

基于集成欠采样的企业财务困境预警方法研究

武珈薇, 范 宏*

东华大学旭日工商管理学院, 上海

收稿日期: 2025年12月22日; 录用日期: 2026年1月14日; 发布日期: 2026年1月26日

摘要

企业财务困境预警(FDP)是防范系统性金融风险、提升资本配置效率的重要工具。然而,现实企业数据中困境样本(ST企业)比例极低,导致模型易受多数类样本支配,出现识别失衡与边界偏置。为应对这一挑战,本文系统探讨了基于集成欠采样的财务困境预警模型,旨在通过样本再平衡策略提升模型对少数类企业的识别能力与跨时间窗口的稳健性。本文基于2010~2022年中国A股上市公司财务数据,构建涵盖偿债能力、盈利能力、成长性与营运效率等维度的综合指标体系,设置四个预测窗口(T-1至T-4)以考察时间跨度对模型性能的影响。以决策树为基学习器,本文对比分析了传统欠采样方法(ENN, TomekLink, NearMiss)与集成欠采样方法(RUSBoost, EasyEnsemble, HUE)的分类性能。实证结果表明,传统欠采样虽能在短期窗口部分改善少数类识别率,但其整体性能随时间延长显著下降;相较之下,集成欠采样方法在识别能力、分类平衡性与跨期稳定性方面均表现更优。其中, HUE模型在T-1窗口下实现Gmean、AUC等指标的最优表现,并在中长期窗口中保持较高的鲁棒性,并通过Friedman显著性检验进一步验证了HUE与EasyEnsemble的性能稳健性。研究结果表明,集成欠采样策略可在极度不平衡条件下显著提升财务困境企业的识别能力与预测稳定性,为复杂金融场景下的不平衡学习提供了可验证的实证依据,也为监管机构和投资者构建更具前瞻性的风险预警体系提供了有益参考。

关键词

财务困境预警, 集成欠采样, 非平衡数据, 机器学习

Research on an Enterprise Financial Distress Prediction Method Based on Ensemble Undersampling

Jiawei Wu, Hong Fan*

Glorious Sun School of Business and Management, Donghua University, Shanghai

Received: December 22, 2025; accepted: January 14, 2026; published: January 26, 2026

*通讯作者。

Abstract

Financial distress prediction (FDP) serves as an important tool for preventing systemic financial risks and improving capital allocation efficiency. However, in real-world corporate data, the proportion of distressed samples (e.g., Special Treatment “ST” firms) is extremely low, which often leads models to be dominated by majority-class samples, resulting in identification imbalance and boundary bias. To address this challenge, this study systematically investigates financial distress early warning models based on integrated under-sampling, aiming to enhance the model’s ability to identify minority-class firms and its robustness across different time windows through sample rebalancing strategies. Based on financial data of China’s A-share listed companies from 2010 to 2022, this paper constructs a comprehensive indicator system covering dimensions such as solvency, profitability, growth, and operational efficiency. Four prediction windows (T-1 to T-4) are set to examine the impact of time span on model performance. Using decision trees as the base learner, this study compares and analyzes the classification performance of traditional under-sampling methods (ENN, Tomek Link, NearMiss) and integrated under-sampling methods (RUSBoost, EasyEnsemble, HUE). Empirical results show that while traditional under-sampling can partially improve the identification rate of minority classes in short-term windows, its overall performance declines significantly over longer time horizons. In contrast, integrated under-sampling methods demonstrate superior performance in terms of recognition capability, classification balance, and cross-period stability. Among them, the HUE model achieves optimal performance in G-mean, AUC, and other metrics under the T-1 window, and maintains high robustness in medium- to long-term windows. The robustness of HUE and EasyEnsemble is further validated by Friedman significance tests. The research findings indicate that integrated under-sampling strategies can significantly enhance the identification ability and predictive stability of financial distress prediction under highly imbalanced conditions. This study provides verifiable empirical evidence for imbalanced learning in complex financial scenarios and offers valuable references for regulators and investors in constructing more forward-looking risk early warning systems.

Keywords

Financial Distress Prediction, Integrated Under-Sampling, Imbalanced Data, Machine Learning

Copyright © 2026 by author(s) and Hans Publishers Inc.

This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

<http://creativecommons.org/licenses/by/4.0/>



Open Access

1. 引言

企业财务困境预警作为公司治理与金融风险管理的重要组成部分，是维护资本市场稳定、促进资源优化配置与保障投资者权益的关键机制。近年来，宏观经济环境复杂性增强、企业债务结构脆弱性上升以及金融监管体系深化，使得及时识别潜在财务困境企业成为监管部门、金融机构及市场参与者广泛关注的重要问题。构建预测性能更优、鲁棒性更强的财务预警模型，对防范系统性金融风险和提升风险管理能力具有重要理论意义与现实价值。特别的是，在我国市场中，被特别处理(ST)机制作为上市公司财务异常的重要标识，已成为识别财务困境的主流标准之一[1][2]。因此，以“ST”事件作为财务困境界定条件已成为近年来国内研究的常用做法。

企业财务困境预警模型经历了从统计模型向机器学习模型再到集成学习与深度学习演进的过程。经典研究如 Fitzpatrick (1932) [3]、Ohlson (1980) [4]奠定了统计判别基础，但难以应对复杂非线性财务数据。

随后，支持向量机、神经网络、决策树等机器学习方法展现出显著优势[5]。近年来，随机森林、梯度提升、CatBoost、Stacking 等集成学习框架进一步提升了模型泛化能力与预测稳定性[6] [7]，已成为金融风控与违约预测的重要技术路线。

尽管方法不断演进，企业财务困境预警仍面临一个主要挑战：样本类别极端不平衡——困境企业样本(如 ST 样本)显著少于正常企业样本。这种不平衡导致模型偏向多数类，使困境企业识别率下降、召回率不理想、误报成本高昂[8]。为解决该问题，研究者提出两类路径：代价敏感学习与数据重采样。代价敏感学习通过设定非对称损失权重降低漏报风险[9]，但要求准确量化损失代价，在金融情景中存在实现困难与参数敏感问题[10]。

相比之下，数据重采样策略因其模型无关性、更强适配性与可解释性，成为不平衡学习的重要方向。过采样技术如 SMOTE 及其改进方法通过合成少数类样本平衡数据分布[11]。与此同时，欠采样策略通过选择性去除多数类样本避免冗余信息与决策边界模糊，如 NearMiss、ENN、聚类欠采样与密度自适应欠采样[12]。与集成学习结合的欠采样方法如 EasyEnsemble [13]等，进一步提升模型稳定性与少数类识别能力。

然而，尽管重采样技术不断演进，现有研究在企业财务困境预测领域仍存在明显不足。一方面，相关研究多将重点放在 SMOTE 等过采样方法上，对欠采样策略的系统性评估仍然有限，导致不同欠采样机制在金融场景中的适用性与效果缺乏充分论证；另一方面，当前研究对于不同类型欠采样算法在高维、异质性特征丰富的企业财务数据下的表现差异尚未形成统一结论。与此同时，对于中国资本市场特有的 ST 制度与财务风险特征背景下的欠采样实证研究亦相对匮乏，难以为中国金融监管与企业风险管理提供针对性的理论与实践参考。

本文针对不平衡问题欠采样策略围绕企业财务困境预警的实际需求开展研究。具体而言，本文通过系统梳理与对比多类欠采样算法及其理论基础，构建统一的实验评价框架，并结合多种机器学习与重采样模型对其性能表现进行全面检验。同时，基于中国上市公司样本开展实证分析，探讨不同欠采样方法在困境企业识别准确性、模型稳定性与边界学习能力方面的差异特征，并进一步分析其在实际风险预警应用中的可行性与优势。本研究旨在为金融不平衡数据建模提供系统化的实证依据与方法论参考，拓展企业财务困境预警的领域理论研究，并为监管部门和金融机构构建精准、稳定且具有可解释性的风险识别体系提供科学支持。

2. 相关模型与方法

2.1. EasyEnsemble

EasyEnsemble 的核心思想是通过自主采样技术，从多数类样本中生成多个子集，每个子集与少数类样本组合形成平衡的训练数据。其具体实现流程如下：

设原始训练集中多数类样本数为 N_{maj} ，少数类样本数为 N_{min} ，算法将进行 N 次迭代，每次迭代包含以下步骤：(1) 从多数类样本中随机抽取 N_{min} 个样本(允许重复抽样)；(2) 将抽取的多数类样本与所有少数类样本组合，形成平衡的训练子集；(3) 在平衡子集上训练一个基分类器 C_i 。最终，通过加权投票法集成所有基分类器，形成最终分类器，其中 α_i 为第 i 个分类器的权重。

$$H(x) = \text{sign} \left(\sum_{i=1}^{N_c} \alpha_i C_i(x) \right)$$

2.2. RUSBoost

RUSBoost 是一种将随机欠采样(RUS)与 AdaBoost.M2 算法相结合的混合集成学习方法，旨在处理类别不平衡问题。其核心过程是在 Boosting 的每一轮迭代中，先对训练数据进行随机欠采样以平衡类别分

布, 再训练基学习器。该算法的具体过程如下:

- (1) 为训练数据集中的每个样本分配初始权重: $D_1(i) = \frac{1}{m}$, 其中 m 为样本总数。
- (2) 迭代训练:
 - (a) 随机欠采样: 根据当前样本权重分布 D_t , 对原始训练集进行随机欠采样, 创建一个临时的平衡数据集 S'_t ;
 - (b) 训练弱分类器: 使用临时数据集 S'_t 及其对应的权重 D'_t 来训练一个弱分类器 h_t ;
 - (c) 计算伪损失: 使用原始数据集和权重来计算弱分类器的伪损失 ϵ_t , 根据伪损失计算本次弱分类器的权重更新参数 $\alpha_t = \frac{\epsilon_t}{1 - \epsilon_t}$;
 - (d) 更新样本权重并归一化: 根据弱分类器 h_t 的表现更新每个样本的权重。对更新后的权重进行归一化处理, 得到新一轮的权重分布。

输出最终模型: 经过 T 轮迭代后, 通过加权投票组合所有弱假设, 得到最终的强分类器。

$$H(x) = \arg \max \sum_{t=1}^T h_t(x, y) \log \frac{1}{\alpha_t}$$

2.3. HUE

财务数据通常具有高维度、多特征间存在复杂非线性关系的特点。在应对财务数据不平衡分类问题时, 欠采样技术因其能直接利用真实数据且计算效率较高而备受关注。然而, 传统的随机欠采样方法存在一个显著缺陷, 即可能因过度删除多数类样本而丢失关键信息, 从而导致模型性能下降。Ng 等人[14]提出了一种创新的基于哈希的欠采样集成模型(HUE)。HUE 算法的执行流程主要包括三个核心步骤: 基于哈希的子空间划分、基于距离的样本选择、以及分类器的训练与集成。

(1) 基于哈希的子空间划分

HUE 首先采用迭代量化(ITQ)[15]这一无监督哈希方法对多数类样本进行处理。ITQ 通过最小化量化误差, 将样本映射到二进制哈希码, 并保留样本间的相似性。其目标函数如下:

$$Q(C, R) = \|C - VR\|^2$$

其中, C 代表哈希码, V 是经过 PCA 投影后的实数特征向量, R 是一个旋转矩阵。通过优化该函数, 获得每个多数类样本的 B 位哈希码。拥有相同哈希码的样本被归入同一个哈希子空间, 子空间总数 $n = 2^B$ 哈希码位数 B 的设定与数据集的不平衡程度相关, 计算公式为:

$$B = \left\lceil \log_2 \frac{3N_{maj}}{N_{min}} \right\rceil$$

(2) 基于距离的样本选择

对于每一个哈希子空间(参考子空间), 模型构建一个训练子集, 包含所有少数类样本和 N_{min} 个从多数类中采样的样本。采样并非随机, 而是采用基于汉明距离的加权策略。样本权重 ω_i 的计算公式如下:

$$\omega_i = \begin{cases} 1, & d_i = 0 \\ \frac{1}{d_i 2^B}, & d_i \neq 0 \end{cases}$$

其中, d_i 表示样本 i 的哈希码与参考子空间哈希码之间的汉明距离。该公式确保了参考子空间内的样本 ($d_i = 0$) 具有最高的权重, 同时邻近子空间 (d_i 较小) 的样本也有机会被选中, 从而在保证子集针对性的同

时, 增强了多样性和鲁棒性。

(3) 分类器的训练与集成

为每一个生成的平衡训练子集训练一个基分类器。最终, 通过多数投票法将所有基分类器集成为最终的强分类器 $H(x)$, 其决策函数为:

$$H(x) = \text{sign}\left(\sum_{i=1}^{N_c} C_i(x)\right)$$

HUE 算法通过哈希函数将多数类样本划分到不同的子空间, 这一机制能够有效处理高维财务特征。同时, HUE 能够保留与少数类样本空间分布相似的健康企业样本, 这些样本往往包含区分财务困境的关键边界信息。对于财务数据中常见的异质性(即不同行业、不同规模企业的财务特征分布差异较大), HUE 的子空间划分能够自适应地形成多个局部平衡的训练集, 从而提升模型对异质财务模式的捕捉能力, 这使其特别适用于企业财务困境预测任务。

3. 实验设计与数据

3.1. 样本与数据来源

本研究基于中国金融困境数据库(CSMAR)提供的上市公司财务数据构建实验样本。数据覆盖 2010~2022 年中国 A 股上市公司, 并以沪深交易所特别处理(ST)制度作为企业财务困境判定标准。ST 标签能够反映企业连续亏损、资产显著缩水或持续经营能力下降等实质性风险, 具有制度严谨性与市场代表性。

为刻画财务风险演化特征, 本文依据企业被实施 ST 前 1 至 4 年公开披露的财务报表数据构建四个预测窗口(T-1 至 T-4), 如表 1。每个数据集均包含 21 项关键财务指标, 涵盖偿债能力、盈利能力、运营效率与发展潜力四大维度, 例如流动比率、资产负债率、总资产增长率等, 能够全面反映企业财务状况及其变化趋势。

随着注册制改革推进, A 股上市公司数量持续增长, 但企业财务健康状态呈长期结构性失衡特征——非 ST 企业占比约 97.1%~97.4%, ST 企业比例不足 3%, 类别极度不平衡。这一现实背景不仅凸显财务困境预测在实践中的挑战性, 也为检验欠采样策略在极端不平衡环境下的有效性提供了实验场景。表 1 给出了各时间窗口数据集的统计特征。

Table 1. Description of the financial distress dataset

表 1. 财务困境数据集描述

周期	样本量	企业数量		比例
		NST	ST	
T-1	54,883	53,441	1442	97.37%:2.63%
T-2	49,168	47,885	1283	97.39%:2.61%
T-3	43,466	42,294	1172	97.30%:2.70%
T-4	37,843	36,738	1105	97.08%:2.92%

为确保模型训练与评估的公平性与稳健性, 各数据集均采用分层抽样, 按 70%:15%:15% 比例划分为训练集、验证集与测试集, 使样本类别分布与整体数据保持一致。验证集与测试集保持等规模设计, 有助于提升超参数调整过程的可靠性, 并确保最终模型性能评估的客观性与无偏性。

3.2. 财务指标体系

本研究构建了一套多维度财务指标体系, 如表 2 所示, 用于企业财务困境的综合预警。该体系从偿债能力、增长能力、盈利能力、营运能力及现金流状况五个关键维度出发, 选取具有代表性的财务比率, 系统反映企业的财务健康状态和潜在风险。各指标定义明确、数据可得, 便于通过趋势分析与异常识别, 有效识别潜在的偿债风险与运营压力。

Table 2. Financial distress indicator system
表 2. 财务困境指标体系

一级指标	二级指标	指标计算
偿付能力	流动比率	流动资产/流动负债
	速动比率	(流动资产 - 存货)/流动负债
	利息保障倍数	息税前利润/利息费用
	资产负债率	负债总额/资产总额
增长能力	资本保值增值率	期末所有者权益/期初所有者权益
	资本积累率	所有者权益增长额/所有者权益
	总资产增长率	本期总资产增长额/期初总资产总额
	营业收入增长率	本期营业收入增长额/上期营业收入总额
营运能力	应收账款周转率	营业收入/应收账款平均余额
	存货周转率	营业成本/存货平均余额
	流动资产周转率	营业收入/流动资产平均余额
	固定资产周转率	营业收入/固定资产平均净额
相对价值指标	总资产周转率	营业收入/总资产平均余额
	营业收入现金含量	经营活动产生的现金流量净额/营业收入
	公司现金流	现金流入 - 现金流出
	企业自由现金流	经营活动现金流量净额 - 资本性支出
盈利能力	资产报酬率	息税前利润(EBIT)/总资产平均余额
	总资产净利润率	净利润/总资产平均余额
	净资产收益率	净利润/股东权益平均余额
	营业利润率	营业利润/营业收入
	成本费用利润率	利润总额/成本费用总额

3.3. 模型评价指标

本研究在财务困境预测(FDP)这类典型的不平衡分类问题中, 仅依赖准确率难以全面评估模型性能。为此, 本文采用 TPR、TNR、Gmean、AUC 和 MCC 五个指标, 从多维度综合评价模型预测能力。

1) TPR

TPR (True Positive Rate)又称灵敏度或召回率, 反映模型对财务困境企业(ST)的正确识别能力。TPR 越高, 表明模型对困境企业的捕捉能力越强。其中 TP 表示真正例(实际为 ST 并预测为 ST), FN 表示假反例(实际为 ST 但预测为 NST)。

$$\text{TPR} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

2) TNR

TNR (True Negative Rate)又称特异度，反映模型对正常企业(NST)的正确识别能力，TNR 越高，说明模型对多数类的判别越准确。其中 TN 表示真反例(实际为 NST 并预测为 NST)，FP 表示假正例(实际为 NST 却预测为 ST)。

$$\text{TNR} = \frac{\text{TN}}{\text{TN} + \text{FP}}$$

3) Gmean

Gmean 衡量模型在少数类与多数类上的平衡识别能力。该指标要求 TPR 和 TNR 同时达到较高水平时才会取得较大值，能够有效避免模型因过度偏向某一类别而导致的评估偏差。在财务困境预测中，Gmean 能够敏感地反映模型对风险企业和正常企业的综合判别质量，特别适用于评估不平衡分类场景下的整体性能。

$$\text{Gmean} = \sqrt{\text{TPR} \times \text{TNR}}$$

4) AUC

AUC 指标即通过计算 ROC 曲线下的面积来衡量模型在不同阈值下的总体区分能力。AUC 值越大，表明模型无论在识别 ST 还是 NST 时均拥有更稳健的判别能力。与单一阈值指标相比，AUC 能较好地概括模型在完整阈值范围内的表现，Jabeur 等人(2021 年)。预测了财务困境，并使用连续三年的企业财务数据的 AUC 度量，评估了 FDP 模型的性能。

5) MCC

MCC (Matthews Correlation Coefficient) 是一种基于混淆矩阵计算的相关系数，用以衡量预测结果与真实标签之间的线性相关程度。

$$\text{MCC} = \frac{\text{TP} \times \text{TN} - \text{FP} \times \text{FN}}{\sqrt{(\text{TP} + \text{FP})(\text{TN} + \text{FN})(\text{TN} + \text{FP})(\text{TN} + \text{FN})}}$$

其值域为 $[-1, 1]$ ，当结果大于 0 为正相关，小于 0 为负相关，等于 0 则表示无相关关系，与其他指标相比，MCC 能在样本类别极度不平衡的情况下依旧保持较好的稳定性与可靠性，能够在不平衡下依旧提供对预测与真实结果关联性的稳健测度，衡量 FDP 算法整体表现。

3.4. 超参数寻优过程

为确保所有对比模型均在最优配置下运行，本研究对所有机器学习模型及欠采样方法进行了系统的超参数优化。本研究采用网格搜索(Grid Search)作为超参数优化的核心策略，模型性能的评估采用 5 折交叉验证进行，并以 Gmean 作为核心优化指标，因其能平衡少数类与多数类的识别性能。各模型的主要调优参数及其搜索空间如下所述：

集成欠采样方法参数空间：(1) EasyEnsemble：主要调优参数为 n_estimators (基分类器的数量，即平行子集的数量)，搜索范围为 $[5, 10, 15, \dots, 95, 100]$ 。其基分类器采用上述优化后的决策树。(2) RUSBoost：调优参数包括 n_estimators 和 learning_rate (学习率，用于控制每轮弱分类器对最终模型的贡献权重，搜索范围 $[0.01, 0.05, 0.1, 0.2, 0.3, 0.4, 0.5, 1.0]$)。基分类器同样为优化后的决策树。(3) HUE：其哈希码位数 B 根据数据集的不平衡比率由 B 动态确定，此为算法内在机制而非可调超参数。主要调优参数为 n_estimators，基分类器亦为优化后的决策树。

传统欠采样方法: 如 ENN、Tomek Links、NearMiss 等, 其核心在于基于规则的样本选择策略, 通常不涉及或仅涉及少量关键参数(如近邻数 k), 且这些参数在相关文献中已有较为公认默认值。为聚焦于比较不同采样策略本身的效果, 并控制实验变量, 本研究对这些方法采用其默认参数设置。经过上述网格搜索与交叉验证流程, 各模型在 T-1~T-4 预测窗口上确定了最优参数组合。

4. 实证结果与分析

4.1. 基准模型性能评估

为建立比较基准, 本研究首先评估了在未采用任何采样处理时, 几种常用分类模型在四个时间窗口(T-1 至 T-4)中的预测性能。**表 3** 汇总了逻辑回归(LR)、随机森林(RF)、梯度提升树(GBDT)以及 LightGBM 等的性能结果。

Table 3. Performance of the baseline model no sampling (T-1~T-4)

表 3. 未采用采样处理时基准模型的性能表现(T-1~T-4)

模型	T-1					T-2				
	AUC	TPR	TNR	Gmean	MCC	AUC	TPR	TNR	Gmean	MCC
LR	0.6882	0.5672	0.6913	0.6261	0.0973	0.6353	0.5556	0.6948	0.6213	0.0946
LDA	0.8132	0.1045	0.9971	0.3228	0.2275	0.8394	0.0171	0.9986	0.1307	0.0632
KNN	0.5360	0.0149	0.9980	0.1221	0.0468	0.5408	0.0085	0.9986	0.0924	0.0311
DT	0.8556	0.2985	0.9946	0.5449	0.4272	0.8355	0.0769	0.9944	0.2766	0.1415
ADA	0.9280	0.2836	0.9936	0.5308	0.3978	0.9339	0.1880	0.9922	0.4319	0.2731
RF	0.9270	0.2388	0.9971	0.4880	0.4072	0.9049	0.0598	0.9969	0.2442	0.1427
GBDT	0.9393	0.2388	0.9968	0.4879	0.4023	0.9315	0.1709	0.9930	0.4120	0.2617
LightGBM	0.9429	0.2687	0.9956	0.5172	0.4123	0.9363	0.0940	0.9930	0.3055	0.1551
XGBoost	0.9400	0.2761	0.9966	0.5246	0.4376	0.9321	0.1709	0.9922	0.4118	0.2522
模型	T-3					T-4				
	AUC	TPR	TNR	Gmean	MCC	AUC	TPR	TNR	Gmean	MCC
LR	0.5476	0.3830	0.6864	0.5127	0.0252	0.4735	0.3085	0.6975	0.0024	0.0946
LDA	0.7537	0.0106	0.9987	0.1031	0.0400	0.7759	0.0319	0.9958	0.0709	0.0632
KNN	0.5237	0.0000	0.9990	0.0000	0.0053	0.5298	0.0000	0.9996	-0.0036	0.0311
DT	0.7921	0.0638	0.9955	0.2521	0.1271	0.7707	0.0213	0.9958	0.0454	0.1415
ADA	0.8628	0.0957	0.9955	0.3087	0.1823	0.8215	0.0426	0.9951	0.0873	0.2731
RF	0.8605	0.0000	1.0000	0.0000	0.0000	0.8477	0.0000	1.0000	0.0000	0.1427
GBDT	0.8571	0.0957	0.9990	0.3093	0.2618	0.8352	0.0426	0.9958	0.0947	0.2617
LightGBM	0.8707	0.1170	0.9981	0.3418	0.2673	0.8414	0.0638	0.9970	0.1552	0.1551
XGBoost	0.8743	0.0745	0.9990	0.2728	0.2224	0.8493	0.0426	0.9992	0.1628	0.2522

总体而言, 集成学习模型(RF, GBDT, LightGBM)在 AUC 指标上表现较优, T-1 窗口的平均 AUC 超过 0.91。然而, 这些模型对少数类样本的识别能力较弱, 其真阳性率(TPR)普遍低于 0.55。与此同时, 真阴性率(TNR)均维持在 0.95 以上, 说明模型在分类过程中倾向于将样本判断为多数类, 呈现出“高特异

度、低灵敏度”的典型类别不平衡偏差。

从综合性能指标来看, Gmean 与 MCC 的整体水平偏低, 在 T-1 时间窗口上, 平均值分别为 0.46 和 0.32, 表明传统分类器在极端不平衡的数据分布下, 难以有效识别困境企业的关键特征, 因此有必要引入采样或代价敏感学习机制以提升模型性能。此外, 随着预测时间窗口从 T-1 延长至 T-4, 各模型的判别能力呈下降趋势, AUC 指标平均降低约 10.9%, 反映出财务指标的时效性对困境企业预测具有重要影响。

因此, 基于基准模型的结果可以得出以下启示: (1) 单纯依靠模型结构优化(如更深的集成树模型)不足以应对极度不平衡分布; (2) 在长期预测中, 信息衰减进一步加剧模型的偏向性。由此可见, 若要提升对少数类企业的识别性能, 需在模型训练阶段引入样本再平衡策略, 以改善类别分布并强化边界学习。

4.2. 传统欠采样与集成欠采样方法性能评估

在基准模型性能受限的基础上, 为进一步提高模型对少数类的识别能力, 本文引入了传统欠采样与集成欠采样两类方法, 并在四个时间窗口下系统比较其分类性能(表 4~7)。所有欠采样方法均基于决策树作为基学习器, 以保证模型结构的一致性, 从而使性能差异主要反映采样机制本身的影响, 而非分类器能力差异。

整体来看, 相比于分类性能较好的基准模型 LR、LightGBM、XGBoost, 欠采样策略更能够显著改善模型对少数类(困境企业)的识别能力(TPR), 但其具体效果随算法机制与预测窗口的变化而有所差异。传统欠采样方法(如 ENN、Tomek Link 和 NearMiss)主要通过删除边界样本、噪声样本或冗余样本来缓解类别分布失衡。实验结果显示, 这类方法在一定程度上提升了模型的少数类识别率, 但改进幅度有限, 且在部分窗口下仍存在明显的性能不均衡现象。

以 T-1 窗口为例(表 4), ENN 与 Tomek Link 在保持较高特异度($TNR \approx 0.995$)情况下, TPR 仅为 0.32, 导致 Gmean 与 MCC 均处于中等水平(约 0.56)。NearMiss 虽然通过优先选择靠近少数类的多数类样本大幅提升了 TPR, 但同时造成 TNR 显著下降(仅 0.1436), Gmean 仅 0.3737, 总体平衡性较差。相比之下, 集成欠采样方法整体表现更为优越。EasyEnsemble 在 T-1 窗口下实现了 $AUC = 0.9427$ 与 $Gmean = 0.8955$ 的较优表现, TPR 达 0.9037, TNR 亦维持在 0.8873, 呈现出明显的均衡特征。HUE 的整体表现进一步提升, Gmean 达 0.9021, 较最优的基线模型提升约 0.28, 说明该方法在捕捉少数类信号的同时保持了良好的多数类判别能力。值得注意的是, HUE 的 AUC(0.9476)与 ENN(0.9500)相当, 但前者在 TPR 上提升近 0.6, 显示其在整体判别结构上更具鲁棒性。

Table 4. Predictive performance of different under-sampling methods in T-1
表 4. T-1 窗口下不同欠采样方法的预测性能

类型	模型	AUC	TPR	TNR	Gmean	MCC
欠采样	ENN	0.9500	0.3211	0.9949	0.5652	0.4399
	Tomek	0.9512	0.3211	0.9954	0.5653	0.4486
	NM	0.7216	0.9725	0.1436	0.3737	0.0536
集成欠采样	RUSBoost	0.7615	0.3394	0.9839	0.5779	0.3348
	EasyEnsemble	0.9427	0.9037	0.8873	0.8955	0.3733
	HUE	0.9476	0.9083	0.8961	0.9021	0.3902

在 T-2 窗口(表 5)中, 传统欠采样方法的识别能力进一步下降。ENN 与 Tomek Link 的 TPR 均低于 0.1, 几乎失去识别功能; NearMiss 虽维持高 TPR, 但 TNR 仅 0.12, 导致预测极度偏向少数类。相比之

下, 集成欠采样方法的表现更为稳健。EasyEnsemble 与 HUE 的 TPR 分别为 0.823 与 0.867, Gmean 达 0.844 与 0.865, AUC 均超过 0.90。尤其是 HUE, 其 Gmean 较 ENN 提升 0.55, 表明其在较长预测期下仍能保持良好的分类平衡性与泛化能力。

Table 5. Predictive performance of different under-sampling methods in T-2**表 5.** T-2 窗口下不同欠采样方法的预测性能

类型	模型	AUC	TPR	TNR	Gmean	MCC
欠采样	ENN	0.9265	0.0985	0.9974	0.3135	0.2163
	Tomek	0.9309	0.0985	0.9971	0.3134	0.2103
	NM	0.6046	0.9507	0.1228	0.3417	0.0369
集成欠采样	RUSBoost	0.8980	0.7980	0.8581	0.8275	0.2928
	EasyEnsemble	0.9085	0.8227	0.8659	0.8440	0.3129
	HUE	0.9306	0.8670	0.8625	0.8648	0.3276

对于 T-3 与 T-4 窗口(表 6、表 7), 随着预测期延长, 财务特征对困境状态的解释力显著下降, 所有模型的整体性能均有所回落。然而, 集成欠采样方法仍保持相对稳健的识别能力。其中, HUE 在 T-3 窗口的 Gmean 达 0.776, TPR 为 0.79, 明显高于其他欠采样模型; 在 T-4 窗口中, HUE 依然保持 Gmean = 0.753、TPR = 0.78 的较高水平, EasyEnsemble 也能维持 Gmean = 0.712 的较优水平, 均优于传统欠采样模型 ENN 等。这表明, 尽管财务信号逐渐衰减, 集成欠采样模型不仅显著提升了模型的识别能力与整体判别平衡性, 还在跨时间窗口的预测中表现出更高的稳健性与泛化性。

Table 6. Predictive performance of different under-sampling methods in T-3**表 6.** T-3 窗口下不同欠采样方法的预测性能

类型	模型	AUC	TPR	TNR	Gmean	MCC
欠采样	ENN	0.8472	0.0323	0.9984	0.1795	0.1032
	Tomek	0.8530	0.0323	0.9991	0.1795	0.1216
	NM	0.6483	1.0000	0.0848	0.2912	0.0513
集成欠采样	RUSBoost	0.8163	0.7527	0.7554	0.7541	0.1931
	EasyEnsemble	0.8296	0.7258	0.7875	0.7560	0.2039
	HUE	0.8493	0.7903	0.7618	0.7759	0.2111

Table 7. Predictive performance of different under-sampling methods in T-4**表 7.** T-4 窗口下不同欠采样方法的预测性能

类型	模型	AUC	TPR	TNR	Gmean	MCC
欠采样	ENN	0.8281	0.0060	0.9995	0.0771	0.0345
	Tomek	0.8222	0.0060	0.9993	0.0771	0.0299
	NM	0.5724	0.9107	0.1697	0.3932	0.0365
集成欠采样	RUSBoost	0.7844	0.7024	0.6785	0.6904	0.1370
	EasyEnsemble	0.7943	0.6607	0.7678	0.7123	0.1689
	HUE	0.8248	0.7798	0.7263	0.7525	0.1892

综合四个时间窗口的结果可得出以下结论: (1) 传统欠采样方法虽能在短期窗口内部分改善 TPR, 但其性能易受样本删减与边界失真影响, 整体稳定性较弱; (2) 集成欠采样方法在各窗口下均表现出显著优势, 尤其在中长期预测场景中依然保持较高的 Gmean 与 MCC; (3) HUE 在远期预测中的表现尤为突出, 其在 T-3 与 T-4 窗口下的分类平衡性明显优于其他模型, 验证了其子空间多样化机制在高噪声与特征退化环境下的有效性。

综上所述, 集成欠采样策略在识别能力、模型稳定性及跨时间窗口的适应性方面均显著优于传统欠采样方法。其通过多样化子空间学习与集成融合机制, 不仅提升了模型对困境企业的识别精度, 也有效降低了样本不平衡所引发的性能波动, 验证了集成欠采样方法在企业财务困境预测任务中的可靠性与实用价值。

4.3. 预警时间窗口效应分析

为系统评估预警模型在不同预测周期下的性能稳定性, 本研究分析了各模型在 T-1 至 T-4 四个时间窗口上的表现。图 1 通过箱线图展示了不同算法在 Gmean、AUC、MCC 与 TPR 四项核心指标上的分布, 揭示了模型性能随预测窗口延长的变化趋势。

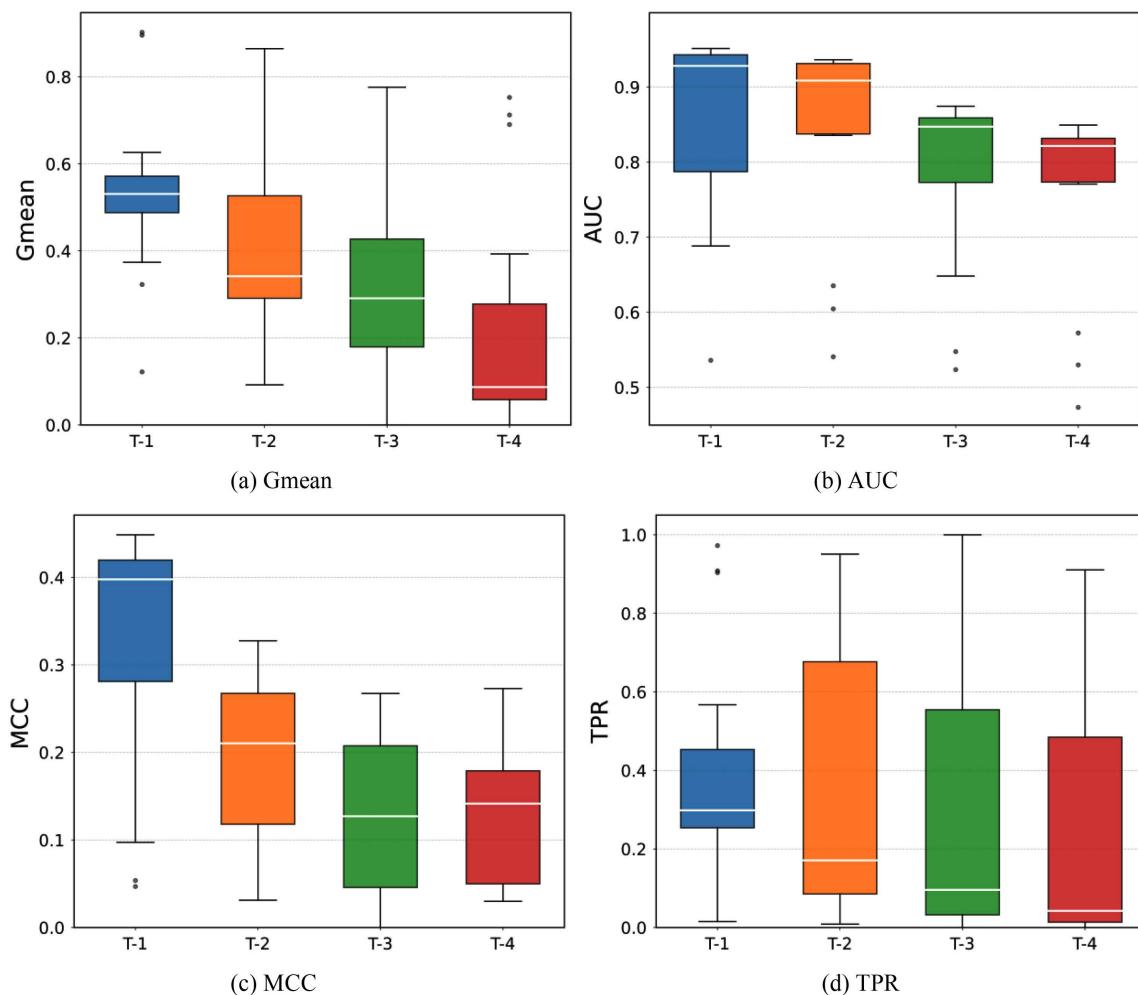


Figure 1. Comprehensive performance comparison of prediction models across T-1 to T-4

图 1. T-1 至 T-4 时间窗口下的预警模型综合性能对比

综合图1与表4至表7的结果,可以得出以下结论:所有模型的性能均呈现显著的时间衰减效应。T-1窗口的各项指标中位数最高,且分布最为集中,表明在财务困境发生前一年,模型判别能力最强且结果最稳定。以Gmean为例,集成欠采样方法在T-1窗口的中位数超过0.5,而在T-4窗口下降至约0.1,反映出财务指标预警能力的自然衰减。上述结果表明,财务困境预警模型的性能高度依赖于预测窗口的选择,T-1为最优预警期。而在中长期预测中,采用集成欠采样策略是维持模型稳健性的有效途径。

4.4. 预警时间窗口效应分析

为进一步比较不同模型的综合性能,本研究采用统计检验方法对基准模型和欠采样模型进行了系统排序。图2展示了基于多个评估指标平均排名的临界差异(Critical Difference)图,并标注了不同显著性水平下的差异区间。

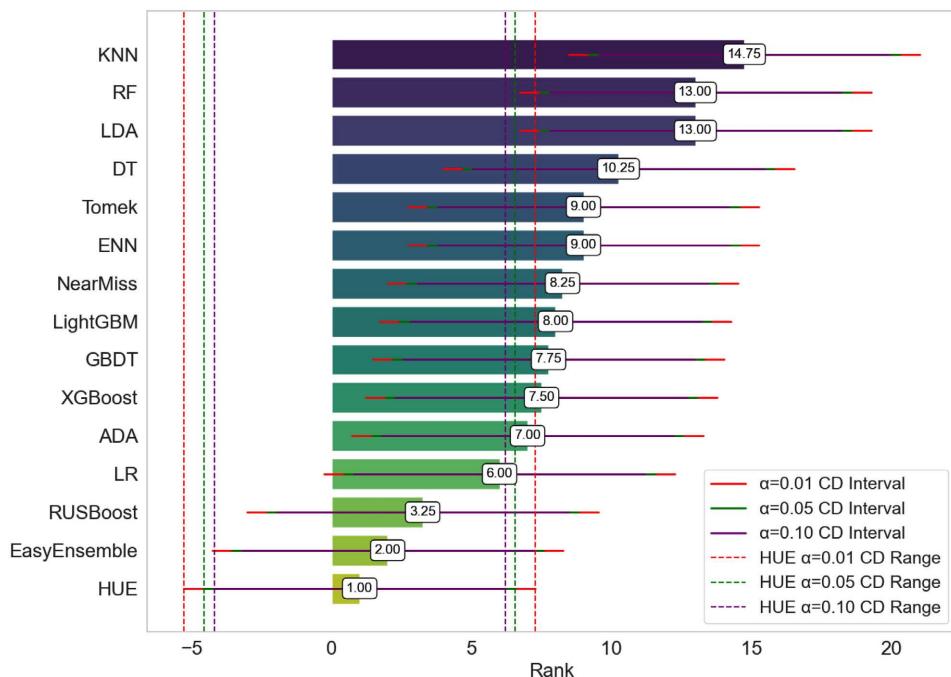


Figure 2. Average ranking of prediction models

图2. 预警模型平均排名

从排序结果来看,HUE、EasyEnsemble、RUSBoost三种集成欠采样方法显著优于其他模型,分别排前三名。传统机器学习方法中,KNN、随机森林与线性判别分析表现相对较好,但整体排名居于中后。在统计显著性方面,在 $\alpha=0.01$ 的显著性水平下,HUE与EasyEnsemble构成的第一梯队与其他模型之间存在明显差异,表明其集成欠采样性能优势具有统计上的可靠性。该检验结果从统计推断角度证实,针对企业财务困境这类高度不平衡的分类问题,专门设计的集成欠采样方法(尤其是HUE)在整体判别效能上显著优于传统分类器及其他欠采样策略,为模型选择提供了实证依据。

5. 研究结论

本研究围绕企业财务困境预警中的类别不平衡问题,系统探讨了欠采样策略的应用效果。通过构建多时间窗口的实验框架,对传统欠采样方法与集成欠采样方法进行了全面对比分析,得出以下主要结论:集成欠采样方法在财务困境预警中展现出显著优势。HUE与EasyEnsemble等集成方法在四个预测窗口

中都保持了最优异的综合性能，不仅显著提升了模型对少数类样本的识别能力，还保持了良好的类别平衡性。特别是在 T-1 时间窗口下，集成方法的性能表现最为突出，这为确定最佳预警时机提供了重要依据。研究还揭示了财务困境预警中明显的时间窗口效应。随着预测时间的延长，所有模型的性能都呈现出系统性下降趋势，但集成欠采样方法表现出了更强的稳健性。这一发现不仅证实了财务指标时效性的*重要性*，也为不同预警需求下的模型选择提供了参考。

基于上述发现，本研究可为实践提供以下具体建议：(1) 对于短期高风险企业筛查，监管机构或投资者在需要进行下一年度的财务风险排查时，可优先采用 HUE 或 EasyEnsemble 等集成欠采样模型，利用其在 T-1 窗口下的高精度进行重点监控。(2) 对于中长期风险趋势判断，若需进行更前瞻性的风险评估，应明确认识到模型性能的衰减，可将模型预测结果作为辅助参考指标，并结合行业宏观环境、公司治理结构等定性信息进行综合判断，而非单独依赖定量模型。(3) 在构建企业财务风险预警系统时，应考虑纳入集成欠采样模块，以应对普遍存在的类别不平衡问题。同时，系统应支持多时间窗口的预测，并明确标注不同窗口下预测结果的可靠性差异。

尽管本研究取得了一定成果，但仍存在一些局限性。未来研究可进一步探索深度学习与欠采样方法的结合，加强模型的可解释性研究，并验证方法在不同市场环境下的适用性。通过这些深入研究，有望推动财务困境预警研究向更精准、更实用的方向发展。

参考文献

- [1] 孙洁, 李辉, 韩建光. 基于滚动时间窗口支持向量机的财务困境预测动态建模[J]. 管理工程学报, 2010(4): 174-180.
- [2] 王宗胜, 尚姣姣. 我国制造业上市公司财务困境预警分析[J]. 统计与决策, 2015(3): 174-177.
- [3] Fitzpatrick, P.J. (1932) A Comparison of the Ratios of Successful Industrial Enterprises with Those of Failed Companies.
- [4] Ohlson, J.A. (1980) Financial Ratios and the Probabilistic Prediction of Bankruptcy. *Journal of Accounting Research*, **18**, 109-131. <https://doi.org/10.2307/2490395>
- [5] Ding, Y., Song, X. and Zen, Y. (2008) Forecasting Financial Condition of Chinese Listed Companies Based on Support Vector Machine. *Expert Systems with Applications*, **34**, 3081-3089. <https://doi.org/10.1016/j.eswa.2007.06.037>
- [6] García, V., Marqués, A.I. and Sánchez, J.S. (2019) Exploring the Synergetic Effects of Sample Types on the Performance of Ensembles for Credit Risk and Corporate Bankruptcy Prediction. *Information Fusion*, **47**, 88-101. <https://doi.org/10.1016/j.inffus.2018.07.004>
- [7] Gao, R., Cui, S., Wang, Y. and Xu, W. (2025) Predicting Financial Distress in High-Dimensional Imbalanced Datasets: A Multi-Heterogeneous Self-Paced Ensemble Learning Framework. *Financial Innovation*, **11**, Article No. 50. <https://doi.org/10.1186/s40854-024-00745-w>
- [8] He, H.B. and Garcia, E.A. (2009) Learning from Imbalanced Data. *IEEE Transactions on Knowledge and Data Engineering*, **21**, 1263-1284. <https://doi.org/10.1109/tkde.2008.239>
- [9] Zheng, W. and Zhao, H. (2020) Cost-Sensitive Hierarchical Classification for Imbalance Classes. *Applied Intelligence*, **50**, 2328-2338. <https://doi.org/10.1007/s10489-019-01624-z>
- [10] Tang, J., Li, J., Xu, W., Tian, Y., Ju, X. and Zhang, J. (2021) Robust Cost-Sensitive Kernel Method with Blinex Loss and Its Applications in Credit Risk Evaluation. *Neural Networks*, **143**, 327-344. <https://doi.org/10.1016/j.neunet.2021.06.016>
- [11] Bej, S., Davtyan, N., Wolfien, M., Nassar, M. and Wolkenhauer, O. (2020) Loras: An Oversampling Approach for Imbalanced Datasets. *Machine Learning*, **110**, 279-301. <https://doi.org/10.1007/s10994-020-05913-4>
- [12] Sun, Z., Ying, W., Zhang, W. and Gong, S. (2024) Undersampling Method Based on Minority Class Density for Imbalanced Data. *Expert Systems with Applications*, **249**, 123328. <https://doi.org/10.1016/j.eswa.2024.123328>
- [13] Liu, X.-Y., Wu, J.X. and Zhou, Z.-H. (2009) Exploratory Under-Sampling for Class-Imbalance Learning. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, **39**, 539-550. <https://doi.org/10.1109/tsmc.2008.2007853>
- [14] Ng, W.W.Y., Xu, S., Zhang, J., Tian, X., Rong, T. and Kwong, S. (2022) Hashing-Based Under-Sampling Ensemble for Imbalanced Pattern Classification Problems. *IEEE Transactions on Cybernetics*, **52**, 1269-1279.

<https://doi.org/10.1109/tcyb.2020.3000754>

- [15] Gong, Y., Lazebnik, S., Gordo, A. and Perronnin, F. (2013) Iterative Quantization: A Procrustean Approach to Learning Binary Codes for Large-Scale Image Retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **35**, 2916-2929. <https://doi.org/10.1109/tpami.2012.193>