Statistics and Application 统计学与应用, 2025, 14(11), 236-248
Published Online November 2025 in Hans. https://www.hanspub.org/journal/sa
https://doi.org/10.12677/sa.2025.1411326

Hans汉斯

机器学习驱动的财务舞弊识别模型

梁楚雯[1]，张慕瑶[2]，谷司琪[3]

[1]北方工业大学理学院，北京
[2]北方工业大学经济管理学院，北京
[3]北方工业大学电气与控制工程学院(无人机学院)，北京

收稿日期：2025年10月19日；录用日期：2025年11月10日；发布日期：2025年11月20日

摘　要

财务舞弊是企业或个人操纵财务信息以谋取私利。本文结合高检测机器学习模型，建立分类能力较强的财务舞弊识别模型，以提高财务审计效率。建模分三部分：第一，数据预处理。本文收集了不同公司2017~2023年的财务报表及舞弊公司数据，进行数据清洗。依财务理论及SPSS的独立样本*T*检验，确定六大财务特征指标。最后对数据进行Z-Score标准化。第二，随机森林模型的建立与评估。本文运用Python构建随机森林模型。并运用混淆矩阵进行模型评估，舞弊样本的召回率为58%，模型AUC值为73%，可知该模型的区分性能较好，但舞弊召回率仍有待提升。第三，XGBoost模型的建立与评估。本文运用Python构建XGBoost模型。用分层抽样方法确定数据集、直方图算法加速建树，并通过带早停的交叉验证训练训练模型并监控性能。运用混淆矩阵进行评估，得到模型舞弊样本的召回率为67%，模型AUC值为79.06%，模型对两类样本的分类能力较好，且舞弊召回率较高。

关键词

财务舞弊，随机森林模型，XGBoost模型

Machine Learning-Driven Financial Fraud Detection Model

Chuwen Liang[1], Muyao Zhang[2], Siqi Gu[3]

[1]School of Science, North China University of Technology, Beijing
[2]School of Economics and Management, North China University of Technology, Beijing
[3]School of Electrical and Control Engineering (School of Unmanned Aerial Vehicle), North China University of Technology, Beijing

Received: October 19, 2025; accepted: November 10, 2025; published: November 20, 2025

文章引用: 梁楚雯, 张慕瑶, 谷司琪. 机器学习驱动的财务舞弊识别模型[J]. 统计学与应用, 2025, 14(11): 236-248.
DOI: 10.12677/sa.2025.1411326

Abstract

Financial fraud refers to the manipulation of financial information by enterprises or individuals to seek personal gains. This paper combines high-detection machine learning models to establish a financial fraud identification model with strong classification capabilities, aiming to enhance the efficiency of financial auditing. The modeling process is divided into three parts: First, data preprocessing. This paper collects financial statements of different companies from 2017 to 2023, along with data on fraudulent companies, and conducts data cleaning. Based on financial theories and the independent samples T-test in SPSS, six key financial characteristic indicators are identified. Finally, the data undergoes Z-Score standardization. Second, the establishment and evaluation of the Random Forest model. This paper employs Python to construct the Random Forest model. The model is evaluated using a confusion matrix, revealing a recall rate of 58% for fraudulent samples and an AUC value of 73%. It can be seen that the model demonstrates good discriminatory performance, yet there is still room for improvement in the recall rate of fraudulent cases. Third, the establishment and evaluation of the XGBoost model. This paper utilizes Python to build the XGBoost model. The dataset is determined using stratified sampling, and the histogram algorithm is employed to accelerate tree construction. The model is trained and its performance is monitored through cross-validation with early stopping. Upon evaluation with a confusion matrix, the model shows a recall rate of 67% for fraudulent samples and an AUC value of 79.06%. The model exhibits good classification capabilities for both types of samples, with a relatively high recall rate for fraudulent cases.

Keywords

Financial Fraud, Random Forest Model, XGBoost Model

1. 引言

1.1. 研究背景

在当今复杂多变的经济环境中，财务舞弊行为已成为一个全球性问题，严重破坏了资本市场的诚信基础，并可能引发系统性金融风险。

传统的财务舞弊识别方法主要依赖于审计师的经验判断和对财务报表的简单分析，该模式在实践的过程中会出现一些矛盾点，如手工审计抽样技术与非结构化数据处理需求间的矛盾日益凸显，可能导致审计周期的延长与边际成本的递增。

财务舞弊的检测数据源聚焦于企业的财务报表，如资产负债表、利润表、现金流量表等。多数研究表明，机器学习在财务舞弊识别中具有较高的准确性和效率，能够从大量的财务数据中快速、准确地发现异常数据，并进行分类、预测和识别。

1.2. 研究目标

目前阶段，财务报表的审核并非完全依靠人工审计，还伴随着技术的深度协同。为了减轻人工审计工作量，提高整体财务数据审核效率，本文致力于建立两个基于机器学习的财务舞弊识别模型，希望模

型能够依据所选择的财务特征指标高准确率地识别出财务舞弊的情况，并不断优化模型的区分性能，为审计人员提供初步审计阶段的有力决策支持。

1.3. 研究方法

本文主要采用了模型建立法、文献调查法和比较研究法等。

① 模型建立法。本文基于经典财务舞弊理论框架，采用机器学习方法构建财务舞弊识别模型。在数据预处理阶段，我们利用 SPSS、Excel 等软件对分析财务报表主表的数据，并根据独立样本 T 检验的原理，合理利用软件筛选出具有显著判断力的财务特征变量。在模型构建阶段，我们分别运用随机森林模型以及 XGBoost 模型，构建出较高判别度的财务舞弊识别模型。

② 文献调查法。根据本文主题，结合中国知网等权威知识库，我们对财务舞弊理论、财务特征指标的选取以及模型理论公式等内容进行查阅，为本文的研究做出更加完整的理论支撑。

③ 比较研究法。本文的比较研究法主要体现在不同机器学习模型作为财务舞弊识别模型的性能优劣，本文依据两种模型在财务舞弊领域的性能优劣不同，进行尝试性比较研究。

1.4. 文献综述

我国对财务舞弊识别模型的研究主要从 2005 年开始。朱敏[1]从实证的角度通过我国上市公司的财务数据分析构建单因素方差分析方法等四种分析模型作为识别模型；刘君[2]采用径向基概率神经网络构建模型，与线性模型相比，提高了精确度；张新民[3]等通过 Logistic 回归技术得到模型总体识别率为 71.3%；贺颖[4]基于偏最小二乘法的有效降维优势以及支持向量机的拟合度高等优势构建模型，使得舞弊公司、非舞弊公司的识别正确率都达到了 80%以上；陈彬[5]建立了 BP 神经网络识别模型等四种模型，并以“期望错误分类成本法”为依据对比四种识别模型，得出 Elman 神经网络技术在财务舞弊识别模型上识别的准确率最高；成雪娇[6]采用 GA-BP 神经网络、支持向量机等三种算法对样本数据进行训练建模，并进一步通过模型叠加构建出了一个新的识别模型；王柳匀[7]基于分类算法构建了一些分类型识别模型，其中随机森林模型的准确率最高，达到 79%；程新尧[8]基于两组舞弊指标，构建了机器学习模型、信息融合模型对舞弊进行识别；赵豪杰[9]对财务指标数据进行整理，构建了 CNN 和带注意力机制的 BiLSTM 融合模型——CNN-BiLSTM，并在纵向的多年度数据对比上，该模型的性能得到了提升；何子英[10]基于集成学习算法与 SHAP 的可解释机器学习模型，准确地识别出了上市公司的财务舞弊行为。

财务舞弊识别模型的发展经历了几个阶段：

- 在早期，财务舞弊识别模型主要通过关键指标的异常波动来识别风险，主要用到的模型有 Altman Z-Score 模型和 Beneish M-Score 模型，Altman Z-Score 使用 5 个财务比率指标预测企业破产风险，Beneish M-Score 通过 8 个财务特征指标检测盈余的操纵。该阶段的模型缺乏系统化的量化分析框架，并且指标的选择和阈值的设定具有较强的主观性。
- 中期阶段是统计学习方法盛行的时期，这一阶段发展了监督学习的核心算法如支持向量机(SVM)和 EM 算法，SVM 通过核技巧将线性分类拓展到高维非线性空间，成为处理文本、图像等复杂数据的核心工具。EM 算法则通过 E 步(期望步)和 M 步(极大步)的交替迭代，解决了含隐变量模型的参数估计问题。然而，该时期的模型高度依赖人工设计的特征，而特征选择的偏差可能遗漏关键舞弊信号，或引入冗余噪声，降低模型准确性和鲁棒性。
- 当前正处于机器学习与深度学习盛行的时期，财务舞弊识别模型已经进入了一个新的发展阶段，随机森林模型和 XGBoost 等作为机器学习的分支，在财务舞弊识别方面也展现出了很强的泛化能力。随机森林等集成学习方法，通过构建多棵决策树并综合其结果，具有很强的抗过拟合能力和泛化能力。

1.5. 假设条件

◆ 任何财务舞弊行为最终都会体现于报表数据的不实。
◆ 舞弊公司与非舞弊公司在特征空间中的分布可通过线性决策边界分离。
◆ 各财务特征指标相互独立。
◆ 历史数据的舞弊模式适用于未来舞弊样本的输入。

2. 数据预处理

2.1. 数据采集与数据清洗

在本文中，我们建模的目标是实现对不同公司的财务舞弊数据进行较高准确度地识别，因此在建模之前，我们需要收集不同公司的财务报表数据。根据 CSMAR 数据库提供的财务报表数据，我们选择从该数据库中下载 2017~2023 年的财务报表主表数据(包含资产负债表、利润表、现金流量表和所有者权益变动表)和舞弊公司的数据(包括证券代码、日期以及是否舞弊信息)作为本文的样本数据集，并把 B 股、北交所、新三板的股票删除，因为 B 股和 A 股是一个公司，而新三板和北交所的股票因挂牌和上市门槛较低、投资者准入门槛较高等因素而不具有代表性。同时，剔除金融业(证监会 2012 版行业分类)的财务报表数据，因为金融业的商业模式与其他行业有显著不同，这使得财务数据的披露以及生成方式也会有所不同，因此若不剔除金融业，则研究结果可能会收到这些差异的干扰，从而影响结果的准确性。为了简化研究的过程，我们只考虑年度财务报表，对于季度、半年度等会计中期不予考虑。数据采集过后，我们对缺失值和异常值进行处理，删除四个主表中的大量缺失值，并对一些明显下载过后异常的数据进行移除，得到一份清洗过后的数据集，数据采集流程见图 1。

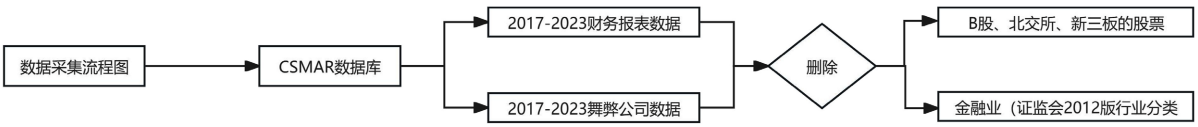

Figure 1. Data collection flowchart
图 1. 数据采集流程图

2.2. 特征指标的构建与分析

为了确定能明显辨别出财务报表数据是否舞弊的财务指标，我们对所收集到的四个主表中的指标进行分析。为了能够比较容易地确定和分析指标，我们利用 Excel 的 Power query 的合并表功能依据两个关键字：证券代码与日期，将四张主表和舞弊公司表合成一张表。最后结合财务理论和会计学等相关知识将资产负债率、流动比率、净资产收益率(ROE)、净利率、经营净现金流/营业总收入、应收账款/销售收入、存货周转率和应收账款周转率作为我们本文研究的八个财务特征指标。

2.3. 特征指标的显著性检验

目前我们选出了八个与财务舞弊相关的指标，但每个指标对财务舞弊影响的显著性程度却未知，而且若财务特征指标过多，可能会造成后续建模的过拟合性。因此，还应对我们挑选的八个指标进行显著性检验，选出对财务舞弊影响最显著的特征指标。

关于财务特征指标的显著性检验的方法，我们选择 SPSS 的独立样本 T 检验。独立样本 T 检验用于比较两组独立样本的均值是否存在显著性差异。原假设为：两组独立样本的均值无显著性差异，而备择假设为：两组独立样本的均值有显著性差异。在 T 检验的过程中，将显著性水平定为 0.05，将公司是否

舞弊(即违规标志)作为分类变量，舞弊为 1，非舞弊为 0，将资产负债率等八个财务特征指标作为检验变量，分析不同指标的显著性差异，得到结果(见图 2)：

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
资产负债率	Equal variances assumed	9.239	.002	-7.159	4978	.000	-.045558538	.0063639235	-.058034632	-.033082444
	Equal variances not assumed			-6.590	2216.618	.000	-.045558538	.0069136649	-.059116476	-.032000601
流动比率	Equal variances assumed	14.320	.000	4.542	4978	.000	.2199190801	.0484206697	.1249932309	.3148449293
	Equal variances not assumed			4.841	2993.664	.000	.2199190801	.0454284514	.1308449383	.3089932219
净资产收益率（ROE)	Equal variances assumed	43.101	.000	3.159	4978	.002	.1372565825	.0434451276	.0520849884	.2224281766
	Equal variances not assumed			2.343	1639.604	.019	.1372565825	.0585760257	.0223648692	.2521482958
净利率	Equal variances assumed	87.352	.000	7.329	4978	.000	.2054370710	.0280316788	.1504826285	.2603915135
	Equal variances not assumed			7.355	2619.390	.000	.2054370710	.0279312902	.1506674405	.2602067015
存货周转率	Equal variances assumed	12.080	.001	1.808	4978	.071	266.8057833	147.5912906	-22.5381824	556.1497490
	Equal variances not assumed			2.829	3795.403	.005	266.8057833	94.32316666	81.87679966	451.7347669
应收账款周转率	Equal variances assumed	6.964	.008	1.783	4978	.075	9.549930992	5.355173022	-.948567885	20.04842987
	Equal variances not assumed			2.314	4702.388	.021	9.549930992	4.127783817	1.457540454	17.64232153
经营净现金流/营业总收入	Equal variances assumed	9.255	.002	-2.226	4978	.026	-.060904872	.0273553077	-.114533329	-.007276414
	Equal variances not assumed			-2.618	3801.289	.009	-.060904872	.0232646878	-.106517345	-.015292398
应收账款/销售收入	Equal variances assumed	147.664	.000	-13.689	4978	.000	-.154007416	.0112506942	-.176063734	-.131951097
	Equal variances not assumed			-11.142	1833.822	.000	-.154007416	.0138227675	-.181117435	-.126897396

Figure 2. Independent T-test results visualization
图 2. 独立样本 T 检验结果图

根据 SPSS 分析的结果图(见图 2)，在这套 *T* 检验的体系中，还包含了 *Levene* 检验。*Levene* 检验是为了评估不同组之间的方差是否存在显著性差异，用来判断方差是否齐性。在 *Levene* 检验中，原假设是方差相等，而备择假设是方差不等。首先看 *Levene's Test* 的 Sig.值，若 Sig.小于 0.05，说明方差不齐，应看 Equal variances not assumed 行的 Sig. (2-tailed)值，若 Sig. (2-tailed)值小于 0.05，则拒绝 *T* 检验的原假设，即财务特征指标在舞弊与非舞弊公司的均值存在显著性差异。若 Sig.大于等于 0.05，则说明方差齐性，应看 Equal variances assumed 行的 Sig. (2-tailed)值，若 Sig. (2-tailed)值小于 0.05，则拒绝 *T* 检验的原假设，即财务特征指标在舞弊与非舞弊公司的均值存在显著性差异。同理，在 *T* 检验中，若 Sig. (2-tailed)值大于等于 0.05，则不拒绝 *T* 检验的原假设，即财务特征指标在舞弊与非舞弊公司的均值无显著性差异。

由独立样本 *T* 检验结果可以看出，所有指标都通过了上述 *Levene* 检验和 *T* 检验的筛选，但并非所有指标都可纳入财务特征指标的体系中。根据存货周转率的 Mean Difference，即舞弊公司与非舞弊公司的均值差，舞弊公司异常高于非舞弊公司(约 226.8)，说明数据的单位可能存在问题，或者有异常值没有及时处理，为了简化问题的分析，我们将此指标删除，不作为本文的财务特征指标。根据应收账款周转率的 Mean Difference，舞弊公司比非舞弊公司高约 9.54，情况复杂，因为应收账款周转率越高，说明这个公司的应收账款回收速度快，这对企业的财务状况和运营效率有着积极影响，而舞弊公司的应收账款周转率高，只能说明舞弊公司利用该指标当作一个虚假的繁荣信号，希望在审计阶段蒙混过关，因此我们需要将这个与舞弊辨认预期情况不符的指标删除。

最终，留下本文研究的六个财务特征指标：资产负债率、流动比率、净资产收益率(ROE)、净利率、

经营净现金流/营业总收入、应收账款/销售收入(见表 1)。

Table 1. Final financial feature indicators table
表 1. 最终财务特征指标表

最终财务特征指标	
资产负债率	净利率
流动比率	经营净现金流/营业总收入
净资产收益率(ROE)	应收账款/销售收入

2.4. 数据标准化

- 本文虽选取的财务特征指标都为比率的形式，但不同指标的比率取值差异较大，因此可将指标转换到相同的尺度，避免因量纲和量级差异导致的模型偏差。
- 标准化后的数据可以使模型更好地适应不同类型和规模的企业数据，提高模型的泛化能力，从而更准确地识别财务舞弊行为。
- 通过标准化，数据的分布更加均匀，可以减少模型训练过程中的资源消耗。
- 财务数据中可能存在大量的异常值和噪声，而标准化可以减少数据中的异常值和噪声对模型的影响，从而降低模型的过拟合风险。

基于以上数据标准化的优势，我们决定将包含六个特征值的财务数据进行 *Z-Score* 标准化，得到一份均值为 0，方差为 1 的标准正态分布数据，公式如下：

$$Z = \frac{X - \mu}{\delta} \tag{1}$$

其中，X 表示原始数据点，μ 表示特征数据的均值，δ 表示特征数据的标准差，X 为原始数据点。

结合 SPSS 统计软件的使用，我们最终得到一份经过 Z-Score 标准化后的数据。

3. 模型建立与评估

3.1. 随机森林模型的建立与评估

首先，我们从 Python 的机器学习库的 model_selection 模块中导入 train_test_split 函数，用于将数据集随机分割为训练集和测试集；接着，从该库的 ensemble 模块中导入 RandomForestClassifier 类，用于构建和训练随机森林模型；同时，导入 accuracy_score、precision_score、confusion_matrix 等，提供分类模型性能的评估指标；并导入 SMOTE 类，调整分类样本不平衡的问题。

导入部分结束后，先让 Python 读取财务数据集，包含 1988 条非舞弊数据与 1416 条舞弊数据。并删除“证券代码”、“证券简称”和“日期”这三个非特征列，将“资产负债率”等六个财务特征指标设置为一个特征矩阵，同时将“违规标志”(包含 0 和 1 两类)设置为目标变量。接着，划分训练集和测试集，将数据集的 20%划定为测试集，其余 80%划定为训练集，令 stratify 与目标变量相等，确保训练集和测试集中各类别的比例与原始数据集中的比例相同，并将随机种子设为 42，保证实验的可重复性。下一步，建立随机森林模型并进行训练，先指定森林中决策树的数量为 100 棵，树的数量过多虽会提升模型准确率，但会增加计算成本，而“100”能在准确率和效率之间取得平衡；并控制每棵决策树的最大深度为 5 层，防止模型过拟合；同时固定随机种子，保证结果可复现。最后，运用混淆矩阵评估模型，输出准确率、精确率等组成的模型评估报告。

在随机森林中，每棵决策树通过不纯度下降选择最优分裂点，通常用信息增益作为其核心指标，信息熵公式如下：

$$Entropy(D) = -\sum_{t=1}^{T} p_t \log_2 p_t \tag{2}$$

其中，D 为当前节点的数据集，p_t 为第 t 类样本在 D 中的占比。

信息增益公式如下：

$$IG = Entropy(D) - \sum_{k=1}^{K} \frac{|D_k|}{|D|} Entropy(D_k) \tag{3}$$

其中，D_k 代表分裂后子节点的数据集，k 表示分裂后生成的子节点的索引。

构建决策树的算法会自动选择信息增益最大的分裂方式，以构建高效且准确的决策树模型。

同时，随机森林通过 Bagging 抽样生成每棵树的训练子集，某个样本在自助采样过程中被抽中的概率如下：

$$p = 1 - \left(1 - \frac{1}{N}\right)^N \tag{4}$$

其中，原始财务数据集的数据量大小为 N。

训练部分结束，最终评估训练好的模型在测试集上的表现，得到模型的预测结果和预测概率，结果如下：

详细分类报告：

	precision	recall	f1-score	support
0	0.73	0.81	0.77	398
1	0.68	0.58	0.62	283
accuracy			0.71	681
macro avg	0.70	0.69	0.69	681
weighted avg	0.71	0.71	0.71	681

Figure 3. Classification evaluation metrics for the random forest model
图 3. 随机森林模型分类评估报告

根据该分类评估报告(见图 3)，可知模型的整体准确率为 0.71，准确性能较好；由非舞弊类样本的精确率为 0.73，舞弊类样本的精确率为 0.68，可知该模型对舞弊与非舞弊的预测精度较高；同时，非舞弊类样本的召回率为 0.81，说明模型对非舞弊企业的识别程度较高，但舞弊类企业的召回率只有 0.58，说明模型对舞弊企业的识别度有待提高；根据两类样本的 f1 分数都高于 0.6，可知模型对两类样本的判别能力较可靠，但仍有提升空间(财务舞弊识别混淆矩阵见图 4)。

根据当前随机森林模型 AUC 的值为 0.73 (见图 5)，说明该模型可作为初步判断财务样本为舞弊或非舞弊的工具，应用于审计的初期，一定程度上可达到缓解审计人员工作量的效果。

为了直观看出本文所选六大指标对财务舞弊识别模型的重要性程度，我们基于随机森林模型输出指标的重要性得分，见图 6：

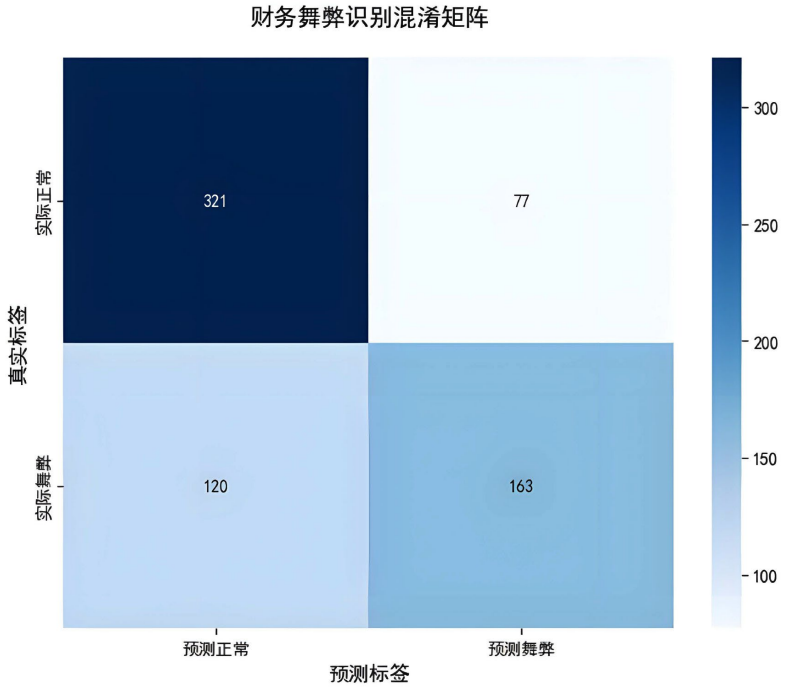

Figure 4. Confusion matrix heatmap for random forest model
图 4. 随机森林模型混淆矩阵热力图

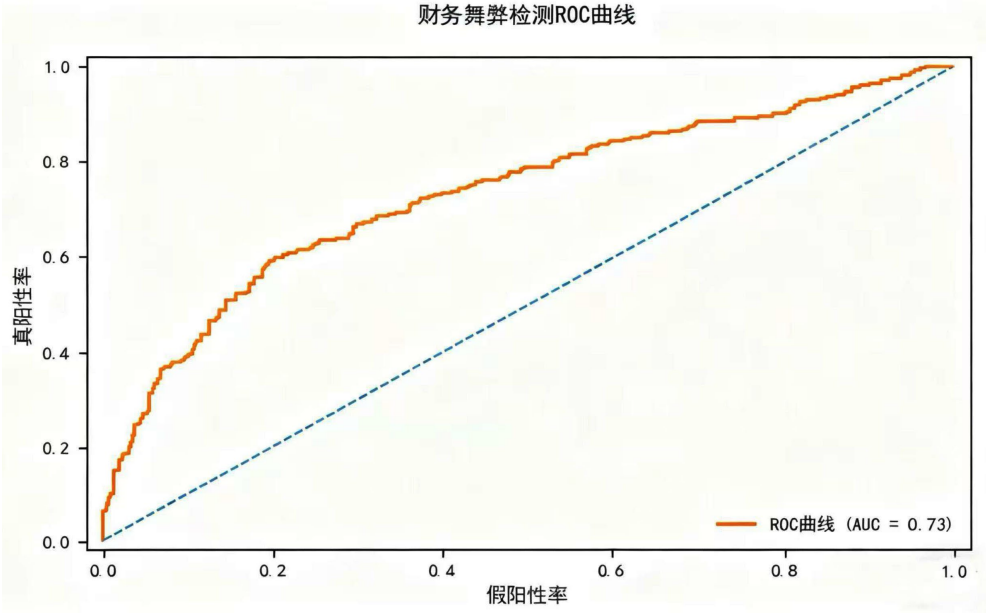

Figure 5. ROC curve plot for random forest model
图 5. 随机森林模型 ROC 曲线图

由财务特征指标重要性排序图可知(见图 7)，“净利率”、“净资产收益率”以及“应收账款/销售收入”三大指标对财务舞弊识别模型的影响最大，说明这三者的分类判断能力最强；而后三者的分类判断能力较弱。因此在今后财务舞弊识别模型的发展过程中，可以将前三者视为对该模型重要性最强、判断力最好的三个指标。

特征重要性：
净利率 0.2742
净资产收益率（ROE） 0.2514
应收账款/销售收入 0.2330
经营净现金流/营业总收入 0.0875
流动比率 0.0825
资产负债率 0.0714

Figure 6. Feature importance score
图 6. 特征重要性得分

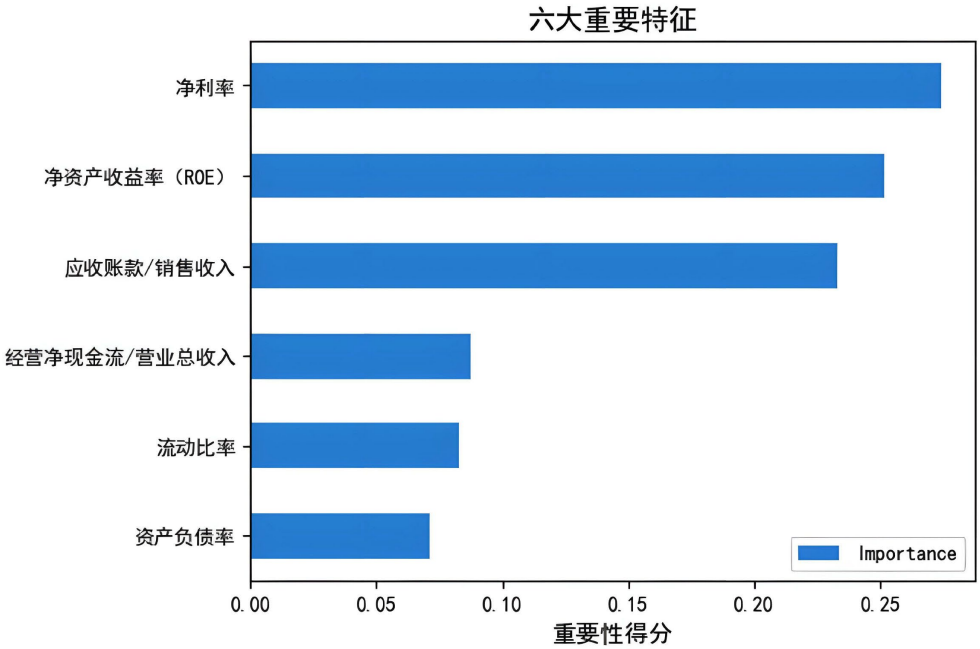

Figure 7. Importance ranking diagram of financial characteristic indicators
图 7. 财务特征指标重要性排序图

3.2. XGBoost 模型的建立与评估

首先，我们从 sklearn.model_selection 模块中导入 train_test_split，用于将样本数据集划分为训练集和测试集，并从 sklearn.metrics 模块中导入 confusion_matrix、classification_report 等多个用于评估的函数和类；接着，从 XGBoost 库中导入 XGBClassifier 类，用于创建 XGBoost 模型。

导入部分结束后，先加载财务数据集，将本文研究的资产负债率等六大财务特征指标设置为一个特征矩阵，同时将违规标志设置为目标变量。接着，对数据进行清洗，将无穷大值转换为缺失值，并过滤部分缺失值，保留 80%完整性样本，再用 Z-Score 异常值过滤的 3σ 原则去掉部分异常值，防止极端值的干扰。在模型建立前本文已经结合 SPSS 软件对数据进行预处理，但二次数据清洗可确保数据质量更加稳定，提升模型的鲁棒性。该阶段结束后，本文使用分层抽样方法确保训练集和测试集中各类别比例与原始数据集保持一致，划分原始数据集的 20%为测试集，其余 80%为训练集，以应对类别不平衡的问题。随后，创建 XGBoost 模型，并动态计算类别权重，处理样本中两类别比例不平衡的问题，提升模型对少数类的识别能力；将模型的迭代次数设置为 1000，即训练 1000 棵树，较大的迭代次数可提升模型的拟合

能力，同时设置初始的较小深度为 3，间接控制叶子节点数，以及更保守的学习率 0.5，有助于模型稳定，防止模型过拟合。在该阶段可体现该模型的目标函数，包括损失函数和正则化项两部分，如下：

$$Object = \sum_{i=1}^{N} L\left(y_i, \hat{y}_i\right) + \sum_{k=1}^{K} \mu\left(f_k\right) \tag{5}$$

该公式中，前一项为损失函数，后一项为正则化项。其中，f_k 表达式如下：

$$f_k = \alpha T + \frac{1}{2}\rho * \omega^2 \tag{6}$$

其中，T 为树的总叶子节点数，ω 表示叶子节点的权重，ρ 为叶子分裂阈值，ω 为正则化系数。

设置每棵树训练时使用的样本比例为 0.7，使用的特征比例同样为 0.7，通过随机选择部分样本和部分特征，可增加模型多样性，减少模型方差，防止模型过拟合，提高模型的泛化能力；设置随机种子数，保证模型可复现；使用直方图算法对模型训练过程进行加速，加速树的构建过程。

直方图算法的加速本质上是通过最大化分裂增益来选择最佳分裂点，公式如下：

$$G = \frac{G_l^2}{H_l + \gamma} + \frac{G_r^2}{H_r + \gamma} - \frac{\left(G_l + G_r\right)^2}{H_l + H_r + \gamma} - \varepsilon \tag{7}$$

其中，G_l、G_r 分别表示左右子节点的一阶导数和以及二阶导数和，γ 为正则化系数，防止模型过拟合，ε 为分裂阈值，当增益大于 0 且超过 ε 才可分裂，否则停止分裂。增益对模型的目标函数有着直接的影响，增益越大，分裂对模型的优化效果越显著，而通过最大化增益可确定最佳分裂点，但是增益过大可能造成模型过拟合，所以正则化系数等参数会对模型进行约束，防止模型过拟合。最终，通过增益与参数间的相互平衡，模型的性能得到优化。

构建完 XGBoost 模型后，本文通过带早停的交叉验证训练方法来训练模型，并在训练过程中监控模型在测试集与验证集上的性能表现。“早停”机制指如果测试集在若干轮内没有提升，模型训练将提前停止，可节省计算资源。

模型训练部分结束，最终基于混淆矩阵得到模型评估报告(见图 8)，结果如下：

优化版分类报告：

	precision	recall	f1-score	support
正常	0.70	0.77	0.73	243
违规	0.74	0.67	0.70	242
accuracy			0.72	485
macro avg	0.72	0.72	0.71	485
weighted avg	0.72	0.72	0.71	485

测试集 AUC: 0.7906

Figure 8. Classification evaluation metrics for the XGBoost model
图 8. XGBoost 模型分类评估报告

根据该分类评估报告，可知模型的整体准确率为 0.72，准确性能较强；由非舞弊样本的精确率为 0.70，舞弊样本的精确率为 0.74，可知该模型对舞弊类样本和非舞弊类样本的预测精度较高；同时，非舞弊类

样本的召回率为 0.77，说明模型对非舞弊类样本的识别能力较强，舞弊类样本的召回率为 0.67，相比于本文研究的前两个模型已有了显著提升，模型对舞弊样本的识别能力增强；由模型对两类样本的 f1 分数均在 0.70 及以上，可知模型对两类样本的判断能力较可靠。根据模型的 AUC 值为 0.7906，可知模型对两类样本的分类性能较强(该模型混淆矩阵见图 9)。

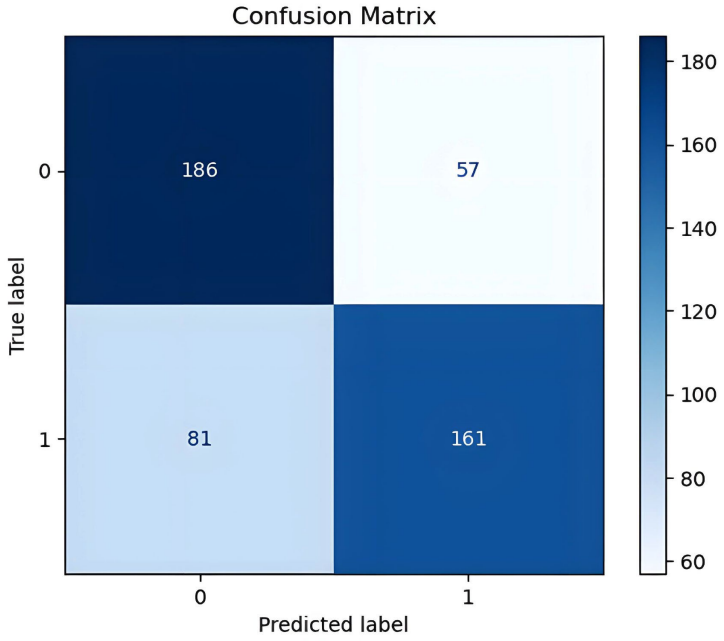

Figure 9. Confusion matrix heatmap for XGBoost model
图 9. XGBoost 模型混淆矩阵热力图

模型在不断迭代训练弱学习器时，测试集与训练集的 AUC 出现了如下变化。

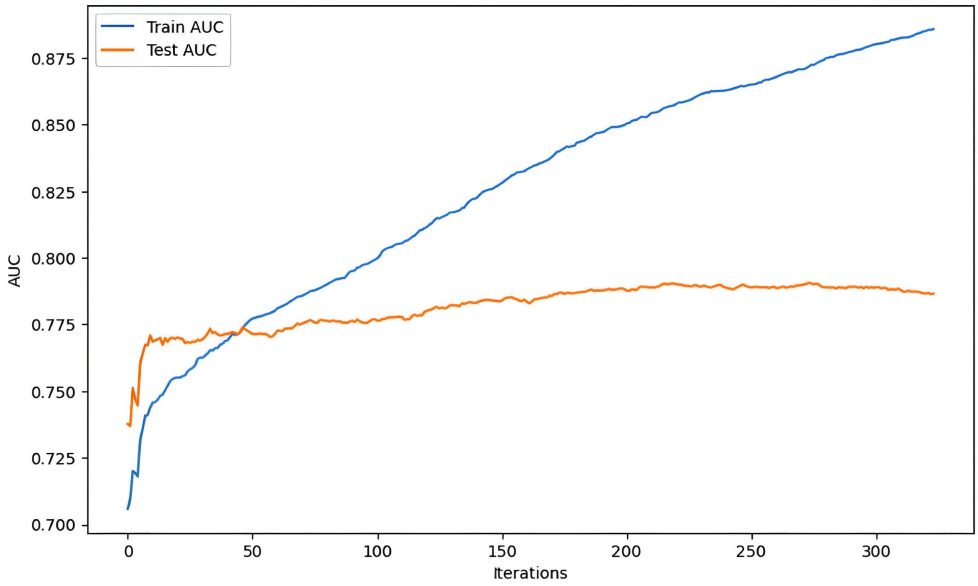

Figure 10. AUC evolution plot during model iteration
图 10. 模型迭代中的 AUC 变化图

模型在训练集上的区分能力随迭代次数的增加逐渐上升(见图 10)，这是由于模型不断增加决策树以优化损失函数；而模型在测试集上的区分能力则是随着迭代次数的增加先上升，随后进入平台期，这是由于模型不断学习训练集，造成过拟合现象，导致了泛化能力的停滞。

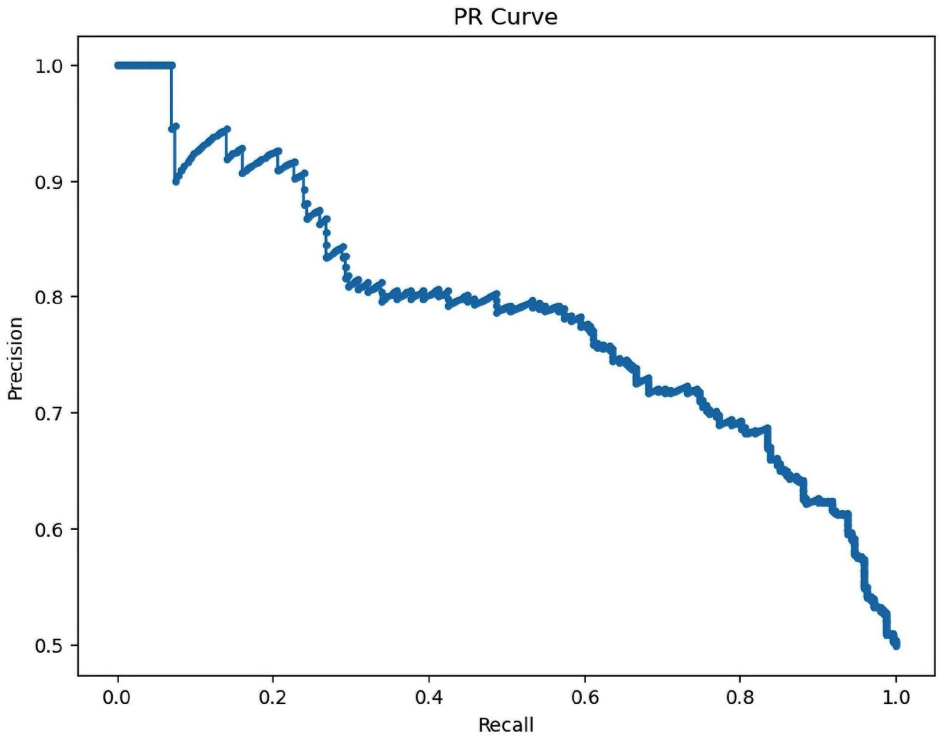

Figure 11. Recall-precision curve for fraud detection using XGBoost
图 11. XGBoost 舞弊类召回率与精确率关系曲线

在保证舞弊精确率的前提下提升舞弊召回率是本文研究的难点，为了找到该难点的突破口，可从这两种指标的关系入手(见图 11)。由该曲线可知，存在舞弊精确率和召回率均在 0.7 及以上的点，这些点可视为舞弊识别的“平衡点”，能较精确地识别出舞弊企业，并对舞弊企业的识别程度较强。因此，如何优化模型使得模型舞弊识别达到“平衡点”为后续研究的下一步目标，优化后的 XGBoost 模型可为金融机构提供更好的财务舞弊识别技术。

4. 结论与展望

4.1. 研究结论

本文基于机器学习分别构建随机森林模型和 XGBoost 模型作为财务舞弊识别模型，打破了传统意义上完全依赖人工经验的财务舞弊识别方式。结合机器学习模型的多方面技术优势，如捕捉非线性关系能力等能力，本研究将其优势与财务舞弊现象相结合，并使用 Python 建立财务舞弊识别模型，三种机器学习模型的整体准确率均达到 70%。该模型可有效应用于财务审计工作的初步筛查阶段，不但能减轻审计人员的工作量，还能为审计智能化转型提供新的思路和方向。

在研究初始阶段，本文从 CSMAR 数据库收集了七年的财务报表主表数据以及舞弊公司相关数据，同时将数据集合并，结合舞弊相关理论以及 SPSS 软件的独立样本 T 检验方法确定了本研究采用的六大财务特征指标。

在随机森林模型的构建阶段，本文运用 Python 完成了模型的建立，并引入 SMOTE 采样技术，该技术可解决舞弊样本与非舞弊样本的数据量不平衡问题，打破了传统采样技术如随机过采样等造成的模型过拟合问题，增强了模型的泛化能力。

在 XGBoost 模型的构建阶段，本文对财务数据集进行了第二次数据预处理，并用 Z-Score 异常值过滤的 3σ 原则去掉部分异常值，防止极端值的干扰。同时设置模型迭代次数等参数，使用带早停的交叉验证方法来训练模型，最后达到提升模型泛化性能与准确性的目的。

在两类模型的评估阶段，我们使用混淆矩阵评估方法对三类模型分别进行评估，可知三类模型的精确率均在 0.7 附近，说明模型对舞弊与非舞弊企业的预测精度较高；XGBoost 模型的舞弊召回率显著高于前两类模型，说明该模型的舞弊识别能力相对更好。

4.2. 研究局限性与未来展望

本研究建立的前两类机器学习模型只适用于财务审计的初期筛选阶段，由于两模型对与舞弊企业样本的召回率均在 0.5~0.6 区间内，说明模型对舞弊企业的识别度还有待提高，无法完全脱离审计人员而独立识别包含舞弊样本的财务数据集。而 XGBoost 模型的舞弊召回率虽接近 0.7，但还存在提升空间。同时，本研究仅考虑了部分财务指标，并且没有引入其他维度的信息，例如股东信息等，而这些维度的数据也可能对财务舞弊造成影响，缺乏一定的客观性。基于此，应引入多模态数据融合的财务舞弊识别模型，其为未来发展的突破方向，本质在于整合财务数据与其他非结构化的多源信息，如图像、文本等，并用自然语言处理技术等对不同类型的数据进行处理，同时需确定采取何种融合策略，将不同类型的数据结合起来，进一步提升财务舞弊识别模型的客观性。

参考文献

[1] 朱敏. 上市公司财务报告舞弊的识别方法及模型研究[D]: [硕士学位论文]. 成都: 四川大学, 2005.

[2] 刘君, 王理平. 基于概率神经网络的财务舞弊识别模型[J]. 哈尔滨商业大学学报(社会科学版), 2006(3): 102-105.

[3] 张新民, 吴革. 财务报告舞弊的特征与识别模型研究[J]. 财贸经济, 2008(12): 36-40.

[4] 贺颖. 基于偏最小二乘法-支持向量机的上市公司财务舞弊识别模型研究[D]: [硕士学位论文]. 石河子: 石河子大学, 2010.

[5] 陈彬. 我国上市公司财务舞弊识别模型对比研究[D]: [硕士学位论文]. 西安: 西北大学, 2012.

[6] 成雪娇. 基于数据挖掘的中国上市公司财务舞弊识别研究[D]: [硕士学位论文]. 重庆: 重庆理工大学, 2018.

[7] 王柳匀. 基于分类算法的财务报表舞弊识别研究[D]: [硕士学位论文]. 北京: 中国财政科学研究院, 2021.

[8] 程新尧. 基于信息融合的财务舞弊识别研究[D]: [硕士学位论文]. 重庆: 重庆理工大学, 2022.

[9] 赵豪杰. 基于 CNN-BiLSTM 模型的上市公司财务舞弊识别研究[D]: [硕士学位论文]. 苏州: 苏州大学, 2023.

[10] 何子英. 基于集成学习与 SHAP 的财务舞弊识别研究[D]: [硕士学位论文]. 南昌: 江西财经大学, 2024.