Pure Mathematics 理论数学, 2026, 16(6), 155-169
Published Online June 2026 in Hans. https://www.hanspub.org/journal/pm
https://doi.org/10.12677/pm.2026.166165

Hans汉斯

基于Transformer-CatBoost融合模型的不平衡数据分类研究

胡译丹*，高　阳#，过子宽

中国石油大学(北京)理学院，北京

收稿日期：2026年4月20日；录用日期：2026年5月26日；发布日期：2026年6月30日

摘　要

在二分类问题中，由于数据失衡处理不当、特征繁杂等问题，捕捉特征之间的复杂相关关系成为挑战。为改善这种现象，建立Transformer-CatBoost融合模型，引入Transformer挖掘用户数据的深层信息与CatBoost抗过拟合实现高效分类。首先，为Transformer编码器增加违约分类头，构建BaseTransformer。然后，将*M*个BaseTransformer集成作为模型的第一层学习器，得到第一层的预测结果与原特征一同输入第二层学习器CatBoost，实现基于Stacking的模型融合。采取多样化的模型评价指标，对比10种数据不平衡处理方法，选择了NCR方法参与实验，随后引入Optuna方法优化模型参数。最后，将模型与各种基准模型比较，借助消融实验证得模型的有效性与可行性，并利用Lending Club数据集证得模型的泛化能力。

关键词

Transformer-CatBoost融合模型，违约识别，召回率，不平衡数据处理，消融实验

Research on Imbalanced Data Classification Based on Transformer-CatBoost Fusion Model

Yidan Hu*, Yang Gao#, Zikuan Guo

College of Science, China University of Petroleum (Beijing), Beijing

Received: April 20, 2026; accepted: May 26, 2026; published: June 30, 2026

Abstract

In binary classification problems, due to problems such as improper handling of data imbalance and

*第一作者。
#通讯作者。

文章引用: 胡译丹, 高阳, 过子宽. 基于 Transformer-CatBoost 融合模型的不平衡数据分类研究[J]. 理论数学, 2026, 16(6): 155-169. DOI: 10.12677/pm.2026.166165

complex features, capturing the correlation between features becomes a challenge. To improve this phenomenon, a Transformer-CatBoost fusion model is established, introducing Transformer to mine the deep information of user data and CatBoost to resist overfitting and achieve efficient classification. Firstly, a default classification head is introduced to the encoder of Transformer to build BaseTransformer. Then, *M* BaseTransformers are integrated as the first layer of the model. The prediction results of first layer together with the original features are input into the second layer CatBoost to build the stacking fusion model. Ten data imbalance processing methods are applied for comparison by using a variety of model evaluation indicators, and the NCR method is selected. Then the Optuna method is introduced to optimize the model parameters. Finally, the model is compared with various benchmark models. The effectiveness and feasibility of the model are verified by ablation experiments. And we prove the generalization ability of the model using the Lending Club dataset.

Keywords

Transformer-CatBoost Fusion Model, Default Identification, Recall, Imbalanced Data Processing, Ablation Experiments

1. 引言

二分类问题广泛存在于金融风控、医疗诊断、工业故障检测、网络入侵识别等诸多领域，然而这些问题常面临同样的困难：数据集中正负样本的分布不平衡且特征间呈现复杂的非线性关系。不平衡数据指各类别的分布不均衡，通常表现为：一种类别样本数量充足，而另一种我们更为关注类别的样本数量却极少，这种“长尾分布”为二分类模型的性能带来挑战，会引发包括模型偏差、泛化能力差以及评估指标产生误导等一系列问题。除此之外，特征间的交互作用与耦合模式也限制了模型的表达能力[1] [2]。

大量研究表明，二分类问题已从单一模型发展为混合模型，大数据分析技术发挥着越来越重要的作用。单一的模型主要分为四种：统计学模型、简单机器学习模型、深度学习模型、集成学习模型。Charizanos等提出了一种基于统计学的二分类模糊逻辑回归框架，为系数输入和输出融入了三角模糊数的组合形式，并能输出清晰的分类结果[3]。Hazarika 等提出了一种新型密度加权孪生 SVM 方法解决不平衡数据的二分类问题，方法在模型训练阶段，根据训练数据点的重要性来赋予权重，并基于现实世界数据集的 F1 分数与 G 均值进行验证，证得有效性[4]。Dawkrajai 等利用加权的 GRU 算法，对 GRU 架构内的输入权重、循环权重和偏置权重这三类权重进行调整，并生成不同不平衡比例的数据集展开测试，证得模型优越性[5]。Akinjole 等对决策树、SVM、XGBoost、AdaBoost、随机森林、MLP 等进行了比较分析并结合多种重采样方法，证得集成学习在违约识别的不平衡数据二分类中具有显著优势[6]。

由于单一模型存在上述局限性，融合模型结合不同算法的优势，准确性与鲁棒性显著提升。Fei 等基于 Stacking 建立了 RF、XGBoost、LightGBM 与逻辑回归的融合模型，通过真实的数据检验与多样评价指标，证得了模型优越性[7]。陈玉沂等基于 Voting 将 LightGBM、XGBoost、CatBoost 进行融合提升了模型性能，并采用多种解释方法进行结果分析[8]。Zohair 等结合 ANN、AdaBoost、随机森林方法，并与逻辑回归构建融合模型，解决了糖尿病患者的二分类、多分类问题与患病严重程度的预测[9]。蔡青松等将 LightGBM、DeepFM 和 CatBoost 选作基模型，CatBoost 作为次模型，利用模型融合提升性能[10]。

虽然融合模型为违约识别带来新思路，但现有模型经常忽视数据失衡的处理，且复杂特征关系使得违约识别的可靠性更具挑战。为改善这类问题，本文与之前研究最大的不同主要体现在以下三个方面：

1) 构造 BaseTransformer：以 Transformer 编码器为基础，应用 GELU 激活函数、简化位置编码、添加少数类输出头与少数偏置，提高对复杂特征的捕捉能力。

2) 基于 Stacking 原理构建 Transformer-CatBoost 融合模型：将 M 个 BaseTransformer 集成作为第一层学习器，输出结果作为新特征与原特征一同输入第二层学习器 CatBoost，得到违约识别结果。

3) 数据不平衡处理与 Optuna 优化：初始参数下对比 10 种方法平衡数据，最终选用 NCR 处理训练集数据不平衡；随后引入 Optuna 进行多维混合空间的参数优化。

文章分为 5 个小节，第 1 部分是引言，阐述研究背景与问题。第 2 部分是相关理论，介绍了研究主要用到的方法，第 3 部分是模型介绍，解释了 Transformer-CatBoost 融合模型的结构与流程，第 4 部分是实验与结果分析，包括数据预处理、特征筛选、数据不平衡处理、参数优化、模型对比、消融实验与模型泛化，第 5 部分是对本文研究的总结。

2. 相关理论

2.1. Transformer

Transformer [11]算法是一种基于自注意力机制的深度学习模型，由谷歌团队在 2017 年提出，主要由自注意力机制(Multi-Head Self Attention)和前馈神经网络组成(FNN)构成。其强大的自然语言处理能力(NLP)与长距离依赖建模能力能更好地捕捉用户数据中不同类型指标的潜在关系，多头自注意力机制支持并行计算，还可以动态调整注意力权重，提高了处理长序列数据时的效率与灵活性。

2.2. CatBoost

CatBoost [12]来自俄罗斯的 Yandex 在 2017 年开源的机器学习库，属于 Boosting 算法。它可以利用 Ordered Target Statistics 数据的顺序信息来计算类别特征的统计量，而无须对非数值型特征进行预处理，避免了 One-hot 编码与 Label 编码的弊端，减少了计算成本。不仅如此，CatBoost 采用对称树的结构使得模型在训练过程中更稳定，降低了借贷违约预测过拟合的风险。

2.3. Neighbourhood Cleaning Rule (NCR)

NCR [13]是基于局部邻域分析来清理多数类样本的欠采样技术。它可以识别并移除可能对分类模型性能产生负面影响的多数类样本，并保留远离决策边界的多数类样本。这使得 NCR 相比随机欠采样，能保留更多的数据信息，相比 SMOTE 等方法，不会引入合成样本可能带来的噪声。

2.4. Optuna

Optuna [14]采用基于序列模型的优化方法，适合高维混合空间参数优化，天然支持多目标和约束的优化场景，兼具可解释性、灵活性与高效性。它可以处理连续、整数和类别变量的复杂优化问题，内置剪枝机制允许在实验过程中提前终止表现差的参数配置，大幅节省计算成本。

2.5. 评价指标

实验令违约(isDefault = 1)为正类，不违约(isDefault = 0)为负类。研究采用多指标综合评价体系，包括核心指标召回率、AUC 与平均成本，以及可视化指标 PR-AUC。该评估框架有利于实现有效的风险控制，

减少违约用户的漏判，并保证模型的综合性能。各评价指标介绍如下：

2.5.1. 召回率(Recall)

召回率[15]是衡量分类模型识别正类能力的评价指标，高召回率可以在一定程度上降低违约用户被漏判的可能，计算见式(1)：

$$\text{Recall} = \text{TP}/(\text{TP}+\text{FN}), \tag{1}$$

其中，真阳性 TP 代表实际违约，预测违约样本，假阴性 FN 代表实际违约，预测不违约样本。

2.5.2. AUC

AUC [15]能够衡量模型对正负样本的排序能力，适用于数据不平衡的情形且不受阈值的影响。AUC 也叫做 ROC 曲线下面积，ROC 曲线的横轴为假正率(FPR)是实际不违约被错误预测为违约的比例，纵轴为真正率(TPR)是实际为违约预测违约的比例，计算见式(2)：

$$\text{FPR} = \text{FP}/(\text{FP}+\text{TN}), \text{TPR} = \text{TP}/(\text{TP}+\text{FN}), \tag{2}$$

其中，真阴性 TN 代表实际不违约，预测不违约样本，假阳性 FP 代表实际不违约，预测违约样本。

2.5.3. PR-AUC

PR-AUC [15]为 PR 曲线下面积，是二分类模型的评估指标。该曲线以召回率为横轴，以精确率为纵轴，通过调整分类阈值得到一些不同的精确率和召回率组合，连接这些点便得 PR 曲线，曲线下面积即为 PR-AUC 值，取值范围[0, 1]，越高表示模型性能越好。它与 AUC 的区别是：PR-AUC 侧重于召回率与精确率的权衡，而 AUC 侧重评估模型对正类和负类样本的区分能力。

2.5.4. 平均成本

平均成本[15]是一种成本敏感指标，其对比模型预测结果与真实标签，并为不同类型错误赋值不同业务成本。一般来说，假阴性会导致贷款本金和利息的损失，成本较高；假阳性会错失优质用户并增加审核流程，成本较低。其计算见式(3)，C_{FN} 为将违约客户预测正常所产生的单位成本，C_{FP} 为将正常客户预测违约所产生的单位成本，N 为样本数：

$$\text{Avg\_cost} = \frac{C_{\text{FN}} \cdot \text{FN} + C_{\text{FP}} \cdot \text{FP}}{N}. \tag{3}$$

3. 模型介绍

3.1. 模型整体结构

本文是基于 Stacking 原理建立的 Transformer-CatBoost 融合模型。在 Transformer 编码器的基础上，为主分类头增加违约分类头，实现双头动态融合输出，构成 BaseTransformer，M 个 BaseTransformer 集成作为模型的第一层学习器。由此生成的概率特征与原始特征一同输入第二层学习器 CatBoost，从而输出模型的最终识别结果。Transformer-CatBoost 融合模型的结构见图 1。

3.2. BaseTransformer 的集成

Transformer 主要由编码器和解码器两部分组成，编码器通过自注意力学习输入数据的全局特征表示，提取表格、文本、图像的整体特征，解码器通过掩码自注意力、编码器－解码器注意力机制，逐步生成输出序列[16]。由于分类任务输出是固定类别且静态的，因此不需要解码器的生成过程，去掉解码器有助于简化模型，提升运行效率，我们将它命名为 BaseTransformer。

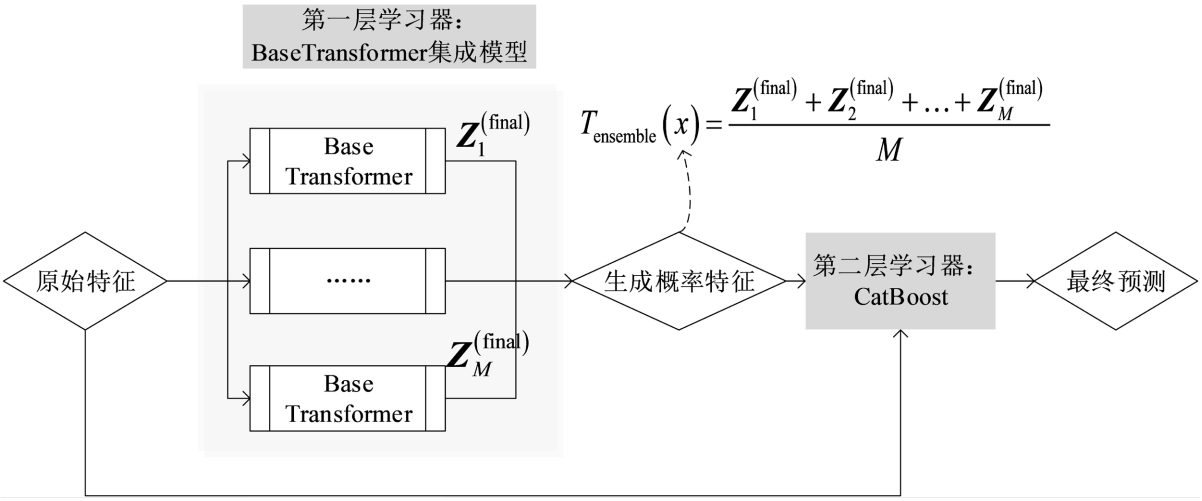

Figure 1. Schematic diagram of Transformer-CatBoost fusion model
图 1. Transformer-CatBoost 融合模型结构图

BaseTransformer 的主要参数有：集成个数 n_model，嵌入维度 d_model，多头注意力机制的头数 nhead，编码器的层数 num_encoder_layer，前馈神经网络的维数 dim_feedforward，dropout 概率，偏置项 pos_weight。图 2 是 BaseTransformer 结构图：

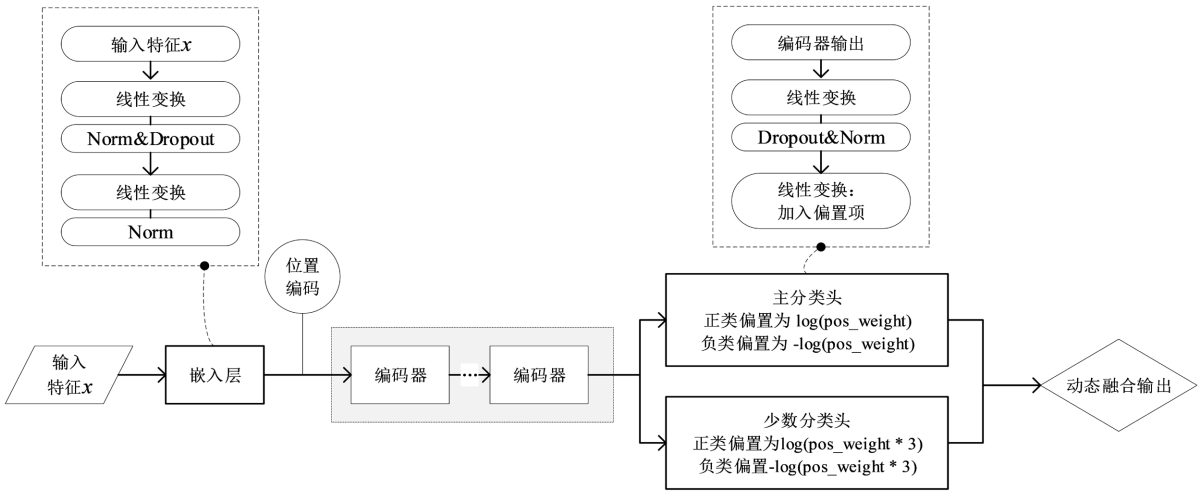

Figure 2. Schematic diagram of BaseTransformer
图 2. BaseTransformer 的结构图

下面介绍 BaseTransformer 的具体流程：

第一步是输入嵌入层(Embedding Layer) [16]。由于表格数据特征之间复杂的非线性关系，传统 Transformer 的单线性变换无法捕捉高阶特征交互，需要深度的非线性嵌入，不仅如此，稀疏特征较多(0 元素为主)需要温和的负值处理，而 ReLU 激活函数对负输入简单归零，因此改用更深的网络结构——非线性激活函数 GELU，进行两次线性变换并添加归一化稳定训练过程，以增强模型对表格数据复杂特征的捕捉能力，得到 $X_e = \text{Embedding}(x)$， $x \in \mathbb{R}^{\text{input\_size}}$ 见公式(4)：

$$X_e = \text{Embedding}(x) = \text{LayerNorm}\left(\text{Linear}_2 \circ \text{Dropout} \circ \text{LayerNorm} \circ \text{GELU} \circ \text{Linear}_1(x)\right). \quad (4)$$

由于表格数据不需要复杂的位置编码，因此将传统正弦(或余弦)函数的位置编码简化，改为简单的可

学习偏移 $X_0 = X_e + P$， P 为可学习参数。

第二步是编码器(Encoder Layer) [16]，使用 GELU 激活函数，编码器第 $l-1$ 层的输出 X_{l-1}，$l = 1,2,\cdots,N$，传递到编码器第 l 层。编码器与多头注意力机制的结构见图 3：

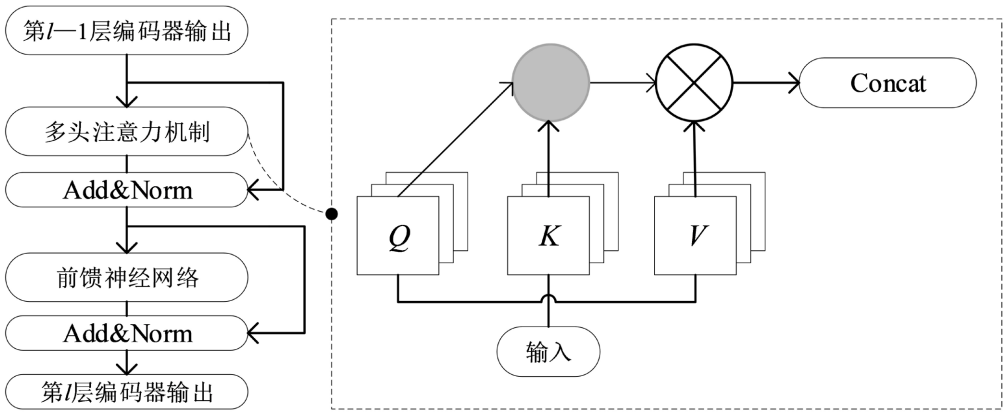

Figure 3. Schematic diagram of the encoder structure
图 3. 编码器结构图

将经过多头注意力机制输出的各个头的结果合并见式(5)，其中 H = nhead，进行残差连接与归一化(Add & Norm)得到 Y_l 见式(6)：

$$\text{MultiHead}(Q,K,V) = \text{Concat}(\text{head}_1, \ldots, \text{head}_H)W^O,\ W^O \in \mathbb{R}^{\text{d\_model}\times\text{d\_model}}, \tag{5}$$

$$Y_l = \text{LayerNorm}\left(X_{l-1} + \text{Dropout}\left(\text{MultiHead}(Q,K,V)\right)\right). \tag{6}$$

然后由双重归一化的前馈神经网络与 Add & Norm 得 X_l 见式(7) (8)，最后编码器的结果记为 X_N：

$$\text{FFN}(Y_l) = W_2 \cdot \text{GELU}\left(\text{LayerNorm}\left(W_1 Y_l + b_1\right)\right) + b_2, \tag{7}$$

$$X_l = \text{LayerNorm}\left(Y_l + \text{Dropout}\left(\text{FFN}(Y_l)\right)\right). \tag{8}$$

第三步是双头分类器输出(Dual-Head Classifier)。构造主分类头与少数类(违约)头动态融合输出，两分类头结构相同，主头负责整体分类，少数头设置更大偏置关注违约数据，具体见公式(9)，主头偏置 b_2^{main} 与少数类偏置 b_2^{minority} 见公式(10)：

$$\begin{aligned} Z^{\text{main}} &= W_2^{\text{main}} \cdot \left\{\text{LayerNorm}\left[\text{Dropout}\left(\text{GELU}\left(W_1^{\text{main}} X_N + b_1^{\text{main}}\right)\right)\right]\right\} + b_2^{\text{main}}, \\ Z^{\text{minority}} &= W_2^{\text{minority}} \cdot \left\{\text{LayerNorm}\left[\text{Dropout}\left(\text{GELU}\left(W_1^{\text{minority}} X_N + b_1^{\text{minority}}\right)\right)\right]\right\} + b_2^{\text{minority}}, \end{aligned} \tag{9}$$

$$b_2^{\text{main}} = \begin{cases} \ln(\gamma), & \text{isDefault} = 1 \\ -\ln(\gamma), & \text{isDefault} = 0 \end{cases},\quad b_2^{\text{minority}} = \begin{cases} \ln(3\gamma), & \text{isDefault} = 1 \\ -\ln(3\gamma), & \text{isDefault} = 0 \end{cases}, \tag{10}$$

其中，γ = pos_weight = 不违约样本数/违约样本数，γ 为模型参数。动态融合输出见公式(11)：

$$Z^{\text{final}} = \sigma(\omega) Z^{\text{main}} + \left(1 - \sigma(\omega)\right) Z^{\text{minority}}, \tag{11}$$

其中，$\sigma(\omega) = \text{sigmoid}(\omega) = 1/\left(1 + e^{-\omega}\right) \in (0,1)$ 是可学习融合权重，可以自适应调整分类头的贡献值。设置它的初始状态为 $\sigma(0.5) = 1/\left(1 + e^{-0.5}\right) \approx 0.6225$，初始倾向主分类头(62.25% vs 37.75%)。

综上，由 M 个 BaseTransformer 得到的结果分别记为 $Z_1^{\text{final}}, Z_2^{\text{final}}, \cdots, Z_M^{\text{final}}$，集成后得到式(12)：

$$T_{\text{ensemble}}(x)=\left(Z_1^{\text{final}}+Z_2^{\text{final}}+\cdots+Z_M^{\text{final}}\right)\Big/M. \tag{12}$$

3.3. 损失函数和优化器

在集成算法的训练中应用加权交叉熵损失函数[17]，主要参数有：模型迭代次数 num_epochs，训练批次大小 batch_size，学习率 lr，L2 范数惩罚项 weight_decay，损失函数权重 weight。损失函数见式(13)：

$$\ell_{loss}=-\frac{1}{N}\sum_{i=1}^{N}\omega_{y_i}\left[y_i\log p_i+(1-y_i)\log(1-p_i)\right], \tag{13}$$

其中，$N=$batch_size 为当前批次的样本总数，y_i 为真实标签(0 或 1)，ω_{y_i} 是类别 y_i 的权重，p_i 是模型预测样本 i 满足 $y_i=1$ 的概率。为提升模型对违约样本的关注，我们给予少数类更多的权重，即 $\omega_1>\omega_0$。

由于 AdamW [18]支持初期大学习率加速收敛，支持正则化与自适应学习率分离，既可避免自适应学习率干扰正则化效果，又抑制大学习率的参数震荡，平衡了稳定性与收敛速度，因此选择 AdamW 动态调整 BaseTransformer 学习率。

3.4. Transformer-CatBoost 融合模型

将 BaseTransformer 集成后得到的概率结果与原特征一同输入第二层学习器 CatBoost。CatBoost 主要参数有：决策树的最大数量 iterations，学习率 learning_rate，单棵树最大深度 depth，违约样本权重 scale_pos_weight，早停指标 eval_metric=AUC，早停耐心值 early_stopping_rounds，目标函数见式(14)：

$$\Phi_{CB}=-\frac{1}{N}\left[\sum_{y_i=0}\log(1-\tau_i)+4\sum_{y_i=1}\log\tau_i\right], \tag{14}$$

其中，y_i 为真实标签(0 或 1)，τ_i 是模型预测样本 i 属于类别 1 (违约)的概率。

4. 实验与结果分析

本文数据来源于阿里云天池的金融风控－贷款违约预测大赛[19]，从中随机抽取 100,000 条样本进行建模分析。数据共包含 47 个字段，其中 46 个为自变量，1 个(isDefault)为因变量。数据集的违约样本占比 19。59%，违约与非违约样本比例约为 1:4。自变量可分为三类：连续变量(如贷款金额、贷款利率等)，离散变量(如贷款等级、就业年限、房屋所有权状况、贷款用途、地区编码等)以及日期变量(如贷款发放月份与信用额度开立月份)。实验的具体流程见图 4：

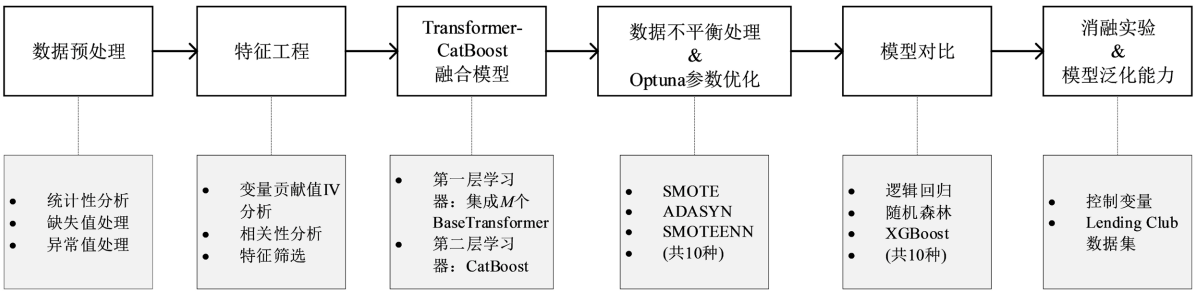

Figure 4. Overall framework of the research
图 4. 研究的总体框架

4.1. 数据预处理

我们先进行数据预处理，主要包括以下三个关键步骤：

1) 统计性分析，分析变量取值范围、极值、标准差等，对文本型离散特征与日期特征标签编码。

2) 缺失值处理，删除含缺失值的 14280 个样本后，剩余完整样本 85,720 个。引入独立样本 t 检验处理连续变量和卡方检验处理离散变量，观察完整组与缺失组在各关键变量上的分布差异。结果显示：虽有部分变量存在统计显著差异($p < 0.05$)，但标准化均值差 SMD 与 Cramér's V 系数均小于 0。1，这表明组间实际差异幅度极小。因此缺失机制可视为近似完全随机缺失(MCAR) [20]，故删除缺失样本不会引起样本选择性偏差，不影响建模的稳健性。因此将缺失样本删除。

3) 异常值处理，由于离散变量长尾分布中的低频取值(非 0)被识别为异常值，而非 0 值恰好是识别违约的关键，所以保留离散低频取值，以维持原始信息的完整性。对于连续变量异常值，采用箱线图法[21]识别，图 5 是异常值比例，可以看出，变量异常值都不超过总样本的 10%。为避免直接删除异常值造成样本量损失和信息失真，本文采用 Winsorization 缩尾方法[22]替换异常值，将识别的异常值替换为相应的上、下边界值(即上四分位数或下四分位数)，从而在减弱极端值影响的同时保留样本原始容量。

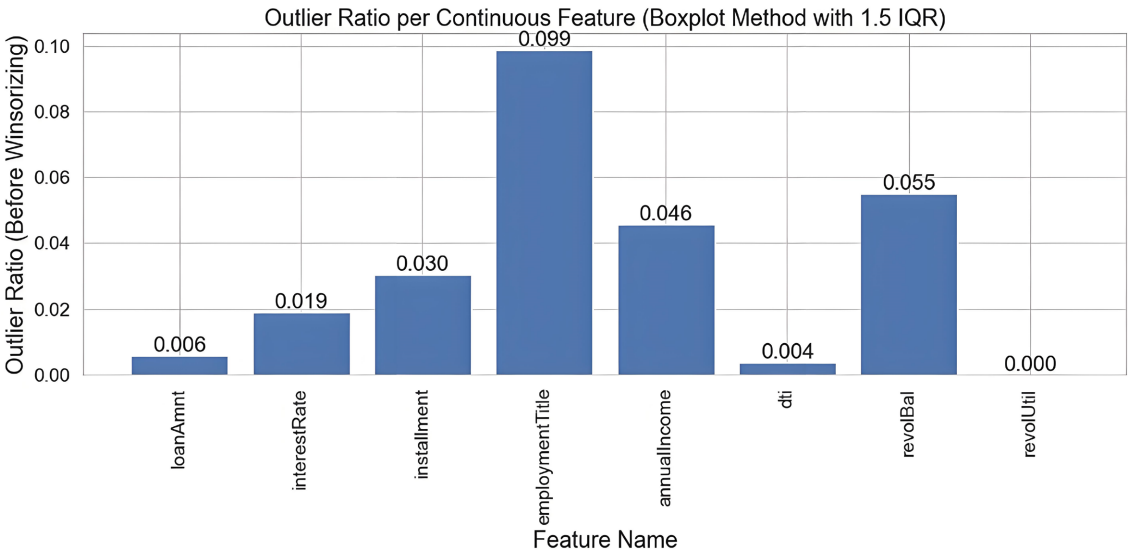

Figure 5. Proportion of outliers
图 5. 异常值比例

4.2. 特征筛选

基于预处理后的数据，我们利用变量贡献值 IV [23]识别出较强区分度变量。设置分箱为 $k \in [5, 6, 7, 8]$，对自变量计算 IV 值(表 1)，得到区分度前五的是：贷款子等级 > 贷款等级 > 贷款利率 > FICO 上限 = FICO 下限：

Table 1. Information value
表 1. IV 值

自变量	$k = 5$	$k = 6$	$k = 7$	$k = 8$	平均值
贷款子等级	0.502207	0.520053	0.522946	0.530975	0.519045
贷款等级	0.490034	0.490034	0.490034	0.490034	0.490034
贷款利率	0.462853	0.481918	0.484914	0.496426	0.481528
FICO 上限	0.11598	0.118853	0.12019	0.121556	0.119145
FICO 下限	0.11598	0.118853	0.12019	0.121556	0.119145

不仅如此，我们引入 SIS 特征筛选常用的三种边际相依度量 Pearson、Spearman、Kendall [24]相关系数(图 6)，其中相关系数超过 0.05 的特征与 IV 值识别出的区分度变量高度重合：

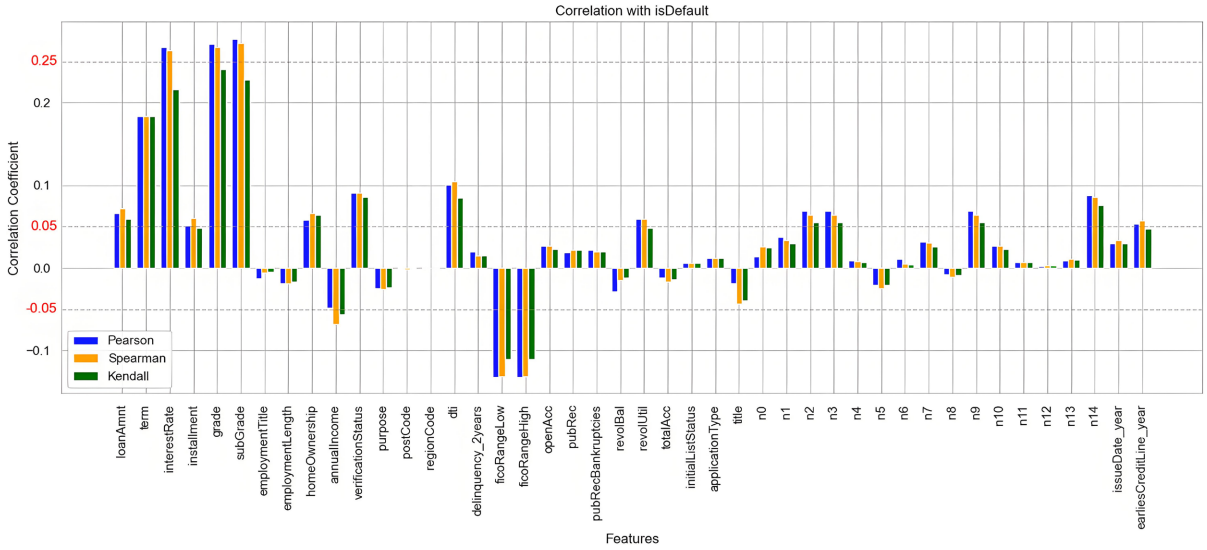

Figure 6. Three correlation coefficient charts
图 6. 三种相关系数图

我们假设相关系数超过 0.05 的变量为真值特征，利用 DC-SIS、χ^2 + MV 混合方法、χ^2 + SIS 混合方法进行特征筛选，其中 χ^2 检验计算离散特征与因变量的相关度，MV 与 SIS 计算连续特征与因变量的相关度。引入评估指标：选中真值特征的比例 SSR、选中特征中真值特征的比例 PSR、逻辑回归预测性能 AUC、选中特征数量 Size、冗余度、稳定性，其中 SSR＝选中真值特征/选中总特征数，PSR＝选中真值特征数/全部真值特征数，冗余度为特征间的平均线性相关度，稳定性衡量特征筛选对数据扰动的鲁棒性。

得到特征筛选性能对比图(图 7)，(a)为评价指标雷达图，可看出 DC-SIS 的 SSR 与 PSR 明显高于另外两种方法，(b)为真值特征筛选情况图，可看出 DC-SIS 可以筛选出全部真值特征，因此我们选择 DC-SIS 进行特征筛选，共筛选出 33 个特征。

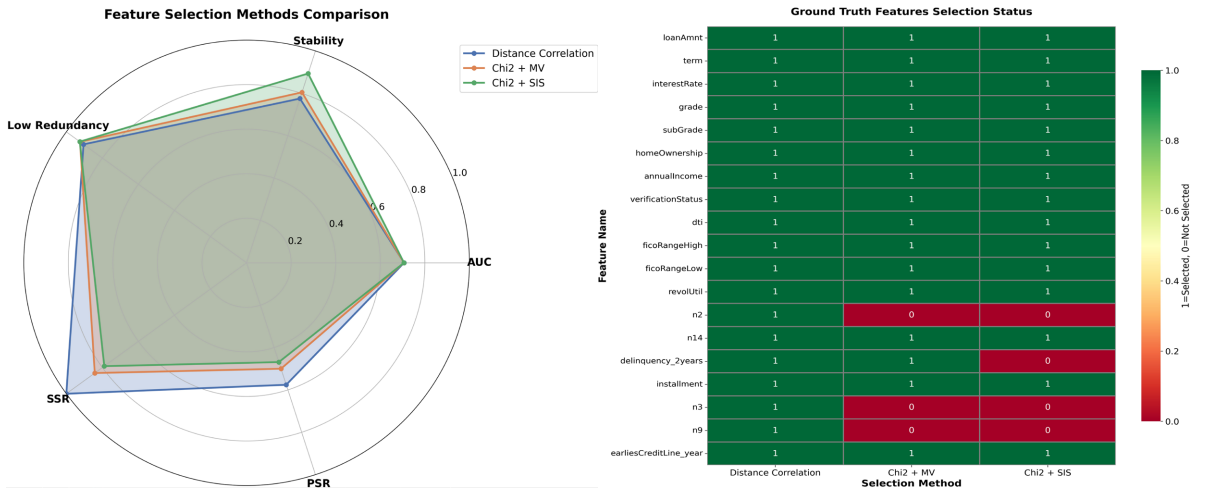

Figure 7. Feature screening performance comparison
图 7. 特征筛选性能对比

4.3. 数据不平衡处理与 Optuna 参数优化

由于数据不平衡会使得模型在预测时倾向于多数类，所以需要进一步处理数据。第一次划分，将数据按 8:2 划分为训练集 A 和测试集 B，为保证模型有效性，只处理 A 的不平衡。第二次划分，将平衡后的 A 按 8:2 划分为训练集 A0 和验证集 A1，将测试集 B 的识别结果作为模型结果。

不同的数据不平衡处理方式会影响模型的识别性能，因此，本文选择 10 种数据不平衡处理方式进行实验，得到初始参数下 Transformer-CatBoost 融合模型的结果见表 2：

Table 2. Handling of data imbalance
表 2. 数据不平衡的处理

方法	召回率%	AUC%	平均成本(元)
Random Oversampling	**89.40**	**71.85**	**471.16**
SMOTE	54.84	67.47	1017.70
SMOTEENN	**78.36**	**68.28**	**638.27**
SMOTETomek	54.58	67.95	1022.08
Borderline-SMOTE	56.20	68.19	989.38
SVM-SMOTE	57.23	69.57	967.28
K-Means SMOTE	42.60	68.02	1225.79
ADASYN	54.08	67.74	1026.22
Tomek Links	39.12	71.81	1264.14
NCR	**62.53**	**71.64**	**868.93**

观察表 2，除 ROS、SMOTEENN、NCR 三种方法，其余方法的召回率都在 60%以下，平均成本都在 900 以上。但由于 ROS、SMOTEENN 在处理数据不平衡的过程中，会生成大量噪声样本，而 NCR 能够有效避免这一问题：NCR 并非简单生成或复制样本，而是通过清理类别重叠区域中的噪声样本进行数据清洗，同时保留原始数据的真实分布结构，使数据集更干净、边界更清晰。由测试损失函数图(图 8)可以看出，SMOTEENN 损失过高，过拟合严重，而 NCR 损失函数取值低，且逐渐降低后趋于稳定，因此利用 NCR 参与实验。

随后我们利用 Optuna [25]进行融合模型的参数优化。建立目标函数见式(15)，并在进行 50 次优化后得到了模型的最优参数：

$$\max\ F\left(\mathrm{AUC},\mathrm{Recall}\right)=F_1-P_{\mathrm{AUC}}-P_{\mathrm{Recall}}$$

$$\text{s.t.}\ \begin{cases} F_1=\dfrac{1}{3}\cdot\left[\dfrac{\mathrm{AUC}+\mathrm{Recall}}{2}+\sqrt{\mathrm{AUC}\cdot\mathrm{Recall}}+\dfrac{2\cdot\mathrm{AUC}\cdot\mathrm{Recall}}{\mathrm{AUC}+\mathrm{Recall}}\right], \\ P_{\mathrm{AUC}}=2\cdot\left(\alpha-\mathrm{AUC}\right)\mathbb{I}_{\{\mathrm{AUC}<\alpha\}},\ \ P_{\mathrm{Recall}}=2\cdot\left(\beta-\mathrm{Recall}\right)\mathbb{I}_{\{\mathrm{Recall}<\beta\}}, \\ \mathrm{Recall}\geq 90\%,\ \ \mathrm{AUC}\geq 70\%, \end{cases} \tag{15}$$

其中，α 与 β 分别为 AUC 最低阈值 0.7 与 Recall 最低阈值 0.9。$\mathbb{I}_{\{\mathrm{AUC}<\alpha\}}$ 为指示函数，当 $\mathrm{AUC}<\alpha$ 为 1，否则为 0。类似的，$\mathbb{I}_{\{\mathrm{Recall}<\beta\}}$ 当 $\mathrm{Recall}>\beta$ 为 1，否则为 0。

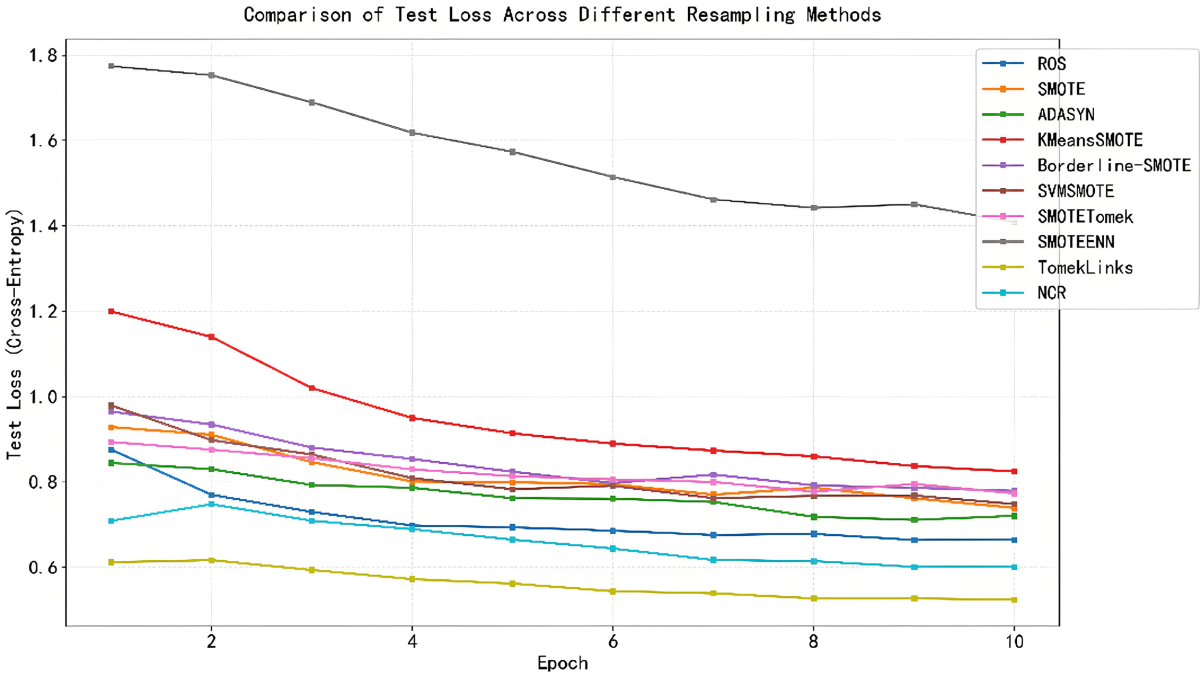

Figure 8. Test loss curve summary chart
图 8. 测试损失曲线汇总图

4.4. 模型对比

为证得本文模型的有效性与可行性，选择了目前在违约预测中表现优秀的模型进行对比，包括 8 种单一模型与 4 种融合模型，如文献[7]中 RF、XGBoost、LightGBM、逻辑回归融合模型(随机森林、XGBoost、LightGBM 为第一层学习器，逻辑回归为第二层学习器)、文献[8]中 LightGBM、XGBoost、CatBoost 的 Voting 融合模型、LightGBM、XGBoost、CatBoost 的 Stacking 融合模型(LightGBM、XGBoost、CatBoost 为第一层学习器，CatBoost 为第二层学习器)，并调整模型参数使得模型表现最优，得到结果见表 3：

Table 3. Comparison of results of different models
表 3. 不同模型的结果对比

模型	召回率%	AUC%	平均成本(元)
逻辑回归	68.94	71.47	716.25
随机森林	69.41	71.30	761.75
SVM	67.03	71.53	751.50
XGBoost	72.89	71.80	705.44
CatBoost	69.94	71.89	749.83
MLP	32.56	71.20	1382.09
RF/XGBoost/LightGBM/LR 融合	46.01	70.49	1150.40
LightGBM/XGBoost/CatBoost 的 Voting 融合	73.83	71.98	690.10
LightGBM/XGBoost/CatBoost 的 Stacking 融合	82.13	71.54	576.18
Transformer-CatBoost 融合模型	**90.05**	**71.63**	**462.84**

由表 3 的结果可以看出，单一模型在召回率上取值较低，可能会带来巨大的漏判损失，风险较大。融合模型中，虽然 RF\XGBoost\LightGBM\LR 融合模型的性能较差，但 LightGBM、XGBoost、CatBoost 的两类融合模型的召回率分别为 73.83%、82.13%，AUC 分别为 71.98%、71.54%，平均成本分别为 690.10 与 576.18，可见选择合适的模型组合与融合方法，对控制漏判率、提高模型综合性能、降低业务平均成本至关重要。相比之下，我们提出的 Transformer-CatBoost 融合模型在二分类任务中优势显著，在各基准模型中召回率最高，平均成本最低，不仅实现了风险最小化，而且保证了二分类模型的综合性能。

不仅如此，我们做出 PR 曲线对比见图 9，可以看出，Transformer-CatBoost 融合模型的 PR-AUC 取值位于各算法第 3 名，且当召回率超过 0.7 后，Transformer-CatBoost 融合模型(深蓝色曲线)的精确率并没有因此降低，反而高于其他算法在精确率上的表现，这进一步说明该模型的性能优势。

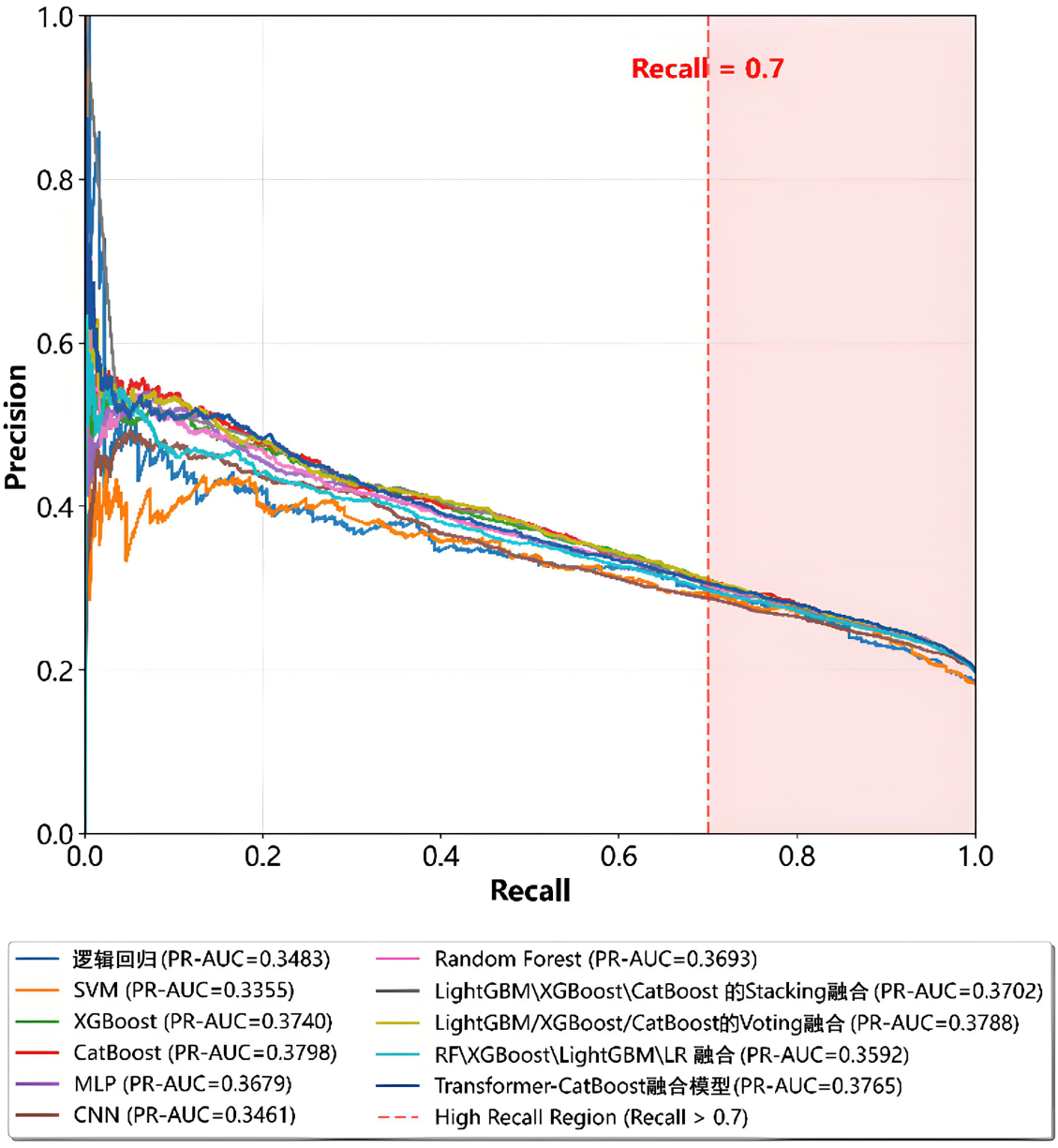

Figure 9. Comparison of PR curves for different algorithms
图 9. 不同算法 PR 曲线对比

4.5. 消融实验

为了验证 Transformer-CatBoost 融合模型各个模块的有效性，控制单一变量对模型进行消融实验[26]。在消融实验中，采用多样化的评价指标，观察不同模块对本文模型的必要作用，分别是：

1) Transformer-CatBoost-A，删除 BaseTransformer 少数分类头，保留 BaseTransformer 的集成与 CatBoost 融合；

2) Transformer-CatBoost-B，删除 BaseTransformer 集成操作，保留 BaseTransformer 少数分类头与 CatBoost 融合；

3) Transformer-CatBoost-A + B，删除 BaseTransformer 少数分类头与 BaseTransformer 集成操作，保留与 CatBoost 融合；

4) Transformer-CatBoost-C，删除 CatBoost，保留 BaseTransformer；

5) Transformer-CatBoost-D，删除 BaseTransformer，保留 CatBoost 用于二分类；

6) Transformer-CatBoost-E，删除 BaseTransformer 集成，保留 BaseTransformer 少数分类头与 CatBoost 融合，并增加单个 BaseTransformer 深度与宽度，使得参数量相当。

7) Transformer-CatBoost-F，更改数据，利用未经过特征筛选的数据进行实验。

得到消融实验结果见表 4：

Table 4. Ablation experiments
表 4. 消融实验

模型	召回率%	AUC%	平均成本(元)
Transformer-CatBoost	**90.05**	**71.63**	**462.84**
Transformer-CatBoost-A	89.84	71.75	467.19
Transformer-CatBoost-B	89.37	71.75	473.11
Transformer-CatBoost-A + B	89.55	71.75	467.48
Transformer-CatBoost-C	65.32	70.99	827.69
Transformer-CatBoost-D	69.94	71.89	749.83
Transformer-CatBoost-E	87.84	71.50	491.86
Transformer-CatBoost-F	89.90	70.73	466.93

观察表 4，BaseTransformer 少数分类头、集成操作提高了模型的召回率，降低了业务平均成本。不仅如此，融合模型召回率取值与单独使用 BaseTransformer 集成或 CatBoost 相比，分别提升 24.73%、20.11%。证得模型的各个成分相互协作以及在二分类任务中的必要性。

4.6. 模型泛化能力

为检验模型的泛化能力与实际应用价值，本文将 Transformer-CatBoost 融合模型应用于 Lending Club 数据集[27]，并与 3.4 节的基准模型对比，得到表 5 结果：

Table 5. Comparison of different model results based on the dataset Lending Club
表 5. 基于数据集 Lending Club 的不同模型结果对比

模型	召回率%	AUC%	平均成本(元)
逻辑回归	81.33	86.89	156.50
随机森林	76.90	91.20	145.25

续表

SVM	64.00	85.60	188.00
XGBoost	34.67	90.11	247.00
CatBoost	79.43	92.67	124.44
MLP	11.08	88.36	352.19
RF/XGBoost/LightGBM/LR 融合	51.90	91.23	190.69
LightGBM/XGBoost/CatBoost 的 Voting 融合	54.11	92.88	182.81
LightGBM/XGBoost/CatBoost 的 Stacking 融合	59.81	92.61	168.81
Transformer-CatBoost 融合模型	**89.33**	**91.06**	**137.75**

观察表 5，Transformer-CatBoost 融合模型在公开数据集 Lending Club 上的性能仍具显著优势，其召回率取值 89.33%，远超其他算法，AUC 取值 91.06%，平均成本为 137.75 元。可见 Transformer-CatBoost 融合模型有一定的泛化能力，具备实用价值。

5. 总结

由于数据不平衡、特征关系不明等原因，提取特征相关关系、改善违约识别模型的性能面临挑战。为改善这一问题，本文提出 Transformer-CatBoost 融合模型。在数据预处理与特征筛选阶段处理了数据缺失与异常，分析 IV 值与相关系数，并对比 DC-SIS、χ^2 + MV、χ^2 + SIS 三种特征筛选方法。在第一层学习器中，对 Transformer 编码器进行 GELU 激活、简化位置编码、添加少数分类头、添加偏置项等构成 BaseTransformer，并集成 BaseTransformer 增加鲁棒性；训练过程选取加权的交叉熵损失函数，采用 AdamW 动态调整模型学习率，并将 CatBoost 作为第二层学习器，输出模型的最终识别结果。在数据不平衡处理中，选择 NCR 处理数据不平衡。在 Optuna 参数优化中，建立目标函数并进行 50 次优化实验。在模型对比中，选择召回率、AUC、PR-AUC、平均成本作为评价指标，将本文模型与各种基准模型对比。实验表明，模型显著优于其他模型。在消融实验中，通过控制变量法证得模型各个成分的有效性与必要性。在模型泛化能力的证明中，将方法迁移到另一公开数据集 Lending Club 上仍表现出色，因此在违约风控等误判代价高、数据不平衡的场景中，模型具有实用价值。

虽然本文在借贷违约预测表现良好，但也存在一些局限性，未来工作将围绕以下几个方面展开：

1) 特征衍生的探索：特征衍生可以带来不同的数据特点，也可以一定程度上改善模型结果；

2) 模型融合的广泛性：Transformer 算法是否可以尝试与多样的模型融合，以达到更好的结果。

基金项目

国家自然科学基金(12171482)。

参考文献

[1] Chiang, J.Y., Lio, Y., Hsu, C.Y., Ho, C. and Tsai, T. (2023) Binary Classification with Imbalanced Data. *Entropy*, **26**, Article 15. https://doi.org/10.3390/e26010015

[2] Zheng, M., Wang, F., Hu, X., Miao, Y., Cao, H. and Tang, M. (2022) A Method for Analyzing the Performance Impact of Imbalanced Binary Data on Machine Learning Models. *Axioms*, **11**, Article 607. https://doi.org/10.3390/axioms11110607

[3] Charizanos, G., Demirhan, H. and İçen, D. (2024) Binary Classification with Fuzzy Logistic Regression under Class Imbalance and Complete Separation in Clinical Studies. *BMC Medical Research Methodology*, **24**, Article No. 145. https://doi.org/10.1186/s12874-024-02270-x

[4] Hazarika, B.B. and Gupta, D. (2022) Density Weighted Twin Support Vector Machines for Binary Class Imbalance

Learning. *Neural Processing Letters*, **54**, 1091-1130. https://doi.org/10.1007/s11063-021-10671-y

[5] Dawkrajai, J., Weerachapichasgul, W., Daosud, W. and Kittisupakorn, P. (2025) Enhancing Fault Diagnosis in Imbalanced Data Using Weighted GRU Architecture. *Engineering Journal*, **29**, 35-44. https://doi.org/10.4186/ej.2025.29.7.35

[6] Akinjole, A., Shobayo, O., Popoola, J., Okoyeigbo, O. and Ogunleye, B. (2024) Ensemble-Based Machine Learning Algorithm for Loan Default Risk Prediction. *Mathematics*, **12**, Article 3423. https://doi.org/10.3390/math12213423

[7] Fei, H. and Huang, H. (2019) Research on Internet Credit Risk Prediction Based on Model Fusion. *Statistics and Application*, **8**, 823-834.

[8] 陈玉沂, 刘高勇, 蔡焕仪. 个人信贷违约预测机器学习模型的解释性方法研究[J]. 现代计算机, 2024, 30(13): 68-72.

[9] Zohair, M., Chandra, R., Tiwari, S. and Agarwal, S. (2024) A Model Fusion Approach for Severity Prediction of Diabetes with Respect to Binary and Multiclass Classification. *International Journal of Information Technology*, **16**, 1955-1965. https://doi.org/10.1007/s41870-023-01463-9

[10] 蔡青松, 吴金迪, 白宸宇. 基于可解释集成学习的信贷违约预测[J]. 计算机系统应用, 2021, 30(12): 194-201.

[11] 张瑶娜, 卓佩妍, 刘自金, 等. 基于 Transformer 编码器和残差网络的信贷违约预测模型[J]. 计算机应用, 2024, 44(S1): 324-329.

[12] Prokhorenkova, L., Gusev, G., Vorobev, A., *et al*. (2018) CatBoost: Unbiased Boosting with Categorical Features. *Advances in Neural Information Processing Systems*, **31**, 1-8.

[13] Otoo, G., Appati, J.K., Yaokumah, W., Soli, M.A.T., Nwolley, S.J. and Ludu, J.Y. (2021) Evaluation of Data Imbalance Algorithms on the Prediction of Credit Card Fraud. *International Journal of Intelligent Information Technologies*, **17**, 1-26. https://doi.org/10.4018/ijiit.289967

[14] Khan, M.S., Peng, T., Khan, M.A., Khan, A., Ahmad, M., Aziz, K., *et al*. (2025) Explainable Automl Models for Predicting the Strength of High-Performance Concrete Using Optuna, SHAP and Ensemble Learning. *Frontiers in Materials*, **12**, Article ID: 1542655. https://doi.org/10.3389/fmats.2025.1542655

[15] Hussein Sayed, E., Alabrah, A., Hussein Rahouma, K., Zohaib, M. and Badry, R.M. (2024) Machine Learning and Deep Learning for Loan Prediction in Banking: Exploring Ensemble Methods and Data Balancing. *IEEE Access*, **12**, 193997-194019. https://doi.org/10.1109/access.2024.3509774

[16] Vaswani, A., Shazeer, N., Parmar, N., *et al*. (2017) Attention Is All You Need. *Advances in Neural Information Processing Systems*, **30**, 1-27.

[17] Ho, Y. and Wookey, S. (2019) The Real-World-Weight Cross-Entropy Loss Function: Modeling the Costs of Mislabeling. *IEEE Access*, **8**, 4806-4813. https://doi.org/10.1109/access.2019.2962617

[18] Zhou, P., Xie, X., Lin, Z. and Yan, S. (2024) Towards Understanding Convergence and Generalization of AdamW. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **46**, 6486-6493. https://doi.org/10.1109/tpami.2024.3382294

[19] 阿里云天池实验室. 违约贷款数据集[DB/OL]. https://tianchi.aliyun.com/dataset/140861, 2022-11-11.

[20] Pham, T.M., Pandis, N. and White, I.R. (2022) Missing Data, Part 2. Missing Data Mechanisms: Missing Completely at Random, Missing at Random, Missing Not at Random, and Why They Matter. *American Journal of Orthodontics and Dentofacial Orthopedics*, **162**, 138-139. https://doi.org/10.1016/j.ajodo.2022.04.001

[21] Fitrianto, A., Wan Muhamad, W.Z.A., Kriswan, S. and Susetyo, B. (2022) Comparing Outlier Detection Methods Using Boxplot Generalized Extreme Studentized Deviate and Sequential Fences. *Aceh International Journal of Science and Technology*, **11**, 38-45. https://doi.org/10.13170/aijst.11.1.23809

[22] 李逸君, 王思淼, 赵沐歌, 等. 具有缺失值及异常值的时间序列处理与再筛选机制[J]. 实验科学与技术, 2025, 23(6): 34-42.

[23] Dastane, O., Goi, C.L. and Rabbanee, F.K. (2023) The Development and Validation of a Scale to Measure Perceived Value of Mobile Commerce (MVAL-SCALE). *Journal of Retailing and Consumer Services*, **71**, Article 103222. https://doi.org/10.1016/j.jretconser.2022.103222

[24] El-Hashash, E.F. and Shiekh, R.H.A. (2022) A Comparison of the Pearson, Spearman Rank and Kendall Tau Correlation Coefficients Using Quantitative Variables. *Asian Journal of Probability and Statistics*, **20**, 36-48. https://doi.org/10.9734/ajpas/2022/v20i3425

[25] Lai, L., Lin, Y., Liu, Y., Lai, J., Yang, W., Hou, H., *et al*. (2024) The Use of Machine Learning Models with Optuna in Disease Prediction. *Electronics*, **13**, Article 4775. https://doi.org/10.3390/electronics13234775

[26] Shipman, A., Mead, D., Feng, Y., Escribano, J., Angeloudis, P. and Demiris, Y. (2022) Novel Trajectory Prediction Algorithm Using a Full Dataset: Comparison and Ablation Studies. 2022 *IEEE* 25*th International Conference on Intelligent Transportation Systems* (*ITSC*), 8-12 October 2022, 2401-2406.

[27] 天池实验室. Lending Club 贷款数据[DB/OL]. https://tianchi.aliyun.com/dataset/19517, 2019-04-15.