Operations Research and Fuzziology 运筹与模糊学, 2025, 15(2), 185-197
Published Online April 2025 in Hans. https://www.hanspub.org/journal/orf
https://doi.org/10.12677/orf.2025.152075

Hans汉斯

基于非平衡数据的机器学习企业财务风险预警模型研究

邢林飞

上海理工大学管理学院，上海

收稿日期：2025年2月20日；录用日期：2025年3月26日；发布日期：2025年4月3日

摘　要

制造业企业一般需要较高的研发投入与供应链协同要求，但由于当前国内整体行业的外部环境愈趋复杂、市场竞争激烈，其所面临的财务风险问题亦愈发严峻，因此构建适用于我国制造业企业的财务风险预警模型十分有必要。当前国内制造业的ST与非ST企业数量占比存在极大差距，所以多数研究运用的是非平衡数据。对此本文除了对选取的ST企业进行1:3匹配同等资产规模的非ST企业外，还利用了SMOTE过采样方法对非平衡的样本数据集进行处理。文中共采用了2020年到2023年的168家中国制造业上市企业数据，并综合12种机器学习模型进行了预警效果综合对比。结果表明Extra Trees模型的效果最优，且数据集在经过平衡处理后该模型的预测准确率得到了18%的提升。本研究期望所做的模型研究对于国内制造业企业在财务风险预警防控上具有实际的参考与应用价值，能够助力维护行业与经济的稳健发展。

关键词

非平衡数据，机器学习，财务风险预警，制造业

Research on Machine Learning Models for Enterprise Financial Risk Warning Based on Imbalanced Data

Linfei Xing

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

Received: Feb. 20th, 2025; accepted: Mar. 26th, 2025; published: Apr. 3rd, 2025

Abstract

Manufacturing enterprises typically require high R&D investment and supply chain coordination.

文章引用: 邢林飞. 基于非平衡数据的机器学习企业财务风险预警模型研究[J]. 运筹与模糊学, 2025, 15(2): 185-197.
DOI: 10.12677/orf.2025.152075

However, due to the increasingly complex external environment and intense market competition in the domestic industry, the financial risks they face are becoming more severe. Therefore, it is crucial to develop a financial risk warning model suitable for manufacturing enterprises in China. Given the significant disparity in the number of ST and non-ST enterprises in the domestic manufacturing sector, most studies utilize imbalanced data. In response, this paper matches selected ST enterprises with non-ST enterprises of the same asset size in a 1:3 ratio and employs the SMOTE oversampling method to address the imbalanced dataset. The study uses data from 168 Chinese manufacturing listed companies from 2020 to 2023 and compares the performance of 12 machine learning models for risk prediction. The results indicate that the Extra Trees model performs the best, with an 18% improvement in prediction accuracy after balancing the dataset. This research aims to provide practical reference and application value for financial risk warning and prevention in domestic manufacturing enterprises, contributing to the stable development of the industry and economy.

Keywords

Imbalanced Data, Machine Learning, Financial Risk Warning, Manufacturing Industry

Copyright © 2025 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

近年来，随着人工智能、大数据和云计算等数字技术飞速进步，我国数字经济正以前所未有的速度蓬勃发展，逐渐成为实体经济转型发展的关键驱动力，倒逼着制造业企业加快转型升级速度。又因 2020 年初全球新冠疫情的爆发，加之制造业企业的研发投入高产出少、产品开发周期长和供应链不稳定，急速加剧了制造业企业的财务风险问题。制造业是实体经济高质量发展的关键领域，而高质量发展是全面建设社会主义现代化国家的首要任务。为完善制造业企业财务风险管理体系建设，研究并构建出适配我国制造业上市企业的财务风险预警模型，对于促进我国制造业行业健康发展、助力国民经济高质量发展至关重要。

对于国内上市企业财务风险预警的研究，传统的 BP 神经网络早已被应用，丁琳(2008) [1]通过 1:1 配对同规模、行业与时间的 ST 与非 ST 公司参与模型训练，在研究中创新地将公司治理指标加入模型中。宋彪、朱建明、李煦(2015) [2]选取 2012、2013 年被沪深两市特别处理的工业制造业企业，将 ST 企业与非 ST 企业按 1:2 的比例进行随机抽样配比，并引入大数据指标，利用支持向量机构建财务风险预警模型进行研究。龙志、陈湘州(2023) [3]选取 2018~2022 年我国 A 股上市企业为研究样本，为平衡样本数据集采用了 SMOTE 过采样方法，构建了 TOPSIS-FCM-CNN 的财务风险预警模型进行研究。马秀颖、郝立宁等(2024) [4]选取了 2018~2022 年中国 A 股非金融上市公司，但对于不平衡数据集并未处理，只是采用对不均衡数据敏感性较高的查全率作为主要判别效果评价指标。王德青、薛守聪等(2025) [5]选取了 40 家 ST 企业按照 1:1 等比例匹配行业、资产规模相近的非 ST 企业，并构建了函数型 Logistic 模型进行了财务风险预警研究。

综上，从现有的多数相关研究中不难发现样本数据集往往是极为不平衡的，多数情况下因为方法的局限性、相关数据特征不符前提要求就使得研究难以进行，只能对研究样本进行 1:1 或 1:2 的对照样本匹配。本文在对选取的 ST 企业样本按照 1:3 比例匹配非 ST 企业的基础上，利用 SMOTE 过采样方法对不

平衡数据集进行处理，并结合机器学习方法构建制造业企业的财务风险预警模型进行研究。本文研究旨在为市场环境复杂化、财务风险多元化下的企业财务风险预警研究提供帮助，也为实现制造业健康、可持续的发展，助力实体经济发展、建设制造业强国提供保障。

2. 财务风险相关概念

财务风险是指融资不当、资本结构不合理和财务状况不佳可能导致利益相关者破产或损失的风险。对于企业而言，财务风险往往由外部环境和企业自身带来，外部环境主要包括社会、技术和经济环境等方面变化的影响，企业自身包括内部控制的失效、财务决策的不当、财务管理制度不适应、战略决策的失误、治理结构的失衡和资本结构的不合理等。

国外学者通常将企业资不抵债以致破产的状况视为财务风险的标志，这是狭义的视角。然而在我国普遍的观点是认为如果一个企业连续两年亏损，那么它就面临着财务风险。基于这一标准，研究者们通常会选择被特别处理(ST)或被退市风险警示(*ST)的企业作为数据收集的重点对象，并将这些企业与正常运营的企业进行对比。从企业资本周转全过程的角度来考虑，可以将财务风险划分为筹资风险、投资风险、营运风险和收益分配风险，本文研究也据此角度来构建合适的财务指标体系。

3. 机器学习模型

机器学习作为人工智能的核心分支，是一门融合计算机科学、统计学和数学优化等多学科理论的交叉学科。其核心机制是通过对海量数据的深度学习和模式识别，自主构建预测模型。当面对新的数据输入时，系统能够基于已有模型进行智能判断，并通过持续学习不断优化模型性能，最终实现类人的学习能力。

3.1. 传统机器学习模型

(1) 支持向量机：

支持向量机(Support Vector Machine, SVM)的核心原理是在特征空间中构造一个最优分离超平面，该超平面能够最大化不同类别数据点之间的分类间隔。在处理线性可分数据时，SVM 通过求解以下优化问题来确定最优超平面：

$$\min \quad \frac{\|w\|^2}{2}+C\sum_{i=1}^{n}\xi_i \tag{1}$$

$$\text{s.t.} \quad y_i\left(w^{\mathrm{T}}x_i+b\right)\geq 1-\xi_i,\ \ \xi_i\geq 0 \tag{2}$$

其中 w 是超平面的法向量，b 是偏置项，ξ_i 是松弛变量，C 是正则化参数。对于非线性可分数据，SVM 采用核技巧将原始特征空间映射到高维特征空间，在该空间中寻找线性可分超平面。

(2) 朴素贝叶斯：

朴素贝叶斯(Naive Bayes)的核心在于利用特征的条件概率进行类别预测，其理论基础是贝叶斯定理：

$$P(Y\mid X)=P(X\mid Y)\cdot P(Y)/P(X) \tag{3}$$

其中，$P(Y\mid X)$ 是后验概率，表示在观测到特征 X 的条件下类别 Y 的概率；$P(X\mid Y)$ 是似然函数，表示在类别 Y 的条件下特征 X 出现的概率；$P(Y)$是先验概率，表示类别 Y 的初始概率；$P(X)$是证据因子，作为归一化常数。

(3) K 最邻近：

K 最邻近算法(K-Nearest Neighbor, KNN)的核心原理是通过计算待测样本与训练集中所有样本的特

征空间距离，选取距离最近的 K 个邻居，根据这些邻居的类别或属性值进行预测决策。

(4) 逻辑回归：

逻辑回归(Logistic Regression, LR)对前提假设的要求相对宽松，无需满足因变量与自变量之间的线性关系假设，亦不要求误差项服从正态分布。逻辑回归公式如下：

$$\mathrm{LR}\left(\frac{P}{1-P}\right)=\beta_0+\beta_1X_1+\beta_2X_2+\cdots+\beta_nX_n \tag{4}$$

其中，X_n 表示第 n 个自变量，$\beta_n\left(n=1,2,\cdots,n\right)$ 为变量的参数估计值。

(5) 决策树：

决策树(Decision Tree, DT)能够从具有复杂结构和潜在无序性的训练数据中，构建出层次化的树形分类模型。其是通过递归地选择最优特征进行数据划分，逐步生成决策规则，从而实现对样本的高效分类。

3.2. 集成学习模型

(1) 随机森林：

随机森林(Random Forest, RF)是通过构建多棵差异性显著的决策树，并依据平均(回归任务)或多数表决(分类任务)的原则进行集成决策，这显著提升了模型的泛化能力和鲁棒性。

(2) 极度随机树：

极度随机树(Extremely Randomized Trees, ExtraTrees)的核心思想在于通过集成多棵随机决策树，并对其预测结果进行聚合来生成最终预测。这种方法不仅提高了模型的稳定性，还增强了其泛化能力。

(3) 多层感知机：

神经网络模型，特别是多层感知器(Multi-Layer Perceptron, MLP)通过将输入特征进行多层次的线性和非线性变换，逐步提取和组合信息，最终生成输出结果。最简单的 MLP 模型公式表达如下：

$$\hat{y}=w[0]\cdot w[1]+w[1]\cdot x[1]+\cdots+w[p]\cdot x[p]+b \tag{5}$$

其中，$x[p]$ 为样本特征值，$w[p]$ 为特征值权重，b 为偏置项，$\hat{y}$ 为线性组合输出。

(4) XGBoost：

XGBoost (Extreme Gradient Boosting)在传统 GBDT 的基础上引入了正则化项以控制模型复杂度，同时采用了二阶泰勒展开对损失函数进行优化，并支持并行化计算与稀疏数据处理。具体的目标函数如下：

$$Obj\left(t\right)=\sum_{i=1}^{n}\left(y_i,\hat{y}_i^{(t-1)}+f_t\left(x_i\right)\right)+\Omega\left(f_t\right)+\text{constant} \tag{6}$$

(5) LightGBM (LGBM)：

LightGBM (Light Gradient Boosting Machine)是采用了基于直方图的决策树算法，将连续特征值离散化为直方图 bins，从而减少了计算复杂度和内存占用。通常 LightGBM 模型的输出结果可以表示为：

$$y=\sum_{t=1}^{T}f_t\left(x\right) \tag{7}$$

x 表示输入数据，T 是决策树数量，y 表示输出结果。

(6) 梯度提升树(GBT)：

梯度提升树(Gradient Boosting Trees, GBT)其核心思想是通过迭代地构建多个决策树，并将它们的结果累加，从而逐步提升模型的预测性能。在每一轮迭代中，GBT 会添加一个新的决策树，该树的目标是

拟合上一轮迭代后模型的残差。通过这种方式，每一棵新树都在尝试修正前一轮模型的错误，从而逐步降低整体的预测误差。

(7) AdaBoost：

AdaBoost (Adaptive Boosting，自适应增强)是通过组合多个弱分类器(如决策树、感知机等)来构建一个强分类器，其核心思想是通过自适应调整样本权重，使模型在每一轮迭代中更加关注那些难以分类的样本，最终通过加权多数表决的方式将所有弱分类器整合为一个强分类器。

3.3. 机器学习的评价标准

本文采用 5 种常用评价指标：精准率、召回率、F1 分数、准确率和 AUC 值，来全面评估模型的整体性能。在具体应用中，ST 企业为正类，非 ST 企业为负类，并通过混淆矩阵中的真正类(*TP*)、真负类(*TN*)、假正类(*FP*)和假负类(*FN*)来计算上述评价指标。混淆矩阵是计算上述指标的基础，其结构如下表 1 所示。

Table 1. Confusion matrix
表 1. 混淆矩阵表

实际分类	预测为 ST 企业	预测为非 ST 企业
ST 企业	*TP*	*FN*
非 ST 企业	*FP*	*TN*

(1) 精准率

精准率(Precision)衡量的是模型预测为正类的样本中实际为正类的比例，即全部预测为 ST 的样本中实际为 ST 的样本量。其计算公式为：

$$P = \frac{TP}{TP + FP} \tag{8}$$

(2) 召回率

召回率(Recall)衡量的是实际为正类的样本中被模型正确预测为正类的比例，即 ST 的样本中有多少被正确预测。其计算公式为：

$$R = \frac{TP}{TP + FN} \tag{9}$$

(3) F1 分数

F1 (F1-Score)分数是精准率和召回率的调和平均数，它综合反映了模型的预测精度和覆盖率，特别适用于不平衡数据集。其计算公式如下：

$$\mathrm{F1} = \frac{2 \times P \times R}{P + R} \tag{10}$$

(4) 准确率

准确率(Accuracy)衡量的是模型预测正确的样本占总样本的比例，它反映了模型的整体预测能力。其计算公式如下：

$$ACC = \frac{TP + TN}{TP + FP + FN + TN} \tag{11}$$

(5) AUC 值

AUC 值是 ROC 曲线下的面积，用于衡量模型在不同阈值下区分正负类的能力。AUC 值越接近 1，模型的分类性能越好；AUC 值为 0.5 时，模型相当于随机猜测。

4. 研究样本的选取

4.1. 数据来源

本文模型构建所需要的数据均来自于锐思(RESSET)数据库，在剔除核心变量缺失的样本后，本研究最终选取了 2020 年至 2023 年间国内 A 股市场中 42 家首次被特别处理(ST)的制造业上市企业以及 126 家正常经营的制造业上市企业作为研究样本。

4.2. 样本选取

本文对研究样本的选取分为两个步骤：第一步，筛选出 A 股市场 2020 至 2023 年间首次被特别处理的制造业上市企业，初步确定有 61 家 A 股制造业企业被特别处理，在排除了 18 家因其他原因被特别处理、历史上曾被 ST、关键数据缺失的企业后，保留了 42 家 ST 企业；第二步，依据证监会 2012 年的行业分类标准，按照 1:3 的比例，匹配了相同行业、相同会计期间且总资产规模相近的 126 家非 ST 制造业上市企业作为对照组。综上，本文共选取了 168 家制造业上市企业作为研究样本。

4.3. 时间选取

证监会规定上市企业连续两年净利润为负是触发“其他风险警示”的条件之一，不少研究文献也表明，越是使用接近企业被 ST 时间点的财务数据，模型的预测准确性越高。因此，本文选择使用企业首次亏损的前一年(第 T 年)作为分析的时间点。以 2023 年作为被预测年为例，那么 2020 年为第 T 年，2021 年为第 T + 1 年，2022 年为第 T + 2 年。本文将使用 2020 年(第 T 年)的数据来做 2023 年(第 T + 3 年)的企业财务风险预测。

5. 预警指标的选取

5.1. 财务指标选取

本文参考财务风险预警相关领域的高质量文献资料，构建了一个适宜的财务指标体系，该体系主要包括了盈利能力、偿债能力、成长能力、营运能力、现金流量和资本结构。下表 2 列示了部分关键一、二级指标，完整的详表因篇幅原因无法完全展开。

Table 2. Financial indicators and calculation formulas
表 2. 财务指标及计算公式

一级指标	二级指标	计算公式
盈利能力	净资产收益率	净利润*2/(期初股东权益 + 期末股东权益)
	资产报酬率	息税前利润*2/(期初总资产 + 期末总资产)
……	……	……
资本结构	流动负债权益比率	流动负债/股东权益
	权益乘数	资产合计/股东权益合计

盈利能力是企业财务绩效的核心指标之一，反映了企业通过经营活动获取利润的能力。偿债能力是

衡量制造业上市公司能否按期偿还债务的关键指标，它对于评估企业能否持续稳定发展至关重要。成长能力是企业发展速度与潜在价值的重要体现，强大的成长能力不仅是企业持续发展的驱动力，也是其增强市场竞争力的核心要素之一。营运能力是评估企业管理和运用资产效率及效果的关键指标，反映了企业在资源配置、运营管理及资金周转方面的综合水平。现金流量是企业财务状况的实时反映，它能够揭示企业持续发展的能力。资本结构指的是上市公司资金来源的组成及其比例分配，主要由长期资本和短期债务构成，资本结构的选择直接影响企业的财务风险。

5.2. 非财务指标选取

引入非财务指标与财务指标相结合，构建多维度的财务风险预警模型能够显著增强模型的性能与效用，以更精准地识别企业财务风险。本文选择了反映股权结构、组织治理结构和研发投入三个方面的指标，这些非财务指标能够从公司治理、创新能力及战略规划等维度补充财务指标的不足，为财务风险预警提供更为全面的分析视角。部分非财务指标及其定义如下表 3 所示，详表因篇幅原因无法完全展开。

Table 3. Non-financial indicators and calculation formulas
表 3. 非财务指标及计算公式

一级指标	二级指标	计算公式
股权结构	股权集中度 1	第一大股东持股数/总股数
	股权集中度 5	前五大股东持股之和/总股数
	……	……
组织治理结构	独立董事比例	独立董事人数/董事会总人数

股权结构是影响企业决策效率、经营稳定性及长期发展能力的关键因素；研发投入是制造业企业发展的关键驱动力。然而在激烈的市场竞争中，研发失败的风险相对较高，会影响产品推广、迭代，继而影响现金流会引发财务风险问题；良好的公司治理结构能够有效降低代理成本、提升决策透明度，并增强企业的市场信誉与竞争力董事会是公司治理的核心，独立董事的引入有助于增强董事会的独立性与决策科学性。

5.3. 数据预处理

本文在样本与指标选取后对数据集进行了缺失值处理，考虑到数据集中包含许多数值较大的数据点，这导致计算复杂度增加，模型不容易收敛。因此，本文对数据进行了归一化处理，将数据值转换到了(0, 1)范围内，以便于后续研究工作的进行。

5.4. 特征重要性分析

本文采用 XGBoost 模型对数据集中的特征进行重要性分析排序，以验证所选指标的适宜性。基于原始数据集对模型进行训练，并计算出 79 个特征预警指标的重要性并排序，其结果的可视化展示如图 1 所示。

图 1 展示了 79 个预警指标在 XGBoost 模型中的重要性得分排序，这些得分是通过对特征重要性进行加权平均计算得出，重要性得分反映了每个特征对模型预测能力的贡献大小。通过观察可以发现所有选定的指标都显示出了一定的重要性，从而证实了这些指标的选择是合理适宜的。指标的得分较高、排

序在前就表明其拥有着更为优异的区分能力。

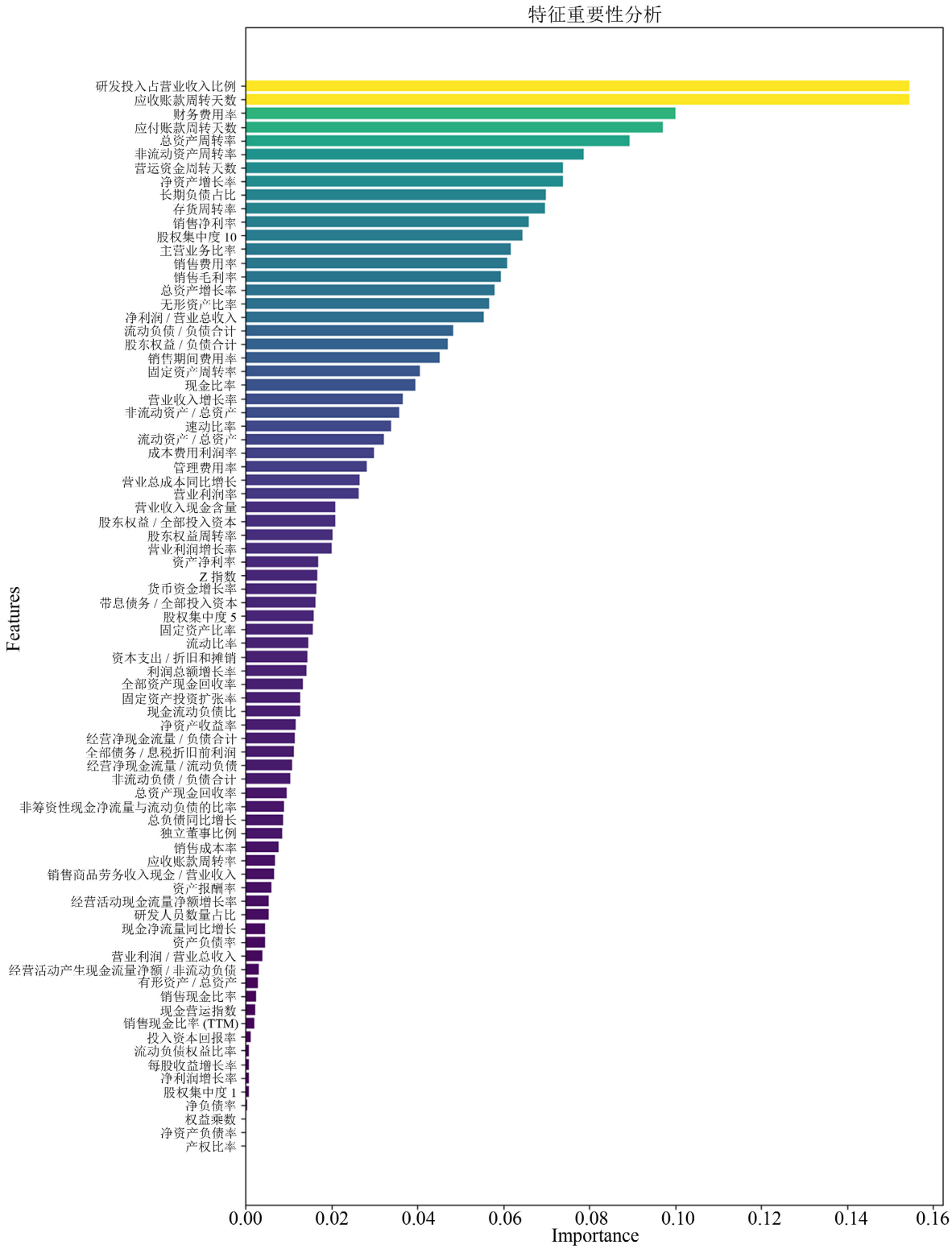

Figure 1. Feature importance analysis
图 1. 特征重要性分析

6. 实验结果与分析

6.1. 预警模型的构建

本文首先构建了一个包含 12 种不同机器学习模型的分类器，将数据集按照 70%训练集和 30%测试集的比例进行划分，再让数据进行特征缩放遍历每个分类器进行模型训练。在训练完成后，使用测试集对模型进行预测，并计算每个模型的精确度、召回率、F1 分数、准确率和 AUC 值，最后输出这些性能指标如表 4 所示，并绘制相应的 ROC 曲线如图 2 所示。

Table 4. Comparison of test results for different models
表 4. 不同模型测试结果对比

Model	Precision	Recall	F1-score	Accuracy
Support Vector Machine	0.45	0.33	0.38	0.69
Random Forest	0.56	0.60	0.58	0.75
Extra Trees	0.50	0.40	0.44	0.71
Naive Bayes	0.50	0.53	0.52	0.71
K Nearest Neighbors	0.57	0.27	0.36	0.73
Logistic Regression	0.41	0.47	0.44	0.65
Decision Tree	0.50	0.60	0.55	0.71
XGBoost	0.58	0.47	0.52	0.75
LGBM	0.50	0.47	0.48	0.71
Multilayer Perceptron	0.40	0.53	0.46	0.63
Gradient Boosting Trees	0.53	0.53	0.53	0.73
AdaBoost	0.46	0.40	0.43	0.69

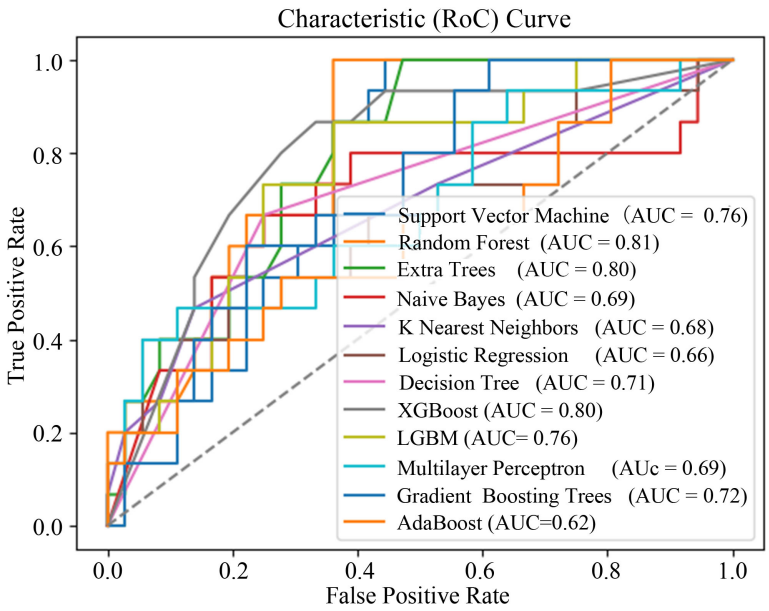

Figure 2. Comparison of ROC curves for different models
图 2. 不同模型 ROC 曲线对比图

6.2. SMOTE 算法原理与不平衡数据处理

考虑到我国市场上 ST 企业与非 ST 企业的数量存在较大差异，这种不平衡的数据分布可能会让预测偏向于多数类，从而影响预警模型的性能。为了解决这个问题，一般研究中通常会采用手动匹配样本或利用算法的方式来减轻类别不平衡对分类结果的影响。本文选择了 SMOTE (Synthetic Minority Over-Sampling Technique)插值采样方法来处理不平衡的数据，它通过测量少数类样本之间的欧几里得距离来确定每个样本的 K 个最近邻，再根据数据不平衡的程度随机选择一定数量的这些近邻样本，与原有的少数类样本结合，创造出新的“少数类样本”。这个过程有效地扩充了少数类样本的规模，使数据集更加平衡。这种方法不仅能够保留原有少数类样本的信息，还能通过生成新的合成样本来提高模型对于少数类的识别能力，从而减少因数据不平衡导致的影响。下图 3 为 SMOTE 算法的原理图，图中蓝色五角星为少数类样本，黑色五边形为多数类样本，蓝色十字星形则为生成的“少数类样本”。

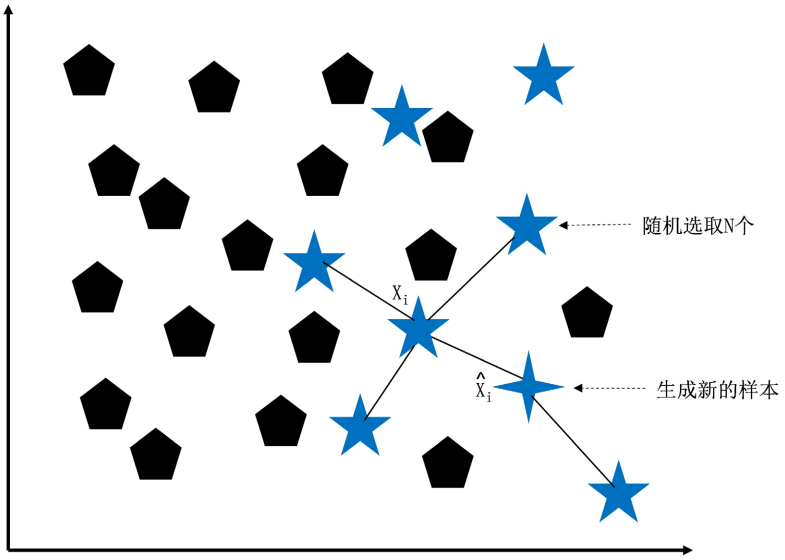

Figure 3. Diagram of the SMOTE algorithm principle
图 3. SMOTE 算法原理图

6.3. 平衡后数据集测试结果与 AUC 可视化分析

在定义了包含 12 种不同机器学习模型的分类器之后，本文对经过 SMOTE 过采样处理的数据集进行了重新划分，训练集和测试集的比例设定为 7:3。再让数据进行特征缩放遍历每个分类器，进行模型训练，在测试集上进行预测，并计算了每个模型的精确度、召回率、F1 分数、准确率和 AUC 值，并绘制了 ROC 曲线以供分析，如表 5 与图 3 所示。

结果如表 5 所示，从评价指标准确率看，依次从高到低为：Extra Trees、Random Forest、K Nearest Neighbors、Multilayer Perceptron、LGBM、Support Vector Machine、Gradient Boosting Trees、Logistic Regression、Decision Tree、AdaBoost、Naïve Bayes。通过综合比较，其中分类效果最好的是 Extra Trees 模型，其精准率、召回率、F1 分数分别为 0.86、0.95、0.90，表明该模型能够更好地应用于财务风险预测。

对比图 2 与图 4 可以发现，经过 SMOTE 处理后的数据集让模型预测的整体效果比之前未平衡数据集的效果有显著的提升。结合图表来看，Extra Trees 模型的 AUC 值达到了 0.97，这一数值在所有模型中是最高的，表明 Extra Trees 模型能够有效地识别出 ST 企业，显示出该模型在制造业上市企业财务风险预警方面具有出色的效果。相较于未对非平衡数据集处理前，经处理后的 ExtraTrees 模型的精准率、召

回率、F1 分数、准确率分别提高了 36%、55%、46%和 18%，进而验证了本文中对非平衡数据集处理方式的科学性与可行性。

Table 5. Comparison of test results for different models on the balanced dataset
表 5. 平衡数据集后不同模型测试结果对比

Model	Precision	Recall	F1-score	Accuracy
Extra Trees	0.86	0.95	0.90	0.89
Random Forest	0.82	0.97	0.89	0.88
K Nearest Neighbors	0.87	0.87	0.87	0.87
Multilayer Perceptron	0.80	0.97	0.88	0.87
LGBM	0.80	0.95	0.87	0.86
Support Vector Machine	0.78	0.95	0.86	0.84
Gradient Boosting Trees	0.77	0.95	0.85	0.83
Logistic Regression	0.76	0.89	0.82	0.80
Decision Tree	0.77	0.87	0.81	0.80
AdaBoost	0.72	0.95	0.82	0.79
XGBoost	0.76	0.82	0.78	0.78
Naive Bayes	0.80	0.53	0.63	0.70

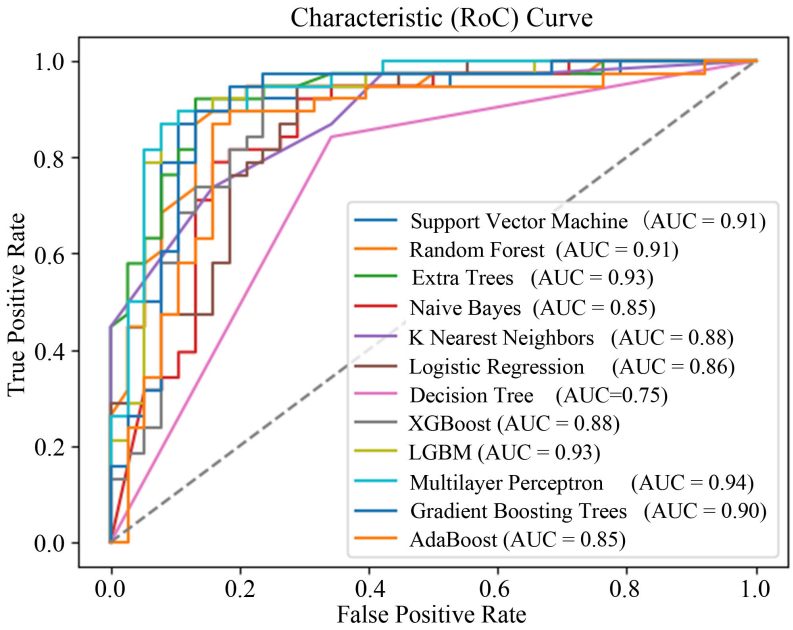

Figure 4. Comparison of ROC curves for different models on the balanced dataset
图 4. 平衡数据集后不同模型 ROC 曲线对比图

6.4. 模型的超参数调优

合理选择超参数能够显著提升机器学习模型的性能。通过对超参数的调整，可以提高模型的训练效率与预测精度，经过良好调优的模型具备更高的鲁棒性与稳定性。本文采用了 GridSearchCV (网格搜索)

进行超参数优化。首先定义了超参数网格，确定 n_estimators、criterion、max_depth、min_samples_split、min_samples_leaf 多个超参数的候选值，用于后续的网格搜索。进行网格搜索时，使用 GridSearchCV 进行 5 折交叉验证(cv = 5)来找到最佳的超参数组合。在经过多轮调整之后得到了最优模型(ExtraTrees)的最佳参数，criterion = entropy、max_depth = 10、min_samples_leaf = 8、min_samples_split = 20、n_estimators = 250。

6.5. 不同数据平衡技术效果对比

本文在构建模型时也选取了其他两种数据平衡技术来进行处理。其一是欠采样处理，设置 frac 为 0.33 减少多数类样本的数量进行平衡；其二是数据增强，对原始数据集添加高斯噪声与插值处理。在 ExtraTrees 模型中其与 SMOTE 处理的效果对比如下表 6 所示。

Table 6. Comparison of model performance with different data processing methods
表 6. 不同数据处理方法的模型效果对比

Method	Precision	Recall	F1-score	Accuracy
SMOTE	0.86	0.95	0.90	0.89
Undersampling	0.72	0.73	0.72	0.73
Data Augmentation	0.79	0.83	0.82	0.80

通过表中数据对比，可以明显看出 SMOTE 差值采样方法的模型效果更好，因此本文在模型研究时最终也选取了 SMOTE 处理的方法。但如果少数类样本中存在噪声，SMOTE 可能会放大这些噪声从而影响模型的性能，且插值会影响过拟合的问题。本文在模型构建时采用了正则化技术、选择了合适的 K 值，且使用集成学习方法也能有效减轻这些偏差，提升模型的泛化能力。

7. 结论

针对现有的企业财务风险预警研究存在研究样本数据集不平衡的问题，本文提出了利用 1:3 比例样本匹配与 SMOTE 过采样结合的方法来改善数据不平衡问题。具体结论为：一，针对样本数据不平衡性问题，本文为研究样本按 1:3 比例匹配对照组，并对数据集采用了 SMOTE 算法进行处理，有效解决了传统随机过采样方法易导致模型过拟合的问题，为机器学习模型的特征学习提供了更优质的数据基础；二，本文系统比较了传统机器学习模型与集成学习模型共 12 种算法的预测性能，实证结果表明 ExtraTrees 模型在主要评价指标上均优于其他算法，展现出更有效的性能，适用于制造业上市企业的财务风险预测；三，对不平衡数据集进行改善处理后，文中 12 种机器学习的模型效能都有显著的提升，其中 ExtraTrees 模型的精准率、召回率、F1 分数、准确率分别提高了 36%、55%、46%和 18%，验证了技术方法的有效性。

然而本文研究仍存在一些局限性，未来可从以下几个方向进一步深化。首先，本文所采用的数据集缺乏时间维度特征，后续研究可进一步整理精细的时序数据，以捕捉企业财务风险的动态演变规律。其次，还可结合宏观经济指标与行业周期性特征，构建更为全面的风险预警体系，从而提升模型的预测精度与适用性。此外，考虑到数据质量对模型预测性能的关键影响，后续研究可引入数据质量评估框架，结合异常检测和文本情感分析等技术，构建具有数据质量保障机制的机器学习财务风险预警模型，从而提高模型的鲁棒性和预测准确性。

参考文献

[1] 丁琳. 基于公司治理的上市公司 BP 神经网络财务预警模型研究[J]. 当代经济, 2008(12): 116-117.

[2] 宋彪, 朱建明, 李煦. 基于大数据的企业财务预警研究[J]. 中央财经大学学报, 2015(6): 55-64.

[3] 龙志, 陈湘州. 企业财务风险预警模型的构建与检验[J]. 财会月刊, 2023, 44(24): 54-61.

[4] 马秀颖, 郝立宁, 木仁, 崔巍. 改进判别分析模型及其在上市公司财务风险预警问题中的应用[J]. 数理统计与管理, 2024(10): 1-25.

[5] 王德青, 薛守聪, 芦智昊, 郭梦霞, 侯伊雯. 基于函数型 Logistic 模型的企业财务困境预警研究[J]. 运筹与管理, 2025(1): 1-8.