳兴, 李兴博, 焦海涵. 基于 GBDT 模型对于金融防欺诈的应用研究[J]. 应用数学进展, 2025, 14(12): 201-214. DOI: 10.12677/aam.2025.1412500

opening up new lending channels for low-income groups, and driving economic growth. However, its rapid expansion has also led to an increase in lending default risks, with risk characteristics differing from those of traditional finance. The industry is in urgent need of innovating default risk management and updating governance strategies. This paper uses machine learning ensemble models to predict borrowers' default risks: first, it examines the development status of internet consumer finance, borrower characteristics, default situations and their causes; second, it analyzes the classification, characteristics and prevention of consumer financial fraud, and outlines anti-fraud models; third, based on data from Kaggle, after preprocessing, it trains XGBoost, LightGBM and CatBoost models to obtain optimal parameters and establishes a comprehensive evaluation model for empirical analysis; finally, it puts forward suggestions for strengthening default risk prediction and management based on the research.

Keywords

Internet Consumer Finance, Machine Learning, GBDT, Fraud Prevention

1. 引言

1.1. 研究背景

随着我国消费机制完善，消费对经济增长的作用愈发凸显，消费金融的发展不仅提升了消费层次，也契合了人们对高品质生活的追求。以 80 后、90 后为代表的年轻群体消费观念先进、意愿强烈，且还款能力与意愿较高，进一步推动了消费增长。互联网消费金融作为新趋势，借互联网技术迅速拓展线上平台，规模急剧增长。据 Wind 数据，2022 年疫情的冲击与行业整改很大程度地掣肘了互联网消费金融规模的增加。2022 年，中国互联网消费金融行业放款规模 21 万亿元，同比增长 4.3%；余额规模 6.2 万亿元，同比增长 7.2%。

本文以互联网消费金融借款人的违约风险预测和管理为研究重点，构建风险评估模型，对借款人的违约概率及其原因进行分析，并通过 Kaggle 网站上的 Credit Card Fraud Detection 数据集进行实证验证，为我国互联网消费金融的违约风险管理提供新的技术支持。

1.2. 研究现状

国内外学界已从多维度研究互联网消费金融借款人违约风险，覆盖内外部影响因素、预测指标及方法转型三大方向。

内源性因素方面，周永圣等(2019) [1]指出信用受用户基础信息、行为与心理三维度影响；外源性因素上，钟肖英(2016) [2]聚焦平台服务，Niinimäki J-P (2019) [3]关注借贷双方信息不对称。预测指标体系以个人、信贷、财务为核心，刘鹏翔(2017) [4]、雷舰(2019) [5]等纳入拓展变量丰富其内涵。方法上，已实现从传统数理模型向机器学习及融合算法的转型，谭中明等(2018) [6]构建 GBDT 模型精准预判违约，Dahiya S 等(2018) [7]通过多机器融合学习器提升特征变量选取准确率。

总体来看，现有研究虽实现多维化、多元化与定量化，但存在预测指标与违约成因结合不紧密、新技术应用不足、新客户违约识别框架不完善等问题。据此，本文对比多种机器学习模型，构建 XGBoost-

LightGBM-CatBoost 融合模型，为违约风险预测提供新路径。

2. 理论基础

2.1. 消费金融欺诈分析

消费金融领域，欺诈问题不容忽视，其线上与线下模式下的表现各有差异。线上业务欺诈多集中于授信申请阶段，涵盖用户主动骗贷、身份信息遭黑客盗用、集体欺诈三种类型，其中集体欺诈专业性与危害性尤甚；线下欺诈则常见于传统信贷业务，涉及销售代表、合作商家等多方，中介机构为高额佣金包装用户申请贷款的现象，给消费金融公司带来严重损失。

消费金融欺诈特点突出。主观上，欺诈者骗贷意愿强烈，行为受个体意图驱动；方式上，随着技术发展，欺诈手段不断翻新，覆盖交易、支付等多个环节；频率上，线上业务因用户活跃、隐私泄露等问题，欺诈案件高发；更新上，欺诈团队紧跟技术潮流，手段迭代迅速，监管难以跟上。

针对这些特点，防范欺诈可从三方面着手：完善身份验证机制，整合多方信息精准识别用户；构建全面用户画像，涵盖多维度行为特征；实施事中干预与事后核查策略，及时阻断欺诈并评估处理风险。

2.2. 消费金融反欺诈模型

2.2.1. 机器学习模型

互联网发展使机器学习算法在金融数据分析与风控中应用增多。其分有监督和无监督学习两类，前者用人工标记样本训练，结果依赖预设参数，如决策树等；后者直接对原始数据学习分类，如 K-means 等，实践中常联合多种算法。与传统评分卡模型比，基于机器学习的模型在反欺诈系统优势明显，结合大数据能构建高效风控系统，助力消费金融机构降本增效、优化服务，推动普惠金融与经济增长。

2.2.2. GBDT 模型

GBDT 即梯度提升决策树算法，也叫 MART。它整合简单学习器(多为决策树)，累积预测结果形成最终输出，融合决策树与集成学习理念。具体流程是学习器依次训练，依前一个成果优化，给误判样本更高权重，评估时聚合各学习器权重结果预测。每次迭代新弱学习器调整残差，以构建最终预测值。

2.2.3. XGBoost 模型

XGBoost 是对 GBDT 的优化，主要用于解决分类和回归问题。它采用二阶泰勒展开优化目标函数，能更精准确定下一步最优树模型，涵盖树结构与叶子节点权重。工作时，先假设树结构，计算分裂节点时的损失减少量以选最优分裂点；接着遍历所有节点分裂方案，选属性值作分裂点，算出叶节点权重，量化分裂损失减少，选最优分裂属性。该过程持续至达到预定停止条件，如最大树深度或损失减少量低于阈值。

2.2.4. LightGBM 模型

LightGBM (Light Gradient Boosting Machine，轻量级梯度提升机)是 GBDT 算法高效化的衍生，其在高效并行训练方面表现突出，主要特点是训练速度快、内存消耗低、准确率高，并且支持处理大规模数据，适用于分布式计算。LightGBM 的提出解决处理大规模数据时 GBDT 每次迭代都需遍历所有的数据难以应对的情况，使其更适应于工业级应用。

2.2.5. CatBoost 模型

CatBoost 是基于 GBDT 框架的 Boosting 族进阶算法，与 XGBoost、LightGBM 相近。它采用对称决策树技术，高效处理类别型特征，可解决梯度偏差与预测偏移问题，兼具无需手动预处理类别特征、默认参数表现优、支持 GPU 并行训练、防过拟合、预测快速等优势。

2.2.6. XGBoost-LightGBM-CatBoost 模型

XGBoost-LightGBM-CatBoost 模型融合了 XGBoost、LightGBM 和 CatBoost 算法的优势，通过 5 折交叉验证确定最佳参数优化性能。实现步骤为：交叉验证选单个模型最佳参数；用最佳参数模型输入测试集；单个模型处理数据输入结果，加权平均(权重依预测性能定)；选加权概率最高类别为最终预测结果。此集成方法结合各模型特点，弥补不足，软投票关注预测置信度，结果更可靠精确，单个模型有差异时效果尤佳。

3. 互联网消费金融借款人违约风险预测模型

本章将使用 Kaggle 网站 Credit Card Fraud Detection 数据集进行分析，基于用户信息和用户交易行为建立反欺诈模型，使用和对比 XGBoost 模型、LightGBM 模型、CatBoost 模型和 XGBoost-LightGBM-CatBoost 模型的性能，实现金融反欺诈识别，模型具体流程图如图 1：

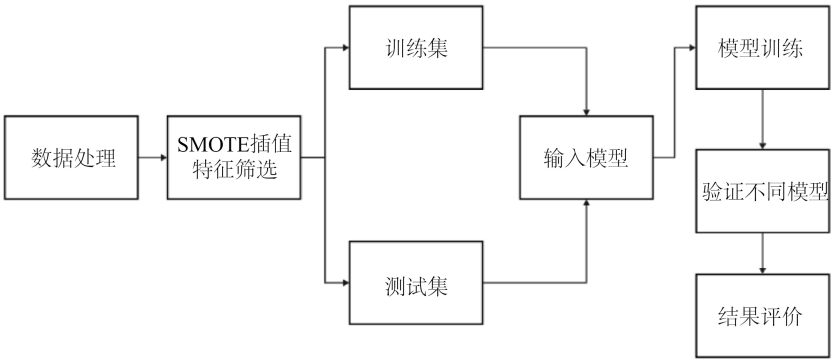

Figure 1. Model flow chart
图 1. 模型流程图

3.1. 数据处理

3.1.1. 数据来源

本文数据来源于 Kaggle 网站 Credit Card Fraud Detection 数据集，共获取了个人信贷申请记录 284,807 条，特征数量 28 个。其中正常还款数据为 284,315 条，违约数据为 492 条，可得其违约率为 0.00173，数据集结果展示如表 1。

Table 1. Dataset results presentation
表 1. 数据集结果展示

样本分析	数量
总样本个数	284,807
违约个数	492
正常个数	284,315
违约率	0.001727485630620034

3.1.2. 数据预处理

(1) 缺失值处理

在数据预处理阶段要进行缺失值处理。经检查缺失值发现数据集不存在缺失值，相关结果如表 2。

Table 2. Results of missing data imputation
表 2. 缺失值处理结果

缺失值处理情况	处理结果
Amount	0
Class	0
正常个数	Int64

(2) 样本均衡

互联网消费金融的信贷业务存在类别不平衡情况。将数据集按 7:3 的比例划分训练集和测试集，其正负样本比值为 0.18%，比例极不均衡，因此要使用算法进行样本均衡处理，本文使用 SMOTE 插值法进行样本均衡处理，见表 3。

Table 3. Balancing class distribution
表 3. 样本均衡处理

前后处理情况	数量
SMOTE 插值前训练集样本数	199364
SMOTE 插值前训练集正样本数	356
SMOTE 插值前训练集负样本数	199008
SMOTE 插值前训练集正负样本比例	0.18%
SMOTE 插值后训练集正负样本数	199008
SMOTE 插值后训练集正负样本比例	50.00%

3.1.3. 数据特征选择

本文使用 Catboost 对特征进行选择，设定阈值为 3%，经筛选后最终保留 13 个特征，基于相关系数的特征筛选见表 4：

Table 4. Feature screening based on correlation coefficients
表 4. 基于相关系数的特征筛选

Feature	Correlation
Time	0.012323
V1	0.101347
V3	0.192961
V4	0.133447
V7	0.187257
V10	0.216883
V11	0.154876
V12	0.260593
V14	0.302544
V15	0.004223
V16	0.196539
V17	0.326481

3.2. 互联网消费金融借款人违约风险预测的实证分析

3.2.1. 模型性能评估指标

为了对比 XGBoost 模型、LightGBM 模型、CatBoost 模型和基于二分类加权软投票的集成模型 XGBoost-LightGBM-CatBoost，本文使用 ROC 曲线、AUC 值、KS 值与准确率、精确率(Precision)、召回率(Recall)、F1-Score 进行对比，其评价指标均可通过混淆矩阵计算得出。

ROC (Receiver Operating Characteristic)曲线是评价分类器性能的一个重要工具。它通过描绘在不同阈值设置下，真正率(TPR, True Positive Rate)和假正率(FPR, False Positive Rate)之间的关系来评估模型性能。TPR 高且 FPR 低的模型性能优异，ROC 曲线越接近左上角，说明模型的区分能力越强。AUC (Area Under Curve)值表示 ROC 曲线下的面积，是判断模型好坏的一个统计量。AUC 值越接近 1，表示模型的预测能力越好，能更有效地区分正负样本。

$$\mathrm{TPR}=\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} \tag{1}$$

$$\mathrm{FPR}=\frac{\mathrm{FP}}{\mathrm{FP}+\mathrm{TN}} \tag{2}$$

在金融反欺诈场景中，正样本极少，负样本占绝大多数，仅使用准确率容易造成模型性能误判。因此，引入精确率(Precision)和召回率(Recall)更为关键：精确率衡量模型预测为欺诈的样本中实际为欺诈的比例，召回率衡量所有真实欺诈样本中被模型成功识别的比例。F1-Score 是二者的调和平均，适用于不平衡分类问题。精确率、召回率和 F1-Score 的计算公式如下：

$$\mathrm{Precision}=\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}} \tag{3}$$

$$\mathrm{Recall}=\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} \tag{4}$$

$$\mathrm{F1}=\frac{2\times\mathrm{Precision}\times\mathrm{Recall}}{\mathrm{Precision}+\mathrm{Recall}} \tag{5}$$

在实际业务中，通常需要在精确率与召回率之间进行权衡：若追求高精确率，可能会漏掉部分欺诈案例；若追求高召回率，则可能将更多正常交易误判为欺诈。因此，可根据业务风险偏好调整分类阈值。

KS (Kolmogorov-Smirnov)值用于衡量模型对正负样本判别能力的指标，它是真正率和假正率之差的最大值，可以直观反映模型区分正负样本的能力。KS 值与模型的判别能力存在正相关，通常 KS 值大于 0.2 即可认为模型具有较好的区分能力。

准确率是指所有被模型正确预测的样本占总样本的比例。准确率较高的模型在整体上具有较好的预测效果。

$$\mathrm{Accuracy}=\frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{FP}+\mathrm{TN}+\mathrm{FN}} \tag{6}$$

3.2.2. 模型训练

在模型开发过程中，参数调整是提升模型预测性能的关键步骤。贝叶斯调优能自动化地调整机器学习模型的超参数以优化模型性能。与传统的网格搜索或随机搜索相比，贝叶斯调优更加高效。贝叶斯优化在开始前，对超参数效果有一个基本假设，通常采用无信息检验，假设超参数在给定范围内均匀分布，这个假设为优化过程提供了起始点。

贝叶斯优化通过采集函数来决定接下来探索的超参数组合，采集函数在决策过程中起着关键作用，它帮助算法在探索和利用之间找到平衡。采集函数包括预期改进(EI)测量一个候选点带来性能提升的预期值，偏好于在模型性能不确定性高的区域探索；概率改进(PI)：计算一个候选点提升性能的概率，偏好于

可能带来性能改进的区域；上限置信界(UCB)：结合均值和方差来选择下一点，权衡已知最优点附近的搜索(利用)和参数空间中不确定性较高区域的搜索(探索)。每次迭代后，利用观察到的性能数据更新超参数效果的后验分布。这种更新利用了贝叶斯规则，综合考虑先验知识和新的观察数据来改进对超参数性能的理解。贝叶斯调优的过程能确保找到最优的超参数组合，且适用性较广。

贝叶斯优化工作具体流程如下：

(1) 初始化：需要选择一组初始的超参数值，这些可以是随机选取的，用来构建初始的后验分布模型。

(2) 迭代过程：使用已构建的后验分布和采集函数来确定下一组待评估的超参数。使用这组超参数运行模型，并对模型性能进行评估。根据获得的新性能数据更新后验分布，这一过程会结合新旧数据不断优化超参数的预测模型。

(3) 终止条件：迭代直到达到预定的最大迭代次数或模型性能改进低于某个阈值。

3.2.3. XGBoost 模型训练与预测

对 XGBoost 进行参数边界设定：学习率(learning_rate)范围在 0.01 到 0.2 之间；树的最大深度(max_depth)范围在 3 到 10 之间；子采样比例(subsample)范围在 0.5 到 0.9 之间；列采样比例(colsample_bytree)范围在 0.5 到 0.9 之间；叶节点最小样本权重和(min_child_weight)范围在 1 到 6 之间；然后调用贝叶斯参数优化函数传入处理后的特征数据、标签数据进行模型训练。经训练后 XGBoost 的最优参数见表 5：

Table 5. Optimal parameter combination of XGBoost
表 5. XGBoost 的最优参数组合

参数(parameter)	Value
目标分数(target)	0.998610
学习率(learning_rate)	0.150079
树的最大深度(max_depth)	5.109427
叶节点最小样本权重和(min_child_weight)	1.947671
子采样比例(subsample)	0.590743
列采样比例(colsample_bytree)	0.553434

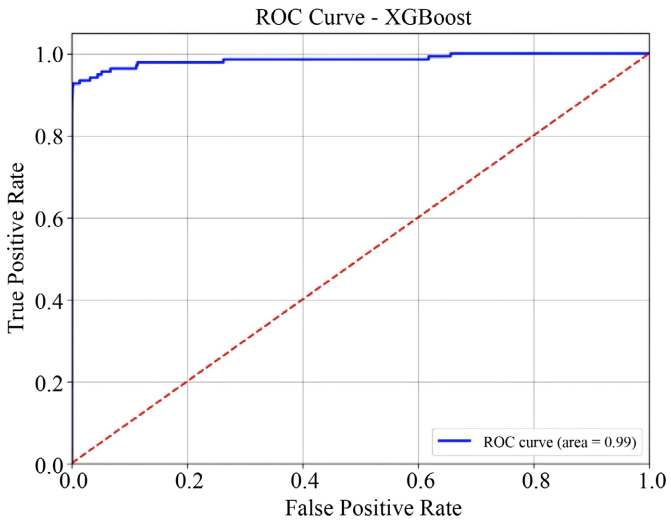

Figure 2. AUC of XGBoost
图 2. XGBoost 的 AUC

使用最优参数组合进行模型预测，得到 XGBoost 的 AUC 为 0.98544，KS 为 0.924196，Accuracy 为 0.999649，Precision = 0.9492，Recall = 0.8235，F1-Score = 0.8816。预测结果如图 2 和图 3。

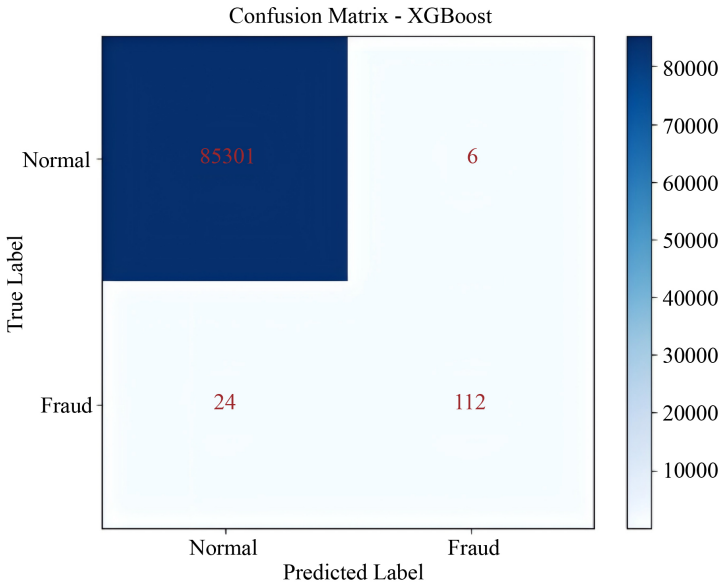

Figure 3. Confusion matrix for the XGBoost model
图 3. XGBoost 混淆矩阵

3.2.4. LightGBM 模型训练与预测

对 LightGBM 进行参数边界设定：学习率(learning_rate)范围在 0.01 到 0.2 之间；叶子节点数量(num_leaves)范围在 20 到 40 之间；叶子节点中的最小数据量(min_data_in_leaf)范围在 10 到 200 之间；树的最大深度(max_depth)范围在 5 到 15 之间；子采样比例(subsample)范围在 0.5 到 0.9 之间；列采样比例(colsample_bytree)范围在 0.5 到 0.9 之间；然后调用贝叶斯参数优化函数传入处理后的特征数据、标签数据进行模型训练，经训练后 LightGBM 的最优参数见表 6。

Table 6. Optimal parameter combination of LightGBM
表 6. LightGBM 的最优参数组合

参数(parameter)	Value
目标分数(target)	0.998611
学习率(learning_rate)	0.065333
树的最大深度(max_depth)	36.503833
叶子节点数量(num_leaves)	25.842892
子采样比例(subsample)	0.646544
列采样比例(colsample_bytree)	0.6727780

使用最优参数构建模型并预测，得到 LightGBM 的 AUC 为 0.958081，KS 为 0.877283，Accuracy 为 0.999462，Precision 为 0.875，Recall 为 0.7721，F1-Score 为 0.82 预测结果如图 4 和图 5。

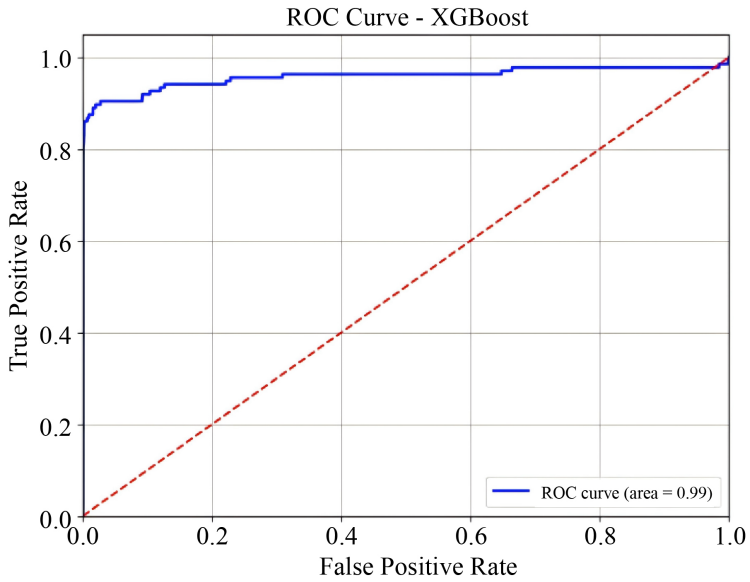

Figure 4. AUC of LightGBM
图 4. LightGBM 的 AUC

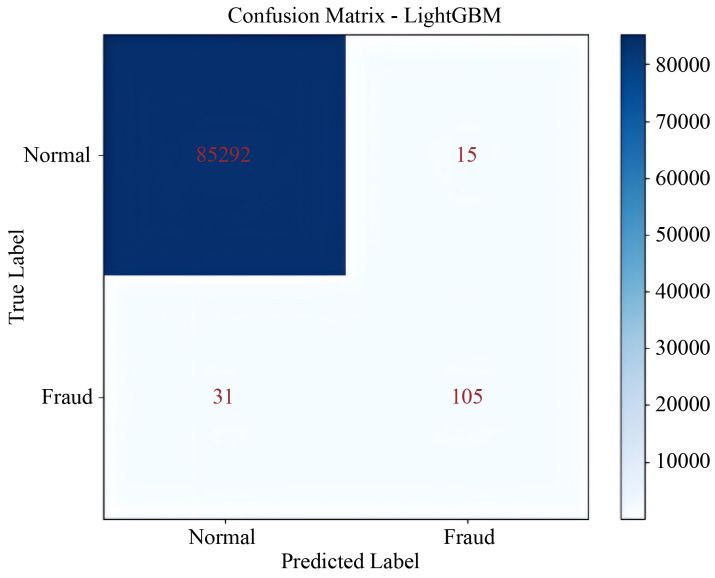

Figure 5. Confusion matrix of the LightGBM model
图 5. LightGBM 的混淆矩阵

3.2.5. CatBoost 模型训练与预测

对 CatBoost 进行参数边界设定：学习率(learning_rate)范围在 0.01 到 0.2 之间；树的深度(depth)范围在 4 到 10 之间；L2 正则化参数(l2_leaf_reg)范围在 1 到 10 之间；特征值边界数(border_count)范围在 5 到 255 之间；正负样本权重平衡系数(scale_pos_weight)范围在 0.01 到 1.0 之间；然后调用贝叶斯参数优化函数传入处理后的特征数据、标签数据进行模型训练，经训练后 CatBoost 的最优参数见表 7。

使用最优参数组合进行模型预测，得到 CatBoost 的 AUC 为 0.980773，KS 为 0.916128，Accuracy 为 0.999482，Precision 为 0.925，Recall 为 0.8162，F1-Score 为 0.8654，预测结果如图 6 和图 7。

Table 7. Optimal hyperparameter combination for CatBoost
表 7. CatBoost 的最优参数组合

参数(parameter)	Value
目标分数(target)	0.998655
特征值边界数(border_count)	44.355738
树的深度(depth)	4.487357
L2 正则化参数(l2_leaf_reg)	25.842892
学习率(learning_rate)	0.014111
正负样本权重平衡系数(scale_pos_weight)	0.232192

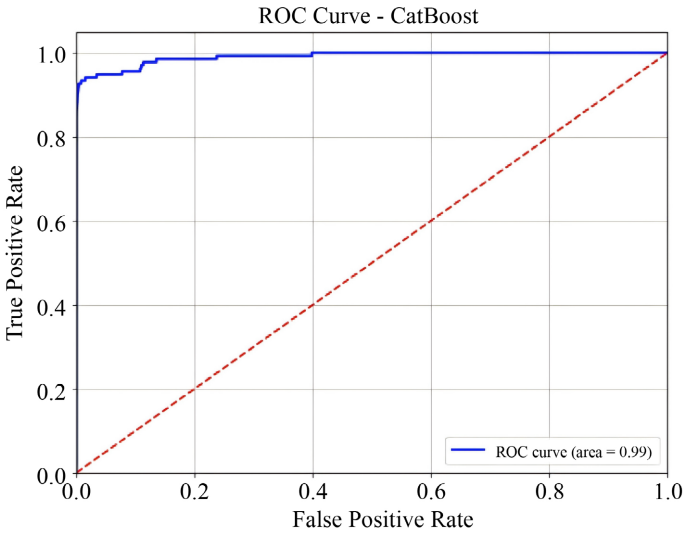

Figure 6. AUC of CatBoost
图 6. CatBoost 的 AUC

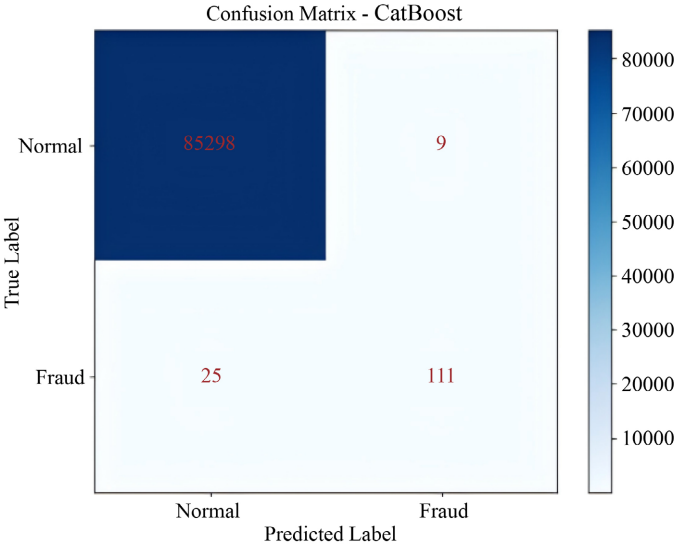

Figure 7. Confusion matrix of the CatBoost model
图 7. CatBoost 的混淆矩阵

3.2.6. XGBoost-LightGBM-CatBoost 模型构建、训练与预测

XGBoost-LightGBM-CatBoost 模型是使用最优参数构建 XGBoost、LightGBM 和 CatBoost 模型通过集成并软投票得来，其 AUC 为 0.990773，KS 为 0.926148，Accuracy 为 0.999602，Precision 为 0.9573，Recall 为 0.8235，F1-Score 为 0.8856，预测结果如图 8 和图 9。

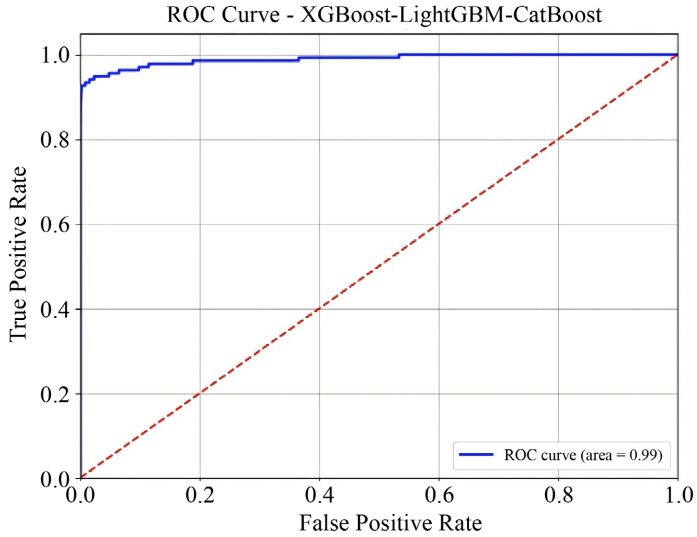

Figure 8. AUC of XGBoost-LightGBM-CatBoost
图 8. XGBoost-LightGBM-CatBoost 的 AUC

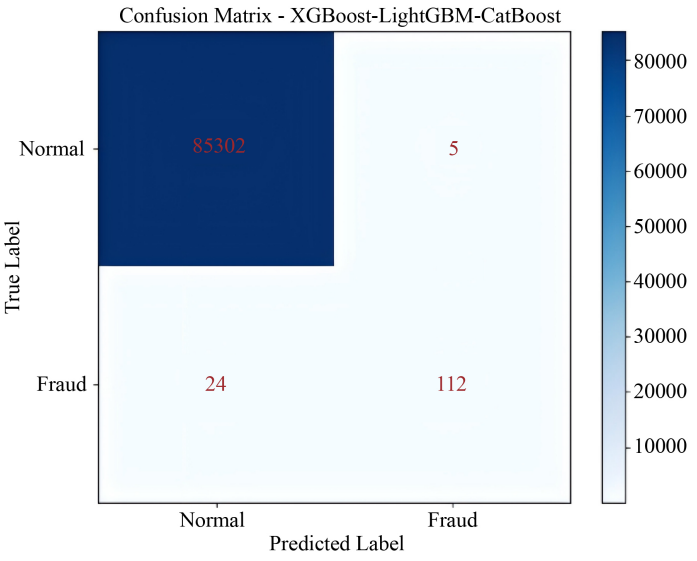

Figure 9. Confusion matrix of the XGBoost-LightGBM-CatBoost ensemble model
图 9. XGBoost-LightGBM-CatBoost 的混淆矩阵

3.2.7. 模型预测结果对比分析

本文比较了单个模型在测试集上的分类效果，各个模型均经过贝叶斯调参达到最优。为保证各单模型分类效果的稳定，所有数据均通过五折交叉验证并取其平均值，预测结果见表 8。

Table 8. Comparison of prediction results among different models
表 8. 不同模型预测结果对比

Model	AUC	KS	Accuracy	Precision	Recall	F1-Score
XGBoost	0.98544	0.924196	0.999649	0.9492	0.8235	0.8816
LightGBM	0.958081	0.877283	0.999462	0.8750	0.7721	0.8200
CatBoost	0.980773	0.916128	0.999482	0.9250	0.8162	0.8654
XLC	0.990773	0.926148	0.999602	0.9573	0.8235	0.8856

经结果对比显示，XGBoost-LightGBM-CatBoost 融合模型在测试集上的准确率为 0.999602，AUC 值为 0.990773，KS 值为 0.926148，精确率和 F1-Score 均优于单一模型，召回率与 XGBoost 相同，但误报更少，表明其在金融欺诈识别任务中具有更好的综合性能。

3.2.8. 特征重要性分析

由上述可知 Catboost 模型对于大规模、多维度的数据处理比其他两个分类器更具有优势，因此本文通过 Catboost 模型对借款人特征进行重要性排序，具体情况如表 9 所示。其列出了各解释变量对模型的贡献情况，将位于前十的各解释变量贡献度由高到低进行排列，其中前十个特征贡献度高达 62.8%。这些指标可以作为预测金融欺诈的关键指标：

Table 9. Feature importance
表 9. 特征重要性

Feature	Importance
V4	12.02496
V14	9.481508
V8	8.919415
V12	6.965279
V1	5.363229
V10	4.580627
V26	4.354876
Amount	3.87673
V17	3.636011
V28	3.620691

3.3. 模型分析结果

互联网消费金融的目标客户以传统金融覆盖不足的低收入、弱信用群体为主，其贷款具有“小额、无抵押、无担保”且用途难追踪的特点。这些特性加剧了平台与借款人的信息不对称，推高逆向选择、道德风险及违约概率，因此需结合用户行为特性，采用区别于传统金融的违约率预测方法。

对比单一模型与 XGBoost-LightGBM-CatBoost 融合模型可见，后者预测精确度更高。该模型整合三种算法核心优势，能更精准捕捉违约行为，实验显示其准确率接近 100%，证明其在分类准确度与实际应用拟合效果上均具备显著优势。

此外，通过使用 CatBoost 模型进行特征筛选，可以进一步提高预测的效率和准确性。CatBoost 能通

过量化特征对预测误差的贡献度，对违约相关指标进行重要性排序，明确核心风险特征，为贷款平台筛选申请者提供清晰参考，助力平台更高效地开展风险评估与管理工作。

4. 加强互联网消费金融违约风险预测管理的建议

4.1. 加强借款人信息采集和处理

互联网提供广泛且实时更新的信息，能有效揭露申请者潜在风险(如在线搜索发现债务或法律问题)，大数据分析可帮金融机构精准评估客户风险与盈利潜力，开展预先审批营销，实现精准定位与高效审批。

依托大数据收集分析信息，可在获客、审批环节形成连贯流程，完善市场策略与风控。因此，收集、筛选并分析借款人信息至关重要，这是优化风险管理与业务效率的核心环节。

4.2. 持续优化预测方法

当下，互联网消费金融借款人违约风险预测是领域焦点与难点。该领域全线上操作、无实物担保、审批快，平台高度依赖风险计量模型管理违约风险，其可综合评估借款人资质、给出客观预测，配合审批流程实现高效操作。目前模型以个人信息、央行信用报告为评估指标，大数据技术推动下，模型整合多维动态指标，预测更精准。未来，风险计量模型将持续创新，通过新技术提升违约预测准确性，降低运营风险。

4.3. 加强征信体系建设

当前，我国除港澳台地区外，个人征信体系仍有较大提升空间。中国人民银行征信中心是唯一规范运作且能有效获取数据的机构，其他规模化个人征信机构较少，导致数据获取渠道、覆盖面受限，难以满足深度与广度需求，且存在数据非法采集、泄露、交易等安全问题。

新型金融让传统信用获取困难人群得以贷款，也要求新型金融机构获取更有效信用数据以降低风险。在此背景下，互联网与大数据推动征信改革：一方面创新业务模式，深化与金融机构合作，依托技术获取全面客户信息，优化风控与服务；另一方面革新数据处理技术，提升处理效率与查询便捷性，助力机构高效用数据开展业务。

5. 结论

互联网消费金融市场的主要风险之一是借款人违约风险，有效预测该风险能减少信息不对称，缓和逆向选择与道德风险，推动市场发展。本文运用 GBDT 算法族构建集成机器融合模型，研究借款人违约状况，得出结论：基于 XGBoost、LightGBM 和 CatBoost 构建的 XGBoost-LightGBM-CatBoost 融合模型，在预测违约风险时，分类性能和预测准确性优于单一模型。为增强违约风险预警与管理，需加强对借款人资料的收集分析以提升风险管理效率，同时开发新预测技术，利用风险计量方法实现全面精准预测。此外，国内应加强征信系统建设，发展信用信息共享，改进征信业务模式与数据处理技术。

鉴于目前互联网消费金融市场数据可得性较差，论文数据来源比较单一，全部是来自 Kaggle 网站 Credit Card Fraud Detection 数据集，单一平台的客户行为相似性可能较大，得出的结论会比较片面，模型的普适性较差。未来的研究方向需要增加客户选取指标的维度，并寻找未经 pca 处理的特征指标。

参考文献

[1] 周永圣, 孙苗苗, 王晶. 互联网消费金融债权信用研究——基于蚂蚁花呗业务模式的分析[J]. 价格理论与实践, 2019(3): 126-129.

[2] 钟肖英. 互联网消费金融中的逆向选择和道德风险研究——基于信号传递和重复博弈的视角[J]. 金融理论与实

践, 2016(12): 59-63.

[3] Niinimäki, J. (2018) Credit Markets under Asymmetric Information Regarding the Law. *The North American Journal of Economics and Finance*, **47**, 380-390. https://doi.org/10.1016/j.najef.2018.05.003

[4] 刘鹏翔. P2P 网贷平台借款人信用风险的影响因素分析——以拍拍贷平台为例[J]. 征信, 2017, 35(3): 71-76.

[5] 雷舰. P2P 网贷借款人信用风险因素分析与对策[J]. 金融理论与实践, 2019(12): 31-39.

[6] 谭中明, 谢坤, 彭耀鹏. 基于梯度提升决策树模型的 P2P 网贷借款人信用风险评测研究[J]. 软科学, 2018, 32(12): 136-140.

[7] Dahiya, S., Handa, S.S. and Singh, N.P. (2018) A Feature Selection Enabled Hybrid-Bagging Algorithm for Credit Risk Evaluation. *Expert Systems*, **34**, e12217. https://doi.org/10.1111/exsy.12217