E-Commerce Letters 电子商务评论, 2025, 14(12), 6542-6553
Published Online December 2025 in Hans. https://www.hanspub.org/journal/ecl
https://doi.org/10.12677/ecl.2025.14124644

Hans汉斯

# 电商企业财务风险的失衡数据预测研究

## ——基于CSA优化SMOTE与随机森林的融合模型

**郭宇新，刘媛华**

上海理工大学管理学院，上海

收稿日期：2025年12月5日；录用日期：2025年12月24日；发布日期：2025年12月31日

## 摘　要

**鉴于财务风险预测对电子商务行业的发展及电商企业的可持续发展具有极其重要的作用，而现有研究对数据不平衡状态下的预测效果仍存在不足，本文以2020~2024年的电商企业数据为样本开展研究。首先，对比不同采样方式对机器学习分类器效果的影响，确定以SMOTE-RF为基础模型，再引入改进的乌鸦搜索算法(CSA)优化，利用融合模型对数据集进行财务风险预测和分析。实证分析发现，在不平衡数据集上，经乌鸦搜索算法优化后的SMOTE-RF组合分类器整体表现尚可，提高了对少数类的识别效果；在引入改进的CSA后，ICSA-SMOTE-RF模型在保持较高特异度的同时，获得了较高的召回率，对财务风险预测效果有着较大幅度的提升。实证结果表明，本文提出的融合模型能够较好地反映出财务指标之间复杂的非线性关系，为电商企业的风险预测提供了可靠的理论研究方法。**

## 关键词

**电商企业财务风险，SMOTE，随机森林，不平衡数据，乌鸦搜索算法**

# Study on Predicting E-Commerce Financial Risk from Imbalanced Data

## —An Integrated Model Based on CSA-Optimized SMOTE and Random Forest

**Yuxin Guo, Yuanhua Liu**

Business School, University of Shanghai for Science and Technology, Shanghai

Received: December 5, 2025; accepted: December 24, 2025; published: December 31, 2025

## Abstract

**In the context of the rapid development of e-commerce industry, financial risk prediction is of great**

文章引用: 郭宇新, 刘媛华. 电商企业财务风险的失衡数据预测研究[J]. 电子商务评论, 2025, 14(12): 6542-6553.
DOI: 10.12677/ecl.2025.14124644

**significance to the sustainable development of e-commerce enterprises. However, the existing financial risk prediction research still demonstrates limitations in the prediction effect under the condition of imbalanced data. This study selects e-commerce enterprises from 2020 to 2024 as research subjects. By comparing the effects of different sampling strategies on the performance of machine learning classifiers, SMOTE combined with Random Forest (RF) was determined as the base prediction model. Furthermore, the improved Crow Search Algorithm (CSA) was introduced to optimize model performance, employing a fusion model to predict and analyze financial risks in the dataset. Empirical findings reveal that on imbalanced datasets, the overall performance of the SMOTE-RF combined classifier optimized by the Crow Search Algorithm is acceptable, improving the recognition effect for minority classes. The ICSA-SMOTE-RF model based on improved CSA effectively increases recall while maintaining high specificity, significantly improving financial risk prediction performance. The empirical findings suggest that the proposed hybrid model in this paper effectively captures the complex nonlinear relationships among financial indicators, providing a reliable theoretical and methodological approach for risk prediction in e-commerce enterprises.**

## Keywords

**E-Commerce Enterprise Financial Risk, SMOTE, Random Forest, Imbalanced Data, Crow Search Algorithm**



## 1. 引言

作为数字经济的重要组成部分，电商企业的运营模式高度依赖互联网技术的驱动与供应链协同的完美对接，这导致在经济环境剧变下，电商企业财务风险的动态演变更为复杂且预测困难。电商行业一般都是轻资产运营，企业的现金流周转快但稳定性较弱，平台佣金、物流成本、广告成本都是企业财政问题的压力来源[1]。例如电商平台产业聚集会造成电商企业之间竞争激烈；行业同质化竞争导致企业相互压价，毛利率降低；消费者需求的不断变化会加快企业存货周转周期，存货更容易发生折扣，从而增加企业的经营成本等。电商企业之间的竞争格局迫使众多中小型电商企业通过持续的市场推广投入和价格补贴策略来维持流量与用户粘性，直接抬高了运营成本并压缩了利润空间[2]。

另外，经济环境变化迅速、电商行业监管趋严，要求企业必须对内部流程改造、系统升级及人员培养等方面进行投资，这种大环境的不可控性会增强企业经营成本，对企业的资金筹划和长期财务弹性都产生一定的挑战[3]。企业若要保持竞争优势，就必须持续投入成本到大数据分析、人工智能应用、物流智能化以及支付安全技术等领域。这类投入往往投资金额大、效益周期长，对企业的资本结构和现金流产生严重影响。若技术更新速度超过企业消化能力或投入未能带来预期效益，极易造成投资损失，削弱企业偿债能力[4]。

这些因素都会影响电商企业的核心财务指标，导致传统财务分析方法难以有效捕捉早期风险信号。此外，电子商务的数据流量具有实时变动与高度异构的特点，这使得其财务风险传播速度远超传统行业。由于资金链的脆弱性可能迅速显现，因此，及时准确地对财务风险进行辨识变得尤为迫切。也可以说，财务风险预测在电商企业的运营管理中具有不可替代的战略价值。

财务风险预测研究的发展历程可大致分为传统统计模型(如逻辑回归[5]和多变量分析[6]等)、机器学

习算法(如神经网络[7]、支持向量机[8]、深度学习[9]等)和智能优化融合[10]三个阶段。

早期研究主要是利用财务比率分析和统计方法，通过筛选关键指标构建判别模型。这类方法简单易懂，但依赖标准正态分布假设，面对复杂的非线性关系或高维数据时，很容易出现维度灾难。随着机器学习技术的出现，研究方法也随之发生变化，如 SVM 利用核函数映射解决非线性问题；随机森林、XGBoost 等方法通过多模型集成协作提高泛化性能，预测结果更安全可靠等。随着研究的深入和深度学习的发展，越来越多的学者尝试让财务风险预测问题从风险企业和正常企业相同或者相差很小的理想数据转变为真实企业数据，研究的重点逐渐聚焦于解决实际应用中面临的财务数据失衡问题。

针对类别不平衡问题，现有研究主要从数据与算法两个层面提出解决路径：在数据层面，可采用合成少数类过采样技术(SMOTE)重构样本分布；在算法层面，则可通过改进损失函数或样本加权机制增强模型对少数类的识别能力。基于上述思路，本文选取多种常用不平衡数据处理方法，以缓解电商企业财务数据中风险样本严重不足的问题，并从中优选表现最佳的 SMOTE 方法，与随机森林模型相结合，并引入具有参数简洁、收敛速度快、全局搜索能力强等特点的乌鸦搜索算法(CSA)，优化模型参数配置，构建了一种融合 SMOTE、CSA 与随机森林的财务风险预测协同框架，以期为电商企业高维、非平衡财务数据的风险预警提供方法参考。

## 2. 基本原理介绍

### 2.1. 不平衡数据处理方法

在机器学习任务中常见的数据通常会出现类别不平衡问题，即一类数据的样本数量远多于其他类别。例如，在欺诈检测中，正常交易远多于欺诈交易；在疾病诊断中，健康人数量要比病人多得多。同样地，财务风险预测研究也面临着数据不平衡的挑战，且不平衡比例通常较高，导致传统模型易倾向于多数类，难以识别风险样本。

从数据处理的角度看，重采样方法是解决类不平衡问题常用的有效技术手段。在 SMOTE 方法[11]提出之前，处理不平衡问题的主要方法为随机欠采样或过采样。欠采样方法是随机地删除一些多数类样本，使数据达到平衡，但这也意味着会丢失多数类中的重要信息；而过采样方法是随机复制一些少数类样本，但因为只是简单复制了已有的样本，没有引入新的信息，很容易导致模型出现过拟合问题。

SMOTE (Synthetic Minority Over-sampling Technique)方法，即合成少数类过采样技术，它的核心思想不是机械地模仿少数类样本，而是制造新的少数类样本来实现过采样。其实现步骤如下：

1) 在原始的少数类样本集合中，对于每一个样本 $x_i$，找到其中 $k$ 个最近的少数类邻居(通常使用欧氏距离)，这个近邻集合称为 $N_k(x_i)$。

2) 从近邻集合 $N_k(x_i)$中，随机选择一个样本 $x_{zi}$，应用下面的公式合成新的少数类样本：

$$x_{new} = x_i + \delta \times \left( x_{zi} - x_i \right) \tag{1}$$

其中，$\delta$ 表示一个在[0, 1]区间内均匀分布的随机数。

上面的公式可以理解为在连接 $x_i$ 和 $x_{zi}$ 的线段上，随机创造出新的数据点，这些新点既不是 $x_i$ 也不是 $x_{zi}$，但具有与 $x_i$ 和 $x_{zi}$ 相似的特征模式，通过在特征空间内进行插值的方法，SMOTE 方法生成了全新的、非重复性的样本，有效地扩充了少数类的决策区域，从而帮助模型学习到更具有鲁棒性的决策边界，避免了简单的过采样带来的过拟合风险。

SMOTE 方法有效缓解了过拟合，能生成更具代表性的决策区域，但在实际应用中，当少数类分布不连续时，可能生成噪声样本；或者在边界区域合成样本时，无视多数类分布而造成类别重叠等问题。也因此后续学者提出了很多改进算法，旨在解决其局限性，比如 Borderline-SMOTE 方法[12]，聚焦于边界

样本的过采样，通过增强决策边界的清晰度提升模型对风险样本的辨识能力；或 ADASYN (Adaptive Synthetic Sampling)方法[13]，通过自适应计算样本合成密度，重点增强分类难度较大的少数类区域数据。本文在后续对模型的性能进行分析时也参考了相关的方法进行对比，并在文中选取的数据方法中选择了表现较佳的 SMOTE 方法进行进一步分析。

## 2.2. 随机森林原理

随机森林(Random Forest)作为一种高效的集成学习框架[14]，该算法通过构建并聚合大量决策树的预测结果，显著优化了模型的综合性能与泛化效果。其设计理念巧妙结合了自助采样思想(Bootstrap Aggregating, Bagging)及随机子空间技术(Random Subspace Method)，引入双重随机机制从而增强模型多样性，并有效缓解过拟合问题。在模型训练环节，采用自助采样独立生成多个训练样本集，再从全部特征中随机选择一部分子集用于每棵决策树的构建，最终通过集成所有树的预测输出，形成具有高鲁棒性的预测模型。随机森林的训练步骤如下：

1) 通过自助采样法从原始训练数据集中有放回地随机抽取多个子样本集(称为 Bootstrap 样本)，生成 T 个大小与原始训练集相同的子训练集。每个子训练集用于训练一棵决策树。

2) 对于每一棵决策树的生长过程，在每次节点分裂时，不从所有 M 个特征中挑选最优特征，而是随机选择一个特征子集(通常大小为 m，且 $m << M$)，从 m 个特征中挑选最佳分裂特征和分裂点。这棵决策树会完全生长，通常不进行剪枝，即直到节点中的样本都属于同一类别或样本数少于预设阈值为止。

3) 对于分类问题(如本文的财务风险预测)，最终的预测结果采用多数投票原则：将待预测样本输入森林中的每一棵树，获得其预测类别标签，最终森林输出的类别是获得票数最多的那个类别。相应的数学公式如下：

对于一个样本 $x_i$，若随机森林的预测类别为 $\hat{y}$，则：

$$\hat{y} = \text{mode}\left\{h_1(x_i), h_2(x_i), \ldots, h_T(x_i)\right\} \tag{2}$$

其中，$h_t(x_i)$表示第 $t$ 棵决策树对样本 $x_i$ 的预测类别，mode 表示返回出现次数最多的类别。

## 2.3. 乌鸦搜索算法原理

乌鸦搜索算法(Crow Search Algorithm, CSA)是 2016 年提出的一种元启发式优化算法[15]，该算法巧妙地模拟了乌鸦的行为模式，重点借鉴了其高效率的藏食策略与精准的偷食技巧。乌鸦作为一种高智商的鸟类，具备极强的位置记忆能力，能够长期记住大量藏食点的坐标。同时，乌鸦还能敏锐地观察同伴行为，通过分析其他乌鸦的活动轨迹来预测和获取食物。这些复杂的认知特性为 CSA 算法的设计提供了核心依据，使其在优化过程中能有效利用这些生物智能特性。

在算法中，采用 $d$ 维搜索空间对优化问题的解空间进行精确描述，其中每只乌鸦的位置对应一个潜在解。假设乌鸦的种群规模为 $N$，最大迭代次数为 $MIT$。第 $i$ 只乌鸦在 $d$ 维空间的位置可以表示为一个向量，代表当前的一个解：

$$x_i^{gen} = \left[x_{i,1}^{gen}, x_{i,2}^{gen}, \ldots, x_{i,d}^{gen}\right], \quad i = 1,2,\ldots,N; gen = 1,\ldots,MIT \tag{3}$$

在迭代过程中，每只乌鸦会记忆其已知的最佳藏食位置，该位置代表了当前个体所发现的最优解。为寻求更优解，乌鸦在搜索空间中通过跟踪同伴的方式探索潜在的更好的位置。也就是算法在每一轮迭代中，将执行以下位置更新机制：

1) 跟随阶段：在该阶段，乌鸦 $i$ 将随机选择种群中另一只乌鸦 $j$ 进行跟踪，试图找到乌鸦 $j$ 的食物藏匿点。位置更新公式如下：

$$x_i^{gen+1} = \begin{cases} x_i^{gen} + r_i \times FL_i^{gen} \times \left(m_j^{gen} - x_i^{gen}\right), & if\ r_j \geq AP_j^{gen} \\ a\ \ random\ \ position, & otherwise \end{cases} \tag{4}$$

其中，$x_i^{gen}$ 是乌鸦 $i$ 在第 $gen$ 次迭代时的位置；$x_i^{gen+1}$ 是乌鸦 $i$ 在第 $gen+1$ 次迭代时的新位置；$m_j^{gen}$ 是乌鸦 $j$ 在第 $gen$ 次迭代时已知的最佳位置(即其藏匿点)；$r_i$ 和 $r_j$ 是范围在[0, 1]内的均匀分布的随机数；$FL_i^{gen}$ 是乌鸦 $i$ 在第 $gen$ 次迭代时的飞行长度，这个参数控制着搜索的步长；$AP_j^{gen}$ 是乌鸦 $j$ 在第 $gen$ 次迭代时的感知概率，它决定了被跟踪的乌鸦 $j$ 是否能够意识到自己被跟踪，从而采取反制措施(即让跟踪者飞到一个随机位置)。

2) 记忆更新阶段：在乌鸦移动到新位置后，需要更新其记忆(即已知的最佳藏匿点)。记忆更新公式如下：

$$m_i^{gen+1} = \begin{cases} x_i^{gen+1}, & if\ fit\left(x_i^{gen+1}\right) is\ better\ than\ fit\left(m_i^{gen}\right) \\ m_i^{gen}, & otherwise \end{cases} \tag{5}$$

其中，$m_i^{gen}$ 是乌鸦 $i$ 在第 $gen$ 次迭代时的记忆(历史最佳位置)；$fit(\cdot)$是适应度函数。对于最小化问题，如果新位置的适应度值更小(更好)，则更新记忆。

## 3. 数据筛选与不平衡方法分析

### 3.1. 数据选择与预处理

#### 3.1.1. 样本选取与数据来源

随着互联网的不断普及和发展，很多企业都会涉及电子商务或线上业务，但企业本质上可能还是属于制造业等行业，为了更精确地对电商企业进行分析，本文只选取了主营电子商务的企业或营业收入大部分来自电商销售的企业，参考巨潮资讯网中的“互联网商务”行业分类以及 CSMAR 数据库上市公司所属行业名称和其经营范围，主要包括互联网和相关服务、零售、批发等行业和部分制造业企业，若其营业收入的 80%以上来源为电子商务活动，则认为该企业为电商企业。经筛选，本文在 CSAMR 数据库中选取 2020 到 2024 年电子商务上市企业共计 121 家。

由于电商活动的开展高度同步于市场动态，本研究在采集财务数据的同时，整合了宏观层面的市场数据(数据主要来源于 CSMAR 数据库和国家统计局公开数据平台)。选择 2020~2024 年作为研究周期，既包含电子商务行业高速发展的成熟期，也覆盖后疫情时代行业动态调整阶段，有助于捕捉不同经济环境下的风险特征演变规律。

样本筛选以证监会发布的《上市公司风险警示规则》为依据，将 ST 及*ST 标记企业归入财务风险组，健康组为同期未触发风险警示的企业。鉴于当期的风险情况是证监会通过分析企业上一期的财务数据所确定的，也就是说用 T − 1 期的数据预测第 T 期的风险情况不符合财务风险预测的适宜和及时性，因此本文采用 T − 2 期的截面数据，即企业被实施 ST 前的两个完整会计年度数据进行后续风险预测分析，以规避风险事件发生前后的数据干扰。

综上，本文共筛选出财务风险组数据 99 个，财务健康组为 2406，不平衡比例约为 1:24。

#### 3.1.2. 指标筛选与数据预处理

财务指标的选取遵循全面性、代表性和可获取性原则，参考国内外相关财务预警研究，从偿债能力、盈利能力、营运能力、发展能力和现金流量五个维度构建指标体系，筛选出一些常用的财务指标。此外，对于宏观指标，本文认为经济环境稳定性和行业的发展情况对电商企业的影响较大，因此选取了国内生产总值－第三产业(批发和零售业)增长率、行业景气指数和通货膨胀率(CPI)进行分析。所有指标如表 1 所示。

**Table 1.** Financial risk prediction indicator system
**表 1.** 财务风险预测指标体系

| 类型 | 指标 |
|---|---|
| 偿债能力 | 速动比率、现金比率、营运资金与借款比、利息保障倍数、资产负债率 |
| 盈利能力 | 净资产收益率(ROE)、营业毛利率、营业净利率 |
| 营运能力 | 应收账款周转率、存货周转率、总资产周转率 |
| 发展能力 | 总资产增长率、营业收入增长率 |
| 现金流量 | 经营活动产生的现金流量净额、投资活动产生的现金流量净额、筹资活动产生的现金流量净额、经营活动产生的现金流量净额/流动负债 |
| 宏观指标 | 国内生产总值－第三产业增长率、行业景气指数、通货膨胀率(CPI) |

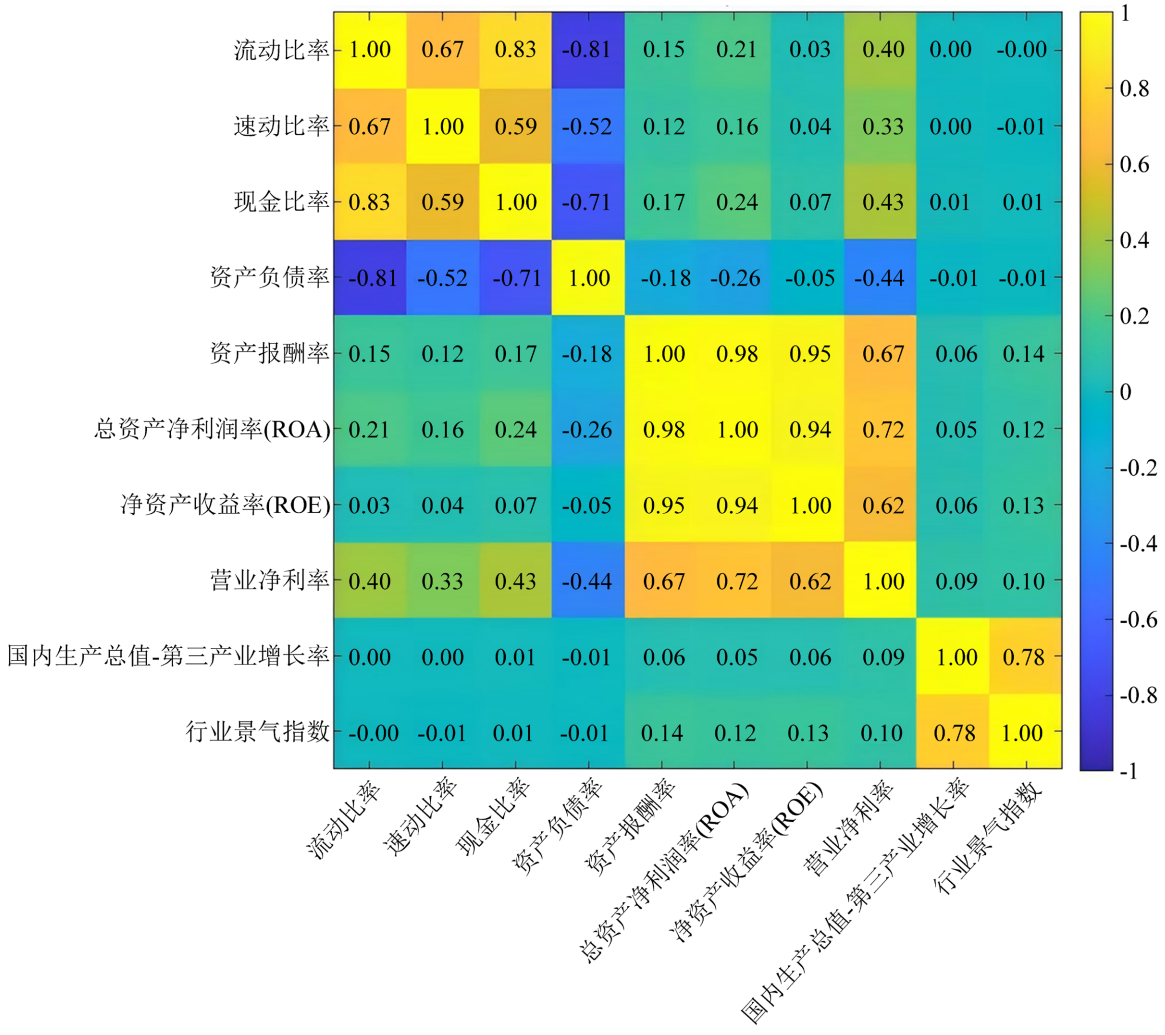


**Figure 1.** Indicator correlation analysis chart (top 10 pairs)
**图 1.** 指标相关性分析图(前 10 对)

表 1 中指标能够从不同角度反映企业的财务状况，为风险预测提供多维度的信息支持。但其中部分指标可能存在相关性，因此首先需要对数据指标进行相关性分析以减少多重共线的可能性。本文运用

MATLAB 对指标进行相关性分析，剔除高相关度的指标(相关性大于 0.8)，结果如图 1 所示。由于指标较多，图 1 中只保留了相关性较高的前 10 个指标。根据相关性分析结果，本文剔除了流动比率、资产报酬率和总资产净利润率(ROA)，最终保留了 20 个分析指标。

由于不同指标的量纲和量级差异较大，在进行实验分析之前，先对原始数据集进行标准化处理能削弱各个指标的值之间存在的大小差异，并能减小计算过程的复杂程度。Z-Score 标准化公式如下：

$$x' = \frac{x - \mu}{\sigma} \tag{6}$$

其中，$\mu$ 和 $\sigma$ 分别是指标的均值和标准差。

### 3.2. 不平衡数据处理

本节选取了常用的几种不平衡数据处理方法，包括过采样、欠采样、SMOTE、Borderline-SMOTE 和 ADASYN 等方法，将处理后的平衡数据集利用机器学习进行分类，以确定最优的数据处理方案。在评估不平衡数据处理效果时，选择合适的评价指标是模型性能分析的关键。由于电子商务企业财务风险预测中的类别不平衡问题，准确率(Accuracy)易受多数类样本干扰，因此需要采用能够综合反映少数类识别能力的指标，如召回率(Recall)、精确率(Precision)、F1 值(F1-Score)等。

Borderline-SMOTE 与 ADASYN 作为经典的过采样方法，数据处理机制存在差异。Borderline-SMOTE 专注于在少数类样本的边界区域生成新样本，通过强化分类边界的模糊区域提升模型对高危样本的捕捉能力。而 ADASYN 通过自适应计算每个少数类样本的合成权重，对难以学习的样本赋予更高采样密度，从而降低分类器对复杂样本的学习难度。两种方法的核心区别在于样本生成策略的侧重点，前者聚焦于优化决策边界，后者强调对困难样本的针对性增强。

本文采用分层抽样方法将数据集按 7:3 的比例划分为训练集与测试集，将总样本的 70%作为训练集，用于模型构建和参数优化；30%作为测试集，用于验证模型最终的泛化能力。实验结果如表 2 所示。

实验结果表明，对原始数据直接进行分类，虽然准确率很高，但召回率和 F1 分数的结果很低，甚至在 XGBoost 上的结果几乎接近于 0，也就是说，分类的高准确性是严重忽略了少数类样本，其他几种机器学习分类最高也就能达到 50%左右，这对判断企业财务风险情况会产生严重的影响，导致评价错误。同样地，过采样方法的整体准确性要略逊色于几种过采样方法，虽然在神经网络上取得了最高的 F1 值(0.6296)，但在 XGBoost 上导致准确率显著下降至 0.8362。使用欠采样方法在几种机器学习方法上都能够获得极高的召回率，但明显是该方法删除了过多的样本数据，以严重牺牲精确度和准确率为代价得到的，可以说其整体性能较差。

综合各项评价指标来看，几种过采样方法都对少数类的识别产生了积极的影响。其中，标准 SMOTE 使随机森林取得了所有实验中的最高准确率(0.9814)和最高 F1 分数(0.6957)，展现出其在特定模型上的强大优势；Borderline-SMOTE 在提升模型整体判别能力上表现优异，它使得神经网络、XGBoost 和随机森林的准确率均达到了 0.9694 以上，同时其 F1 分数也分别达到了 0.6230、0.4324 和 0.6667，说明其在保持较高准确率的同时，对少数类的识别能力也有显著改善；ADASYN 和传统过采样方法相比上面的两种方法整体性能表现一般。

从模型整体的判别能力来看，SMOTE 与 Borderline-SMOTE 均能稳定提升或维持较高的分类准确率。其中，SMOTE 方法使随机森林模型取得了所有实验配置中的最高准确率，召回率、F1 值虽然在少数类略低于 Borderline-SMOTE，但也能达到 60%以上。本文希望在准确识别风险企业的同时，也能在整体上判断财务健康企业的情况，因此本文选择了整体识别准确率较高的 SMOTE 结合随机森林的风险预测组合进行后续模型构建。

**Table 2.** Performance comparison of imbalanced data handling methods
**表 2.** 不平衡处理方法性能对比

| Accuracy 比较 | | | | |
|---|---|---|---|---|
| Method | 神经网络(NN) | XGBoost | 随机森林(RF) | 支持向量机 SVM |
| Original | 0.9627 | 0.9627 | 0.9774 | 0.9694 |
| SMOTE | 0.9627 | 0.9680 | **0.9814** | **0.9720** |
| Borderline-SMOTE | 0.9694 | **0.9720** | 0.9800 | **0.9720** |
| ADASYN | 0.9720 | 0.9654 | 0.9787 | 0.9694 |
| 过采样 | **0.9734** | 0.8362 | 0.9747 | 0.9614 |
| 欠采样 | 0.7750 | 0.7577 | 0.8029 | 0.9587 |
| Precision 比较 | | | | |
| Method | NN | XGBoost | RF | SVM |
| Original | 0.5152 | **1.0000** | 0.9286 | **0.7500** |
| SMOTE | 0.5135 | **1.0000** | **0.9412** | **0.7500** |
| Borderline-SMOTE | 0.5938 | **1.0000** | 0.9375 | 0.7000 |
| ADASYN | 0.6667 | **1.0000** | 0.8824 | 0.6875 |
| 过采样 | **0.6800** | 0.1781 | 0.8125 | 0.0000 |
| 欠采样 | 0.1392 | 0.1340 | 0.1638 | 0.4706 |
| Recall 比较 | | | | |
| Method | NN | XGBoost | RF | SVM |
| Original | 0.5862 | 0.0345 | 0.4483 | 0.3103 |
| SMOTE | 0.6552 | 0.1724 | 0.4828 | 0.4138 |
| Borderline-SMOTE | 0.6552 | 0.2759 | 0.5172 | 0.4828 |
| ADASYN | 0.5517 | 0.1034 | 0.5172 | 0.3793 |
| 过采样 | 0.5862 | 0.8966 | 0.4483 | 0.0000 |
| 欠采样 | **0.9310** | **0.9655** | **1.0000** | **0.5517** |
| F1 比较 | | | | |
| Method | NN | XGBoost | RF | SVM |
| Original | 0.5484 | 0.0667 | 0.6047 | 0.4390 |
| SMOTE | 0.5758 | 0.2941 | **0.6957** | 0.5333 |
| Borderline-SMOTE | 0.6230 | **0.4324** | 0.6667 | **0.5714** |
| ADASYN | 0.6038 | 0.1875 | 0.6522 | 0.4889 |
| 过采样 | **0.6296** | 0.2971 | 0.5778 | 0.0000 |
| 欠采样 | 0.2422 | 0.2353 | 0.2816 | 0.5079 |

## 4. 财务风险预测融合模型的构建

### 4.1. 乌鸦搜索算法的改进

乌鸦搜索算法虽然具有结构简单、参数较少、全局搜索能力强等优点，但该算法也存在一定的局限性，例如在处理复杂高维问题时，后期收敛速度慢，容易陷入局部最优；固定的飞行长度和感知概率难以平衡算法的探索与开发能力，影响性能；迭代后期，种群趋同，多样性下降，影响收敛速度和最终解的质量。针对上面的问题，本文在传统 CSA 的基础上进行了一些细微的改进，让该算法能更好地适应财务风险预测问题。改进的方法主要包括：

1) 使用混沌映射初始化种群，替代随机初始化，生成分布更均匀、遍历性更好的初始种群，为全局

搜索奠定良好基础。Logistic 映射的公式如下：

$$z_{k+1} = \mu \cdot z_k \cdot (1 - z_k) \tag{7}$$

其中，$z_k \in (0, 1)$是第 $k$ 次迭代的混沌值，$\mu$ 是控制参数。

对于一个 $d$ 维的优化问题，利用 Logistic 映射迭代生成 N 个混沌向量 $Z_1, Z_2, \ldots, Z_N$，将这些混沌向量映射到解空间：

$$x_{i,j} = LB_j + Z_{i,j} \cdot (UB_j - LB_j) \tag{8}$$

其中，$x_{i,j}$ 是第 $i$ 只乌鸦在第 $j$ 维的位置，$LB_j$ 和 $UB_j$ 是该维度的下界和上界。

2) 动态调整感知概率 $AP$ 和飞行长度 $FL$，通过将固定的 $AP$ 和 $FL$ 改进为随迭代次数和种群多样性动态调整，使算法在早期注重探索，后期注重开发，并适配高维特征下的复杂搜索。动态调整公式如下：

$$AP(gen) = AP_{\min} + (AP_{\max} - AP_{\min}) \cdot \frac{gen}{MIT} \tag{9}$$

其中，$gen$ 是当前迭代次数，$MIT$ 是最大迭代次数，$AP_{\min}$ 和 $AP_{\max}$ 是 $AP$ 的最小和最大值。

$$FL(gen) = FL_{\max} \cdot \exp\left(c \cdot \frac{gen}{MIT}\right) \tag{10}$$

其中，$c = \ln(FL_{\min} / FL_{\max})$。

3) 在迭代过程中加入高斯扰动，以一定概率 pmut 对个体(特别是当前最优个体)施加微小扰动，帮助算法跳出局部最优陷阱。公式如下：

$$x_i^{new} = x_i^{old} + \eta \cdot N(0, \sigma^2) \tag{11}$$

其中，$N(0, \sigma^2)$ 是均值为 0、标准差为 $\sigma$ 的高斯随机数，$\eta$ 是一个随着迭代次数增加而减小的权重。

### 4.2. ICSA-SMOTE-RF 财务风险预测融合模型

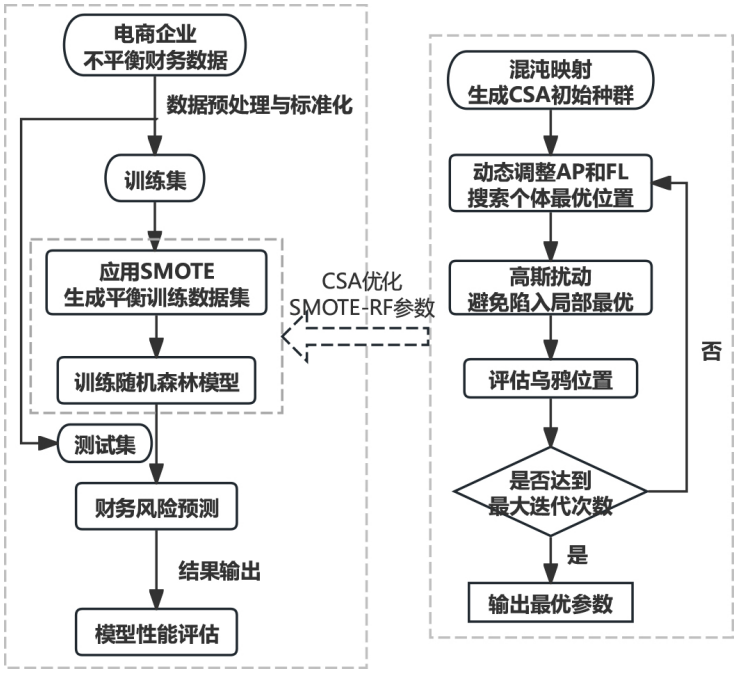


**Figure 2.** Flowchart for solving the e-commerce enterprise financial risk prediction model based on ICSA-SMOTE-RF
**图 2.** 基于 ICSA-SMOTE-RF 的电商企业财务风险预测模型求解流程图

融合模型针对电商企业财务风险预测的求解流程如图 2 所示。

## 5. 实证分析

### 5.1. 性能分析指标

本文第三部分对不平衡的二分类问题(风险企业占少数)，选取了准确率、召回率、精确率、F1 分数等指标进行综合性能分析，为了进一步评价本文提出的财务风险融合模型的性能，本节增加了特异度(Specificity)、受试者工作特征曲线下面积(Area Under the ROC Curve, AUC)以及 G-mean 指标，尝试构成一个更全面、更可靠的评估框架。

特异度衡量的是模型正确识别“低风险”企业(多数类)的能力。在财务风险评估中，高特异度意味着模型能够最大限度地减少误报，即避免将健康的公司误判为高风险。AUC 值提供了一个与阈值无关的、对模型整体分类效力的单一标量评估。它描绘了模型在所有可能的分类阈值下，区分“高风险”与“低风险”电商企业的能力。而 G-mean 可以用于评估模型在不平衡数据集上的综合性能。G-mean 的取值越大，代表算法的分类性能越好，识别水平越高，表达式如下：

$$G-mean = Sensitivity \times Specificity \tag{12}$$

### 5.2. 实验结果与分析

为了更好地对比融合模型的性能，本文除了基础的 SMOTE-RF 模型，还系统性地比较了多种优化策略，包括基于遗传算法 GA 和粒子群算法 PSO 优化 SMOTE-RF 的模型进行性能对比。实验结果如表 3 所示。

**Table 3.** Experimental results of financial risk prediction for e-commerce enterprises
**表 3.** 电商企业财务风险预测实验结果

| 模型 | Accuracy | Recall | Specificity | F1_Score | AUC | G-mean |
|---|---|---|---|---|---|---|
| SMOTE-RF | **0.9814** | 0.5517 | **0.9986** | 0.6957 | 0.9806 | 0.7422 |
| GA-SMOTE-RF | 0.9707 | 0.5172 | 0.9890 | 0.5769 | 0.9766 | 0.7152 |
| PSO-SMOTE-RF | 0.9734 | 0.4828 | 0.9931 | 0.5833 | 0.9790 | 0.6924 |
| CSA-SMOTE-RF | 0.9800 | 0.6552 | 0.9931 | **0.7170** | 0.9789 | 0.8066 |
| ICSA-SMOTE-RF | 0.9720 | **0.7586** | 0.9806 | 0.6769 | **0.9836** | **0.8625** |

综合分析表中数据，可以分析各类别不平衡学习模型的性能存在显著差异，其中 ICSA-SMOTE-RF 模型在关键指标上表现最为出色。该模型在保持较高准确性和特异性的同时，实现了所有模型中最高的召回率，达到了 0.7586。高特异性和召回率的组合优势使其 G-mean 值达到 0.8625，显著优于其他对比模型。这表明 ICSA 优化算法有效提升了 WRF 模型对少数类样本的识别能力，而 AUC 值最高，为 0.9836，进一步证实了模型整体分类性能的优越性。

相比之下，传统 SMOTE-RF 模型虽然准确率最高，但其召回率相对较低，反映出模型对少数类的识别能力不足。比较意外的是，在加入遗传算法和粒子群算法后模型效果没有显著提升，甚至下降，可能是因为优化算法的表现严重依赖于其搜索空间的定义，且少数类样本的不足，导致了优化效果被限制在一个局部最优的区域，对最终性能影响甚微甚至产生有害的参数组合。而 CSA-SMOTE-RF 在召回率和 F1 分数方面表现均衡，可作为实际应用中的可靠备选方案。

综上，乌鸦搜索算法与 SMOTE 技术的结合显著提升了不平衡数据分类效果，其中 ICSA-SMOTE-RF

通过更好的参数优化，在召回率与特异性之间找到了最佳平衡点，为处理类别不平衡问题提供了最有效的解决方案，对实际应用中需要同时兼顾多数类和少数类识别准确性的场景具有重要的指导意义。

## 6. 结论与展望

本文通过有目的地搭建乌鸦搜索算法改进的 SMOTE 和随机森林的融合模型，增强了电子商务公司财务风险预测的精度和可靠性。由于数据本身的偏斜性质，进一步比较了 SMOTE、Borderline-SMOTE 和 ADASYN 的改进效果，SMOTE、Borderline-SMOTE 对消除偏斜都发挥了重要作用，且 SMOTE 融合随机森林的模型效果较好，为训练模型提供了均衡的数据集。将乌鸦搜索算法与 SMOTE-RF 相结合，运用 CSA 和乌鸦搜索算法的局部搜索能力改进 SMOTE-RF 的参数，增强对于少数类的识别。结果表明，模型 ICSA-SMOTE-RF 在 Recall、G-mean、AUC 等性能方面优于基准模型。特别需要指出的是，模型在具有高特异性(健康企业识别能力)的同时，提高了高风险企业的召回率，即模型更善于识别潜在的财务风险讯号。高风险企业的高召回率和高特异性是高风险企业作为风险预警模型的优势之一，为电商企业提供更安全的风险识别。此外，CSA-SMOTE-RF 的实证结果也是均衡的，可以作为备选简化方案。

当然，本文所做研究存在局限性，在后续需要继续扩大。样本从 CSMAR 数据库和国家统计局获取，在样本范围和样本跨度上有待提高，使得模型能够对更长的经济周期、更多数量企业规模的模型进行拟合，使得乌鸦搜索算法在收敛速度以及收敛能力方面对超大规模实时性要求更高的应用场景存在进一步提升的空间。初始设置以及搜索空间对模型的参数寻优效果存在一定影响，存在一定的主观性。在总体上存在进一步优化的空间，如自适应参数调整后的乌鸦搜索算法改进，加入到其他元启发式的算法中进一步改进提升模型的收敛速度和模型鲁棒性。另外，怎样使模型能够同时兼顾到收敛性以及预测精度，寻找到关键财务风险的驱动因子和作用机理，提升模型的解释度，应用到更加细化的风险分级预测或者别的非财务数据融合预测，如：舆情数据、供应链数据等，构建更为动态全面的风险评价体系等将是后续的研究课题。

## 参考文献

[1] 宋佳. 跨国并购的风险控制分析——以汤臣倍健跨国并购 LSG 为例[J]. 经营与管理, 2021(1): 18-24.

[2] 胡霞. 数字化转型对电商企业存货管理效率的影响研究[J]. 财会通讯, 2025(12): 44-48.

[3] 刘尚希, 程瑜, 李成威, 等. 以风险视角透视新发展阶段的企业成本特征及我们的建议——2021 年企业成本调研总报告[J]. 财政研究, 2022(4): 8-28.

[4] 苏美文, 杨文爽, 李博文, 等. 推动人工智能与实体经济深度融合加快发展新质生产力[J]. 工业技术经济, 2025, 44(4): 32-59.

[5] Li, Q., Cai, D. and Wang, H. (2012) Study on Network Finance Risk on the Basis of Logit Model. In: Tan, H., Ed., *Advances in Intelligent Systems and Computing*, Springer, 213-220. https://doi.org/10.1007/978-3-642-27711-5_29

[6] Canbas, S., Cabuk, A. and Kilic, S.B. (2005) Prediction of Commercial Bank Failure via Multivariate Statistical Analysis of Financial Structures: The Turkish Case. *European Journal of Operational Research*, **166**, 528-546. https://doi.org/10.1016/j.ejor.2004.03.023

[7] 严莉红. 基于 LSTM 神经网络的饲料企业财务风险预警模型构建[J]. 饲料研究, 2023, 46(3): 130-134.

[8] 李祥飞, 张再生, 刘珊珊. 改进布谷鸟搜索 SVM 在财务风险评估中的应用[J]. 计算机工程与应用, 2015, 51(23): 218-225.

[9] 向有涛, 王明, 曹琳. 基于多目标深度学习模型的财务风险预测方法[J]. 统计与决策, 2022, 38(10): 184-188.

[10] 王文雯, 姚欣. 注意力机制结合卷积神经网络的医院财务风险预警方法研究[J]. 现代科学仪器, 2023, 40(2): 166-173.

[11] Chawla, N.V., Bowyer, K.W., Hall, L.O. and Kegelmeyer, W.P. (2002) SMOTE: Synthetic Minority Over-Sampling Technique. *Journal of Artificial Intelligence Research*, **16**, 321-357. https://doi.org/10.1613/jair.953

[12] 马贺, 宋媚, 祝义. 改进边界分类的 Borderline-SMOTE 过采样方法[J]. 南京大学学报(自然科学), 2023, 59(6): 1003-1012.

[13] 李志强, 余炫朴. 基于 ADASYN 的跨境电商小微企业信用风险模型优化研究[J]. 江西师范大学学报(哲学社会科学版), 2023, 56(2): 118-127.

[14] Cutler, A., Cutler, D.R. and Stevens, J.R. (2012) Random Forests. In: Zhang, C. and Ma, Y. Eds., *Ensemble Machine Learning*, Springer, 157-175. https://doi.org/10.1007/978-1-4419-9326-7_5

[15] Askarzadeh, A. (2016) A Novel Metaheuristic Method for Solving Constrained Engineering Optimization Problems: Crow Search Algorithm. *Computers & Structures*, **169**, 1-12. https://doi.org/10.1016/j.compstruc.2016.03.001