电 子 科 技 大 学

UNIVERSITY OF ELECTRONIC SCIENCE AND TECHNOLOGY OF CHINA

硕士学位论文

MASTER THESIS

论文题目 **基于机器学习的财务舞弊识别方法研究**

学科专业	金融工程
学　　号	**201711110110**
作者姓名	翟　聪
指导教师	刘　波　教授
学　　院	经济与管理学院

分类号 F830 密级 公开

UDC [注1] 336

学 位 论 文

基于机器学习的财务舞弊识别方法研究

（题名和副题名）

翟聪

（作者姓名）

指导教师 刘波 教 授

电子科技大学 成 都

（姓名、职称、单位名称）

申请学位级别 硕士 学科专业 金融工程

提交论文日期 2023 年 4 月 20 日 论文答辩日期 2023 年 5 月 26 日

学位授予单位和日期 电子科技大学 2023 年 6 月

答辩委员会主席

评阅人

注 1：注明《国际十进分类法 UDC》的类号。

Research on Financial Fraud Identification Method Based on Machine Learning

A Master Thesis Submitted to
University of Electronic Science and Technology of China

Discipline **Financial Engineering**

Student ID **201711110110**

Author **Zhai Cong**

Supervisor **Prof. Liu Bo**

School **School of Management and Economics**

摘　要

近年来，机器学习特别是集成学习在财务领域得到了广泛应用。Bagging（装袋法）和 Boosting（提升法）作为集成学习的两种重要代表，虽然其应用场景获得较大拓展，但在不同场景下的预测和检验表现却存在较大差异，以致对于两种方法孰优孰劣的评判并不一致。财务舞弊的识别是资本市场关注的重大应用场景，集成学习对财务舞弊的识别效果如何、特别是 Bagging 和 Boosting 两种方法的具体表现，引起了学术界的广泛关注。本文采用 2013-2018 年深圳证券交易所 A 股上市公司的相关数据，系统性地对比分析了基于 Bagging 和基于 Boosting 策略的共 5 种集成学习模型在财务舞弊识别上的实际表现。

在集成学习众多模型中，基于 Bagging 策略的随机森林模型，表现优于基于 Boosting 策略的 GBC，Adaboost，Xgboost，LightGBM。具体表现在 NDCG@k，Precision@k，Recall@k 等排序指标上。这一结果可以从财务舞弊场景的数据特征以及 Bagging 策略的设计特点来解释。财务舞弊场景的数据规模较小，特征维度较高，天然的适合使用随机森林这种基于 Bagging 策略的集成学习来处理。且 Bagging 策略充分发挥“三个臭皮匠顶一个诸葛亮”的朴素的集成思想，模拟了将大量市场参与者基于不同信息集做出的判断进行归纳整合的过程。

其次，本文通过基于特征重要度的模型解释性分析，发现虽然不同集成学习关注的重要特征排序存在一定差异，部分重要的特征始终会引起各个模型的注意，其中盈利能力与资产质量类的特征，对于识别财务舞弊行为更加重要，其中当存在收入大幅下降、连续亏损等表现时，通常预示着舞弊概率的进一步提升。

本文揭示了集成学习策略在财务舞弊识别中发挥作用的机理，拓展了集成学习在财务领域的相关研究。

关键词：机器学习，财务舞弊，集成学习，装袋法，提升法

ABSTRACT

In recent years, the machine learning, especially ensemble learning, has been widely used in the financial field. Bagging (bagging method) and Boosting (boosting method) are two important representatives of ensemble learning. Although their application scenarios have been greatly expanded, there are large differences in prediction and inspection performance in different scenarios, so that the judgment of which method is better or worse between the two is not consistent. The identification of financial fraud is a major application scenario that the capital market pays attention to. The effect of ensemble learning on the identification of financial fraud, especially the specific performance of the two methods of Bagging and Boosting, has attracted widespread attention in the academic community. This thesis uses the relevant data of A-share listed companies on the Shenzhen Stock Exchange from 2013 to 2018 to systematically compare and analyze the actual performance of five ensemble learning models based on Bagging and Boosting strategies for identifying financial fraud.

Among the many models of ensemble learning, the random forest model based on the Bagging strategy outperforms the GBC, Adaboost, Xgboost, and LightGBM models based on the Boosting strategy. It is specifically manifested in the ranking indicators such as NDCG@k, Precision@k, and Recall@k. This result can be explained from the data characteristics of the financial fraud scene and the design characteristics of the Bagging algorithm. The data scale of the financial fraud scene is small and the feature dimension is high, which is naturally suitable for the ensemble learning based on the Bagging strategy of the random forest. And the Bagging strategy gives full play to the simple integration idea of “three cobblers are worth one Zhuge Liang”, simulating the process of inductively integrating the judgments made by a large number of market participants based on different information sets.

Secondly, through the model explanatory analysis based on the importance of features, this paper finds that although there are certain differences in the ranking of important features that different ensemble learning focuses on, some important features will always attract the attention of each model. Among them, the features of profitability and asset quality are more important to identify financial fraud. When there are sharp drops in income, continuous losses, etc., it usually indicates a further increase in the

probability of fraud. This thesis reveals the mechanism of ensemble learning strategy in the identification of financial fraud, and expands the related research on ensemble learning in the financial field.

Keywords: Machine Learning, Financial Statement Fraud, Ensemble learning, Bagging, Boosting

目 录

表目录

第一章 绪论

1.1 研究背景与意义

财务舞弊行为（Financial Statement Fraud，或 Account Fraud）是指“上市公司所犯的，通过使用重大误导性财务报表，造成投资者和债权人遭受损害和伤害的一种故意的、非法的行为”（Rezaee, 2005）[1]。在中国 A 股上市公司中，康得新、康美药业、獐子岛等公司的财务舞弊等案例已经引起市场投资者和监管层的高度关注。短期来看，财务舞弊公告后通常伴随着股价下跌，舞弊行为公告之后的 3 天 CARs 显著为负（Xu et al., 2022）[2]，同时高管和董事会也会名誉受损（Fich and Shivdasani, 2007）[3]；从长期来看，如果不能及时发现财务舞弊，最终可能产生更严重的退市和破产现象，影响整个市场的稳定运行和健康发展。Beasley et al.（2010）[4]发现在首次被报道进行了财务舞弊的公司事后股价平均下跌 16.7%，并且有 47%的财务造假公司最终退市。Abbasi（2012）[5]发现美国历史上十大破产案中有四起也与重大财务欺诈相关。在此背景下，研究寻找更准确的财务舞弊识别模型并应用于监管，已经成为监管者、公司管理层、审计师和投资者共同关注的重要问题（Ugrin and Odom, 2010）[6]。

机器学习的核心优势之一是利用大量的历史数据挖掘数据规律，寻找数据中的复杂结构和模式来辅助预测（Hastie et al., 2009[7]；Zhou, 2021[8]）。现有基于机器学习的研究，分别尝试了基于逻辑回归（Persons, 1995[9]; Beneish, 1999[10] ;Dechow, 2011[11];Perols, 2011[12]）、神经网络（Green and Choi, 1997[13]；Fanning and Cogger, 1998[14], Kirkos et al., 2007[15]）、决策树（Kirkos et al., 2007[15]）、支持向量机（Cecchini et al., 2010[16]；Perols, 2011[12]）等经典机器学习方法建立财务舞弊识别模型。作为机器学习的子领域之一，集成学习的优势已在实验和理论上分别得到了证明。Hansen and Salamon（1990）[17]通过实验证明了一组分类器的集成可产出比其中最优个体分类器更精准的预测。Dietterich（2000）[18]从理论上解释了集成学习方法成功的 3 个基本原因: 统计、计算和代表性。目前 Bagging 和 Boosting 是两种主流的集成学习策略。Schapire（1990）[19]证明了通过 Boosting 方法可以将弱分类器组合成一个强分类器，并进一步在 Freund and Schapire（1995）[20]中提出了基于 Boosting 策略的 Adaboost 算法。Breiman（1996）[21]首次提出 Bagging 策略，其后，Breiman（2001）[22]在 Bagging 基础之上，通过引入随机特征抽样的改进处理，提出了 Random Forest 随机森林算法。

Bagging 和 Boosting 作为集成学习的两种重要代表，虽然其应用场景获得较大拓展，但在不同场景下的预测和检验表现却存在较大差异，以致对于两种方法孰优孰劣的评判并不一致。比如，在计算机领域，Dietterich（2000）[23]通过实验数据发现当数据噪声较少时，Boosting 策略明显更有效。Banfield et al.（2007）[24]基于 57 个公开数据集发现 Bagging 策略整体效果更好；在医学领域，Fraz et al（2012）[25]针对视网膜血管识别问题、Lee（2010）[26]针对倾向性评分问题，发现 Boosting 效果更好。而 Wu et al.（2003）[27]针对卵巢癌分类问题，Khalilia et al.（2011）[28]针对疾病风险预测问题，则发现 Bagging 效果更好。在财务舞弊识别领域。Bao et al（2020）[29]引入了考虑样本不均衡特性的 RUSboost 进行财务舞弊识别分析，并发现 Boosting 集成策略效果显著超过 Cecchini（2010）[16]和 Dechow（2011）[11]模型；Wang et al.（2020）[30]和 Xu et al.（2022）[2]发现基于 Bagging 策略的随机森林表现超过其余模型。目前尚未见到有相关研究，基于财务舞弊场景，重点比较 Bagging 和 Boosting 两种集成策略的具体表现何种更优，以及结合场景数据分析如何产生最优的机理。对此问题进行研究有助于丰富集成学习策略在经济与管理领域的研究参考，并为集成学习策略在财务舞弊领域的应用提供普遍性的证据。

1.2 研究内容与框架

本文的研究目标为，结合财务舞弊识别这一重要场景，研究 Bagging 或 Boosting 这两种集成学习策略，哪一种能更有效地提升识别效果，并研究探索其发挥作用的统计和经济机理。围绕上述研究目标，本文采用实证分析的方式进行研究，首先从舞弊理论、舞弊特征、舞弊识别模型、以及包括集成学习在内的多种机器学习方法中总结现有成熟研究经验，并对应支撑到具体的实证分析设计环节。

在实证分析环节，首先，本文对财务舞弊数据及基于三表的基础财务数据情况进行了分析，确定了舞弊样本的定义与筛选方式，并根据可得的黑样本范围确定了对应的研究区间。其次，本文根据同行业对比、区间范围对比等思路，对原始的基础财务数据进行了特征衍生，从资产质量、盈利能力、现金流量、营运能力、行业环境、偿债能力六个方面选取并构建了 122 个特征变量，形成了特征指标库，在建模数据处理中，本文采用了针对黑白样本不平衡问题的重采样，并对行业内样本数量较少的行业进行了行业合并。在模型构建中，本文对比分析了基于 Bagging 策略的随机森林模型和基于 Boosting 策略的 GBC，Adaboost，xgboost，LightGBM 模型，分别在传统的评估指标和创新的评估指标上的具体表现，并总结了 Bagging 策略在财务舞弊场景中更加适用的经济逻辑。最后，本文采用基于特征重要度的模型

解释性分析，对不同集成学习关注的重要特征进行分析，进一步增强研究工作的现实意义。本文的主要研究框架参考图 1-1 所示。

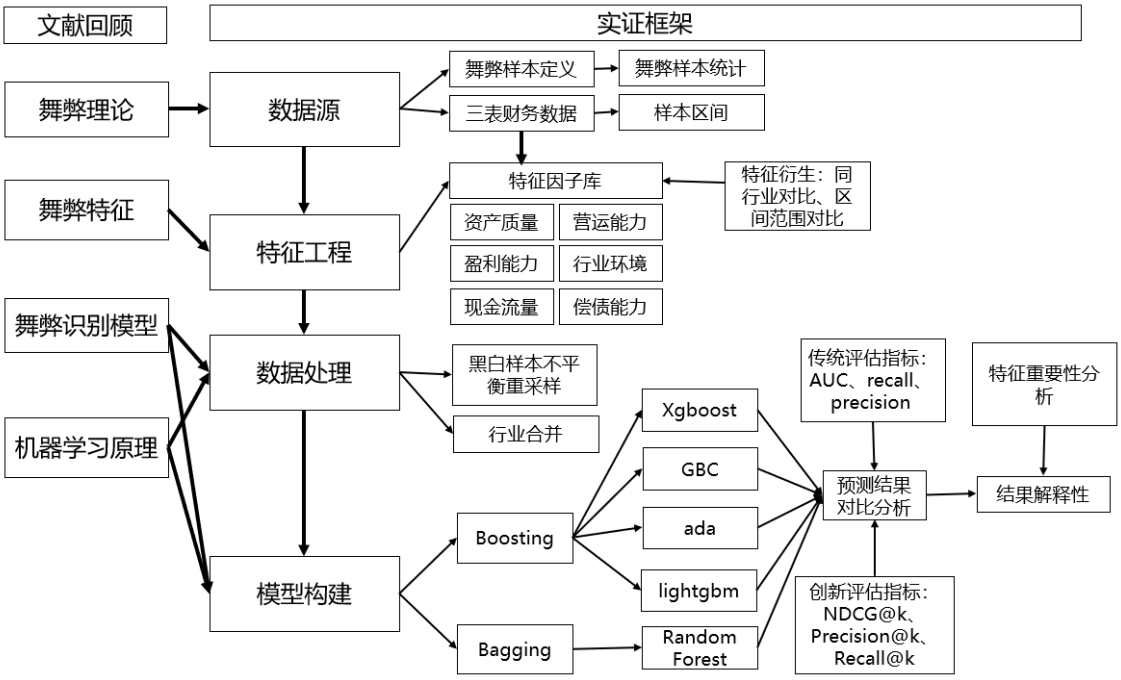

图 1-1 研究框架

1.3 主要创新点

本文在指标体系上结合实用性角度进行了创新，纳入了维度更多且符合业务理解和使用的指标。具体来说，本文结合了已有文献与实务经验，从资产质量、盈利能力、现金流量、营运能力、行业环境、偿债能力六个方面归纳提取了 122 个财务舞弊行为风险监测指标，深入刻画了影响财务舞弊行为的潜在方面，形成了较为全面丰富的财务舞弊特征体系。

本文在机器学习的算法选取视角和对比分析指标上进行创新。本文聚焦于现有 SOTA（State-of-the-Art）模型效果对比，针对性地引入 5 种集成学习模型，其中包括基于 Bagging 策略的随机森林，基于 Boosting 的 4 种模型：Adaboost、GBDT、Xgboost、LightGBM，分别构建了财务舞弊识别模型，并就实际效果进行了深入的对比分析。通过引入更加能反映排序预测质量的创新指标 NDCG@k、Precision@k，Recall@k，发现基于 Bagging 的随机森林模型，在测试集上的表现优于基于 Boosting 策略的其他集成学习模型。且这一发现与财务舞弊数据特点以及 Bagging 策略特点具有逻辑一致性。本文提供了对于 Bagging 集成策略适用场景的机理证据，丰富了集成学习策略在经济与管理领域的研究参考。虽然针对财务舞弊场景，但由于该

场景的本身数据特性具有普遍性，因此为相关结论应用于其他类似领域也提供了证据。

本文在对模型输出结果的解释性分析上进行创新，通过引入特征重要性的概念，进一步探索了哪些特征以及这些特征所属类别对识别财务舞弊更加重要，并发现盈利能力与资产质量在识别财务舞弊中发挥了更重要的作用。本文引入的模型的特征重要性分析，进一步打开了机器学习识别过程的黑箱，将机器学习的输出线索与现实资本市场监管行为联系起来，增强了研究的实际意义。

1.4 结构安排

本文接下来的结构安排如下：第二部分 回顾已有研究；第三部分介绍研究设计；第四部分介绍实证分析结果，最后第五部分对全文进行总结。

第一章绪论。阐述本文研究工作的背景与意义，提出本文的主要贡献与创新，明确本论文的研究框架。

第二章文献回顾。从财务舞弊识别问题的相关理论、财务舞弊识别所需的基本特征、财务舞弊识别的主要模型、基于机器学习的财务舞弊识别模型研究四个方面对已有文献进行回顾与总结，并针对现有研究中存在的短板提出针对性的研究问题和解决方案。

第三章研究设计。本章旨在详细解释和说明本文的实证分析步骤。首先本文对数据来源和数据处理方式进行说明；然后详细定义了舞弊样本的选择方式，并根据资产质量、盈利能力、现金流量、营运能力、行业环境、偿债能力六个方面选取并构建了 122 个特征变量，进行了详细说明。同时，从机器学习角度说明了集成模型的选择过程以及用于模型评估的指标含义。

第四章实证分析。本章首先对样本进行描述性统计分析，并对舞弊样本与非舞弊样本的特征在均值上是否存在显著差异进行了检验。接着对 Bagging 和 Boosting 这两种集成学习策略的表现进行了对比，最后本文基于随机森林这一算法，进一步探索了不同特征和特征类别对识别财务舞弊的贡献程度。

第五章全文总结与展望。本章总结得出：在集成学习众多模型中，基于 Bagging 策略的随机森林模型，表现优于基于 Boosting 策略的 GBC，Adaboost，xgboost，LightGBM。具体表现在 NDCG@k，Precision@k，recall@k 等排序指标上。其次，本文通过基于特征重要度的模型解释性分析，发现虽然不同集成学习关注的重要特征排序存在一定差异，部分重要的特征始终会引起各个模型的注意，其中盈利能力与资产质量类的特征，对于识别财务舞弊行为更加重要，其中当存在收入大幅下降、连续亏损等表现时，通常预示着舞弊概率的进一步提升。最后展望后续工作。

第二章 文献回顾

对于财务舞弊识别这一重要问题，国内外学者分别从理论与实证两个方向进行了大量研究。本章对此进行总结与归纳，以便为全文的研究场景与具体问题的提出做好铺垫。

2.1 财务舞弊相关理论

目前，国内外针对财务舞弊现象已提出的理论解释主要包括：舞弊三角理论、舞弊四因素理论 (即 GONE 理论)等。

舞弊三角理论是 Albrecht（1995）[31]提出的一个著名理论，它用来解释财务舞弊犯罪的发生机制。该理论指出，只有当舞弊主体同时具备动机/压力、机会和合理化借口这三个基本条件时，才会实施舞弊行为。具体而言，动机（Incentive/Pressure）是指舞弊主体为了实现某种目标或面临某种困境而产生的欺诈动机或压力。例如，舞弊者可能想要获取经济利益，如奖金、股票期权、提升公司股价等；机会（Opportunity）是指舞弊主体能够利用其职务或职位上的权力、机会和资源来制造虚假的财务信息。例如，舞弊者可能通过篡改公司财务报表、虚构收入、虚增支出、隐瞒负债等手段来进行欺诈；合理化（Rationalization）是指舞弊主体对自己的欺诈行为进行心理辩解或道德合理化。例如，舞弊者可能告诉自己他们是为了公司或个人的利益，或者他们认为这是一种“正当”的行为。这样，舞弊者就可以减轻自己的罪恶感和焦虑感。对于管理层和审计人员来说，在理解和防范财务犯罪时需要注意这三个方面，并制定有效的控制措施来降低欺诈风险。

虽然舞弊三角理论是财务欺诈领域广泛接受的概念，但它存在一些问题。首先，这个理论过于简化，实际的欺诈行为可能由更多因素导致。其次，确定欺诈者的动机和合理化是困难的。最后，舞弊三角理论缺乏充分的实证证据来支持其有效性。因此，需要更多研究来改进和完善这个理论：Bologua，Lindquist and Wells（1993）[32]提出的 GONE 理论，对舞弊三角理论做了进一步的完善，丰富了对舞弊发生原因的解释。该理论认为财务报告舞弊由四个因素构成，分别是贪婪（Greed）、机会（Opportunity）、需要（Need）和暴露（Exposure）。与舞弊三角理论相比，GONE 理论将三角理论中的动机分为需要和贪婪两个方面，将合理化借口视为与个人心理相关的贪婪因素，并增加了暴露这一外部环境因素。这四个因素中，需要和贪婪是个别风险因素，它们从组织和个人角度说明了利益需求期望差引起的舞弊压力（或动机），是舞弊行为发生的内在原因; 机会和暴露是一般风险因素，它

们从内部环境和外部环境角度说明了制度机制缺失造成的舞弊机会，是舞弊行为发生的外在条件。这四个因素相互作用，相互影响，形成了导致财务报告舞弊行为发生的关键因素。在对舞弊三角理论和 GONE 理论进行继承和对比的基础上，Bologana and Lindquist(1995)[33]提出了一个更为完善和系统的舞弊风险因子理论。该理论认为，影响舞弊行为发生的风险因子可以分为两类：一类是组织或机构能够通过内部控制或管理措施来调节或减少的一般风险因子，包括提供舞弊机会的内部环境、影响舞弊被发现可能性的外部环境以及决定舞弊后果严重程度的处罚力度；另一类是组织或机构难以直接干预或改变的个别风险因子，主要涉及当事者本身具有的道德水平和动机水平。当这些风险因子共同作用，并且当事者在权衡利弊后认为舞弊收益大于代价时，就会产生并实施舞弊行为。

国内的研究主要采用实证分析的方法来对上述财务舞弊经典理论在中国市场进行验证。在对舞弊三角理论的验证研究中，韦琳,徐立文,刘佳（2011）[34] 选取了2000—2009 年发生财务舞弊的 A 股上市公司及其对应的非舞弊公司作为样本，基于舞弊三角理论，对 25 个反映压力、机会和理性化这三个因素的指标进行了检验。发现这些指标与舞弊可能性呈正相关，他们建立的模型能够以 93.7%的准确率识别出舞弊公司；而在对 GONE 理论的研究中，洪荭,胡华夏,郭春飞（2012）[35] 选取了 2006～2009 年因财务舞弊被处罚的上市公司作为样本，实证检验了四个方面与财务舞弊之间的关系，分析了这四个方面对财务舞弊动机、条件、需求和风险的影响机制，并从企业治理和内部控制等方面提出了防范措施。

财务舞弊相关的理论文献为本文的研究提供了重要的理论支持，帮助本文更好地理解和认识舞弊行为发生的场景，为后续业务解读和解释奠定了坚实的基础。

2.2 财务舞弊基本特征研究

在理论指导和业务实践总结之后，国内外学者开始对上市公司的舞弊模式特征进行分析和研究，研究方法上多采用规范研究、案例研究等方法，实证研究相对较少。

国外学者将财务报告舞弊特征称为“红旗（Red Flag）”，并提出了多套符合业务逻辑的红旗模型。例如在早期的研究中，Albrecht and Romney(1986)[36]首次以问卷调查的方式，证实了 86 个“红旗”可作为公司财报舞弊的征兆。而根据 Albrecht, Wernz and Williams（1995）[31]的进一步研究，通过对财务报告的分析，可以发现一些用于解释欺诈迹象，形成了新的“红旗”指标。这些迹象包括财务报告中无法解释的变化、经营危机、迫切需要报告有利收益、异常大的交易、收益质量下降、高负债或其他利益负担以及现金流问题。此外，还包括费用增长快于收入，依赖单

一产品或大型诉讼，频繁更换外部审计师或管理层，关联交易以及与客户或供应商的异常关系。Beneish (1997)[37]则提供了另外一些重要的“红旗指标”，他们通过比较舞弊公司和未舞弊公司，发现一些初步判断财报舞弊风险的因素。这些因素包括公司历史、财务杠杆程度、增长速度以及股价表现。此外，如果一个公司出现应收款项大幅增加、产品毛利率异常变动、资产质量下降、销售收入异常增加或应计利润率上升等情况，则也可能意味着该公司存在舞弊行为。Lee, Ingram and Howard (1999)[39]的研究则进一步揭示了应计部分（即盈余减去经营活动现金流量之差）在财报舞弊中的重要作用。他们发现，财报舞弊与高水平的应计部分密切相关。如果盈余减去现金流量的值为正，则可能是潜在舞弊的一个信号。此外，如果能够结合对存货、应收账款等项目的分析，则可以得到更好的效果。此外，他们还发现，与非舞弊公司相比，舞弊公司通常具有更低的自由现金流、更多的权益类证券发行、更高的财务杠杆、更多的应收账款余额和更高的销售增长率。同时，舞弊公司相对于其资产而言也具有更高的市场回报和市场价值，但其资产和销售绝对额通常较小。Summers and Sweeney (1998)[38]的研究是从内幕交易与财报舞弊之间的关联程度进行分析，同样可以得到类似“红旗”的一些指标。他们发现在舞弊公司中，通常在舞弊发生前一年就会出现存货周转率增高、销售增长加快和总资产报酬率提高等特征。以上这些在传统会计实证领域被提取和发现的“红旗”指标，为本文针对性地构建财务舞弊识别特征集，提供了重要参考。

国内研究则更加贴合 A 股市场的实际情况，同样发现了一些重要的财务舞弊特征。从资产质量和财务特征占总体的比例角度，阎达五和王建英（2001）[40]对可能存在利润操纵的上市公司进行了总体财务指标特征分析，他们发现可以通过分析应收账款周转率、毛利率、资产质量、销售额增长、折旧率、费用率、资产负债率以及应计项目占总资产比例等指标来判断上市公司是否存在财报舞弊；而从上市公司的盈利能力来看，陈信元等（2001）[41]发现如果一个公司在前两年连续亏损，并且当年的业绩没有得到显著改善（为避免被 ST 处理），或者一个公司在前两年的平均净资产报酬率达到 10%，可能意味着该公司存在财报舞弊。

其他一些国内研究则侧重于对单个案例或者财务舞弊类型特点的归纳。例如，章美珍(2002)[42]通过对银广厦舞弊案的详细案例分析，发现行业政策频繁变化和盈余减去经营活动所产生的现金流量的差值指标为负数都会进一步增加公司舞弊的可能性。财务舞弊的实际类型和对应的特征也可能存在一定差异，因此朱锦余和高善生（2007）[43]对 2002-2006 年证监会处罚公告中涉及的财务舞弊特征和舞弊类型进行了总结。他们发现，舞弊上市公司的舞弊类型包括虚假利润表、虚假披露、虚

构销售业务、虚增资产和隐瞒对外担保等。这些公司通常同时采用多种舞弊方法，并且舞弊行为持续时间长达两年以上，最长的甚至达到 9 年。

2.3 财务舞弊识别模型研究

财务舞弊识别问题的研究核心，是建立财务舞弊预测识别的模型。国内外已有一些研究人员致力于财务舞弊识别问题的研究，并取得了相关成果。

国外研究的一个重要方向，是根据传统统计理论和财务会计理论，建立并优化财务舞弊预测模型，侧重于单个模型的调优。在过去的十几年里，基于传统理论，最重要且具有影响力的模型是 Mscore 造假预测模型。这个模型是由 Beneish(1999)[10]提出的，它基于 1982-1992 年被美国证监会(SEC)查处的 74 家财务造假公司，并通过行业和年度配比了 2332 个控制样本来估计各个特征变量的系数。该模型利用应收账款指数、毛利率指数和资产质量指数等八项公司指标来判断造假的可能性。它曾在安然事件爆发之前成功地预测了安然公司造假而轰动一时。由于这个模型是完全利用财务数据建立，为我们提供了一个有效的工具来识别潜在的财务报表舞弊行为。Dechow (2011)[11]在 Mscore 基础上建立了 Fscore 财务造假预测模型， 他们收集了 1982-2005 年美国证监会（SEC）发布的 Accounting and Auditing Enforcement Releases（AAERs）中被判定为财务造假的 676 个公司样本。他们从应计项、财务指标、非财务指标、表外业务和市场信息五个方面全面检验了造假公司的特征指标。结果发现，只包括财务报表指标的模型在判断财务造假方面能力最强，准确度达到 69%。

除了 Mscore 和 Fscore 这类较为经典的模型，其他研究也尝试使用了多种常见的统计模型。Persons（1995）[9]采用逐步逻辑模型发现，财务杠杆、资本周转率、资产构成和公司规模是与欺诈性财务报告相关的重要因素，该研究还考虑了不同类型错误的相对成本水平，他们所得到的模型在所有类型 I 和类型 II 错误的相对成本水平上都优于将所有公司分类为非欺诈公司的简单策略。Green and Choi（1997）[13]则引入人工神经网络（ANN）技术构建了财报舞弊判别模型。他们在模型中使用了三种不同的计算期望的方式，通过对分离模型结果的组合，他们得到更优的模型，并建议审计师在审计初始阶段使用该模型。Eining et.al.（1997）[45]使用了 logit 模型，并发现当同时使用 logit 模型和专家辅助系统时，能够有效帮助审计人员识别财务舞弊问题。Beneish（1999）[10]使用了基于 8 个财务指标的 Probit 回归预测模型来预测财务舞弊，具体来说，他以 1987～1993 受美国证监会处罚的 74 家公司为黑样本，其他非舞弊的上市公司为正常样本，得到的预测准确率达到 75%。

国外研究的另一类范式侧重于对比多类识别模型的有效性。Kirkos et al.（2007）[15]以 76 家希腊制造业公司（包括欺诈和非欺诈公司）为样本，比较了决策树、神经网络和贝叶斯网络等数据挖掘分类技术在识别会计欺诈报告方面的有效性。结果显示，贝叶斯方法的判定率最高。在 10 折交叉验证程序中，贝叶斯信念网络模型正确分类了 90.3%的样本，取得了最佳性能。神经网络模型和决策树模型的准确率分别为 80%和 73.6%。贝叶斯信念网络揭示了欺诈行为与杠杆率、盈利能力、销售业绩、偿付能力和财务困境等方面紧密相关。Phua（2010）[47]比较并概述了 2000 年至 2010 年关于欺诈检测技术和应用的文章，他们研究的欺诈不仅包括财务欺诈，还包括信用卡欺诈。针对这类具有明确数据标签的分类问题，他们发现神经网络和支持向量机的应用范围和频率较高。

国内学者则基于统计理论，建立部分财务舞弊识别模型。例如，陈亮和王炫（2003）[48]以 1991 年至 2003 年间，41 家因操纵营业利润而被公开处罚的上市公司为样本，从经验分析的角度出发，运用单因素方差分析模型构建了一个针对营业利润操纵的识别模型，该模型对会计欺诈公司和正常公司的识别率分别达到了 80％和 93％；部分学者则对经典的 Mscore 和 Fscore 理论进行了验证，钱苹和罗玫 (2015)[49]以 1994－2011 年中国 A 股沪深两市财务造假上市公司为样本，检验了表征财务造假和盈余质量问题的特征指标，发现其他应收款、是否亏损、经营应计项、现金销售率、股票换手率波动率、股权集中度、机构投资者持股比率、是否再融资和股市周期是鉴别中国上市公司造假的关键变量，用这些指标建立的综合模型不仅具有简单易懂的实用性，而且在辨别国内造假公司的能力方面显著优于常用的 Mscore 和 Fscore 模型。

2.4 基于机器学习的财务舞弊识别模型研究

随着人工智能技术快速发展，以及大数据时代的到来，数据可得性不断提高，机器学习算法不断向各个研究领域渗透发展，Green and Choi (1997)[13]首次将数据挖掘与机器学习引入到财务舞弊研究中。目前机器学习算法在财务舞弊识别领域的应用研究思路可以归纳为以下几种：

(1) 对比分析不同的机器学习算法在财务舞弊识别领域的效果，以及是否能够取得超越传统统计模型的效果。根据机器学习理论中的“没有免费午餐定理” (No Free Lunch Theorem) (Wolpert, 1997)[50]，由于数据与目标变量等选择的差异，通常不能预知哪个算法会取得最好的预测效果，因此会采用对比分析的方式。例如 Albashrawi and Lowell (2016)[51]对 2006-2015 年的 40 篇财务舞弊识别相关模型进行了综述，发现从数据挖掘所使用的方法来说，逻辑回归、决策树、神经网络以及

贝叶斯网络已经被超过半数的研究使用。从时间趋势来看，2009 年-2012 年的相关研究成果占到近 10 年的 50%。Kotsiantis et al.(2007)[52] 认为具体哪种算法更优则尚无定论，结果通常取决于数据集、特征集以及算法的组合；

(2) 结合财务领域知识，深化单一机器学习算法的应用。Cecchini et al. (2010)[16]使用了 SVM 来进行财务数据的舞弊检测，其中重点是使用了隐含的基于财务指标的核，从而进行非线性的点映射。该研究正确标记了 80%的舞弊样本和 90.6%的非舞弊样本，验证了基于公开的定量财务特征，结合支持向量机和财务内核，能够有效区分舞弊与非舞弊公司。Perols (2017)[53]在 1998 到 2005 年区间上，使用 SVM 算法进行了研究，提出了三种数据下采样预处理的方法，即考虑从随机抽取白样本匹配全量黑样本、随机抽取特征指标、以及考虑舞弊类型进行特征指标抽取，经过数据预处理的模型识别舞弊的效率得到了有效提升。

近年来，集成学习已被应用于多个领域，例如计算机视觉 (Viola and Jones，2004)[54] ，恶意代码检测 (Kolter and Maloof， 2006) [55]信用卡欺诈检测 (Panigrahi et al., 2009)[56],破产预测 (West et al.,2005)[57]等。在 (2009-2011) 连续三年的 KDD Cup 竞赛中，获奖冠亚军都使用了集成学习 (Zhou, 2012)[58]。

集成学习成为一个主要的学习范式，得益于以下两项工作，Hansen and Salamon (1990) [17]通过实验证明了一组分类器的集成可产出比其中最优个体分类器更精准的预测。Schapire(1990)[19]从理论上证明了通过 Boosting 方法可以将弱分类器组合成一个强分类器。Dietterich(2000)[23]从理论上解释了集成方法成功的 3 个基本原因: 统计、计算和代表性。目前 Bagging 和 Boosting 是两种主流的集成学习策略。Freund and Schapire(1995)[20]中提出了基于 Boosting 策略的 Adaboost 算法。Breiman(1996)[21]首次提出 Bagging 策略。Breiman(2001)[22]在 bagging 基础之上，通过引入随机特征抽样的方式，进一步提出了 Random Forest 随机森林算法。

从 Bagging 和 Boosting 的特点来说，Dietterich(2000)[23]基于多个公共数据集对比发现，当数据噪音比例较少时，Boosting 效果更好，AdaBoost 能够得到较好的结果。当数据噪音较多时，反而 Bagging 效果更好，扛干扰能力强，此时 Adaboost 可能会发生过拟合。当考虑样本规模时，Skurichina et al(2002)[59]发现随着训练样本的增大，bagging 集成学习的多样性降低。在 boosting 中，当训练对象的数量增加时，集合中的分类器变得更加多样化。说明 Bagging 比 Boosting 更适合用于中小型数据集。

得益于集成学习的发展，已有研究将集成学习引入财务舞弊领域。在应用 Boosting 策略上，Bao et al.(2020)[29]以 1991-2008 年美国上市公司为样本，构建了一套基于机器学习集成学习的财务舞弊识别模型。该文针对不平衡样本问题引入

了 adaboost 方法的改进版本 RUSboost 进行分析，同时也重点对比与基于财务指标和逻辑回归的 Dechow (2011)[11],以及基于 SVM 和原始财务信息 Cecchini et al. (2010)[16]两篇基准文献的识别效果，发现在使用了集成学习和原始财务信息的情况下，识别效果更优。在应用 Bagging 策略上，Xu et al.(2022)[2]以中国 A 股上市公司为样本，采用了基于 GONE 理论的特征，对比分析了包括随机森林在内的 6 种机器学习模型对于财务舞弊识别的性能，发现随机森林模型的效果最好，暴露因子在所有特征中占比最高。

2.5 文献评述

综观国内外的相关研究成果，现有研究通常专注于使用单一算法，缺乏对财务场景数据特点与具体算法之间适用性的详细分析，对于结论是否能够扩展其他场景缺乏证据。结合实际场景数据和现有文献，本文发现财务舞弊数据通常表现出以下几个方面的特征：

财务舞弊数据整体样本规模不大。由于年度报表相关数据的经过审计的相对权威性，大部分财务舞弊研究都采用了 firm-year 频率的数据，研究区间通常为数年到数十年，最终纳入分析的样本规模，通常不会很大。在近年的研究中，Dechow (2011) [11]最大采用了 293 条舞弊样本 79358 条非舞弊样本进行训练。Kirkos et al. (2007)[15]由于采用了匹配黑白样本的方式，样本数量仅为 76 条。已有研究表明 Skurichina et al.(2002)[59]通过实验数据发现，相比于 bagging，boosting 对于训练集规模大小更敏感，说明 Bagging 更适合处理中小型数据集。相比于深度学习或计算机视觉领域动辄上千万的数据，财务舞弊数据规模并不大，可能 bagging 策略更加适合。

财务舞弊数据可能面临维度灾难诅咒 (Curse of Dimensionality) 。已有研究中，通常采用专家筛选的方式挑选特定报表指标，例如 Cecchini et al. (2010)[16] 采用 40 个报表特征、Perols (2011)[12] 采用 42 个财务及公司治理特征、Bertomeu et al. (2021)[60]采用 100 个以上特征、Dutta et al. (2017)[61] 采用 116 个特征。当特征数量超过训练样本时，可能会遭遇维度灾难，进而导致模型更容易过拟合。Breiman (2001)[22]提出的随机森林算法基于 bagging 策略，加入了随机特征采样的方式，一定程度上限制了每次训练的特征数，增大了基学习器的多样性，一定程度上降低了维度灾难的可能性，减少了过拟合风险。而 Boosting 算法在不经过特征筛选的情况下，特征过多更可能导致出现过拟合问题。

从研究现状来看，虽然 Bao et al (2020)[29] 以及 Xu et al.(2022)[2]分别对基于 Boosting 和基于 Bagging 策略的集成学习分别基于中美两国数据进行了验证，但对

于 Bagging 策略和 Boosting 策略到底哪一种更适合中国市场财务舞弊识别，尚未形成一致意见。因此，本文重点对 Bagging 和 Boosting 两种集成策略的对比研究能够在一定程度上弥补对这一问题的理解，为引入集成学习，建立科学、高效的财务舞弊识别模型提供经验和思路。

另一方面，考虑到财务舞弊场景的数据特征具有普遍性，因此本文的研究成果也适用拓展到类似场景。例如基于公司层面数据的企业信用评分（Wang et al., 2011）[80]、企业破产（Kim et al., 2010）[81]等问题。甚至部分跨领域问题也存在跟财务舞弊场景类似的高维度过拟合情形，例如 Saeys et al.（2007）[82]提出在医学诊断行业存在临床病例较少，而刻画病情的数据维度较多的情况。Zhang（2014）[83]认为地理信息领域高光谱数据的应用也存在样本较少而维度过高的问题。

2.6 集成学习算法原理

本文主要研究不同策略的集成学习算法在财务舞弊领域的性能，因此本节主要对本文所使用的不同策略的集成学习算法的基本原理及相关研究进行归纳。

2.6.1 随机森林

Breiman（1996）[21]首次提出 Bagging 策略。该作者在原论文基础之上，通过引入随机特征抽样的方式，进一步提出了随机森林算法（Breiman, 2001）[22]。

Bagging 是一种典型的并行集成学习框架，它的核心思想是自助采样。给定一个包含 m 个样本的数据集，可以通过有放回的随机抽样，每次抽取一个样本放入采样集中，重复 m 次，得到一个与原始数据集大小相同的采样集。可以重复这个过程 T 次，得到 T 个采样集，然后基于每个采样集训练出一个基学习器，最后将这些基学习器组合起来。图 2-1 给出了 Bagging 策略的并行框架示意。

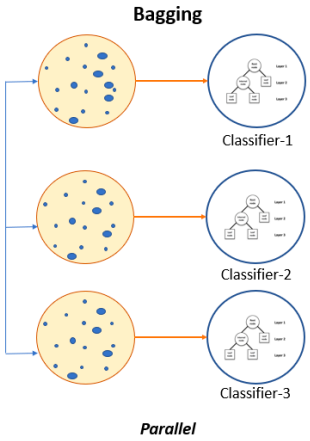

图 2-1 Bagging 策略并行框架

图片来源：pluralsight 网站

随机森林算法在 Bagging 策略的基础上进一步增加了特征的随机性。它的具体步骤如下：首先，假设有 M 个样本，有放回地随机选择 M 个样本。其次，在决策时每个节点需要分裂时，从 N 个特征中随机选取 n 个特征（满足 n<<N），然后从这 n 个特征中选择特征进行节点分裂。最后，基于抽样的 M 个样本和 n 个特征按照节点分裂的方式构建决策树。按照这些步骤构建大量决策树组成随机森林，然后将每棵树的结果进行综合。

2.6.2 自适应增强树 Adaboost

Schapire（1990）[19]最早提出 Boosting 的思想 Freund and Schapire（1997）[20]实现了基于 Boosting 思想的 Adaboost 算法早期版本。Friedman et al.（2000）[71]，进一步基于统计思想，形成了目前常用的 Adaboost 算法。

Boosting 策略一般过程都是从一个弱分类器开始，在训练过程中不断改变样本的概率分布，使得下一次训练时算法更加关注上一轮的错误，并最终组合多个弱分类器的结果。图 2-2 给出了 Boosting 策略的串行框架示意。

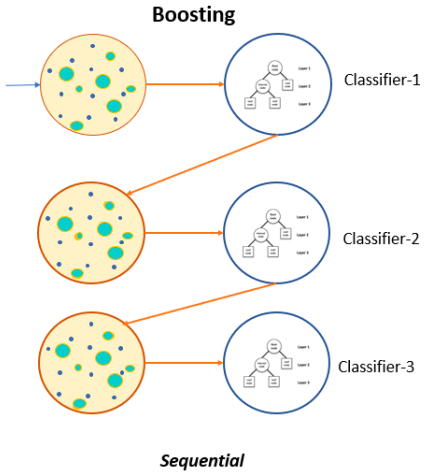

图 2-2 Boosting 算法串行框架

图片来源：pluralsight 网站

具体在自适应增强树（Adaboost）中，是先从初始训练集中训练出一个基学习器，再根据基学习器的表现对训练样本分布进行调整，使得先前基学习器做错的训练样本在后续受到更多关注，然后基于调整后的样本分布来训练下一个基学习器；如此重复进行，直至基学习器数目达到事先指定的值 T，最终将这 T 个基学习器进行加权结合。

2.6.3 梯度提升树 GBDT

Friedman (2001)[71]最早提出了 GBDT 算法。梯度提升树（GBDT）首先构造出可以计算梯度的损失函数。该方法通过多轮迭代,每轮迭代产生一个弱分类器，每个分类器在上一轮分类器的残差基础上进行训练。对弱分类器的要求一般是足够简单，并且是低方差和高偏差的。训练过程中通过降低偏差来不断提高最终分类器的精度，最终的总分类器是将每轮训练得到的弱分类器加权求和得到的。与 Bagging 的方法相比，Boosting 的方法能更好的拟合分类面，但是有过拟合的风险。因此在使用 Adaboost 和 GBDT 的时候需要限制决策树的个数和树的深度，必要时需要采用剪枝和 earlystopping 策略。

2.6.4 极限梯度提升树 XGBoost

Chen et al. , (2016)[72]提出的极限梯度提升树（XGBoost）是对 GBDT 算法的进一步优化增强，并提供了并行计算的处理框架。具体而言其优势在于

（1）在选择每棵树的分裂点时采用了并行化策略，大大提高了模型的速度;

（2）无需对数据归一化处理，算法本身能对缺失值进行有效处理。

（3）模型增加了对树结构复杂度的惩罚项，通过引入正则项和列抽样提高了模型的鲁棒性、能够有效避免过拟合的问题.

（4）模型迭代式生成树，并提供增益、覆盖、权重等多个选项和维度描述特征重要性，具有较强变量解释性。

（5）XGBoost 既可以做回归也可以做分类问题，应用的场景非常广泛。目前 XGBoost 算法已经发展成一套可扩展的开源机器学习系统。

2.6.5 轻度梯度提升树 LightGBM

轻度梯度提升树（LightGBM）是微软亚洲研究院（Ke et al,2017[73]）公布的一个开源、快速、高效的基于决策树算法的提升(GBDT)框架，被用于排序、分类、回归等多种机器学习的任务，支持高效率的并行训练。LightGBM 对 Xgboost 中导致复杂度较高的因素进行了优化。具体来说，XGBoost 通过预排序的算法来寻找特征的最佳分裂点，虽然预排序算法能够准确的找出特征的分裂点，但该方法占用空间的代价太大，在数据量和特征量都比较多的情况下，会严重影响算法性能。即 XGBoost 寻找最佳分裂点的算法复杂度可以简化表示为：复杂度=**特征数量*特征分裂点的数量*样本数量**。LightGBM 就分别从这三个角度提供了优化方法。

为了**减少特征分裂点的数量**和更加高效寻找最佳特征分裂点，LightGBM 区别于 XGBoost 的预排序算法，采用 Histogram 直方图的算法寻找最佳特征分裂点。其基本想法是将连续的浮点特征值进行离散化为 k 个整数并构造一个宽度为 k 的直方图。对某个特征数据进行遍历的时候，将离散化后的值用为索引作为直方图的累积统计量。遍历完一次后，直方图便可累积对应的统计量，然后根据该直方图寻找最佳分裂点。直方图算法本质上就是一种数据离散化和分箱操作，可以大幅降低计算代价和内存占用。

从**减少样本角度**进行优化，LightGBM 引入了单边梯度抽样算法(GOSS，Gradient-based One-Side Sampling)，主要思路是将训练过程中大部分梯度较小的样本剔除，减少每次迭代中的样本数量，仅对剩余样本数据计算信息增益。

从减少特征数量角度，LightGBM 对稀疏特征采用互斥特征捆绑算法(EFB，Exclusive Feature Bundling)，通过将两个互斥的特征捆绑在一起，合为一个特征，在不丢失特征信息的前提下，减少特征数量，从而加速模型训练。

除了 Histogram、GOSS 和 EFB 算法之外，LightGBM 还提出了区别于 XGBoost 的按层生长的叶子节点生长方法，即带有深度限制的按叶子节点生长(Leaf-Wise)的决策树生成算法，精度更高且更有效率，能够节约不必要的计算开销，同时为防止某一节点过分生长而加上一个深度限制机制，能够在保证精度的同时一定程度上防止过拟合。

2.7 本章小结

本章从财务舞弊识别问题的相关理论、财务舞弊识别所需的基本特征、财务舞弊识别的主要模型、基于机器学习的财务舞弊识别模型研究四个方面对已有文献进行回顾与评述，并针对现有研究中存在的短板提出针对性的研究问题和解决方案。最后，对于本文涉及的主要集成学习的算法原理进行了介绍。

第三章　研究设计

3.1 数据来源与数据处理

本文原始财务数据采集自 China Stock Market and Accounting Research (CSMAR)数据库。该数据库已被广泛应用于与中国资本市场有关的相关研究中（Wang et al.,2022[62]; Leippold et al.;2021[63]; Geng et al., 2015[64]）。综合考虑合理性和可行性，并参考研究中常见的做法，本文进行了以下数据样本的提取与筛选处理：

（1）根据报表类型为合并报表，在 2000 年-2020 年共提取到 50812 条“公司-年度观测值”。（2）考虑到金融企业在经营模式方面与一般企业存在显著差异，导致其报表科目的定义以及核算方式等方面与一般企业也存在较大差异，为了保证各项特征与指标在内涵与逻辑上的一致性，本文将金融行业的样本进行剔除，即一级行业分类代码为 J 的样本，共计剔除 912 条。（3）由于在计算特征时需要依据所属行业进行横向对比，因此进一步剔除行业分类信息缺失的样本，共计剔除 3969 条。（4）根据舞弊样本的可得性和时间范围，本文在上述经过剔除的数据中，选取了 2013-2018 年属于深圳证券交易所交易的上市公司的年度报告，共计 10954 条“公司-年度观测值”。

Bertomeu et al. (2021)[60]认为在使用面板数据进行机器学习研究时需要注意，由于面板数据在跨周期内存在相关性，并不存在完全干净的独立样本。但是也还是可以通过选择样本区间，来获得样本外结果，从而避免显著的样本损失。因此，本文参考 Xu et al (2022) [2]、Bao et al.(2020)[29]、Bertomeu et al. (2021)[60]的训练集和测试集划分做法，根据年份的先后顺序，对样本进行训练集和测试集的划分，而非进行全样本的随机抽样按比例划分。其中 2013-2017 年的 8850 条样本作为训练集，2018 年的 2104 条样本作为测试集，在采样前的训练集与测试集的大小比约为 4.2:1。

数据处理方式的不同会直接影响到特征的构建以及样本的分布。本文进行了以下数据预处理：

（1） 行业分类归并处理

基于证监会行业分类，计算行业对比值。考虑到不同行业的数量与规模差异，以及部分细分行业的可比公司数量较少，进而可能导致行业值的不稳定或者缺失，因此本文将证监会行业分类为制造业的，根据二级行业分类并计算行业对比值，其余行业按照一级行业分类区分并计算对比值。

在行业归并之后，本文在表 3-1 中统计了 2013 年-2018 年每个行业内的公司数量以及相关统计指标。具体来说，表中第一列表示具体行业名称，第二列列示了

不同行业中，在 2013-2018 年区间的舞弊样本数量，第三列年份数量则具体统计了在 2013-2018 年这 6 年中，某个行业是否有对应的样本数据，如果没有覆盖到具体公司，则年份数量小于 6。最后第四至第六列为 2013-2018 年内每个行业覆盖的公司数量的对应统计值。第五列为样本时间范围内每年每个行业的公司数量的标准差。第六列为每个行业的公司数量的最小值，后续分别表示 25%分位数，中位数，75 分位数以及最大值。可以看出，样本时间范围内共有 43 个行业分类，其中原分类数量最多的制造业，在修改分类后上市公司数量在每年的分布更为分散和均匀，减少了行业参考值为不合理异常值的可能性。

（2） 异常值处理

本文所用特征主要为构造的规则特征，在计算中主要涉及滞后一期、二期计算，分子分母的比值计算等，因此若前一年数值缺失或为 0，导致增长率计算过程中分母为 0，此时计算结果作为缺失处理。

（3） 缺失值处理

由于目前涉及所有变量均为 01 类型的指标，若存在缺失则视为该指标并不满足，因此填充为 0。（4）增加独热编码变量：本文基于训练集中涉及到的行业与年份变量，通过独热编码的方式，进行了特征衍生。增加年份变量 5 个，行业变量 43 个，共计 48 个新增特征。加上原始的 122 个特征，进入模型的实际有 170 个特征。（5）特征筛选：为了尽量保留特征的原始效果，本文计算了训练集特征的方差，剔除掉方差等于 0 的无关特征。（6）样本类别不平衡问题处理：为了避免样本类别不平衡问题带来的问题，本文参考 He and Garcia（2009）[65]的做法采用随机下采样（Random Undersampling）的方式处理样本类别不平衡问题，具体来说，是从训练集的非舞弊样本中随机抽取与舞弊样本数量相同的样本，构成新的数据集，以保证黑白样本的比例处于相对均衡的状态。同时为了稳健性起见，本文也重复多次进行随机下采样处理，并基于多次重复的结果进行模型效果的对比分析。

表 3-1 各行业上市公司数量分布与舞弊样本数量（2013-2018 年）

行业名称	舞弊样本数量	年份数量	均值	标准差	最小值	25%分位	中位数	75%分位	最大值
A_农林牧渔业	17	6	27.5	1.64	26	26	28	29	29
B_采掘业	28	6	27.17	1.72	25	26	27	29	29
C13_农副食品加工业	2	6	31.33	1.75	30	30	30	32	34
C14_食品制造业	8	6	18.17	2.93	14	16	18	21	21
C15_酒、饮料和精制茶制造业	1	6	16.33	0.52	16	16	16	17	17
C17_纺织业	11	6	24.17	1.33	22	24	24	25	26
C18_纺织服装、服饰业	2	6	21.33	1.03	20	21	21	22	23
C19_皮革、毛皮、羽毛及其制品和制鞋业	0	6	3	0	3	3	3	3	3
C20_木材加工及木、竹、藤、棕、草制品业	0	6	5	0.89	4	4	5	6	6
C21_家具制造业	0	6	5.17	2.71	3	3	4	7	9
C22_造纸及纸制品业	7	6	14.5	0.84	14	14	14	15	16
C23_印刷和记录媒介复制业	0	6	6	1.1	5	5	6	7	7
C24_文教、工美、体育和娱乐用品制造业	0	6	10.5	1.22	9	10	10	12	12
C25_石油加工、炼焦及核燃料加工业	0	6	7.5	0.84	6	7	8	8	8
C26_化学原料及化学制品制造业	27	6	136	12.99	119	127	136	147	151
C27_医药制造业	29	6	113.67	22.49	89	95	111	133	141
C28_化学纤维制造业	2	6	15.67	1.51	14	14	16	17	17
C29_橡胶和塑料制品业	18	6	46.83	6.37	40	42	46	53	54
C30_非金属矿物制品业	15	6	48.33	2.58	44	48	48	49	52
C31_黑色金属冶炼及压延加工业	0	6	10.67	0.52	10	10	11	11	11
C32_有色金属冶炼及压延加工业	6	6	36.33	3.83	32	33	36	40	41
C33_金属制品业	2	6	38	3.35	33	36	39	41	41
C34_通用设备制造业	17	6	86	5.02	79	83	86	90	91

表 3-2 各行业上市公司数量分布与舞弊样本数量（2013-2018 年）（续）

行业名称	舞弊样本数量	年份数量	均值	标准差	最小值	25%分位	中位数	75%分位	最大值
C35_专用设备制造业	37	6	119.17	15.99	97	107	122	132	136
C36_汽车制造业	12	6	56.67	6.56	49	52	55	61	66
C37_铁路、船舶、航空航天和其它运输设备制造业	4	6	16.33	3.44	13	14	16	19	21
C38_电气机械及器材制造业	37	6	138.5	14.8	118	129	139	150	156
C39_计算机、通信和其他电子设备制造业	36	6	204	39.44	165	172	194	236	256
C40_仪器仪表制造业	7	6	32.5	6.8	23	28	33	37	41
C41_其他制造业	7	6	13.5	1.05	12	13	14	14	15
C42_废弃资源综合利用业	1	6	1.83	0.98	1	1	2	3	3
D_电力、煤气及水的生产和供应业	9	6	38.17	5.91	32	33	38	44	45
E_建筑业	15	6	43.67	7.94	33	37	46	49	52
F_批发和零售业	12	6	64.67	2.16	62	63	64	66	68
G_交通运输、仓储和邮政业	2	6	26.17	2.86	24	24	25	28	31
H_住宿和餐饮业	3	6	8.17	1.83	6	6	8	10	10
I_信息传输、软件和信息技术服务业	48	6	151.5	41.9	107	116	147	186	204
K_房地产业	13	6	60.83	3.54	57	58	61	63	66
L_租赁和商务服务业	20	6	23.33	8.8	15	16	22	30	35
M_科学研究和技术服务业	4	6	18	7.85	11	13	15	22	31
N_水利、环境和公共设施管理业	6	6	23.17	4.49	19	21	21	25	31
O_居民服务、修理和其他服务业	0	2	1	0	1	1	1	1	1
P_教育业	0	3	1	0	1	1	1	1	1
Q_卫生和社会工作业	1	6	4.5	1.87	2	3	4	6	7
R_文化、体育和娱乐业	10	6	23.5	7.53	13	18	24	30	32
S_综合行业	2	6	7.5	0.84	7	7	7	8	9

注释：表格报告了样本区间内，不同行业的上市公司数量分布的统计信息。其中，第一列为 2012 版证监会行业分类名称，其中制造业按照 2 级行业名称展开统计，其余行业按照一级

行业名称。第二列为 2013 年-2018 年该行业内的舞弊样本数量。第三列为年份数量，显示了在样本时间范围内某个行业是否存在数据的情况。第四列为样本时间范围内每个行业的公司数量的多年平均值，第五列为样本时间范围内每年每个行业的公司数量的标准差。第六列为每个行业的公司数量的最小值，后续分别表示 25%分位数，中位数，75 分位数以及最大值。

3.2 定义舞弊样本

在国外研究中，通常是以 SEC 公布的“会计审计监管系列文告（Accounting and Auditing enforcement releases，AAERs）”，作为认定财务报表舞弊的标准来源（Dechow et al., 2011[11]; Bao et al., 2020[29]）。而在中国，主要以相关监管部门如证监会、交易所的处罚公告作为认定财务报表舞弊的标准（洪荭等, 2012[35]; 钱苹和罗玫, 2015[49]; 张成浩等, 2021[66]）。

参考已有文献，本文所用舞弊样本来自以下两个来源：

来源一：基于证监会官网公布的由证监会或证监局出具的行政处罚决定，通过人工阅读行政处罚公告的方式，根据正文中对上市公司及高管违规行为的描述，例如含有“虚减虚增净利润”、“虚减虚增成本”，被监管机构认定存在严重的财务信息的“虚假记载、误导性陈述或重大遗漏”，性质上存在信息披露违法，导致直接影响到具体某段时间的年度报告财务数据的案例作为舞弊标签依据，将涉及到的年度报告标记为舞弊样本。

来源二：深圳证券交易所作为上市公司一线监管部门，其针对上市公司年报问题出具的纪律处分决定通常披露了上市公司在信息披露合规方面的问题，其中也会对财务舞弊以及重大违规行为进行处分。因此本文根据深交所官网的监管信息公开中披露的纪律处分，下载并阅读纪律处分正文，同样根据与来源（1）相同的判定逻辑，标记舞弊样本标签。

若在两个来源均判定为舞弊标签，则只计算来源一的标签。

表 3-3 样本时间分布（2013-2018 年）

年份	非舞弊样本	舞弊样本	样本总数	舞弊样本占比（%）
2013	1525	29	1554	1.87
2014	1578	36	1614	2.23
2015	1682	50	1732	2.89
2016	1793	90	1883	4.78
2017	1918	149	2067	7.21
2018	1980	124	2104	5.89

注释：表格展示了舞弊样本与非舞弊样本在训练集和测试集上分年份的数量分布情况。第一列为年份，第二列为非舞弊样本的数量，第三列为舞弊样本的数量，第四列为舞弊样本与非舞弊样本的数量之和，第五列为舞弊占比，表示舞弊样本数量占当年所有样本数量的比例，以百分数表示。

在表 3-2 中给出了舞弊样本和非舞弊样本在时间上的分布情况统计。可以发现，首先在 2013 年至 2018 年舞弊样本的数量整体趋势上是逐渐上升的，这说明随着市场对舞弊问题的日益关注，监管机构对财务舞弊的稽查力度也是逐年增大，进行财务舞弊的成本与门槛也会变得更高。（2）舞弊样本占每一年所有样本的数量占比最高为 2017 年的 7.2%，最低为 2013 年的 1.87%，均低于 10%，舞弊样本占比远远小于非舞弊样本，存在样本类别不平稳问题。本文这一结果高于现有研究（Bao et al., 2020）[29]对于美国公司的占比，这是由于本文在舞弊样本的筛选条件和来源中，为了尽量合理的增加舞弊样本数量，提高模型潜在的可用信息和表现力，加入了纪律处分等来源，而并非单一根据证监会立案调查结果来判断。需要注意的是，在训练集数据进行下采样处理之后，每一次参与训练的非舞弊样本数量为 354，与训练集中 2013-2017 年的所有舞弊样本数量相等。

表 3-1 的第二列列示了 2013 年-2018 年每个行业内舞弊样本的累计数量。从舞弊行为的行业分布来看，数量最多的是制造业，具体在专用设备制造业、电气机械及器材制造业、计算机、通信和其他电子设备制造业、化学原料及化学制品制造业、医药制造业等二级行业舞弊案例较为频发。此外本文样本也覆盖了传统会计观点认为容易发生舞弊的农林牧渔业。值得注意的是，信息传输、软件和信息技术服务业的舞弊发生概率仅次于制造业，这与该行业具有的轻资产以及收入确认手段复杂等特点有直接关系。

3.3 构建特征集

本文主要参照现有文献（Beneish, 1999[10]；Cecchini et al., 2010[16]；Abbasi et al., 2012[5]）中常用的财务舞弊衡量维度，从资产质量、盈利能力、现金流量、营运能力、行业环境、偿债能力六个方面选取并构建了 122 个特征变量。由于详细的变量定义内容角度，具体请参见附录中的附表 1。

这几个维度的具体构建过程如下：

资产质量：资产质量主要分析了财务舞弊公司在实施财务舞弊过程中，由于资金的不合理流转，导致的在资产负债表各科目的不合理变动问题。已有研究中，Beneish（1999）[10]采用剔除固定资产的非流动性资产占总资产的比例变动作为资

产质量指数，并发现舞弊公司的资产质量存在下降。部分研究（Cecchini et al. 2010[16]; Dechow et al. 2011[11]）采用存货增长率，来识别舞弊公司是否通过减少成本，增加存货进而提升收益的舞弊手段。本文采用了基于资产负债表的重要会计科目计算的 41 项指标，分别从结构占比合理性、增长合理性以及同行业对比合理性方面衡量了应收款项、其他应收款、长期应收款、预付账款、存货、商誉、净资产、货币资金、在建工程等科目的影响。

盈利能力：舞弊行为的一项重要目标，就是提升利润水平。以此来维持市场的良好印象和股价水平。Perols and Lougee(2011)[67]认为糟糕的业绩和盈利能力会增加通过财务舞弊改善财务状况的压力。Cecchini et al（2010）[16]发现当舞弊行为发生时，通常伴随的较高的营业收入增长率。本文从占比合理性、增长合理性以及行业水平合理性方面，衡量了毛利率、净资产收益率、收入变动、净利润变动、销售费用、非经常性损益等项目的具体表现，构建了 47 项指标。

现金流量：已有研究（Beneish 1999[10]; Dechow et al. 2011[11]）认为净经营现金流与净利润的差值，衡量了应计项目对财务报表的影响，且当在实施收入造假的时候，该指标通常为正。本文主要关注了经营现金流勾稽异常的五种表现，以及存在经营现金流的波动异常，共计 6 项指标。

偿债能力：Dechow（2011）[11]采用了债务杠杆 leverage 来衡量长期债务占总资产的比例，认为杠杆较高的公司，更有动机通过采用舞弊的方式，提升财务与市场表现，进而达到现有债务合约的条款约束，或者以更优的利率发行新的债券。本文主要从资产负债率、短期偿债能力（例如流动比率、速动比率、现金比率）、有息负债（包括长期借款、应付债券等中长期债务）的增长变动以及行业水平对比，来衡量公司的偿债能力，共计 21 项指标。

营运能力：已有研究从总资产周转率、应收账款周转天数变化率等方面（Cecchini et al.，2010[16]；Abbasi et al.，2012[5]）研究舞弊公司的营运能力。Abbasi et al.（2012）[5]认为当发生盈利舞弊时，营业收入迅速增加，并导致了较高的总资产周转率。考虑到与资产质量本身的匹配性，以及在盈利能力模块已经包括了营业收入相关指标，本文主要从存货周转、应收账款周转方面衡量公司的营运能力，共计 5 项指标。

行业环境：Rezaee and Riley (2010)[68]发现当经济衰退期间，舞弊行为会增加。换句话说，当行业环境恶化时，部分公司可能为了继续维持业绩水平达到避免 ST 或者避免退市的目的实施舞弊。因此本文加入同行业平均收入和毛利率下滑两项指标，来刻画这一趋势。

变量定义如附录中的附表 1 所示。根据变量定义进行计算，若满足变量定义，则变量赋值为 1，若数值未满足阈值或基础数值存在缺失，则默认为 0。

3.4 集成模型选择

本文将采用横向对比的方式研究 5 种集成学习算法的实证表现。所用模型定义如表 3-3 所示。为了简便起见，在下文中出现的具体模型，均采用模型缩写指代。

本文主要从代表性和多样性的角度出发选择这些模型：Breiman（1996）[21]首次提出 Bagging 策略。该作者在原论文基础之上，通过引入随机特征抽样的方式，进一步提出了随机森林算法（Breiman, 2001）[22]。根据谷歌学术的检索结果，该算法目前已被引用近十万次。Sagi and Rokach（2018）[69]基于 Web of Science 论文检索数据以及部分技术网站的发帖数据，对 2014-2017 年常用的机器学习算法使用频率进行了统计，发现随机森林的使用度最高。Fernandez-Delgado et al.（2014）[70]基于 121 个不同领域的数据集，实证对比了 179 种分类器上的表现，发现随机森林的准确率最高。无论从使用频率还是实际效果，随机森林都显示在 Bagging 策略中的优势，属于公认的基于 Bagging 策略的 SOTA 模型，因此本文将随机森林作为研究 Bagging 策略集成代表。

Schapire（1990）[19]最早提出 Boosting 的思想。Boosting 策略一般过程都是从一个弱分类器开始，在训练过程中不断改变样本的概率分布，使得下一次训练时算法更加关注上一轮的错误，并最终组合多个弱分类器的结果。Freund and Schapire（1997）[20]实现了基于 Boosting 思想的 Adaboost 算法早期版本。Friedman et al.（2000）[71]，进一步基于统计思想，形成了目前常用的 Adaboost 算法。其他学者分别基于 Boosting 策略，分别提出了 GBDT（ Friedman , 2001[71]），Xgboost（ Chen et al. , 2016[72]）, LightGBM（Ke et al,2017）[73]等算法。

虽然这四种算法都是基于 Boosting 思想，在 Boosting 策略中哪一种才是真正的 SOTA 模型仍存在争议，因为这四种算法**在具体实现上存在不同的优化方向**：Adaboost 侧重于改变每一轮训练的样本权重来调整样本的概率分布；与 Adaboost 不同，GBDT 的学习目标是通过对前一棵树的残差进行学习进而降低损失函数。并且 GBDT 会利用梯度迭代的方法使得每次迭代产生的损失函数最小,使模型变得越来越精确； XGBoost 在 GBDT 的基础上进一步优化增强，增加了对树结构复杂度的考量。LightGBM 是 GBDT 的一种工程实现，主要从提升优化效率和内存使用出发。这四种算法在样本权重修改、残差学习、树结构复杂度、内存优化等方面的差异，本质上都依赖于实现 Boosting 策略的具体过程，而在 Bagging 策略实现过程中并不存在对应环节或公认成熟的模型。例如 Bagging 策略中所有样本均为有

放回抽样产生，不同数据子集之间不涉及调整样本权重等手段。因此本文将 Adaboost、GBDT、Xgboost、LightGBM 这四种算法均作为研究 Boosting 策略集成的代表，而 Bagging 策略则只选择了随机森林。

在具体实现上，本文算法实现均基于 scikit-learn 机器学习库（Pedregosa et al., 2011）[74]。本文采用在训练集中进行分层 5 折交叉验证进行模型调参。为了避免随机参数搜索导致的局部最优问题，本文采用网格搜索（GridSearchCV）进行模型超参数搜索。为了避免调参范围太大导致的计算资源需求过高，本文首先绘制每种模型的每个参数对应的验证曲线，并根据验证曲线的趋势拐点及参数合理性，最终确定网格调参参数范围，具体见表 3-4 给出了所有模型的网格调参参数范围与备选数量。

表 3-4 模型定义表

序号	集成学习类别	模型缩写	模型全称	常用名
1	Bagging	rf	Random Forest Classifier	随机森林
2	Boosting	ada	Ada Boost Classifier	自适应增强树
3	Boosting	gbc	Gradient Boosting Classifier	梯度提升树
4	Boosting	xgboost	Extreme Gradient Boosting	极限梯度提升树
5	Boosting	lightgbm	Light Gradient Boosting Machine	轻度梯度提升器

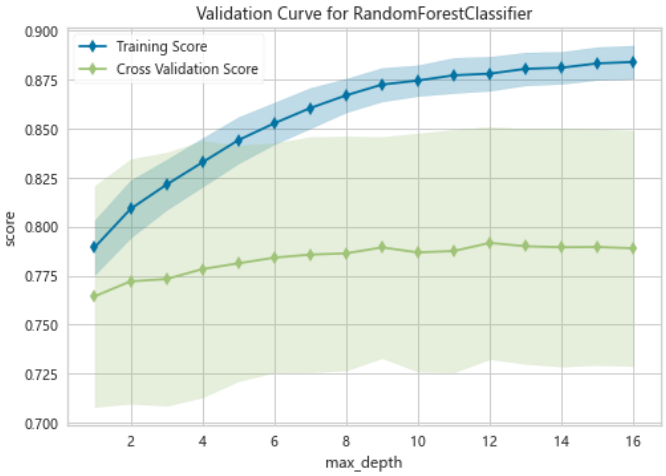

图 3-1 随机森林 max_depth 参数验证曲线

例如，图 3-1 提供了随机森林 max_depth 参数验证曲线,随着 max_depth 的增加，随机森林中决策树的树深度限制不断放宽，模型复杂度进一步增加，训练集效果提升但泛化误差不断增加，验证集上的模型效果逐渐达到瓶颈，因此选择 6 至 12 的范围作为 max_depth 调参选择，其他算法的参数范围选取方式类似。

表 3-5 网格调参参数范围与备选数量

模型简称	参数名称	参数备选范围	备选数量	参数组合总数
rf	n_estimators	[50, 100]	2	1152
rf	max_depth	[6, 8, 10, 12]	4	1152
rf	min_impurity_decrease	[0.0001, 0.0002, 0.001]	3	1152
rf	max_features	[1.0, 'sqrt', 'log2']	3	1152
rf	bootstrap	[True, False]	2	1152
rf	criterion	['gini', 'entropy']	2	1152
rf	min_samples_split	[3, 7]	2	1152
rf	min_samples_leaf	[3, 7]	2	1152
ada	n_estimators	[100, 200, 300]	3	36
ada	learning_rate	[0.01, 0.1, 0.2, 0.3, 0.4, 0.5]	6	36
ada	algorithm	['SAMME', 'SAMME.R']	2	36
gbc	n_estimators	[50, 100, 150]	3	2592
gbc	learning_rate	[0.01, 0.1, 0.2]	3	2592
gbc	subsample	[0.2, 0.5, 1.0]	3	2592
gbc	min_samples_split	[3, 7]	2	2592
gbc	min_samples_leaf	[3, 7]	2	2592
gbc	max_depth	[4, 6, 8, 10]	4	2592
gbc	min_impurity_decrease	[0.0001, 0.001]	2	2592
gbc	max_features	[1.0, 'sqrt', 'log2']	3	2592
xgboost	learning_rate	[0.01, 0.1]	2	864
xgboost	n_estimators	[100, 200]	2	864
xgboost	subsample	[0.2, 0.3, 1]	3	864
xgboost	max_depth	[4, 6]	2	864
xgboost	colsample_bytree	[0.5, 1]	2	864
xgboost	min_child_weight	[2, 4]	2	864
xgboost	reg_alpha	[1e-06, 0.01, 0.1]	3	864
xgboost	reg_lambda	[1e-06, 0.01, 0.1]	3	864
lightgbm	num_leaves	[50, 60]	2	5184
lightgbm	learning_rate	[1e-06, 0.01, 0.1]	3	5184
lightgbm	n_estimators	[100, 200]	2	5184
lightgbm	min_split_gain	[0.4, 0.5]	2	5184
lightgbm	reg_alpha	[1e-06, 0.01, 0.1]	3	5184
lightgbm	reg_lambda	[1e-06, 0.01, 0.1]	3	5184
lightgbm	feature_fraction	[0.5, 0.9, 1]	3	5184
lightgbm	bagging_fraction	[0.4, 0.7]	2	5184
lightgbm	bagging_freq	[2, 4]	2	5184
lightgbm	min_child_samples	[1, 3]	2	5184

注释：表格报告了机器学习模型进行网格调参的参数备选范围。其中第一列为机器学习模型名称，第二列为机器学习模型对应的超参数名称，第三列为参数范围，均为离散值，第四列为根据参数范围计算的单个参数的备选数量，最后一列为某个算法所有参数备选数量的乘积，也代表了需要进行搜索的次数。

3.5 模型评估指标

3.5.1 传统评估指标

财务舞弊识别问题可以归类为机器学习的二元分类问题，即区分财务舞弊样本和非财务舞弊样本，因此同样也适用于常用的分类模型效果评估指标。已有研究(Dechow, 2011[11]; Bao et al.,2020[29]; Cecchini et al.，2010[16])常用指标包括 Accuracy、AUC (Throckmorton et al. ,2015[75]; Hajek and Henriques,2017[76]) 、Recall、F1 等。

本节进一步对常见的分类模型效果评估指标的基本原理与构造方式进行介绍。

财务舞弊样本对应于正例（Positive，也称阳性），非财务舞弊样本对应于负例(Negative，也称阴性)。进一步来说，对于二分类任务，可将所有样例根据其真实所属类别与模型分类结果组合分为真正例(True Positive)，假正例(False Positive)，真反例(True Negative)，假反例(False Negative）四种情况，令 TP，FP，TN，FN 分别表示其对应的样例数，则可以给出如下的混淆矩阵(Confusion Matrix)。

表 3-6 混淆矩阵

真实情况	预测结果	
	正例	反例
正例	TP（真正例）	FN（假反例）
反例	FP（假正例）	TN（真反例）

基于混淆矩阵的结果，可以计算如下常见评估指标：

准确率(Accuracy)是评估分类模型效果最基本的指标，准确率取值在 0 至 1 之间，表示了模型分类正确的样本占总样本数的比例，反应了模型整体的分类结果，既包括将财务舞弊样本正确分类，也包括了将非财务舞弊样本分类正确的结果。而业务中通常更关心财务舞弊样本的正确分类，而非财务舞弊样本分类正确的问题，因此本文并未将该指标纳入分析范围。

$$\textbf{准确率：}Accuracy = \frac{TP + TN}{TP + TN + FN + FP} \qquad (3-1)$$

精确率(Precision)又称为精确率，通常是指真实为舞弊的数量，在预测结果中预测为舞弊样本的数量。例如有 10 家公司舞弊，但是模型预测输出有 20 家公司，那么此时的精确率为 50%。从公式可以看出，分母为预测输出为舞弊的样本数量。

$$\textbf{精确率：}Precision = \frac{TP}{TP + FP} \qquad (3-2)$$

召回率（Recall）又称查全率以及敏感度（Sensitivity），计算所有真实的舞弊样本中被正确分类的比例，取值在 0 至 1 之间，越大表示对阳性样本的识别越完全。

$$召回率：Recall = Sensitivity = \frac{TP}{TP + FN} \quad (3-3)$$

精确率与召回率是一对矛盾的变量，一般来说精确率越高召回率则越低，而召回率越高精确率则越低。以精确率为纵轴，召回率为横轴作图就可以得到精确率-召回率曲线，也成为“P-R 图”，该图可以直观的对多个模型进行比较，若模型 A 对应的 P-R 曲线完全“包住”模型 B 对应的曲线，则可以说该模型 A 的性能优于模型 B。

混淆矩阵的计算依赖于模型的预测标签，而预测标签依赖于阈值的选择。在机器学习模型在进行预测时，会对每一个测试样本产生一个预测值，然后将这个预测值与一个给定的阈值进行对比。因此，给定的阈值越高，则预测结果就越不容易被识别为正例，判定舞弊的条件更加严格，但同时也放弃了大量真实舞弊样本。但若给定的判别阈值过高，则会导致纳入大量的非舞弊样本。而判别阈值选取本身是一个较为主观的操作，为了避免这一影响，综合考虑在多个判别阈值下模型的真实表现，引入 ROC 曲线与 AUC 指标。

ROC 全称是“受试者工作特征”(Receiver Operating Characteristic）曲线。与 P-R 曲线的绘制过程类似，即根据模型输出的所有样本按照预测存在财务舞弊的概率降序排列后，逐个将排序靠前的公司判定为存在财务舞弊进行预测，然后更新计算累积的真正率 TPR 和假正率 FPR，每一次计算两个指标作为横纵坐标作图，并连接所有结果，即可得到 ROC 曲线。

$$真正率：TPR = \frac{TP}{TP + FN} \quad (3-4)$$

$$假正率：FPR = \frac{FP}{TN + FP} \quad (3-5)$$

真正率 TPR 代表了真实存在财务舞弊的公司中，模型判定存在财务舞弊的公司比例。假正率 FPR 代表了真实无财务舞弊的公司中被模型判定为存在财务舞弊的公司的比例。当不断调整分割阈值，将更多的预测存在财务舞弊概率不断减小的公司逐渐归类为存在财务舞弊时，TPR 与 FPR 的值都在不断增加，因此绘制出的曲线应该是一个在[0,1]*[0,1]上单调递增的曲线。

本文将 ROC 曲线包裹的面积定义为 AUC（Area Under ROC Curve）。当模型的 A 的 ROC 曲线包住模型 B 时，就可以判定模型 A 有着更好的分类效果以及泛化性能，在财务舞弊识别领域，已有 Hajek and Henriques (2017)[76]采用 AUC 作为评估指标。

综上所述，本文使用 AUC、Recall、Precision 指标初步评判分类模型的分类效果。

3.5.2 创新评估指标

部分文献提出传统分类评估指标在实际使用中，可能存在一定问题。例如，Fawcett (2006)[77]认为 AUC 仅代表了随机选择的真实舞弊样本排序高于随机选择的真实非舞弊样本排序的概率。Lobo et al., (2008)[79]提出了 AUC 指标由于存在多种问题，例如该指标忽略了预测输出的概率值和模型实际拟合优度之间的关系，此外 ROC 的绘图空间中，其实存在左上角和右上角这类假正率或假负率极高的极端区域，而 ROC 曲线无法进行单独分析。此外，AUC 指标对于黑白样本的遗漏影响都是等价考虑的。这些问题事实上与财务舞弊场景的特点并不一致。

在财务舞弊场景下，受限于监管资源和监管人员有限的注意，监管人员只会更加关心预测排名靠前的部分公司，AUC 是对所有阈值划分情况下的统一评价，并无法全面覆盖实际业务逻辑。此外，召回率 Recall 和精确率 Precision 都是针对默认阈值为 0.5 进行判定，这样基于固定数值的判断，同样也不符合财务舞弊场景。因此，为了准确刻画不同机器学习模型对于财务舞弊的识别效果，需要进行评估指标的创新，创新的思路主要是从评价角度上引入能够对排名靠前的公司以及排名顺序进行刻画的评价指标。

本文参考 Xu et al (2022) [2]和 Bao et al.(2020)[29]的做法，根据模型预测舞弊概率排名最高的 K 家公司，分别计算了 NDCG@k、Precision@k、Recall@k 三个指标。NDCG@k 是用于衡量推荐算法排名特性的常见指标。其主要思想是，如果舞弊样本的预测排名越靠前，说明模型效果越好，且排名越靠前带来的增益越高，排名越靠后则增益效果递减，最后累计计算前 k 个样本的累计增益。具体来说，首先需要计算$DCG@k$，即

$$DCG@k = \sum_{i=1}^{k} \frac{(2^{rel_i} - 1)}{log_2(i+1)} \tag{3-6}$$

其中 k 表示在舞弊预测概率排名最高的 k 名样本。如果排名第 i 名的样本是真实舞弊样本，那么rel_i等于 1，否则为 0。将前 k 个样本的 DCG 指标进行累计，则得到$DCG@k$指标。$NDCG@k$的定义为

$$NDCG@k = \frac{DCG@k}{idealDCG@k} \tag{3-7}$$

其中$idealDCG@k$表示在最理想的情况下，DCG@k 指标能达到的最高数值，即所有真实舞弊样本的排名都处于最高。因此 NDCG@k 指标可以看做是将$DCG@$进行均一化，充分考虑了舞弊样本的真实分布情况。在具体实现中，不同的研究采用了不同的k，通常会参考黑样本的实际占比以及样本总数。例如 Bao et al. (2020)[29]采用了样本数量的 1%，而 Xu et al(2022)[2]采用了样本数量的 12%。考虑到在实际

应用中，监管层更加关注疑似风险最高的样本列入清查名单，因此本文采用了固定数值，即 k=50，约占本文测试集样本数量的 2.3%，介于上述研究之间且不失一般性。进一步的，本文把根据前 k 名样本计算的精准率指标定义为 Precision@k，前 k 名样本的召回率定义为 Recall@k，作为对模型排名效果的补充分析。

3.6 本章小结

本章旨在详细解释和说明本文的实证分析步骤。首先本文对数据来源和数据处理方式进行说明；然后详细定义了舞弊样本的选择方式，并根据资产质量、盈利能力、现金流量、营运能力、行业环境、偿债能力六个方面选取并构建了 122 个特征变量，进行了详细说明。同时，从机器学习角度说明了集成模型的选择过程以及用于模型评估的指标含义。

第四章　实证分析

4.1 描述性统计

本文参考 Dechow(2011)[11]的做法，在附录中的附表 2 中报告了训练集中的舞弊样本和非舞弊样本的所有变量的分组描述性统计，并对舞弊样本与非舞弊样本的特征在均值上是否存在显著差异进行了双侧 T 检验。首先，从样本数量对比来看，2013 年至 2017 年训练集中，舞弊样本有 354 条，而非舞弊样本有 8496，样本之间存在严重不平衡，因此需要进行非平衡样本的重采样处理。其次，舞弊样本和非舞弊样本在本文所用的输入变量上面存在显著差异，基于均值差异检验的结果，有 72 个特征变量在 1%置信水平下显著，10 个特征变量在 5%置信水平下显著，这种差异越大，在模型训练过程中，每个特征最优切分点的选取就更明确，进而可以提升建模的效果。最后，本文基于业务经验和文献基础构建的舞弊风险预测特征，大部分都是属于正向预测舞弊概率的。这一点也从表中可以得到验证，即从均值差异的符号来看，没有均值差异为负且在 1%水平下显著的变量。具体从部分指标来看，舞弊样本组的 V1_应收款项占比高于同行业的指标均值显著高于非舞弊组，与部分研究（Dechow et al.,2011[11]）认为的舞弊行为为了掩饰虚构的利润，通常采用增加应收账款的操纵手法逻辑一致。

4.2 Bagging 与 Boosting 集成策略效果对比

4.2.1　基于传统指标的对比

表 4-1 比较了 5 种集成学习模型在传统的 AUC、Recall、Precsion 指标上的具体表现。

表 4-1 模型效果对比 (传统指标)

模型	AUC	Recall	Precision
rf	**0.846**	**0.831**	0.151
lightgbm	0.846	0.823	0.139
ada	0.844	0.798	**0.161**
gbc	0.839	0.807	0.140
xgboost	0.837	0.815	0.146

注释：表格报告了 5 种集成学习模型在测试集上的效果。表格按照评估指标 AUC 进行由高到低排序。第一列为模型名称，第二列及之后为模型性能评估指标，分别包括 AUC、Recall、Precison。详细定义参见研究设计部分。该表结果基于单轮下采样，随机下采样种子参数 seeds=0。

从 AUC 指标来看，随机森林和 LightGBM 的数值最高，为 0.846。从召回率指标来看，随机森林的效果最优，为 0.831，说明当以默认阈值 0.5 来判定舞弊与非舞弊样本时，随机森林算法能够将 83.1%的真实舞弊样本成功识别；从精确率来说，Adaboost 的效果最优，为 0.161，说明以默认阈值 0.5 来判定舞弊与非舞弊样本时，Adaboost 算法识别的舞弊样本中，约有 16%是真实舞弊样本。从 Precison 指标来看，当使用默认阈值划分舞弊与非舞弊时，结果的质量并不会很高。

图 4-1 给出了 5 中集成学习算法的 ROC 曲线的具体形态对比，从图中可以看出，这 5 种集成学习的 AUC 水平都相对比较高，其中随机森林和 LightGBM 的 ROC 曲线更偏向图形的左上角，但相比其他算法的差异并不明显。说明基于 Bagging 和基于 Boosting 策略的集成学习算法，难以通过在 ROC 曲线上进行对比判断优劣。

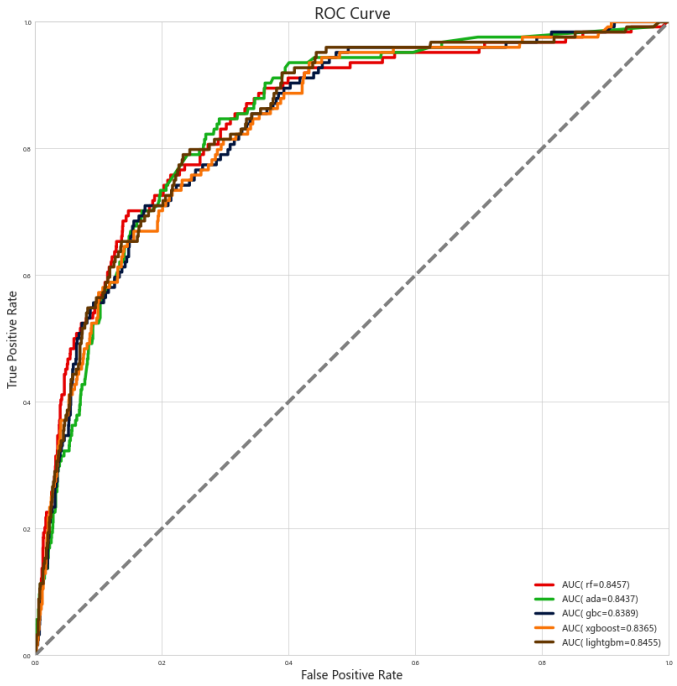

图 4-1 ROC 曲线汇总

由于精确率与召回率是一对矛盾的变量，在不同的判定阈值下，精确率越高则召回率则越低，而召回率越高精确率则越低。以精确率为纵轴，召回率为横轴作图就可以得到精确率-召回率曲线，也成为“P-R 图”，该图可以直观的对多个模型进行比较，若模型 A 对应的 P-R 曲线完全从右侧“包住”模型 B 对应的曲线，则可以说该模型 A 的性能优于模型 B。AP（Average Precision）是指平均精确率。它是 PR 曲线下的面积，用来衡量分类模型的性能。AP 越高表示模型的性能越好。

图 4-2 给出了 5 中集成学习算法的 P-R 曲线的具体形态对比，从图中可以看出，随机森林的 P-R 曲线更靠右，在 Recall 达到一定水平的情况下，随机森林算

法对应能得到更高的 Precision 值，随机森林的平均精确率 AP 为 0.2995，也是这五种算法中最高的，算法效果更优。

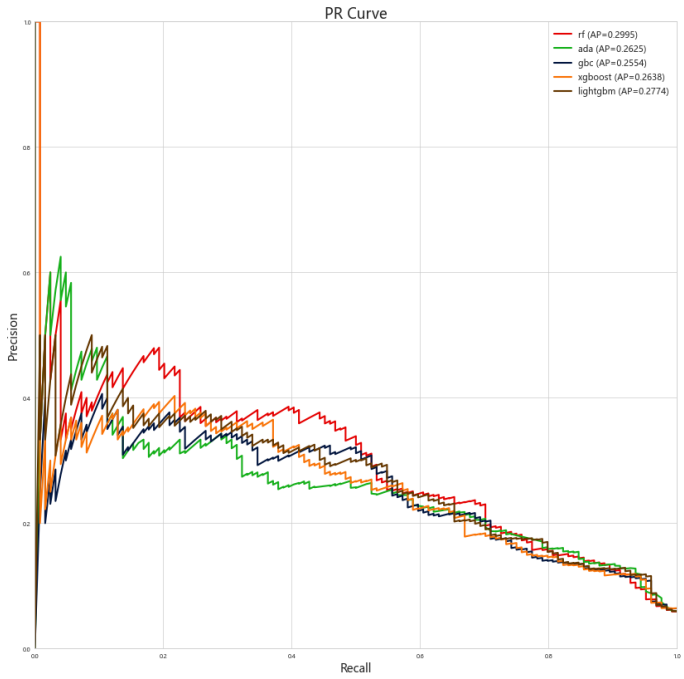

图 4-2 P-R 曲线汇总

从这三种传统的分类指标效果来看，随机森林算法虽然在 AUC 指标上与 LightGBM 算法差异不大，但效果仍然领先于其余三种基于 Boosting 的算法。此外，随机森林算法的召回率也是最高的，因此本文认为随机森林算法所代表的 Bagging 策略，相比基于 Boosting 策略的集成学习具有一定的优势。

4.2.2 基于创新指标的对比

表 4-2 比较了 5 种集成学习模型在创新的 NDCG@k，Precsion@k 以及 Recall@k 指标上的具体表现。

表 4-2 模型效果对比 (创新指标)

模型	NDCG@k	Precision@k	Recall@k
rf	0.482	0.480	0.194
lightgbm	0.380	0.380	0.153
ada	0.357	0.340	0.137
xgboost	0.356	0.340	0.137

gbc	0.312	0.340	0.137

注释：表格报告了 5 种集成学习模型在测试集上的效果。表格按照评估指标 NDCG@k 进行由高到低排序。第一列为模型名称，第二列及之后为模型性能评估指标，分别包括 NDCG@k、Precision@k、Recall@k。详细定义参见研究设计部分。该表结果基于单轮下采样，随机下采样种子参数 seeds=0。

NDCG 指标衡量了预测结果排序的质量，舞弊样本的预测排名越靠前，则对应的评估指标越高。本文设置的 k=50，即重点关注预测结果中疑似舞弊概率最高的 50 个样本的表现情况。根据 NDCG@k 指标排名来看，随机森林的指标值为 0.482，高于其他算法。GBDT 的 NDCG@k 指标最差，仅为 0.312。随机森林的 Preciosn@k 指标值为 0.48，高于其他模型，说明随机森林识别出的前 50 名中公司中，有 48%是真实舞弊公司。随机森林的 Recall@k 指标值为 0.191，高于其他模型，说明有 19%的舞弊公司，通过随机森林提示的前 50 名可以识别出来。这些结果说明随机森林在基于对舞弊概率最高的 50 个样本进行的分析中，在衡量排序性能的 NDCG@k、Precision@k、Recall@k 上，随机森林都显著高于其他模型，说明基于 Bagging 策略的集成模型，针对舞弊样本能够给出排名更靠前的预测结果，提升了财务舞弊预测识别的效率。

4.3 基于随机下采样的模型效果对比

为了解决建模过程中面临的黑白样本不平衡问题，本文采用了随机下采样的方式进行处理。作为对比，部分研究（Hajek and Henriques，2017[76]）采用样本匹配的方式解决样本不平衡问题，本质上加入了过多的人为选择因素，即可比样本的选择受到行业划分基准的因素的影响；Xu et al.（2022）[2]采用通过调整 classweight 参数进而改变模型对于黑白样本的权重影响建模结果，由于不是所有模型均具有 class_weight 参数，因此并不适用于所有模型。较少有研究采用上采样即多次重复生成黑样本，使得黑白样本比例平衡的方式，可能导致生成到现实中并不存在的黑样本，且进一步增大了建模的数据量，计算资源需求更高，因此参考 Perols（2017）[53]以及 He and Garcia（2009）[65]的做法，本文采用在训练集上进行随机下采样的方式进行处理。

由于建模样本会随着采样种子的设定而发生改变，导致不同算法调参建模之后得到的算法对比结果发生改变而不唯一，出于稳健性考虑，本文进一步重复进行了随机下采样种子为 0-20 共 21 次建模。在每一次建模中，均根据当前随机下采样种子得到的训练集，重新训练所有集成学习模型，并再次基于上文提到的参数范围，

采用网格搜索重新寻找最优的模型参数，并在相同的测试集（2018 年样本集）上进行验证，最终对重复多次随机下采样之后建模的结果进行平均处理。

表 4-3 比较了 5 种集成学习模型在不同评估指标上的平均结果。

表 4-3 集成模型效果对比 (基于随机下采样 21 次平均)

模型	AUC	Precision	Recall	NDCG@k	Precision@k	Recall@k
rf	0.836	0.150	0.813	**0.390**	**0.375**	**0.151**
lightgbm	**0.837**	0.144	**0.828**	0.373	0.360	0.145
ada	0.828	**0.166**	0.755	0.332	0.341	0.137
gbc	0.829	0.148	0.804	0.331	0.331	0.134
xgboost	0.813	0.150	0.763	0.296	0.297	0.120

注释：表格报告了 5 种集成学习模型在测试集上的效果。表格按照评估指标 NDCG@k 进行由高到低排序。第一列为模型名称，第二列及之后为模型性能评估指标，分别包括 AUC、Precision、Recall、NDCG@k、Precision@k、Recall@k。详细定义参见 3.5 节模型评估指标部分。

从平均结果来看，在传统评估指标中，随机森林与 LightGBM 的 AUC 相比其他算法更高，而 LightGBM 指标的平均召回率比随机森林更高，但是平均精准率更低。GBDT 和 Xgboost 在三个传统指标上的表现均较差。因此从传统指标的平均结果来看，同样难以判断基于 Bagging 和基于 Boosting 算法的具体优劣。

在创新评估指标中，根据 NDCG@k 指标排名来看，随机森林的指标平均值为 0.39，高于其他算法。Xgboost 的 NDCG@k 指标最差，平均值仅为 0.296。随机森林的 Preciosn@k 指标平均值为 0.375，高于其他模型，说明随机森林识别出的前 50 名中公司中，平均有 37.5%是真实舞弊公司。随机森林的 Recall@k 指标平均值为 0.151，高于其他模型，说明有 15%的舞弊公司，通过随机森林提示的前 50 名可以识别出来。

图 4-3 给出了这 21 轮不同下采样随机种子设置下，这五种集成学习算法所对应的 NDCG@k 指标表现。从图中可以看出，在 21 轮结果中，随机森林的 NDCG@k 指标值相对较高，对应的位置连接线整体偏上，而其余基于 Boosting 策略的算法的 NDCG@k 则整体趋势相对偏下，lightgbm 的位置相对靠前。由于不同下采样随机种子对应的 Recall@k 以及 Precision@k 对比图与 NDCG@k 存在类似的表现，故不再重复列示。

这些结果说明随机森林在基于对舞弊概率最高的 50 个样本进行的分析中，在衡量排序性能的 NDCG@k、Precision@k、Recall@k 上，随机森林都显著高于其他模型，说明基于 Bagging 策略的集成模型，针对舞弊样本能够给出排名更靠前的预测结果，提升了财务舞弊预测识别的效率，前述分析结论不变。

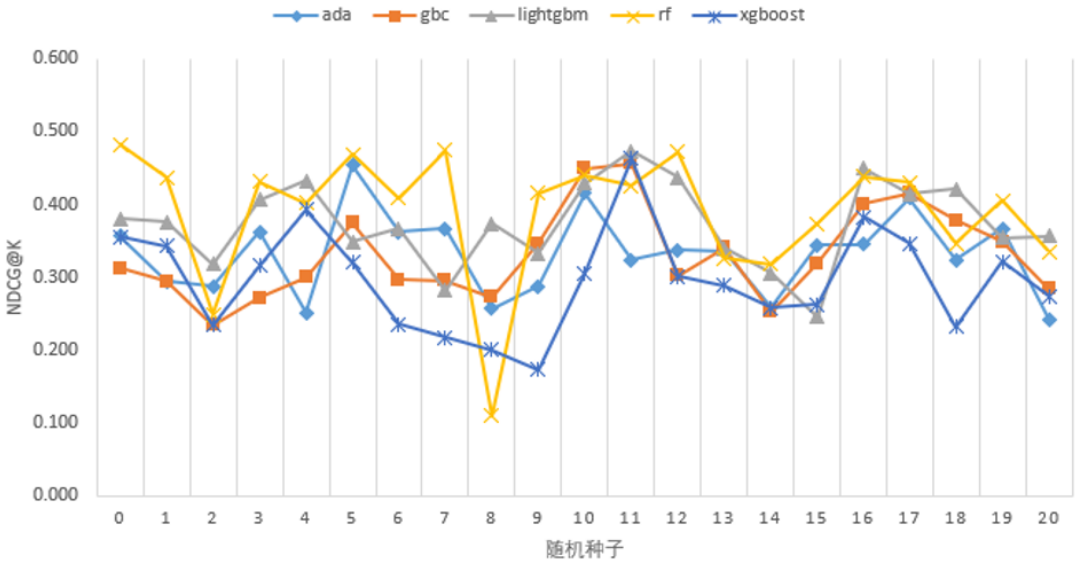

图 4-3 不同下采样随机种子对应 NDCG@k 表现

4.4 Bagging 策略效果更优的机理解释

在财务舞弊问题中，Bagging 策略为何比 Boosting 策略更有效？本文从数据规模和算法设计特性两个角度给出具体的机理解释。

首先，从数据维度和数据规模来看，财务舞弊数据可能面临维度灾难诅咒(Curse of Dimensionality) 。已有研究中，通常采用专家筛选的方式挑选特定报表指标，例如 Cecchini et al. (2010)[16] 采用 40 个报表特征、Perols (2011)[12] 采用 42 个财务及公司治理特征、Bertomeu et al. (2021)[60]采用 100 个以上特征、Dutta et al. (2017)[61] 采用 116 个特征。当特征数量超过训练样本时，可能会遭遇维度灾难，进而导致模型更容易过拟合。随机森林算法基于 bagging 策略，加入了随机特征采样的方式，一定程度上限制了每次训练的特征数，增大了基学习器的多样性，一定程度上降低了维度灾难的可能性，减少了过拟合风险。而 Boosting 算法在不经过特征筛选的情况下，特征过多更可能导致出现过拟合问题。

另一方面，从算法设计所针对的数据特性来看，本文所用的基于证监会处罚的财务舞弊样本，来自监管部门的行政处罚与交易所的纪律处分信息。正如 Gepp（2021）[78]基于美国市场数据做出的类似提醒，由于监管成本的限制，监管机构肯定会漏掉了一些舞弊行为。因此选择的非舞弊样本有可能是未被监管机构纳入调查的潜在舞弊样本，也就是说在这一场景下始终存在数据的噪声，因此需要考虑算法对噪声的处理能力。

从算法设计的特点来看，单一决策树由于只采用了单一训练数据，容易受到训练集中的极端数据或噪声的影响，导致出现过拟合，很少单独使用。Bagging 通过

重采样方法从原始训练集中有放回的采样得到多个训练子集，由于各个训练子集相互独立，降低了基分类器的方差，改善了泛化误差，随机森林在 Bagging 基础上进一步加入了随机特征采样，进一步增大了各个训练子集之间的独立性。Dietterich（2000）[23]通过模拟实验论证了当训练集和测试集的样本标签存在不一致时导致的噪音时，Bagging 的性能更稳定。

形象来说，Bagging 策略充分发挥了“三个臭皮匠顶一个诸葛亮”的朴素的集成思想，模拟了将大量市场参与者基于不同信息集做出的判断进行归纳整合，进而得到更优结果的过程。Bagging 策略中的每一个训练子集就是一个掌握了部分信息（有放回抽样）的专家，随机森林中的每一个训练子集则更是一个只精通于某一个领域（特征子集）的部分信息的专家，正是由于他们判断问题视角的极大差异，使得他们投票共同做出的决定更加收敛与稳健。

对比来说，Boosting 在每轮训练中使用的训练集不变，但训练集中每个样例会根据上一轮的学习结果进行调整，使新学习器针对已有学习器判断错误的样本进行学习，Boosting 类似于让一群专家依次来对上一名专家做出的结论进行反复修正，只要能够得到比上一次更好的结论，那么这种判断就是有效的，显然这种方法能够显著提高弱学习器的学习效果，但很容易受到噪声的影响产生过拟合现象，当样本量不足或特征维度过多时，会使得专家的意见过于偏斜于特定的特征或者样本，导致过拟合。

4.5 基于特征重要度的模型解释性分析

为进一步探索哪些特征对预测识别财务舞弊更加重要，根据 Bertomeu et al(2021)[60]的做法，本文使用 Scikit-learn 的“Feature Importance”功能输出了各个模型的特征重要性。①

例如图 4-4 即表示根据随机森林输出的指标特征重要性前 10 名所绘制的柱形图，柱形长度表示特征的重要性程度，可以直观看出涉及的重要指标。

① 基于单个树模型的特征重要性，通常是指计算单个特征的平均信息增益（Mean information gain）来衡量。即使用某个特征能够使得模型的信息增益提高更多，那么这个特征就更重要。对于集成学习来说，则是考虑特征在多棵树上的平均信息增益。

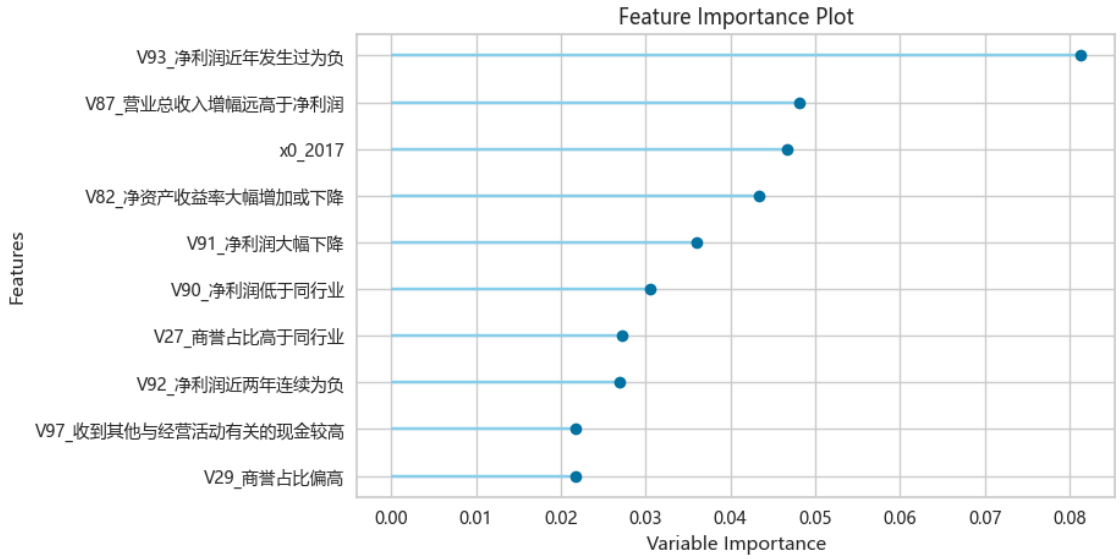

图 4-4 随机森林输出特征重要性示意

本文将每个模型的特征重要性排名前 20 的特征定义为最重要特征，并汇总统计了重要特征被不同算法的使用情况，表 4-4 截取了被算法使用最多的 20 个重要特征。从该表可以看出：商誉占比高于同行业、净资产收益率大幅增加或下降、三费占比降幅高于同行业这三个特征被 5 种集成学习模型视为重要特征，而商誉占比偏高、在建工程勾稽异常、有息负债率高于同行业、收入变动趋势偏离行业、净利润低于同行业、净利润近年发生过为负这 7 项特征被至少 4 种集成模型视为重要特征。主要涉及到 5 项盈利能力特征与 3 项资产质量类特征。

为了进一步分析不同特征对于模型的贡献比例，本文以在上述分析中表现最好的集成学习模型随机森林 RF 为例，来说明不同特征具体是如何影响模型识别的。

表 4-5 报告了在根据随机森林模型中各变量的特征重要性占比大小进行排名，得出的排名靠前的 20 个指标以及这些指标所属类别和累计贡献的重要性。

表 4-4 重要性排名前 20 特征

变量	模型覆盖度	rf	ada	gbc	xgboost	lightgbm	所属类别
V27_商誉占比高于同行业	5	7	10	13	5	12	资产质量
V82_净资产收益率大幅增加或下降	5	4	14	1	2	16	盈利能力
V111_三费占比降幅高于同行业	5	15	7	20	7	11	盈利能力
V29_商誉占比偏高	4	10	9	7	3	22	资产质量
V36_在建工程勾稽异常	4	13	15	10	38	1	资产质量
V78_有息负债率高于同行业	4	16	20	19	26	14	偿债能力
年份虚拟变量 2017	4	3	8	3	23	2	年份虚拟变量
V89_收入变动趋势偏离行业	4	11	3	12	8	32	盈利能力
年份虚拟变量 2013	4	17	33	9	19	3	年份虚拟变量
V90_净利润低于同行业	4	6	27	5	4	9	盈利能力
V93_净利润近年发生过为负	4	1	5	2	1	52	盈利能力
V87_营业总收入增幅远高于净利润	3	2	106	11	20	38	盈利能力
V97_收到其他与经营活动有关的现金较高	3	9	23	17	24	8	现金流管理
V99_经营现金流降幅较大	3	12	31	18	75	10	现金流管理
行业虚拟变量_采掘业	3	29	2	52	17	13	行业虚拟变量
行业虚拟变量_房地产业	3	58	17	57	9	4	行业虚拟变量
V96_支付其他与经营活动有关的现金较高	3	14	29	6	83	5	现金流管理
V92_净利润近两年连续为负	2	8	109	16	56	63	盈利能力
V91_净利润大幅下降	2	5	108	4	77	58	盈利能力
年份虚拟变量 2014	2	34	30	22	13	6	年份虚拟变量

注释：表格报告在不同的机器学习模型中，均提示存在异常的特征。其中第一列为变量名称，第二列为该变量的模型覆盖度。数量越大，说明当前变量被更多的模型所考虑和关注。即首先统计每种模型的特征重要性排名，若某个特征在模型中的排名在前 20 名以内，则该变量的模型覆盖度加一，得到每个变量的模型覆盖度。最后根据模型覆盖度降序排序，本表报告了模型覆盖度最高的 20 个变量。第三列之后为变量具体在某个模型上的特征重要性实际排名，数值越小说明特征重要性排名越靠前。

表 4-5 基于随机森林算法（RF）的特征重要性排名前 20

排名	变量	重要性占比	累计重要性占比	变量类别
1	V93_净利润近年发生过为负	8.13%	8.13%	盈利能力
2	V87_营业总收入增幅远高于净利润	4.81%	12.94%	盈利能力
3	年份虚拟变量 2017	4.67%	17.62%	年度控制变量
4	V82_净资产收益率大幅增加或下降	4.33%	21.95%	盈利能力
5	V91_净利润大幅下降	3.60%	25.55%	盈利能力
6	V90_净利润低于同行业	3.05%	28.61%	盈利能力
7	V27_商誉占比高于同行业	2.73%	31.33%	资产质量
8	V92_净利润近两年连续为负	2.69%	34.03%	盈利能力
9	V97_收到其他与经营活动有关的现金较高	2.18%	36.20%	现金流管理
10	V29_商誉占比偏高	2.17%	38.37%	资产质量
11	V89_收入变动趋势偏离行业	1.98%	40.36%	盈利能力
12	V99_经营现金流降幅较大	1.98%	42.34%	现金流管理
13	V36_在建工程勾稽异常	1.80%	44.15%	资产质量
14	V96_支付其他与经营活动有关的现金较高	1.76%	45.90%	现金流管理
15	V111_三费占比降幅高于同行业	1.75%	47.65%	盈利能力
16	V78_有息负债率高于同行业	1.74%	49.39%	偿债能力
17	年份虚拟变量 2013	1.56%	50.95%	年度控制变量
18	V88_营业总收入增幅远低于净利润	1.38%	52.33%	盈利能力
19	V65_毛利率近一年大幅降低	1.22%	53.55%	盈利能力
20	V6_应收款项近一年增速高于同行业	1.18%	54.74%	资产质量
	其余变量（N=149）重要性		**Cumulative Importance=45.26%**	

注释：表格报告了在随机森林模型中，各变量的特征重要性占比。第一列为排名信息，第二列为变量名称，第三列为按比例表示的特征重要性，并按照降序排列。第四列表示特征重要性的累积值。最后一行表示除上述特征以外剩余特征贡献的重要性累计。下采样随机种子 Seeds=0。

可以进一步得出以下结论：首先，不同重要性的特征对于模型解释的贡献大小存在明显差异。前 10 个主要特征累计提供了超过 38%的解释力，单个特征可以提供至少 2%的解释力。前 20 个主要特征累计提供了超过 54%的解释力，单个特征至少可提供超过 1%的解释力。而前 20 名以外的其他 149 个特征累计只提供了 45.26%的累计解释力；其次，从涉及的变量类别来看，这些特征主要涉及盈利能力、资产质量以及现金流管理维度。其中盈利能力的变动至关重要。在前 20 名特征中，有 10 个特征属于盈利能力，4 项属于资产质量。从实际应用来看，当出现大幅的亏损，或者营业总收入增幅远高于净利润，净资产收益率波动异常，都通常预示着该公司存在舞弊的概率较高。

表 4-6 报告了分类别对特征重要性统计累计占比的结果，盈利能力的解释力占比超过 43%，资产质量的解释力占比超过 21%，说明在可预测的财务舞弊信号中，盈利能力的变化以及资产质量变动是最重要的识别视角。

表 4-6 基于随机森林算法（RF）的特征重要性分类别占比

特征类别	累计特征重要性
盈利能力	43.64%
资产质量	21.36%
年度控制变量	8.65%
偿债能力	8.35%
行业控制变量	7.48%
现金流管理	7.46%
营运能力	2.87%
行业环境	0.17%

注释：表格报告了在随机森林模型中，分类别统计各变量的特征重要性占比。

4.6 本章小结

本章首先对样本进行描述性统计分析，并对舞弊样本与非舞弊样本的特征在均值上是否存在显著差异进行了检验。接着基于传统的 AUC、recall、precsion 指标分析了对 Bagging 和 Boosting 这两种集成学习策略的表现，并发现难以有效区分两种策略的优劣，因此进一步地，结合实际业务经验，通过引入基于排序特性的 NDCG@k、Recall@k、Precision@k 等创新指标进行了对比，并发现随机森林所表的 Bagging 算法更有优势，并根据文献和算法特点，给出了具体的机理解释。最后，为了进一步与业务逻辑对应，找到需要查证或关注的具体的重要指标作为解释，本文基于随机森林这一算法和特征重要性概念，进一步探索了不同特征和特征类别对识别财务舞弊的贡献程度。

第五章 总结与展望

5.1 总结

集成学习作为机器学习的一个重要研究方向，近年已被广泛应用于各个领域。而财务舞弊识别模型的研究也从最初的基于传统逻辑回归等统计模型，发展到逐步引入包括集成学习在内的前沿技术。但在集成学习众多策略之中，特别是经典的 Bagging 和 Boosting 这两种集成策略，在什么场景或数据中效果更优一直存在争论。本文从财务舞弊数据场景特点出发进行分析，以 2013-2018 年 A 股市场中深交所上市公司为样本，从资产质量、盈利能力、偿债能力、营运能力、现金流分析、市场环境六个维度构建了 122 个特征变量，系统性的对比分析了基于 Bagging 和基于 boosting 策略的 5 种集成学习模型在财务舞弊识别问题上的实际表现，并分析了不同特征变量的重要性，实证发现：

首先，在集成学习众多模型中，基于 Bagging 策略的随机森林模型，表现优于基于 Boosting 策略的 GBC，Adaboost，xgboost，LightGBM。具体表现在 NDCG@k，Precision@k，recall@k 等排序指标上。这一结果可以从财务舞弊场景的数据特征以及 Bagging 策略的设计特点来解释。财务舞弊场景的数据规模较小，特征维度较高，天然的适合使用随机森林这种基于 Bagging 策略的集成学习来处理。且 Bagging 策略充分发挥“三个臭皮匠顶一个诸葛亮”的朴素的集成思想，模拟了将大量市场参与者基于不同信息集做出的判断进行归纳整合的过程。

其次，本文通过基于特征重要度的模型解释性分析，发现虽然不同集成学习关注的重要特征排序存在一定差异，部分重要的特征始终会引起各个模型的注意，其中盈利能力与资产质量类的特征，对于识别财务舞弊行为更加重要，其中当存在收入大幅下降、连续亏损等表现时，通常预示着舞弊概率的进一步提升。

本文提供了对于 Bagging 集成策略适用场景的机理证据，丰富了集成学习在经济与管理领域的研究参考。同时由于财务舞弊场景数据的特征也具有普遍性，因此本文的结论同样适用类似场景，例如基于公司层面数据的企业信用评分、企业破产等问题，甚至部分例如医学诊断、地理信息等跨领域问题，由于同样存在高维度过拟合的特点，也可参考本文结论；本文引入了模型的贡献度分析，进一步打开了机器学习识别过程的黑箱，将机器学习的输出线索与现实监管行为联系起来，增强了研究的实际意义。

5.2 后续工作展望

财务舞弊的有效识别是一个颇有难度的研究问题。受时间与研究精力所限，本文的研究内容主要存在以下不足：

本文所用特征主要基于财务报表三大报表的原始特征并在此之上构建了 122 个具有一定业务含义的指标，特征范围主要局限于财务指标，但特征范围到底如何考虑仍有诸多争议。从实际应用过程来看，财务指标的数据较为容易获得与处理，对于模型结果的扩展和新时间区间数据的纳入有一定好处，但是非财务数据，例如公司治理指标或者审计特征也可能带来一些增量效果。另一方面，本文没有考虑基于舆情或者年报文本等特征，这是由于当前对这类文本信息的信息含量高低仍存在一定争议，但是文本信息的引入可以借鉴自然语言处理等机器学习领域的其他方法，包括使用前沿的预训练语言模型，这可能提供了新的研究思路。

参考文献

[1] Rezaee Z. Causes, consequences, and deterence of financial statement fraud [J]. Critical Perspectives on Accounting, 2005, 16(3): 277-298.

[2] Xu X, Xiong F, An Z. Using machine learning to predict corporate fraud: evidence based on the gone framework [J]. Journal of Business Ethics, 2022: 1-22.

[3] Fich E M, Shivdasani A. Financial fraud, director reputation, and shareholder wealth [J]. Journal of Financial Economics, 2007, 86(2): 306–336.

[4] Beasley M S, Hermanson D R, Carcello J V, Neal T L. Fraudulent financial reporting: 1998-2007: An analysis of US public companies [R].

[5] Abbasi A, Albrecht C, Vance A, Hansen J. Metafraud: a meta-learning framework for detecting financial fraud [J]. MIS Quarterly: Management Information Systems, 2012, 36(4): 1293–1327.

[6] Ugrin, J.C., and M.D. Odom. Exploring Sarbanes–Oxley's effect on attitudes, perceptions of norms, and intentions to commit financial statement fraud from a general deterrence perspective [J]. Journal of Accounting and Public Policy, 2010, 29(5): 439–458.

[7] Hastie T, Tibshirani R, Friedman J H. The elements of statistical learning: data mining, inference, and prediction [M]. Vol. 2. New York: Springer, 2009: 1-758.

[8] Zhou Z H. Machine learning [M]. Springer Nature, 2021.

[9] Persons O S. Using financial statement data to identify factors associated with fraudulent financial reporting [J]. Journal of Applied Business Research (JABR), 1995, 11(3): 38-46.

[10] Beneish M D. Incentives and penalties related to earnings overstatements that violate GAAP [J]. The Accounting Review, 1999, 74(4): 425-457.

[11] Dechow P M, Ge W, Larson C R, Sloan R G. Predicting material accounting misstatements [J]. Contemporary Accounting Research, 2011, 28(1): 17-82.

[12] Perols J. Financial statement fraud detection: An analysis of statistical and machine learning algorithms [J]. Auditing: A Journal of Practice & Theory, 2011, 30(2): 19-50.

[13] Green B P, Choi J H. Assessing the risk of management fraud through neural network technology [J]. Auditing, 1997, 16: 14-28.

[14] Fanning K M, Cogger K O. Neural network detection of management fraud using published financial data [J]. Intelligent Systems in Accounting, Finance & Management, 1998, 7(1): 21-41.

[15] Kirkos E, Spathis C, Manolopoulos Y. Data mining techniques for the detection of fraudulent financial statements[J]. Expert Systems with Applications, 2007, 32(4): 995-1003.

[16] Cecchini M, Aytug H, Koehler G J, et al. Detecting management fraud in public companies[J]. Management Science, 2010, 56(7): 1146-1160.

[17] Hansen L K, Salamon P. Neural network ensembles[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1990, 12(10): 993-1001.

[18] Dietterich T G. Ensemble methods in machine learning[C]. International Workshop on Multiple Classifier Systems. Springer, Berlin, Heidelberg, 2000: 1-15.

[19] Schapire R E. The strength of weak learnability[J]. Machine Learning, 1990, 5(2): 197-227.

[20] Freund Y, Schapire R E. Experiments with a new boosting algorithm[C]. ICML. Vol. 96. 1996: 148-156.

[21] Breiman L. Bagging predictors[J]. Machine Learning, 1996, 24(2): 123-140.

[22] Breiman L. Random forests[J]. Machine Learning, 2001, 45(1): 5-32.

[23] Dietterich T G. An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting, and randomization[J]. Machine Learning, 2000, 40(2): 139-157.

[24] Banfield R E, Hall L O, Bowyer K W, et al. A comparison of decision tree ensemble creation techniques[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2007, 29(1): 173–180.

[25] Fraz M M, Remagnino P, Hoppe A, et al. An ensemble classification-based approach applied to retinal blood vessel segmentation[J]. IEEE Transactions on Biomedical Engineering, 2012, 59(9): 2538-2548.

[26] Lee B K, Lessler J, Stuart E A. Improving propensity score weighting using machine learning[J]. Statistics in Medicine, 2010, 29(3): 337-346.

[27] Wu B, Abbott T, Fishman D, et al. Comparison of statistical methods for classification of ovarian cancer using mass spectrometry data[J]. Bioinformatics, 2003, 19(13): 1636-1643.

[28] Khalilia M, Chakraborty S, Popescu M. Predicting disease risks from highly imbalanced data using random forest[J]. BMC Medical Informatics and Decision Making, 2011, 11(1): 1-13.

[29] Bao Y, Ke B, Li B, et al. Detecting accounting fraud in publicly traded U.S. Firms using a machine learning approach[J]. Journal of Accounting Research, 2020, 58(1): 199–235.

[30] Wang R, Asghari V, Hsu S C, et al. Detecting corporate misconduct through random forest in China's construction industry[J]. Journal of Cleaner Production, 2020, 268: 122266.

[31] Albrecht W S, Wernz G W, Williams T L. Fraud: bring the light to the dark side of business[M]. New York Irwin Inc, 1995: 15-52.

[32] Bologua G J, Lindquist R J, Wells J T. The accountant's handbook of fraud and commercial crime[M]. John Wiley and Sons Inc, 1993.

[33] Bologna G J, Lindquist R J. Fraud auditing and forensic accounting: new tools and techniques[M]. Wiley, 1995.

[34] 韦琳, 徐立文, 刘佳. 上市公司财务报告舞弊的识别——基于三角形理论的实证研究[J]. 审计研究, 2011(2): 98-106.

[35] 洪荭, 胡华夏, 郭春飞. 基于 GONE 理论的上市公司财务报告舞弊识别研究[J]. 会计研究, 2012(8): 84-90.

[36] Albrecht W S, Romney M B, Cherrington D J, et al. Red-flagging management fraud: A validation[J]. Advances in Accounting, 1986, 3: 323-333.

[37] Beneish M D. Detecting GAAP violation: Implications for assessing earnings management among firms with extreme financial performance[J]. Journal of Accounting and Public Policy, 1997, 16(3): 271-309.

[38] Summers S L, Sweeney J T. Fraudulently misstated financial statements and insider trading: An empirical analysis[J]. Accounting Review, 1998: 131-146.

[39] Lee T A, Ingram R W, Howard T P. The difference between earnings and operating cash flow as an indicator of financial reporting fraud[J]. Contemporary Accounting Research, 1999, 16(4): 749-786.

[40] 阎达五，王建英. 上市公司利润操纵行为的财务指标特征研究[J]. 财务与会计，2001, 10.

[41] 陈信元，张田余，陈冬华. 预期股票收益的横截面多因素分析：来自中国证券市场的经验证据[J]. 金融研究, 2001, 6.

[42] 章美珍. 财务报告舞弊端倪甄别及治理对策[J]. 当代财经，2002, 5.

[43] 朱锦余,高善生. 上市公司舞弊性财务报告及其防范与监管——基于中国证券监督委员会处罚公告的分析[J]. 会计研究, 2007, 241(11): 17-23, 95.

[44] 张筱,高培亮. 我国上市公司审计失败监管——对证监会处罚公告研究的文献综述[J]. 中国内部审计, 2014, 183(9): 93-96.

[45] Eining M M, Jones D R, Loebbecke J K. Reliance on decision aids: An examination of auditors' assessment of management fraud[J]. Auditing: A Journal of Practice & Theory, 1997, 16(2).

[46] Bell T B, Carcello J V. A decision aid for assessing the likelihood of fraudulent financial reporting[J]. Auditing: A Journal of Practice & Theory, 2000, 19(1): 169-184.

[47] Phua C, Lee V, Smith K, et al. A comprehensive survey of data mining-based fraud detection research [EB/OL]. arXiv preprint arXiv:1009.6119, 2010.

[48] 陈亮,王炫.会计信息欺诈经验分析及识别模型[J].证券市场导报，2003, 8: 52-56.

[49] 钱苹,罗玫.中国上市公司财务造假预测模型[J].会计研究,2015,333(7):18-25,96.

[50] Wolpert D H, Macready W G. No free lunch theorems for optimization[J]. IEEE Transactions on Evolutionary Computation, 1997, 1(1): 67-82.

[51] Albashrawi M, Lowell M. Detecting financial fraud using data mining techniques: A decade review from 2004 to 2015[J]. Journal of Data Science, 2016, 14(3): 553-569.

[52] Kotsiantis S B, Zaharakis I, Pintelas P. Supervised machine learning: A review of classification techniques[J]. Emerging Artificial Intelligence Applications in Computer Engineering, 2007, 160(1): 3-24.

[53] Perols J L, Bowen R M, Zimmermann C, et al. Finding needles in a haystack: Using data analytics to improve fraud prediction[J]. The Accounting Review, 2017, 92(2): 221-245.

[54] Viola P, Jones M J. Robust real-time face detection[J]. International Journal of Computer Vision, 2004, 57(2): 137-154.

[55] Kolter J Z, Maloof M A. Learning to detect and classify malicious executables in the wild[J]. Journal of Machine Learning Research, 2006, 7(12).

[56] Panigrahi S, Kundu A, Sural S, et al. Credit card fraud detection: A fusion approach using Dempster–Shafer theory and Bayesian learning[J]. Information Fusion, 2009, 10(4): 354-363.

[57] West D, Dellana S, Qian J. Neural network ensemble strategies for financial decision applications[J]. Computers & Operations Research, 2005, 32(10): 2543-2559.

[58] Zhou Z H. Ensemble methods: foundations and algorithms[M]. CRC Press, 2012.

[59] Skurichina M, Kuncheva L I, Duin R P. Bagging and boosting for the nearest mean classifier: Effects of sample size on diversity and accuracy[C]. International Workshop on Multiple Classifier Systems. Springer, Berlin, Heidelberg, 2002: 62-71.

[60] Bertomeu J, Cheynel E, Floyd E et al. Using machine learning to detect misstatements[J]. Review of Accounting Studies.2021;26(2):468–519.

[61] Dutta I, Dutta S, Raahemi B. Detecting financial restatements using data mining techniques[J]. Expert Systems with Applications, 2017, 90: 374-393.

[62] Wang Y, Yu M, Gao S. Gender diversity and financial statement fraud[J]. Journal of Accounting and Public Policy, 2022, 41(2): 1–43.

[63] Leippold M, Wang Q, Zhou W. Machine learning in the Chinese stock market[J]. Journal of Financial Economics, 2022, 145(2): 64-82.

[64] Geng R, Bose I, Chen X. Prediction of financial distress: An empirical study of listed Chinese companies using data mining[J]. European Journal of Operational Research, 2015, 241(1): 236-247.

[65] He H, Garcia E A. Learning from imbalanced data[J]. IEEE Transactions on Knowledge and Data Engineering, 2009, 21(9): 1263-1284.

[66] 张成浩，李磊，吴德胜. 上市公司财务异常分析[J]. 计量经济学报，2021(3): 706-718.

[67] Perols J L, Lougee B A. The relation between earnings management and financial statement fraud[J]. Advances in Accounting, 2011, 27(1): 39-53.

[68] Rezaee Z, Riley R. Financial Statement Fraud-Prevention and Detection[M]. Hoboken: John&Wiley Sons Inc., New Jersey, 2010.

[69] Sagi O, Rokach L. Ensemble learning: A survey[J]. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 2018, 8(4): e1249.

[70] Fernández-Delgado M, Cernadas E, Barro S et al. Do we need hundreds of classifiers to solve real world classification problems? [J]. Journal of Machine Learning Research.2014;15:3133–3181.

[71] Friedman J H. Greedy function approximation: a gradient boosting machine[J]. Annals of Statistics, 2001: 1189-1232.

[72] Chen T, Guestrin C. Xgboost: A scalable tree boosting system[C]. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2016: 785-794.

[73] Ke G, Meng Q, Finley T, et al. Lightgbm: A highly efficient gradient boosting decision tree[J]. Advances in Neural Information Processing Systems, 2017, 30.

[74] Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: Machine learning in Python[J]. The Journal of Machine Learning Research, 2011, 12: 2825-2830.

[75] Throckmorton C S, Mayew W J, Venkatachalam M et al. Financial fraud detection using vocal, linguistic and financial cues[J]. Decision Support Systems.2015;74:78–87.

[76] Hajek P, Henriques R. Mining corporate annual reports for intelligent detection of financial statement fraud–A comparative study of machine learning methods[J]. Knowledge-Based Systems, 2017, 128: 139-152.

[77] Fawcett T. An introduction to roc analysis[J]. Pattern Recognition Letters, 2006, 27(8): 861–874.

[78] Gepp A, Kumar K, Bhattacharya S. Lifting the numbers game: identifying key input variables and a best-performing model to detect financial statement fraud[J]. Accounting and Finance.2021;61(3):4601–4638.

[79] Lobo J M, Jiménez‐Valverde A, Real R. AUC: a misleading measure of the performance of predictive distribution models[J]. Global Ecology and Biogeography, 2008, 17(2): 145-151.

[80] Wang G, Hao J, Ma J et al. A comparative assessment of ensemble learning for credit scoring[J]. Expert Systems with Applications.2011;38(1):223–230.

[81] Kim M J, Kang D K. Ensemble with neural networks for bankruptcy prediction[J]. Expert Systems with Applications, 2010, 37(4): 3373-3379.

[82] Saeys Y, Inza I, Larrañaga P. A review of feature selection techniques in bioinformatics[J]. Bioinformatics, 2007, 23(19): 2507–2517.

[83] Zhang L, Suganthan P N. Random forests with ensemble of feature spaces[J]. Pattern Recognition, 2014, 47(10): 3429-3437.

附录

附表 1 变量定义表

变量类别	编号	变量名称	变量定义
资产质量	V1	应收款项占比高于同行业	最近一年年末应收款项占流动资产比例处于同行业前 5%的水平，且最近一年占比超过 50%
资产质量	V2	应收款项占比偏高	最近一年年末应收款项账面价值占流动资产比例 60%至 80%
资产质量	V3	应收款项占比极高	最近一年年末应收款项账面价值占流动资产比例大于 80%
资产质量	V4	应收款项近一年增加较快	最近一年年末应收款项较上年末增长率超过 150%,且最近一年年末应收款项占流动资产比例超过 30%
资产质量	V5	应收款项近三年增加较快	最近三年年末应收款项持续增长，且每年增长率均超过 60%,最近一年年末应收款项占流动资产比例超过 30%
资产质量	V6	应收款项近一年增速高于同行业	最近一年年末应收账款较上年末增长率处于同行业前 5%的水平，且应收账款同比增长率超过 80%,且最近一年年末应收款项占流动资产比例超过 30%
资产质量	V7	应收账款占营业总收入比较高	最近一年年末应收账款占同期营业总收入比重超过 100%
资产质量	V8	应收账款占营业总收入比高于同行业	最近一年年末应收账款占同期营业总收入比重处于同行业前 5%的水平，且最近一年占比超过 80%
资产质量	V9	应收账款与营业总收入变动关系异常	最近一年年末应收账款较上年末增长率超过同期营业总收入增长率情况处于同行业前 5%的水平，且最近一年应收账款同比增长超过 100%,且是同期营业总收入同比增长的两倍以上，且两者差额在 10%以上
资产质量	V10	应收账款增加而营业收入下降	最近三年年末应收账款账面余额持续增加而最近三年营业总收入持续下降
资产质量	V11	应收票据占比较高	最近一年年末应收票据占流动资产比例超过 30%

附表 1 变量定义表（续）

变量类别	编号	变量名称	变量定义
资产质量	V12	应收票据占比高于同行业	最近一年年末应收票据占流动资产比例处于同行业前 5%的水平，且最近一年占流动资产比例超过 20%
资产质量	V13	其他应收款占比高于同行业	最近一年年末其他应收款占流动资产比例处于同行业前 5%的水平，且最近一年占流动资产比例超过 10%
资产质量	V14	其他应收款占比较高	最近一年年末其他应收款占流动资产比例超过 20%
资产质量	V15	长期应收款占比高于同行业	最近一年年末长期应收款占非流动资产比例处于同行业前 5%的水平，且最近一年占非流动资产比例超过 10%
资产质量	V16	长期应收款占比较高	最近一年年末长期应收款占非流动资产比例超过 15%
资产质量	V17	预付款项占比极高	最近一年年末预付款项占期末流动资产的比例超过 20%
资产质量	V18	预付款项占比高于同行业	最近一年年末预付款项占期末流动资产的比例处于同行业前 5%的水平，且最近一年占流动资比例超过 10%
资产质量	V19	预付款项占比偏高	最近一年年末预付款项占期末流动资产的比例为 15%至 20%
资产质量	V20	预付款项近三年持续增加	最近三年年末预付款项持续增长，且每年增长率均超过 30%,且最近一年年末预付款项占流动资产比例超过 10%
资产质量	V21	预付款项增加高于同行业	最近一年年末预付款项较上年末增长率处于同行业前 5%的水平，且最近一年年末预付款项占流动资产比例超过 10%
资产质量	V22	预付款项一年大幅增加	最近一年年末预付款项占流动资产比例超过 10%,且预付款项较上年末增长率超过 50%
资产质量	V23	存货占比高于同行业	最近一年年末存货占流动资产比例处于同行业前 5%的水平，且占比超过 30%
资产质量	V24	存货占比偏高	最近一年年末存货账面价值占流动资产比例为 50%至 60%

附表 1 变量定义表（续）

变量类别	编号	变量名称	变量定义
资产质量	V25	存货占比极高	最近一年年末存货账面价值占流动资产比例超过 60%
资产质量	V26	存货增加而营收下降	最近两年年末存货金额持续增加而最近两年营业总收入持续下降，且最近一年年末存货占流动资产比例超过 10%
资产质量	V27	商誉占比高于同行业	最近一年年末商誉占净资产比例处于同行业前 5%的水平，且商誉占比超过 30%
资产质量	V28	商誉占比极高	最近一年年末商誉账面价值占净资产比例超过 80%
资产质量	V29	商誉占比偏高	最近一年年末商誉账面价值占净资产比例为 50%至 80%
资产质量	V30	商誉近两年增加较快	最近两年年末商誉持续增长，且年均增长率超过 50%,且最近一年年末商誉占净资产比例超过 30%
资产质量	V31	商誉近一年大幅增加	最近一年年末商誉较上年末增长率超过 10%或上一年商誉为 0 最近一年新增商誉，且最近一年年末商誉占净资产比例超过 30%
资产质量	V32	近三年内发生资不抵债	最近三年内存在归属于母公司股东净资产为负值的情况
资产质量	V33	近一年内发生资不抵债	最近一年归属于母公司股东净资产为负值
资产质量	V34	净资产大幅下降	最近一年归属于母公司股东净资产较上一年下降 50%以上
资产质量	V35	净资产明显偏低	公司最近一期末归属于母公司股东净资产小于 5000 万元
资产质量	V36	在建工程勾稽异常	最近一年年末预付款项占年度营业总收入比例超过 10%,且年末和下一季度预付款项占在建工程的比例均超过 50%
资产质量	V37	货币资金占比较高且大幅增加	最近一年年末货币资金较去年同期增加超过 90%,且最近一年年末货币资金占流动资产比例超过 30%

附表 1 变量定义表（续）

变量类别	编号	变量名称	变量定义
资产质量	V38	货币资金占比较高且大幅减少	最近一年年末货币资金较去年同期减少超过90%,且最近一年年末货币资金占流动资产比例超过 30%
资产质量	V39	货币资金与短期债务比例异常	最近一年年末货币资金占短期债务比例（短期借款+一年内到期的非流动负债）小于 50%,且最近一年年末短期债务占流动负债比例超过 20%
资产质量	V40	存贷双高	最近一年年末有息负债（短期借款+长期借款+应付债券+年内到期的非流动负债）及货币资金占总资产比例均超过 20%,且利息费用占息税前利润的比例超过 10%
资产质量	V41	递延所得税占比较高	最近一年末递延所得税资产超过净资产的 5%
偿债能力	V42	资产负债率较高	除房地产行业和金融行业外，最近一年年末资产负债率大于 80%
偿债能力	V43	资产负债率高于同行业	最近一年年末资产负债率处于同行业前 5%的水平且最近一年资产负债率大于 70%
偿债能力	V44	资产负债率大幅上涨	最近一年年末资产负债率同比上升超过 30 个百分点，且最近一年资产负债率大于 60%
偿债能力	V45	资产负债率增速高于同行业	最近一年年末资产负债率同比上升处于同行业前 5%的水平且最近一年资产负债率大于 50%
营运能力	V46	应收账款周转低于同行业	最近一年年末应收账款周转率处于同行业后 5%的水平，且最近一年应收账款周转率小于 1.5
营运能力	V47	应收账款周转较慢	最近一年年末应收账款周转率小于 1
营运能力	V48	应收账款周转大幅下降	当期应收账款周转率与上年同期相比下降幅度在 30%以上
偿债能力	V49	流动比率较低	最近一年流动比率低于 0.6
偿债能力	V50	速动比率较低	最近一年速动比率低于 0.5

附表 1 变量定义表（续）

变量类别	编号	变量名称	变量定义
偿债能力	V51	现金比率较低	最近一年现金比率低于 0.1
偿债能力	V52	流动比率低于同行业	最近一年年末流动比率处于同行业后 5%的水平，且最近一年年末流动比率低于 0.8
偿债能力	V53	速动比率低于同行业	最近一年年末速动比率处于同行业后 5%的水平，且最近一年年末速动比率低于 0.4
偿债能力	V54	现金比率低于同行业	最近一年年末现金比率处于同行业后 5%的水平，且最近一年年末现金比率低于 0.2
偿债能力	V55	流动负债率较高	最近一年年末流动负债占总负债比例大于 50%,且资产负债率大于 50%
偿债能力	V56	流动比率大幅下降	最近一年流动比率同比下降超过 0.5,且最近一年年末流动比率低于 1
偿债能力	V57	速动比率大幅下降	最近一年速动比率同比下降超过 0.5,且最近一年年末速动比率低于 0.5
偿债能力	V58	现金比率大幅下降	最近一年现金比率同比下降超过 0.5,且最近一年年末现金比率低于 0.2
偿债能力	V59	流动比率降幅超过同行业	最近一年流动比率比上年同期相比下降程度处于同行业前 5%的水平，且最近一年年末流动比率低于 1
偿债能力	V60	速动比率降幅超过同行业	最近一年速动比率比上年同期相比下降程度处于同行业前 5%的水平，且最近一年年末速动比率低于 0.5
偿债能力	V61	现金比率降幅超过同行业	最近一年现金比率比上年同期相比下降程度处于同行业前 5%的水平，且最近一年年末现金比率低于 0.2
盈利能力	V62	毛利率由负变正	最近一年毛利率由负变为正
盈利能力	V63	毛利率由正变负	最近一年毛利率由正变为负
盈利能力	V64	毛利率近一年大幅提高	最近一年毛利率比上年同期相比提高 15 个百分点以上
盈利能力	V65	毛利率近一年大幅降低	最近一年毛利率比上年同期相比降低 15 百分点以上

附表 1 变量定义表（续）

变量类别	编号	变量名称	变量定义
盈利能力	V66	毛利率增幅远超同行业	最近一年毛利率比上年同期相比提高程度处于同行业前 5%的水平
盈利能力	V67	毛利率降幅远超同行业	最近一年毛利率比上年同期相比下降程度处于同行业前 5%的水平
盈利能力	V68	毛利率为负	综合毛利率为负
盈利能力	V69	毛利率近一年较低	最近一年毛利率低于 10%
盈利能力	V70	毛利率近三年持续较低	最近三年毛利率均低于 15%,且最近三年毛利率低于行业平均值 20 个百分点以上
盈利能力	V71	毛利率近一年远低于同行	最近一年毛利率处于同行业后 5%的水平
盈利能力	V72	毛利率近一年较高	最近一年毛利率超过 80%
盈利能力	V73	毛利率近三年持续较高	最近三年毛利率均高于 60%,且最近三年净利率均超过 30%,且最近三年毛利率高于行业平均值 20 个百分点以上
盈利能力	V74	毛利率近三年长期高于同行业	最近一年毛利率处于同行业前 5%的水平
营运能力	V75	存货周转较慢	存货周转率处于同行业后 5%水平，且最近一年年末存货账面价值占流动资产比例超过 30%
营运能力	V76	存货周转大幅下降	当期存货周转率与上年同期相比下降幅度在 30%以上
偿债能力	V77	有息负债率较高	最近一年年末有息负债（短期借款+长期借款+应付债券+一年内到期的非流动负债）占总资产的比例超过 50%
偿债能力	V78	有息负债率高于同行业	最近一年年末有息负债（短期借款+长期借款+应付债券+一年内到期的非流动负债）占总资产的比例处于同行业前 5%的水平，且该比例大于 20%

附表 1 变量定义表（续）

变量类别	编号	变量名称	变量定义
偿债能力	V79	有息负债率大幅上涨	最近一年年末有息负债（短期借款+长期借款+应付债券+一年内到期的非流动负债）占总资产的比例同比上升超过 10 个百分点，且最近一年该比例大于 30%
偿债能力	V80	有息负债率上涨幅度高于同行业	最近一年年末有息负债（短期借款+长期借款+应付债券+一年内到期的非流动负债）占总资产的比例同比上升处于同行业前 5%的水平，且该比例大于 20%
盈利能力	V81	净资产收益率与行业偏离	最近一年净资产收益率比行业平均水平偏离超过 20 个百分比
盈利能力	V82	净资产收益率大幅增加或下降	最近一年净资产收益率同比增长超过 80%或同比下降超过 50%
盈利能力	V83	营业总收入下滑超过同行业	最近一年营业总收入较上年下滑比例处于同行业前 5%的水平
盈利能力	V84	营业总收入近一年大幅下滑	最近一年营业总收入较上年下滑比例超过 60%
盈利能力	V85	营业总收入近三年持续下滑	最近三年营业总收入持续下滑，且每年下滑幅度超过 20%
盈利能力	V86	营业总收入水平极低	最近一年营业总收入在 5000 万元以下
盈利能力	V87	营业总收入增幅远高于净利润	最近一年营业总收入增幅高于净利润增幅 100 个百分点以上
盈利能力	V88	营业总收入增幅远低于净利润	最近一年营业总收入增幅低于净利润增幅 100 个百分点以上
盈利能力	V89	收入变动趋势偏离行业	最近一年营业总收入变动率高于行业平均变动率 80 个百分点
盈利能力	V90	净利润低于同行业	最近一年净利润处于同行业后 5%的水平（扣非后孰低）
盈利能力	V91	净利润大幅下降	最近一年净利润同比下降超过 80%(扣非后孰低）
盈利能力	V92	净利润近两年连续为负	最近两年净利润连续为负（扣非后孰低）
盈利能力	V93	净利润近年发生过为负	最近三年中存在超过一年净利润为负
现金流管理	V94	近一年经营现金流严重偏离净利润	最近一年经营现金流净额为负，净利润为正，且差额超过净利润的 200%(扣非后孰低）
现金流管理	V95	近三年经营现金流连续偏离净利润	经营活动现金流量净额连续三年为负且净利润连续三年为正（扣非后孰低）

附表 1 变量定义表（续）

变量类别	编号	变量名称	变量定义
现金流管理	V96	支付其他与经营活动有关的现金较高	最近一年支付其他与经营活动有关的现金占经营活动现金流出的比例超过 15%
现金流管理	V97	收到其他与经营活动有关的现金较高	最近一年收到其他与经营活动有关的现金占经营活动现金流入的比例超过 15%
现金流管理	V98	经营现金流大幅波动	最近两年经营活动现金流量净额中一年为负一年为正，且两年的差额超过最近一年经营活动现金流量净额绝对值的 300%
现金流管理	V99	经营现金流降幅较大	最近一年经营活动现金流量净额同比去年下降超过 80%
盈利能力	V100	投资收益占比较高	最近一年投资收益占净利润的比例超过 80%,且绝对值超过 5000 万
盈利能力	V101	投资收益占比高于同行业	最近一年投资收益占净利润的比例处于同行业前 5%的水平且投资收益超过 5000 万（扣非后孰低）
盈利能力	V102	近三年收入下滑而利润增长	公司最近三年净利润（扣非后孰低）持续增长，但收入持续下滑
盈利能力	V103	近三年收入增长而利润下滑	公司最近三年净利润（扣非后孰低）持续下滑，但收入持续增长
盈利能力	V104	巨额亏损	最近一年亏损金额超过 5 亿元（扣非后孰低）
盈利能力	V105	销售费用占比高于同行业	最近一年销售费用占营业总收入的比例处于同行业前 5%的水平，且销售费用率在 10%以上
盈利能力	V106	三费占比较高	最近一年销售、管理和研发费用之和占营业总收入的比例超过 50%
盈利能力	V107	管理费用占比高于同行业	最近一年管理费用占营业总收入的比例处于同行业前 5%的水平，且管理费用率在 10%以上
盈利能力	V108	管理费用占比大幅波动	当期管理费用占营业总收入的比例与上年同期相比的变动 10%以上，且管理费用率在 10%以上
盈利能力	V109	销售管理占比大幅波动	当期销售费用占营业总收入的比例与上年同期相比的变动 10%以上，且销售费用率在 10%以上
盈利能力	V110	三费占比增幅高于同行业	最近一年销售、管理和研发费用之和占营业总收入的比例同比上升幅度处于同行业前 5%的水平，且费用占比超过 20%
盈利能力	V111	三费占比降幅高于同行业	最近一年销售、管理和研发费用之和占营业总收入的比例同比下降幅度处于同行业前 5%的水平，且费用占比低于 20%

附表 1 变量定义表（续）

变量类别	编号	变量名称	变量定义
盈利能力	V112	利息费用占比较高	最近一年财务费用中的利息费用占息税前利润（净利润+利息费用+所得税）的比例超过 50%且息税前利润大于 5000 万元
盈利能力	V113	利息费用发生额较大	最近一年财务费用中的利息费用大于息税前利润（净利润+利息费用+所得税）亏损额的绝对值，且利息费用大于 1000 万元
盈利能力	V114	减值损失占比较高	最近一年减值损失金额占当期营业总成本的比例超过 50%
盈利能力	V115	减值损失占比超过同行业	最近一年减值损失金额占当期营业总成本的比例处于同行业前 5%的水平，且最近一年减值损失金额占当期营业总成本的比例超过 30%
盈利能力	V116	非经常性损益占比极高	当期非经常性损益占净利润的比例在 100%以上，且绝对值超过 5000 万
盈利能力	V117	非经常性损益占比高于同行业	最近一年非经常性损益占净利润的比例处于同行业前 5%的水平且金额超过 5000 万元
盈利能力	V118	非经常性损益占比偏高	当期非经常性损益占净利润的比例为 50%至 100%,且绝对值超过 5000 万
行业环境	V119	同行业平均收入下滑	同行业平均收入连续两年下滑，且下滑总幅度>=20%
行业环境	V120	同行业平均毛利率下滑	同行业平均毛利率连续两年下滑，且下滑总幅度>=50%
盈利能力	V121	近年亏损且业绩微利	去年净利润小于 0 且前年净利润大于 0，最近一年净资产收益率在 0-2%区间。
盈利能力	V122	多年微利	最近三年净资产收益率处于 0-2%区间

注释：表格报告了所用变量的定义与构造规则。其中第一列为变量序号，从 V1 到 V122 进行编列。第二列为变量名称。第三列为变量的详细定义。第四列为变量所属的业务分析类别。

附表 2 变量描述性统计

变量	舞弊样本（N=354)		非舞弊样本(n=8496)		双侧 T 检验	
	均值	标准差	均值	标准差	均值差异	P 值
V1_应收款项占比高于同行业	0.0876	0.2831	0.0505	0.219	0.0371	0.0021***
V2_应收款项占比偏高	0.0791	0.2703	0.0563	0.2304	0.0228	0.0698*
V3_应收款项占比极高	0.0198	0.1394	0.0028	0.0531	0.0169	0.0***
V4_应收款项近一年增加较快	0.0621	0.2418	0.0213	0.1444	0.0408	0.0***
V5_应收款项近三年增加较快	0.0085	0.0918	0.0055	0.0742	0.0029	0.4693
V6_应收款项近一年增速高于同行业	0.0734	0.2612	0.0279	0.1647	0.0456	0.0***
V7_应收账款占营业总收入比较高	0.0706	0.2566	0.0246	0.1549	0.046	0.0***
V8_应收账款占营业总收入比高于同行业	0.0847	0.2789	0.033	0.1785	0.0518	0.0***
V9_应收账款与营业总收入变动关系异常	0.0452	0.208	0.0324	0.177	0.0128	0.1848
V10_应收账款增加而营业收入下降	0.0085	0.0918	0.0168	0.1286	-0.0084	0.2265
V11_应收票据占比较高	0.0169	0.1293	0.0193	0.1376	-0.0024	0.7519
V12_应收票据占比高于同行业	0.0198	0.1394	0.032	0.1761	-0.0122	0.1966
V13_其他应收款占比高于同行业	0.1215	0.3271	0.0328	0.1782	0.0886	0.0***
V14_其他应收款占比较高	0.065	0.2468	0.0165	0.1273	0.0485	0.0***
V15_长期应收款占比高于同行业	0.0395	0.1952	0.0293	0.1687	0.0102	0.2663
V16_长期应收款占比较高	0.0424	0.2017	0.0314	0.1745	0.0109	0.2506
V17_预付款项占比极高	0.0452	0.208	0.021	0.1432	0.0242	0.0023***
V18_预付款项占比高于同行业	0.0932	0.2912	0.0464	0.2103	0.0468	0.0001***
V19_预付款项占比偏高	0.0169	0.1293	0.0199	0.1396	-0.0029	0.6968
V20_预付款项近三年持续增加	0.0085	0.0918	0.0033	0.0573	0.0052	0.1061

附表 2 变量描述性统计（续）

变量	舞弊样本（N=354)		非舞弊样本(n=8496)		双侧 T 检验	
	均值	标准差	均值	标准差	均值差异	P 值
V21_预付款项增加高于同行业	0.0254	0.1576	0.0108	0.1035	0.0146	0.0113**
V22_预付款项一年大幅增加	0.0621	0.2418	0.0335	0.1801	0.0286	0.004***
V23_存货占比高于同行业	0.0424	0.2017	0.0497	0.2173	-0.0073	0.5347
V24_存货占比偏高	0.0537	0.2257	0.0321	0.1764	0.0215	0.0262**
V25_存货占比极高	0.0367	0.1883	0.0481	0.2141	-0.0114	0.3234
V26_存货增加而营收下降	0.0254	0.1576	0.0161	0.126	0.0093	0.1784
V27_商誉占比高于同行业	0.1356	0.3428	0.043	0.2028	0.0926	0.0***
V28_商誉占比极高	0.0282	0.1659	0.0056	0.075	0.0226	0.0***
V29_商誉占比偏高	0.0932	0.2912	0.022	0.1467	0.0712	0.0***
V30_商誉近两年增加较快	0.0169	0.1293	0.0095	0.0972	0.0074	0.1659
V31_商誉近一年大幅增加	0.1215	0.3271	0.0539	0.2258	0.0676	0.0***
V32_近三年内发生资不抵债	0.0198	0.1394	0.0074	0.0858	0.0124	0.0101**
V33_近一年内发生资不抵债	0.0056	0.0751	0.0021	0.046	0.0035	0.1705
V34_净资产大幅下降	0.0339	0.1812	0.0064	0.0795	0.0275	0.0***
V35_净资产明显偏低	0.0141	0.1182	0.0055	0.0742	0.0086	0.0382**
V36_在建工程勾稽异常	0.1977	0.3989	0.0931	0.2906	0.1046	0.0***
V37_货币资金占比较高且大幅增加	0.0989	0.2989	0.084	0.2775	0.0148	0.326
V38_货币资金占比较高且大幅减少	0	0	0	0	0	NA
V39_货币资金与短期债务比例异常	0.1921	0.3945	0.1091	0.3118	0.083	0.0***
V40_存贷双高	0.0932	0.2912	0.0632	0.2433	0.03	0.0242**
V41_递延所得税占比较高	0.0847	0.2789	0.0448	0.207	0.0399	0.0005***

附表 2 变量描述性统计（续）

变量	舞弊样本（N=354)		非舞弊样本(n=8496)		双侧 T 检验	
	均值	标准差	均值	标准差	均值差异	P 值
V42_资产负债率较高	0.0819	0.2746	0.0291	0.168	0.0528	0.0***
V43_资产负债率高于同行业	0.0989	0.2989	0.0292	0.1683	0.0697	0.0***
V44_资产负债率大幅上涨	0.0339	0.1812	0.0117	0.1073	0.0222	0.0002***
V45_资产负债率增速高于同行业	0.0311	0.1738	0.0154	0.1232	0.0157	0.0216**
V46_应收账款周转低于同行业	0.096	0.2951	0.0363	0.1869	0.0598	0.0***
V47_应收账款周转较慢	0.0621	0.2418	0.0199	0.1396	0.0423	0.0***
V48_应收账款周转大幅下降	0.1667	0.3732	0.1194	0.3242	0.0473	0.0075***
V49_流动比率较低	0.048	0.2141	0.0388	0.1932	0.0092	0.3833
V50_速动比率较低	0.1497	0.3573	0.107	0.3091	0.0427	0.0114**
V51_现金比率较低	0.0989	0.2989	0.0387	0.1929	0.0601	0.0***
V52_流动比率低于同行业	0.0593	0.2366	0.0294	0.169	0.0299	0.0014***
V53_速动比率低于同行业	0.0565	0.2312	0.0215	0.1452	0.035	0.0***
V54_现金比率低于同行业	0.065	0.2468	0.0174	0.1308	0.0476	0.0***
V55_流动负债率较高	0.3983	0.4902	0.2728	0.4454	0.1255	0.0***
V56_流动比率大幅下降	0.0113	0.1058	0.0112	0.1052	0.0001	0.9835
V57_速动比率大幅下降	0.0169	0.1293	0.0115	0.1068	0.0054	0.3544
V58_现金比率大幅下降	0.0847	0.2789	0.0427	0.2023	0.042	0.0002***
V59_流动比率降幅超过同行业	0.0226	0.1488	0.0125	0.111	0.0101	0.098*
V60_速动比率降幅超过同行业	0.0141	0.1182	0.0093	0.096	0.0048	0.3589
V61_现金比率降幅超过同行业	0.0593	0.2366	0.0177	0.1317	0.0417	0.0***
V62_毛利率由负变正	0.0226	0.1488	0.0066	0.0809	0.016	0.0005***

附表 2 变量描述性统计（续）

变量	舞弊样本（N=354)		非舞弊样本(n=8496)		双侧 T 检验	
	均值	标准差	均值	标准差	均值差异	P 值
V63_毛利率由正变负	0.0198	0.1394	0.0066	0.0809	0.0132	0.0038***
V64_毛利率近一年大幅提高	0.2486	0.4328	0.1643	0.3706	0.0843	0.0***
V65_毛利率近一年大幅降低	0.274	0.4466	0.1543	0.3613	0.1197	0.0***
V66_毛利率增幅远超同行业	0.0847	0.2789	0.0458	0.209	0.039	0.0007***
V67_毛利率降幅远超同行业	0.1215	0.3271	0.0514	0.2209	0.07	0.0***
V68_毛利率为负	0.0254	0.1576	0.0105	0.1018	0.0149	0.0085***
V69_毛利率近一年较低	0.1328	0.3398	0.0851	0.279	0.0477	0.0018***
V70_毛利率近三年持续较低	0.0085	0.0918	0.0032	0.0563	0.0053	0.093*
V71_毛利率近一年远低于同行	0.1045	0.3064	0.0468	0.2113	0.0577	0.0***
V72_毛利率近一年较高	0.0113	0.1058	0.0134	0.1151	-0.0021	0.7335
V73_毛利率近三年持续较高	0	0	0.0013	0.036	-0.0013	0.4982
V74_毛利率近三年长期高于同行业	0.0706	0.2566	0.0631	0.2431	0.0075	0.5688
V75_存货周转较慢	0.1073	0.31	0.0372	0.1892	0.0702	0.0***
V76_存货周转大幅下降	0.1836	0.3877	0.1283	0.3344	0.0553	0.0025***
V77_有息负债率较高	0.0763	0.2658	0.0254	0.1574	0.0508	0.0***
V78_有息负债率高于同行业	0.1384	0.3458	0.0443	0.2057	0.0942	0.0***
V79_有息负债率大幅上涨	0.1299	0.3367	0.0661	0.2486	0.0638	0.0***
V80_有息负债率上涨幅度高于同行业	0.0311	0.1738	0.0158	0.1246	0.0153	0.0263**
V81_净资产收益率与行业偏离	0.1864	0.39	0.082	0.2744	0.1044	0.0***
V82_净资产收益率大幅增加或下降	0.5932	0.4919	0.3104	0.4627	0.2828	0.0***
V83_营业总收入下滑超过同行业	0.096	0.2951	0.0468	0.2113	0.0492	0.0***

附表 2 变量描述性统计（续）

变量	舞弊样本（N=354)		非舞弊样本(n=8496)		双侧 T 检验	
	均值	标准差	均值	标准差	均值差异	P 值
V84_营业总收入近一年大幅下滑	0.0141	0.1182	0.0071	0.0837	0.0071	0.1274
V85_营业总收入近三年持续下滑	0	0	0.0036	0.0603	-0.0036	0.255
V86_营业总收入水平极低	0.0169	0.1293	0.0062	0.0787	0.0107	0.0152**
V87_营业总收入增幅远高于净利润	0.2712	0.4452	0.0897	0.2858	0.1815	0.0***
V88_营业总收入增幅远低于净利润	0.1808	0.3854	0.1196	0.3245	0.0612	0.0006***
V89_收入变动趋势偏离行业	0.1017	0.3027	0.0357	0.1855	0.066	0.0***
V90_净利润低于同行业	0.2486	0.4328	0.0891	0.2849	0.1595	0.0***
V91_净利润大幅下降	0.3051	0.4611	0.1223	0.3276	0.1828	0.0***
V92_净利润近两年连续为负	0.2486	0.4328	0.0979	0.2972	0.1507	0.0***
V93_净利润近年发生过为负	0.5932	0.4919	0.2577	0.4374	0.3356	0.0***
V94_近一年经营现金流严重偏离净利润	0.1215	0.3271	0.0977	0.2969	0.0238	0.1416
V95_近三年经营现金流连续偏离净利润	0.0452	0.208	0.0302	0.1713	0.0149	0.111
V96_支付其他与经营活动有关的现金较高	0.4492	0.4981	0.3084	0.4619	0.1408	0.0***
V97_收到其他与经营活动有关的现金较高	0.2034	0.4031	0.0816	0.2737	0.1218	0.0***
V98_经营现金流大幅波动	0.3503	0.4777	0.2275	0.4193	0.1228	0.0***
V99_经营现金流降幅较大	0.3277	0.47	0.1877	0.3905	0.1399	0.0***
V100_投资收益占比较高	0.0311	0.1738	0.0411	0.1985	-0.01	0.3505
V101_投资收益占比高于同行业	0	0	0	0	0	NA
V102_近三年收入下滑而利润增长	0	0	0.0012	0.0343	-0.0012	0.5184
V103_近三年收入增长而利润下滑	0.0198	0.1394	0.0137	0.1161	0.0061	0.3352
V104_巨额亏损	0.0847	0.2789	0.0139	0.117	0.0709	0.0***

附表 2 变量描述性统计（续）

变量	舞弊样本（N=354)		非舞弊样本(n=8496)		双侧 T 检验	
	均值	标准差	均值	标准差	均值差异	P 值
V105_销售费用占比高于同行业	0.0621	0.2418	0.0563	0.2304	0.0059	0.6385
V106_三费占比较高	0.0791	0.2703	0.0471	0.2118	0.032	0.0059***
V107_管理费用占比高于同行业	0.1328	0.3398	0.054	0.2261	0.0787	0.0***
V108_管理费用占比大幅波动	0.2486	0.4328	0.1733	0.3785	0.0753	0.0003***
V109_销售管理占比大幅波动	0.1384	0.3458	0.0738	0.2615	0.0646	0.0***
V110_三费占比增幅高于同行业	0.0876	0.2831	0.0314	0.1745	0.0561	0.0***
V111_三费占比降幅高于同行业	0.0932	0.2912	0.0391	0.1938	0.0541	0.0***
V112_利息费用占比较高	0.113	0.317	0.069	0.2534	0.044	0.0015***
V113_利息费用发生额较大	0.0282	0.1659	0.0105	0.1018	0.0178	0.0018***
V114_减值损失占比较高	0.0226	0.1488	0.0014	0.0376	0.0212	0.0***
V115_减值损失占比超过同行业	0	0	0	0	0	NA
V116_非经常性损益占比极高	0.0932	0.2912	0.0441	0.2054	0.0491	0.0***
V117_非经常性损益占比高于同行业	0.065	0.2468	0.0285	0.1664	0.0365	0.0001***
V118_非经常性损益占比偏高	0.0198	0.1394	0.0364	0.1872	-0.0166	0.0992*
V119_同行业平均收入下滑	0.048	0.2141	0.0264	0.1602	0.0217	0.0142**
V120_同行业平均毛利率下滑	0	0	0.0001	0.0108	-0.0001	0.8383
V121_近年亏损且业绩微利	0.0254	0.1576	0.0138	0.1165	0.0117	0.0698*
V122_多年微利	0.0028	0.0531	0.0117	0.1073	-0.0088	0.1237