



南京大學
NANJING UNIVERSITY

研究生毕业论文 (申请硕士学位)

论 文 题 目 基于机器学习的财务舞弊识别模型研究
作 者 姓 名 冯嘉鑫
专 业 名 称 金融
研 究 方 向 金融工程
指 导 教 师 李心丹

2021 年 5 月 8 日

学 号 : MF1815012

论文答辩日期 : 2021 年 5 月 8 日

指导教师 : 

(签字)

**Research on financial fraud identification model
based on machine learning**

BY
Feng Jiaxin

Supervised by
Professor Li Xindan

School of Management Science and Engineering
Nanjing University

May 2021

南京大学学位论文原创性声明

本人郑重声明：所呈交的学位论文，是本人在导师的指导下，独立进行的研究工作所取得的成果。尽我所知，除文中已经注明引用的内容外，本论文不包含任何其他个人或集体已经发表或撰写过的作品成果。对本文的研究做出重要贡献的个人和集体，均已在文中以明确方式说明并且表达了谢意。本人完全意识到本声明的法律结果由本人承担。

研究生签名: 冯嘉鑫 日期: 2021年5月20日

南京大学学位论文使用授权声明

本学位论文作者同意学校保留并向国家有关部门或机构送交学位论文的复印件和电子文档，可以采用影印、缩印或扫描等复制手段保存论文。本文电子文档的内容和纸质论文的内容相一致。除在保密期内的论文外，允许论文被查阅和借阅，可以公布(包括刊登)论文的全部或者部分内容。论文的公布(包括刊登)授权南京大学研究生院办理。

导师签名: 李冉 研究生签名: 冯嘉鑫 日期: 2021年5月20日

南京大学研究生毕业论文中文摘要首页用纸

毕业论文题目： 基于机器学习的财务舞弊识别模型研究

金融 专业 2018 级硕士生姓名： 冯嘉鑫

指导教师（姓名、职称）： 李心丹 教授

摘要

财务报告是投资者了解上市公司经营情况的重要来源，所以金融市场对财务报告质量有更高的要求。上市公司一旦有动机通过掩饰财务报表的手段欺瞒投资者并且付诸行动，舞弊就伴之而来。财务舞弊带给债权人和投资人的损失众目昭彰。完善审计和会计制度是各国监管机构重点采取的措施，我国监管机构也在不断地完善相关法律法规。然而财务舞弊仍然层出不穷，所以仍是监管部门和广大投资者重点关注的问题。

自 2019 年至 2020 年 4 月，在 A 股市场上 22 家上市公司涉嫌财务舞弊，其中，千亿市值中药行业龙头康美药业对 500 亿元的资产进行了造假，成为 A 股史上金额规模最为庞大的财务造假舞弊案。2015 年到 2019 年康得新被查出虚增利润总额累积达 110 多亿人民币，成为 A 股史上涉案利润最多的造假案。对上市公司财务舞弊的研究，无论对于公司经营治理、资本市场的稳健运行，监管职能提升，还是在投资者权利保护方面都存在着明显的益处，综合运用多种方法构建一个高效及时的财务舞弊识别模型就尤为重要。

本文以 2000 年 1 月至 2020 年 12 月因财务舞弊行为受到违规处罚舞弊上市公司年度财务报告做为舞弊样本，拟进行以下主要方面的研究：(1) 多维度的财务舞弊识别指标构建。使用分层的分析方法，具体来说，分为两个维度，第一个维度建立财务报告可信度指标体系，以经典理论舞弊三因素理论为框架构建指标体系，衡量公司发生舞弊的可能性；第二个维度建立财务异常指标体系，从公司会计信息质量、营运能力、盈利能力、偿债能力、变现能力、财务风险水平和其他财务指标 7 个方面构建，衡量公司财务健康程度。(2) 本文选取了 Logistic 回归模型、SVM 支持向量机、CART 决策树和 LightGBM 作为最终财务舞弊识

别模型的候选，并且用非线性主成分分析法对数据降维，对上市公司的财务舞弊行为进行预测与识别，分别在连续舞弊数据集和首次舞弊数据集上对机器学习模型能力、财务报告可信度模型和财务异常模型对比分析，并进而分析在预测过程中较为重要的公司特征。

最终得出以下结论：(1) 机器学习模型能力对比：支持向量机模型是综合财务舞弊识别模型的最佳选择，逻辑回归模型在基于连续舞弊样本的财务报告可信度模型中表现较好。(2) 模型对比：财务报告可信度模型是预警上市公司首次舞弊行为的更优选择。综合模型是甄别上市公司连续财务舞弊行为的更优选择。(3) 指标重要性：在识别公司财务舞弊行为时，可以重点关注舞弊前应计方向、舞弊前三年曾被标出非标意见、资产负债率这三个指标，对于连续舞弊行为和首次舞弊行为的识别都具有较好的效果。

关键词：财务舞弊、识别模型、舞弊三角理论、机器学习

南京大学研究生毕业论文英文摘要首页用纸

THESIS : Research on Early Warning of Financial Fraud Risk Based on
Machine Learning

SPECIALIZATION: Finance

POSTGRADUATE: Feng Jiaxin

MENTOR: Professor Li Xindan

Abstract

Since the financial report is the medium of transmitting economic information, the listed companies constantly require the integrity of information disclosure, fraud has been accompanied by it. The losses caused by financial fraud to investors and creditors are obvious to all. Although the supervision and governance institutions and other relevant departments of various countries have issued various policies and regulations for many times, and constantly improve the accounting and auditing standards, financial fraud is still a problem that attracts the attention of regulatory authorities and investors.

From 2019 to April 2020, 22 listed companies in the A-share market were suspected of financial fraud. Among them, Kangmei pharmaceutical, the leader of the traditional Chinese medicine industry with a market value of 100 billion yuan, fabricated 50 billion of assets, becoming the largest financial fraud case in the history of A-share. Kangdexin has been investigated for four consecutive years since 2015, with a total profit of more than 11 billion, making it the largest profit fraud case in the history of a shares. Therefore, the research on financial fraud of listed companies has obvious benefits for corporate governance, stable operation of capital market, improvement of regulatory functions and protection of investors' rights. In this case, it is particularly important to construct an efficient and timely financial fraud identification model by using a variety of methods.

This paper takes the annual financial report of listed companies subject to the punishment of financial fraud from January 2000 to December 2020 as the fraud sample.

The following main aspects are proposed: (1) The construction of multi-dimensional financial fraud identification index. The first dimension establishes the financial report reliability index system, and constructs the index system based on the classic theory of fraud theory, which measures the possibility of fraud; The second dimension establishes the financial abnormal index system, which measures the financial health of the company from seven aspects: the quality of accounting information, the operation ability, the profitability, the solvency, the liquidity, the financial risk level and other financial indicators.(2)This paper selects Logistic regression model, SVM, Cart decision tree and LightGBM as candidates for the final financial fraud identification model. The paper uses the nonlinear principal component analysis method to reduce the dimension of data, forecasts and identifies the financial fraud behavior of listed companies, and respectively predicts and identifies the machine learning model capability in the continuous fraud data set and the first fraud data set. The financial report reliability model and financial anomaly model are compared and analyzed, and then the more important corporate characteristics in the process of forecasting are analyzed.

Finally, the following conclusions are drawn: (1) Comparison of machine learning model ability: support vector machine model is the best choice for comprehensive financial fraud identification model, and logistic regression model performs better in the financial report credibility model based on continuous fraud samples.(2)Model comparison: the financial report credibility model is a better choice for early warning the first fraud of listed companies. The comprehensive model is a better choice to screen the continuous financial fraud of listed companies.(3)Importance of indicators: when identifying corporate financial fraud, we can focus on three indicators: accrual direction before fraud, non-standard opinions marked three years before fraud, and asset liability ratio, which have good effect on the identification of continuous fraud and first-time fraud.

Keywords: Financial fraud, identification model, fraud triangle theory, machine learning

目录

摘要.....	I
Abstract.....	III
第一章 绪论.....	1
1.1 研究背景.....	1
1.2 研究意义.....	2
1.3 研究内容和框架.....	3
1.4 研究创新点和不足.....	4
第二章 理论及研究综述.....	5
2.1 财务舞弊概念界定及成因理论.....	5
2.1.1 财务舞弊概念界定.....	5
2.1.2 财务舞弊的成因理论.....	5
2.2 财务舞弊的特征研究.....	8
2.2.1 国外研究综述.....	8
2.2.2 国内研究综述.....	9
2.3 财务舞弊识别模型研究.....	10
2.3.1 国外研究综述.....	10
2.3.2 国内研究综述.....	12
2.4 文献评述.....	13
第三章、实证研究方案设计.....	14
3.1 实证研究整体思路.....	14
3.2 样本选取.....	14
3.3 指标选取.....	16
3.3.1 财务报告可信度指标构建.....	16
3.3.2 财务异常指标构建.....	21
3.4 模型介绍.....	23
3.4.1 Logistic 回归模型	24
3.4.2 支持向量机.....	25
3.4.3 CART 决策树	26
3.4.4 LightGBM.....	27
3.4.5 非线性主成分.....	28
第四章 实证分析.....	29

4.1 评价指标的构建.....	29
4.1.1 总精度指标（accuracy）	29
4.1.2 查准率指标（precision）	30
4.1.3 查全率指标（recall）	30
4.1.4 F 度量指标（F-score）	30
4.2 特征多重共线性检验.....	31
4.2.1 财务报告可信度指标.....	31
4.2.2 财务异常指标.....	32
4.2.3 综合指标.....	33
4.3 主成分分析及非线性指标的构建.....	35
4.4 模型实证结果.....	38
4.4.1 基于首次舞弊样本.....	39
4.4.2 基于连续舞弊样本.....	42
4.4.3 模型中的重要指标.....	45
4.5 模型实证结果分析.....	48
第五章 研究结论及展望.....	49
5.1 研究结论.....	49
5.1.1 机器学习模型识别能力.....	49
5.1.2 财务报告可信度模型、财务异常模型和综合模型对比.....	49
5.1.3 财务舞弊预警模型中的特征重要性.....	50
5.2 展望.....	51
参考文献.....	52
致谢.....	56

第一章 绪论

1.1 研究背景

20世纪80年代，证券交易市场在中国诞生。在股市经历了几轮牛熊市后，2000年经济迸发新活力，人们也越来越关注股票市场，研究和预估股票市场的热情从未消退过。在股票市场上投资时我们除了考虑宏观因素外，更应该关注公司自身的价值，尤其是企业的盈利能力，而这在财务报表中得到了充分的体现。财务报表汇总反映了企业在过去一段时间内的整体经营成果和财务状况，为投资者决策提供信息依据。根据会计准则的规定，财务报表附注应当披露可能对公司预估价值产生消极影响的消息，对公司估值产生正向影响的消息谨慎披露以免对投资者造成误导。但在这种情况下，诚实守信可能导致公司声誉受损、股价大幅下跌、公司高级管理人员财富流失以及员工失业等问题。在这种情形下，上市公司为了维护自己的声誉，保证自己的股价稳定不下跌或为了通过融资获得现金，对财务报表数据修饰造假是最有效的手段。这种失信行为会使财务报表使用者获取虚假信息，做出错误判断，最终蒙受巨大损失。

2020年4月，瑞幸咖啡宣布承认财务造假，虚构交易20多亿元人民币，最终遭受“18个月”退市的命运，触发多次熔断机制而导致停止交易，给投资者带来了惨重损失。随后受到了中美相关监管机构的巨额罚款。2020年，瑞幸咖啡从纳斯达克正式退市，虚构交易细节随后曝光，让人细思极恐。为了驱逐利益，开启了长达一年的系统性造假工程，与这起财务造假案件相关联的企业多达四十家。

诸如此类的财务造假案件绝非个例。据证监会报告，自2019年至2020年4月，在A股市场上22家上市公司涉嫌财务舞弊，证监会对18个典型案件进行了禁止上市处罚，6起被移送公安机关处理。其中千亿市值中药行业龙头企业康美药业对500亿元的资产进行了造假，成为A股史上规模最为庞大的财务造假舞弊案。康得新被查出自2015年以来连续四年虚增利润总额累积达110多亿，成为A股史上涉案利润最多的造假案。

尽管我国资本市场的相关法律和监管不断完善，但是根本上杜绝财务舞弊事件的发生还是空中楼阁，尤其是在近些年我国宏观经济整体承压，资本市场运行低迷的情况下更是如此。为了在企业舞弊早期发掘舞弊迹象，尽早的规避财务舞弊风险，避免投资者财产安全受到威胁，全世界的学者根据行业情况的变化选择不同的变量指标，试图建立更加有效地财务舞弊甄别模型。财务舞弊识别模型作为一种辅助工具是对传统人工识别方法补充，可以避免过度依赖审计人员和专家判断，提前预警财务舞弊风险，避免演化成财务舞弊危机，起到防患于未然的作用。

1.2 研究意义

在经济下行压力加大、股票市场不稳定性增强的趋势下，上市公司更容易出现财务造假行为，而这种违规行为将扰动资本市场的规范运行，损害投资者的自身权益。因此，研究上市公司财务舞弊识别模型对于整个市场具有重要的现实意义和理论意义。

(1) 现实意义

其一，夯实上市公司规范发展。可以促进上市公司健全内部法制机制，减少有意或无意的财务造假行为，使企业在资本市场上良性健康发展。

其二，正确反映公司状况，提高资本市场效率。上市公司真实信息披露对整个市场的信息传导效率具有重要影响，减少信息不对称情况发生。对财务舞弊行为的规范是资本市场诚信制度中重要的一环，可以降低资本市场预期风险。

其三，提升监管效果。以往对财务舞弊的认定往往是基于审计人员的主观经验判断，审计人员在对三份财务报表之间逻辑关系分析和财务指标对标分析中，找出矛盾和可疑点，从而对财务舞弊进行甄别。但是，这种方法也存在一些缺点，比如消耗时间长、审计成本高、错误和遗漏率高、严重依赖个人经验和判断、缺乏统一标准等。因此，引入更高效、更准确的算法和设计，建立更有效的财务舞弊模型有利于各监管机构对公司财务造假的监管和预警，为投资者在股票市场的投资保驾护航。

其四，有利于保护报表信息使用者的合法权益。对于市场参与者来说，投资者投资决策最重要的依据就是公司公告、市场信息和公司财务报表数据，篡改的

财务信息往往使公司看起来更有前景。由于信息不对称，再加上投资者自身的专业水平和局限性，导致投资者往往只能看到这些表面信息而忽视企业财务报表存在的问题，造成投资损失。对于上市公司而言，有效控制财务舞弊现象在市场上的蔓延对所有企业是公平的，如果无法控制财务舞弊公司的蔓延，将会导致市场引发“劣币驱逐良币”，对各家公司都是损害。同时造假信息也会使公司内部决策产生偏差。

(2) 理论意义

本论文拟运用机器学习的方法对舞弊识别模型进行优化，即采用优化信息质量和分类技术的改进应用的双重视角对模型进行优化，从而对上市公司财务舞弊行为更准确的预测和识别，并在预测识别过程中分析对财务舞弊行为的重要财务和非财务特征。在分层模型的基础上逐一优化并判别财务舞弊识别效率，并建立综合舞弊识别模型进行优化是对国内文献的补充，为后续学者进一步研究提供一定的参考。

1.3 研究内容和框架

本文共分为五章，各章主要内容如下：

第一章为引言。主要介绍本文的研究背景、研究意义、研究内容和框架以及论文的创新点和不足。

第二章为财务舞弊相关理论及国内外文献评述。首先，对与财务舞弊相关的概念进行辨析；其次，介绍在经济理论和财务理论中对于解释财务舞弊行为的经典成因理论；再者，梳理国内外相关文献，对财务舞弊特征、财务舞弊模型进行文献评述。

第三章为实证研究方案设计。主要包括样本选取、指标选取和模型介绍三部分。本章会先介绍样本选取和筛选的过程，呈现最终样本筛选结果，并按照特定的筛选条件将舞弊样本与非舞弊样本进行配对；其次，对财务报告可信度指标以及财务异常指标筛选归类，综合考虑财务及非财务指标、静态指标和动态指标。再对实证分析中使用的 Logistic 回归模型、SVM 支持向量机、决策树、LightGBM 和非线性主成分作简要介绍。

第四章为实证分析部分。在本章中，通过对指标多重共线性的检验，使用非线性主成分分析法对数据降维，然后基于机器学习构建财务报告可信度模型和财务异常模型，分别对连续舞弊样本集和首次舞弊样本集进行训练和测试，训练和测试完后收集样本实验结果，通过实现构造的评价指标对模型识别能力比较，最后构架综合指标模型，观察是否可以通过指标的综合提高识别舞弊风险的能力。

第五章为本文的总结和相关建议。本章中先对本文的机器学习模型识别能力、模型对比进行分析，并进而分析在预测过程中较为重要的公司特征。

1.4 研究创新点和不足

本文以 2000 年至 2020 年我国 A 股因财务舞弊（虚构利润、虚增资产）的上市公司为样本，搭建指标体系，对机器学习对舞弊的识别能力以及指标特征作出分析，主要存在以下创新点：

其一，不同于以往直接将全部指标纳入模型中的研究方法，将分层分析方法运用到构建财务舞弊指标体系中，并分别检验每一层次的识别效果。将指标分为财务报告可信度指标和财务异常指标，分别识别财务舞弊可能性和财务报告健康程度。在指标选取时综合考虑财务指标和非财务指标、静态指标和动态指标。

其二，不局限于以往只研究连续舞弊样本集，将连续样本集与首次舞弊样本集分别研究，探索模型的侦查效果和预警效果，并且采用 2000-2020 年最新的财务舞弊数据。

由于笔者能力、研究思路及实际条件的限制，本文也在创作中也存在着不足之处。首先，由于我国对于上市公司财务舞弊从舞弊行为发生到实际查处的时滞较长，平均时滞为 3 年，因此，数据库中不可避免的存在一部分样本数据的缺失，尤其是近些年发的舞弊行为；其次，简单粗暴地将两个指标混合在一起的综合模型有些生硬，在后续研究中，如何将财务报告可信度模型和财务异常模型深度融合，提升综合模型识别效率是下一阶段研究重点。

第二章 理论及研究综述

在本章会对财务舞弊相关理论及国内外文献评述。首先，对财务舞弊概念进行界定；其次，介绍在经济理论和财务理论中对于解释财务舞弊行为的经典成因理论；再者，梳理国内外相关文献，对财务舞弊特征、财务舞弊模型进行文献评述。

2.1 财务舞弊概念界定及成因理论

2.1.1 财务舞弊概念界定

财务舞弊行为是公司为了达到某种目的有意识地财务欺诈，是财务工作中有预谋的活动，从而导致财务报表中的财务数据被篡改或被掩饰，从而影响财务报表使用者的投资判断。国内与国外对财务舞弊定义表达方式多样，但是侧重点基本一致，通过总结，我们可以发现财务舞弊具有以下几个特点：（1）财务舞弊的实施者为企业管理层连同企业员工或第三方；（2）财务舞弊的实施对象为财务报表数据；（3）财务舞弊是公司为达到某种目的的故意、有计划的欺诈行为。舞弊行为主要包含：会计记录和凭证的变造、伪造；资产的侵占；对财务信息做虚假报告；虚增利润；虚构资产；偷逃税金；蓄意误用会计原则等。《独立审计具体准则第8号》文件中也指明了舞弊的概念，认为其指的是有意使得会计报表虚假反映的行为，其是法律层面的广义概念，审计师对于是否发生舞弊行为的判断没有硬性要求，但是要求其关注导致财务报表舞弊的重大问题。

2.1.2 财务舞弊的成因理论

二十世纪九十年代，国外专家主要立足舞弊产生的原因对财务报表舞弊进行研究，基于此形成了几大具有代表性的经典成因理论，并且根据因素数量分为四大成因理论，即二因素论（冰山理论）、三因素论（舞弊三角理论）、GONE理论（四因素论）和风险因子理论（多因素理论）。

(1) 舞弊冰山理论

冰山理论，也叫做二因素理论，是 G.Jack Bologna 和 Lindquist Robert J.一起建立的，他们指出，财务舞弊就好比海中冰山，并没有完全的从海面下透露出来，就好像是危险和威胁并未真正的显现。该理论依据舞弊的结构和行为，划分了其海平面上下部分，海平面下的部分是舞弊行为考虑部分，是导致舞弊产生的内在原因，难以发觉，当被主体有意识的隐瞒、掩藏时，其更是很难被发现。海平面上的部分是舞弊结构考虑部分，是舞弊发生的表层原因，包括了经营状况、财务状况、内部控制、投融资状况、治理结构、股票市场表现等，往往是公开披露可以获取的信息。其涉及的因素大多为主观层面的，比如，管理层持有的理念、行为特征、道德观念、贪婪程度等。公司结构层面出现问题易识别，而如果产生涉及主体主观情绪和观念的行为部分更容易被掩饰，反而存在更加巨大的潜在威胁。因此，结合这一理论，行为方面的原因发挥主要作用，是根本动因。所以，在对舞弊原因进行分析时，要充分考虑结构和行为要素，并重点关注行为因素。

(2) GONE 四因素理论

GONE 理论也叫做四因素理论，1993 年，Bologua 建立了这一理论，在这一理论中，四个字母分别代表了舞弊产生的一项原因，“G”代表的是“Greed”，“O”代表的是“Opportunity”，“N”代表的是“Need”，“E”代表的是“Exposure”，分别代表贪婪、机会、需要和暴露四个因子。这些因子共同决定了舞弊行为发生与否，也就是说当企业存在“上市”或“债券融资”某种想法且希望隐藏企业的真实财务状况来吸引更多的投资时，并且舞弊者本身具有贪欲，只要有机会，例如内部控制制度不完善或者外部监管松懈等，只要成本不高且被发现的可能性很小时，舞弊风险就会加大。暴露主要是指在财务报表审计中信息披露和监管机构对潜在违规者的处罚力度，如果舞弊行为被暴露的可能性低或者处罚力度较弱，那可能对潜在违规者无法构成威胁，会增加舞弊风险。贪婪就是指潜在违规者道德水平低下的主观行为会导致舞弊风险的增加。

随后，Bologua 等建立了舞弊风险理论，从组织架构出发，将舞弊风险因子划分为个别风险因子和一般风险因子，前者受到动机和道德品质的影响，后者受到暴露因子和机会的影响。娄权（2004）丰富和发展了 GONE 理论，建立了和

我国实际情况相符的舞弊四因素理论。洪荭（2012）结合其他理论，进一步完善了基于 GONE 理论财务舞弊行为的形成机理。

（3）舞弊三角理论

1995 年，史蒂文·阿伯雷齐特(W.Steve Albrecht)建立了舞弊三因素理论。这一理论指出，之所以会产生财务舞弊，主要是借口、机会、压力因素彼此作用之后导致的。没有上述因素的话，企业就不会有舞弊行为。过高的利润指标设置或获取利益驱使，迫使行为主体产生压力，可能包括经济压力、工作压力和其他压力等。机会是指客观环境与条件，内部控制的缺失为员工留了钻空子的机会而没有良好的外部监督环境，增加了实施舞弊成功的可能性。

Wilk 和 Zimbelman (2004) 通过对被审计单位的舞弊行为进行分析，通过各大会计师事务所对舞弊公司进行评估，实证分析结果显示压力因素是解释企业财务舞弊行为的重要动因之一。戴丹苗和刘锡良（2017）当企业面临资金压力和退市风险时，只要环境状况不稳定就会导致强烈的舞弊动机。Murphy 和 Dacin(2011) 研究表明舞弊者的心暗示是发生舞弊行为的主要成因之一，就是舞弊者在舞弊过程中认为可以将舞弊行为掩盖过去或者认为无法被发现的心理。Cohen (2012) 实证证明财务舞弊借口因素会受到生活态度和价值观的影响。崔东颖，胡明霞（2019）研究表明，利益的驱使、市场监管薄弱以及公司不完善的治理管理机制都可能会导致公司舞弊行为的发生，在面临市场竞争激烈时尤其如此。

（4）舞弊风险因子理论

舞弊风险因子理论是 G. Jack. Bologana & Lindquist Robert 在“GONE 理论”的基础上进行的拓展。该理论认为，风险因子类别为一般和个别。后者很难被外界影响和掌控。前者可以被外界影响和掌控，包含受罚力度和性质等。舞弊行为的成因往往是复杂的，不仅是上述理论中的成因因素，还包括其他因素，比如说各类的心理因素、环境因素和制度因素。所以我们在构建舞弊甄别模型时要考虑多种因素。

2.2 财务舞弊的特征研究

2.2.1 国外研究综述

(1) 财务指标

对于财务舞弊识别影响因子的分析，国外学者在 20 世纪 80 年代就开始了研究。其中最著名的指标体系之一就是 Albrecht (1986) 构建的红旗指标，是作者通过调查问卷的方法实证检验 87 个财务指标之后得到的结论，其研究表明，其中有些财务指标能够对财务舞弊行为进行有效识别。这是首次全面整理了影响财务舞弊的因素，在此之后，有些专家继续使用了红旗理论的结构，对影响财务舞弊的财务指标进行进一步总结和分析。Beaver (1996) 通过测试 30 多个财务指标，实证结果显示对现金流量指标和总负债的监测是甄别财务舞弊行为的有效方法。这是国外学者第一次使用现金流量指标来甄别财务舞弊行为。Leet al(1996) 表示财务舞弊行为能够被盈余现金流量指标识别出来，并且与应收账款周转率存货周转率等其他的财务指标结合研究，是能够有效的提高识别效率。Benish(1997) 指出，通过比较财务舞弊和非财务舞弊的公司数据资料能够了解到，财务杠杆程度可以作为预警企业是否有财务舞弊情况的参考指标。

(2) 非财务指标

初期文献基本上都是围绕财务因素开展研究，但是当内部控制制度逐渐普及和全面发展滞后，识别财务舞弊行为逐渐引入审计结构、监事董事会特征、公司的内控制度、股权分红结构、管理层基本情况等指标。Beasley (1996) 认为公司内部审计制度和审计机构的建立对于降低舞弊行为发生可能性并没有显著效果，而独立董事制度可以显著降低舞弊行为发生的可能性。而 Dechow (1996) 的研究则与之持相反观点，他认为公司内部审计制度和审计机构可以显著减少公司财务舞弊行为的发生。Bell and Carcello (2000) 研究表明，公司成长迅速但是盈利能力欠缺、内部控制制度不完善、制定过高的利润指标、诚信缺失、公司性质等都是影响财务舞弊的重要因素。Anderson (2004) 研究公司治理层面指出，影响财务舞弊行为的因素还包含董事会的持股规模、董事会会议，董事会持股比例和董事会会议的频率，如果在一定阶段内比较高的话，公司财务舞弊发生的概率都能够大大减少。Persons (2006) 的研究表明董事会规模较大，独立程度较高，董

事制度较为完善时，公司发生财务舞弊行为的可能性显著降低。Bruno Frey(2013)再次强调独立董事制度对于财务舞弊行为控制的重要性，独立董事会人数较多，任期长且独立性较强的公司财务舞弊的可能性较低。Bahram(2014)通过研究欧美财务舞弊公司，例如安然、皇家阿霍德、帕玛拉特等企业，发现企业所有权结构、公众媒体评价、治理结构和监管制度、法律制度等对财务舞弊行为有显著影响。Dechow(2011)在统计财务舞弊和非财务舞弊公司的特点之后了解到，在美国1982–2005年被证监会处罚的896家舞弊公司中，一般都是高层管理者操控了舞弊行为，占比50%以上，审计师占比较低，约为15%。

2.2.2 国内研究综述

(1) 财务指标

我国研究公司识别财务舞弊和发达国家相比较而言，开始较晚。国内专家在参考国外研究资料的基础上，主要注重研究挖掘财务指标来识别舞弊。方军雄(2003)研究财务舞弊指标的起步比较早，通过对上市公司的公开信息研究分析，发现识别公司舞弊的财务指标主要包含八个，其中，发挥较大影响作用的是速动比率、权益乘数。许存兴(2013)研究了现金流状态对财务舞弊的影响。通过建立与现金流相关的几个财务指标，将舞弊公司与非舞弊公司1:1配比研究发现，该公司库存现金(持有现金)越多，越容易发生财务舞弊，这是由于持有现金多存在舞弊机会，可舞弊的空间大，不易被察觉。通过分析企业的财务数据，对企业的外在环境、发展潜力和盈利状况进行描述，并根据此分类归纳可以甄别舞弊的财务指标。王泽霞和谢冰(2010)研究指出，通过企业的偿还债务能力、盈利能力、营运能力等指标，能够有效识别其舞弊行为，其中，对舞弊行为最为敏感的指标是营运能力，除此之外，存货周转率、应收账款周转率等指标也有较为明显的影响。李清和任朝阳(2016)在总结上市公司三十个财务造假风险指标之后，通过对这些风险指标归纳分析，发现上市公司如果具有较大的盈利压力或缺乏流动现金的话，舞弊行为就可能出现，这些公司通常在经营现金流量、存货、应收账款、业务收入和成本上面存在较大的舞弊动机。

(2) 非财务指标

我国针对非财务指标有以下研究。杨薇和姚涛（2006）通过研究指出，公司财务舞弊会受到大股东利益一致性、管理层独立性、股权集中度的影响。闵亮，陈婷（2007）主要研究管理层财务造假，发现管理层财务舞弊的原因大多是为了提高持股比例和提升薪酬待遇，增加股东持股比例和提高管理层薪资待遇会减少财务舞弊发生的可能性。王泽霞和谢冰（2008）在 1998–2008 年中国证监会违规事件中，发现审计质量与财务舞弊的内在联系，在审计质量中，审计意见作为审计报告的最终结果，曾出现非标意见的公司财务造假的可能性显著提高。周继军和张旺峰（2011）实证发现，属于内部环境因素的内部控制制度得到完善可以有效抑制管理层财务舞弊，作为内部控制一环的治理结构的完善健全对财务舞弊的控制更加有效。但是，程安林等（2013）的研究对内部控制和财务舞弊是否存在联系产生新的解释，通过对它们之间存在关系的实证研究表明，财务舞弊与内部控制制度的完善程度并不存在显著相关，内部控制制度的完善并不能有效遏制财务舞弊行为的发生。熊方军和张龙平（2016）财务指标和非财务指标都是财务舞弊识别的关键因素，同时也认为，指标的波动性是识别财务造假的有效信息。因此，在审计过程中，需考量这一指标的波动情况。

2.3 财务舞弊识别模型研究

2.3.1 国外研究综述

国外的财务舞弊识别的方法和技术目前已经趋于成熟，识别财务舞弊的模型十分成熟。早在二十世纪八十年代左右，就有学者使用描述性统计、调查问卷等方式识别财务舞弊行为。随着网络技术的全面发展，逻辑回归模型、数据挖掘模型、多元分析模型、神经网络模型运用广泛。尤其神经网络技术以及 Logistic 回归模型运用于财务舞弊识别范围最广。

Loebbecke 和 Willingham（1988）搭建了 L/W 舞弊风险分析模型，其能够分析企业的舞弊动机。Loebbecke、Eining 和 Willingham（1989）在上述模型的基础上，开展了实证研究，得到的研究结论和前人是相符的，他们结合各个舞弊风险因素出现于舞弊案例中的次数，对模型因子对财务舞弊的辨别敏感度进行试

验，发现大部分舞弊诱因都在这一模型的三类诱因中。鉴于此，逻辑回归模型被广泛使用。Bell, Szykowny and Willingham(1991)又实证验证了 L/W 模型。在此之后，Persons 加入更多样本数据研究指出，有效识别舞弊行为的指标主要包含：流动资产的占比、财务杠杆的高低、资本流通率的利用情况等。Beasley(1996)运用逻辑回归模型进行分析，得出在舞弊企业和非舞弊公司之间，独立董事比例、持股比例、任期和内部控制部门的职能有效性方面存在显著差异。Eining、Jones 、Loebbeck (1997)在 Loebbecke and Willingham(1988)的 L/W 模型基础上，创建了基于审计人员交流的逻辑回归模型的整合专家系统。发现以下识别效率排序：整合专家系统模型>逻辑回归模型>风险因素清单。Summers 和 Sweeney(1998)应用分层逻辑回归分析研究了企业财务舞弊和内部交易的关系，并构建了模型。其中包含三个逻辑回归模型估计，它们彼此是独立的：第一个逻辑回归模型通过审计师变动、ROA、Z 积分等财务报表变量对公司是否舞弊进行区分；第二个逻辑回归模型通过总交易次数、舞弊期间内部人销售行为金额等内部交易、净销售数量和净销售额、股票次数和数量变量来对公司是否舞弊进行划分。实证研究表明，内部交易变量的加入强化了解释舞弊行为发生概率的力度，也提升了区分的精准性。Bell and Carcello (2000) 研究指出，显著影响财务舞弊的因素包含营运能力、经营业绩、治理结构等。其中逻辑回归分类更为精准。

随着网络技术的全面发展，神经网络、文本识别技术、数据挖掘技术运用于财务舞弊识别的研究也不断增多，Kirkos (2008) 等选择英国的 338 家企业作为数据样本，利用支持向量机计算方法、反馈式神经网络计算方法、数据挖掘技术中的决策树构建模型，对比各个模型的舞弊识别准确率，最终发现运用决策树算法建立的模型识别准确率高于其它模型。

Choi (1997) 选取了 5 个财务比率和 3 个账户变量，利用神经网络建构模型，识别财务舞弊，结果表明神经网络相对其他检测模型更加有效。Becker(2003) 通过选取财务指标，建立了以人工神经网络和模糊神经网络为基础的识别财务舞弊模型，证实了模糊神经网络比人工神经网络的预测准确度更高。PetrHajek(2017)通过对各个财务舞弊识别模型进行比对，发现神经网络的预测准确度高于其它模型。

2.3.2 国内研究综述

我国目前对于财务舞弊模型的构建以统计方法为主，包括逻辑回归模型。赵英林和陈素华（2007）、岳殿民等（2012）应用 Logistic 模型，并选择了相应的财务指标来识别财务舞弊。杨清香等（2009）利用 Logistic 回归分析比较指出，有控制变量的模型比无控制变量识别财务舞弊的效果更好。房琳琳（2013）建立的逻辑回归预警模型预警精准度达到 91.7%，有效度达到 88.7%，预警效果较好。大部分使用逻辑回归的专家都是以上市公司各个时间段的数据作为研究对象，将通过假设检验的指标应用到 Logistic 回归模型中，能够达到识别财务舞弊的研究目的。

除此之外，专家在研究过程中也利用了主成分研究方法，丰富了识别财务舞弊的方法，强化了识别效果。钱萍（2015）将其研发的 C-score 综合模型和发达国家的 M-score 和 F-score 识别财务舞弊模型实施了分析和比较，说明符合我国国情的财务舞弊识别模型更有效。姚宏等（2007）在主成分分析法的基础上构建了识别财务舞弊的模型，但是只是依据财务指标实施定量研究，无法将上市公司利润控制的方式精准地、全面地展现出来。

随后，我国专家也开始使用数据挖掘、机器学习、自然算法语言等相关方法处理识别财务舞弊的具体问题。蒙肖莲等（2009）建立了以概率神经网络为基础的识别虚假财务报告模型，主要研究了平滑参数、选择变量等问题。陈庆杰（2012）、夏明等人（2015）结合上市公司内部和外部的各项指标展开研究，其指出，构建 RBF 神经网络并对其优化，可以使得识别财务舞弊的能力得到全面提升。但是神经网络方法也是有不足的，其是在经验风险最小化的基础上构建的，因此在实现计算方法时，可能局部会出现极小的问题，很难获得最优全局解。在神经网络法之后，邓庆山和梅国平（2009）指出支持向量机在预测经济发展、识别机器学习文本和模式等方面应用较广。阚宝奎等（2012）全面优化了 SVM 计算方法，结合了谱聚类方法，构建了全新的算法，与传统神经网络计算模式有显著差别。

2.4 文献评述

在分析财务舞弊的识别和治理中，了解到舞弊产生的动因是分析的重要环节。其中，相关动因理论体系中包含风险因子理论、GONE 球论、三因素理论、二因素理论等。从现阶段的研究情况来分析，大多数专家都是以以上主流研究理论为研究基础，进而在这一领域进行探索。但也有一些专家在从多元化的视角解读和研究财务舞弊问题。

对于实施财务舞弊的上市公司所具有的特征，众多学者也给出了相应的答案，总体来看，第一类是与公司自身财务指标相关的特征，其中具体包含资产质量、现金流量、营运能力、偿还债务能力、盈利能力等。当内部控制制度全面应用之后，有专家开始在识别财务舞弊中引入股权结构、董事会特点、监事会特点等内容。

从识别舞弊方法的研究结果来分析，国外的识别方法整体来讲比较多，已经建立了较为完善的实证研究体系。而我国的研究基本上都是利用国外已有的识别方法，特别是针对数据挖掘方法的应用，还处在探索和研究的过程中。其中，国内外利用较多的数据挖掘方法和统计方法包含：支持向量机模型、决策树、随机森林、神经网络模型、Logistic 回归等。通过全面研究国内外的文献资料能够了解到，在我国现阶段的研究中，识别财务舞弊行为存在的问题包括：样本比较少、识别指标不全、方法单一等。

第三章、实证研究方案设计

3.1 实证研究整体思路

本章对下一章实证分析部分提供模型和数据基础，分别从样本选取、指标选取和模型介绍三方面展开。

在指标构建方面，分别从财务报告可信度和财务异常指标方面梳理。根据我们前面的文献综述部分可知舞弊三角理论作为识别财务舞弊行为的经典理论可以有效的识别出财务报表舞弊的可能性，所以我们从舞弊机会、舞弊动机、舞弊借口三个维度归类财务报告可信度指标。参照国泰安数据库中公司研究的分类方法，同时加入识别财务舞弊的特有分类，从公司会计信息质量、营运能力、盈利能力、偿债能力、变现能力、财务风险水平和其他财务指标 7 个维度构建财务异常指标，衡量企业财务报告的健康度。

本文选取了在 2000–2020 年间多次出现财务舞弊的公司，将样本上市以来且在样本期内第一次财务舞弊的年份作为首次舞弊样本集，将连续舞弊的每个年份作为连续舞弊样本集。

为了对上市公司的财务舞弊行为进行预测与识别，并进而分析在预测过程中较为重要的公司特征。通过参考以往学者的研究，本文选取了 Logistic 回归模型、SVM 支持向量机、决策树和 LightGBM 作为最终财务舞弊预警模型的候选，并且用非线性主成分分析法对数据降维。用降维后的数据作为输入变量，分别对财务报告可信度模型、财务异常模型和综合模型在两类样本集上的识别效果进行评估。

3.2 样本选取

本文参照以往学者有关财务舞弊的实证经验和理论，基于国泰安数据库公司研究中“违规处理”选取 2000 年 1 月至 2020 年 12 月由于财务舞弊受到中国证监会违规处罚的 A 股上市公司作为舞弊样本，并对选取非舞弊样本。在指标

的构建上，主要来自于国泰安数据库（CSMAR）公司研究数据库，并通过巨潮资讯网（www.cninfo.com.cn）、锐思数据库（RESSET）、Wind 金融终端及各个公司官网补充数据，均以年度报告数据为选取对象。

表 3-1 国泰安数据库企业违规行为分类表

违规编码	违规行为	违规编码	违规行为
P2501	虚构利润	P2509	擅自改变资金用途
P2502	虚列资产	P2510	占用公司资产
P2503	虚假记载（误导性陈述）	P2511	内幕交易
P2504	推迟披露	P2512	违规股票买卖
P2505	重大遗漏	P2513	操纵股价
P2506	披露不实（其他）	P2514	违规担保
P2507	欺诈上市	P2515	一般会计处理不当
P2508	出资违规	P2599	其他

为了保证实证数据时效性和准确性，样本选取 2000-2020 年之间被中国证监会明确做出违规处罚决定的财务舞弊上市公司，对于连续舞弊公司，每个舞弊年度均作为一个样本，对于首次舞弊样本，将样本上市以来且在样本期内第一次财务舞弊的年份作为首次舞弊样本集。国泰安（CSMAR）数据库对上市公司违规分类共 16 种，具体分类如表 3-1 所示。由于上市公司违规类型非常多，在考虑到虚增利润、虚列资产等是上市公司财务报表舞弊最可能的目的，并且结合数据库中对违规类型的说明以及本文的研究目的，同时考虑到这些违规行为对投资者的投资活动及资本市场的运行带来的影响最大。因此对在违规处理总表中筛选出虚增利润和虚列资产违规类别作为舞弊样本的数据集。

在为建立财务舞弊识别模型的舞弊样本选取控制样本时，应当综合考虑行业、数据完整性、审计意见等因素，本文的选择标准如下：

- (1) 行业分布情况大体与舞弊样本相似
- (2) 与舞弊年度样本相似年度
- (3) 三年均获得标准审计意见
- (4) 剔除 ST 公司

(5) 剔除关键数据缺失公司

经筛选，连续舞弊样本共计 625 条，对应控制样本 1273 条；首次舞弊样本共计 75 条，对应控制样本 73 条。

3.3 指标选取

在指标选取和模型构建参考厦门大学会计学院中国财务舞弊中心牵头构建的智能财务舞弊识别模型，以及 Summer Sweeney (1998) 的分层分析方法。通过结合研究舞弊发生可能性及舞弊财务异常来侦查舞弊行为。具体而言，在指标构建方面，可分为两个部分，即财务报告可信度指标和财务异常指标。

财务报告可信度指标从舞弊机会、舞弊动机、舞弊借口三个维度对企业画像，筛选出对企业财务状况构成威胁的变量，通过对市场信息的跟踪，及时预警企业舞弊行为，发出舞弊信号。财务异常指标主要以财务报表为依据，根据财务指标异常变动来识别舞弊行为，财务异常指标的分类参照国泰安数据库中公司研究的分类方法，同时加入识别财务舞弊的特有分类，从会计信息质量、营运能力、盈利能力、偿债能力、变现能力、财务风险水平和其他财务指标 7 个方面构建指标。

3.3.1 财务报告可信度指标构建

通过对财务报告可信度指标的构建，来衡量上市公司发生舞弊的可能性。舞弊三因素理论为本文财务报告可信度指标构建的基本框架提供了重要思路，根据我们前面的文献综述部分可知舞弊三角理论作为识别财务舞弊行为的经典理论可以有效的识别出财务报表舞弊的可能性，并且舞弊机会、舞弊动机、舞弊借口变量都是可以作为舞弊发生可能性的指标归类。以下会对这三类指标分别介绍。

(1) 舞弊机会指标

公司治理结构一般分成内部治理结构和外部治理结构，股东大会、董事会、经理层、监事会四者作为主体相互制约构成公司内部治理结构，外部治理主体包括市场、政府、社会三个主体，其中第三方审计是外部治理结构的关键环节。内部控制是保证公司合法合规的关键环节。运行失灵的治理环境和失效的内部控制会为公司财务舞弊行为提供土壤。所以我们考虑将公司治理结构和内部控制制度作为舞弊机会指标。

表 3-2 舞弊机会指标

指标代码	指标名称	指标说明	指标类型
NOBM	董事会规模	董事会成员的人数	连续型
NBMD	董事会议次数	本年召开董事会议的次数	连续型
NOBS	监事会规模	监事会成员的人数	连续型
NBSD	监事会议次数	本年召开监事会议的次数	连续型
POID	独立董事比例	独立董事总人数/董事会成员人数	连续型
POCS	流通股比例	流通股总数/总股数	连续型
SMNE	董事长兼任总经理	哑变量, 如果董事长兼任总经理, 取值为 1, 否则为 2	离散型
OCRH	股权集中度	公司前 10 位大股东持股比例之和	连续型
S	第一大股东控制度	公司第一大股东持股比例	连续型
Z	Z 指数	公司第一大流通股股东与第二大流通股股东持股比例的比值	连续型
REL	前十大股东是否存在关联	1=不存在关联。2=存在关联, 3=不能确定	离散型
SUG	内部控制审计报告意见	标准无保留意见=1; 其他=0	离散型
PUB	是否披露内控审计报告	是=1; 否=2	离散型
PUD	内部控制是否有效	是=1; 否=2	离散型
PUV	是否采取整改措施	是=1; 否=2	离散型
AUDI	负责审计事务所规模	哑变量, 如果上市公司由大型会计师事务所负责审计则取值为 1; 否则为 0	离散型
DOUB	是否接受双重审计	如果是, 则取值为 1; 否则为 0	离散型

指标说明:

董事会规模 (NOBM)、董事会议次数 (NBMD)。Persons (2006) 的论据结果显示董事会独立程度以和规模会影响财务舞弊行为的产生。Anderson (2004) 研究公司治理层面指出, 董事会持股比例和董事会议的频率如果在一定阶段内

比较高的话，会大大减少公司发生舞弊的行为。

监事会规模（NOBS）、监事会会议次数（NBSD）。乔珊（2017）论明在财务造假公司和非财务造假公司之间，监事会规模和会议次数存在明显差异。监事会是公司监督和管理的常设机构，监事会规模大也不一定监督效果好，可能存在监管不力的情况；在一般情况下，会议次数越多说明公司对监事机构的监察越重视。

独立董事比例（POID）。刘桂林,陈美芳（2021）在舞弊与非舞弊公司独立董事会特征和监事会特征中发现独立董事和监事比例的提高都有助于减少上市公司财务舞弊行为发生。Bruno Frey（2013）强调独立董事制度对于财务舞弊行为控制的重要性，独立董事会人数较多且独立性较强的公司财务舞弊的可能性较低。

流通股比例（POCS）。吴丽霞（2020）实证分析表明流通股比例与财务舞弊之间关系显著，流通股虽然未直接参与公司治理且持有份额较小，但是通过在二级市场上流通股的买卖，会对公司治理产生间接影响，帮助公司信息透明化、公开化。

董事长兼任总经理（SMNE）、董事会成员兼任其他单位（SMNU）。两职兼任会提高决策效率，更好的传达董事会的意见，但是可能会导致四方制衡的治理结构失衡；董事会成员兼任其他单位会影响参与公司事务的积极性，涉及利益冲突时，也容易做出错误的决定。

股权集中度（OCRH）。余玉苗和吕凡（2010）通过建立逻辑回归模型，发现股权集中度可以有效识别公司财务造假。刘立国（2012）对公司治理效率和股权集中度深入研究，适度的股权集中有利于公司治理体制发挥更大的作用。

第一大股东控制度（S）。任海芝,刘雪,张瑞雪（2020）实证分析中发现第一大股东持股比例与财务舞弊显著负相关。

Z 指数（Z）。如果公司最大股东的持股比例显著高于其他股东，那么他会对公司运作以及股价的市场表现有很大影响。

前十大股东是否存在关联（REL）。股东之间存在关联有可能会为财务舞弊行为的发生创造舞弊机会。

内部控制审计报告意见（SUG）、是否披露内控审计报告（PUB）、内部控制是否有效（PUD）、是否采取整改措施（PUV）。建立有效的内部控制机制对防范财务舞弊极为关键。理论上来讲，严格执行相关法律法规，财务舞弊现象就不会

发生。

负责审计事务所规模 (AUDI)、是否接受双重审计 (DOUB): 这两个指标定义的大型会计师事务所特指国际四大会计师事务所和国内八大内资所, 审计质量体现了外部治理结构质量。一般而言, 审计质量高的公司发生财务舞弊的可能性小。

(2) 舞弊动机和压力指标

基于虚增利润的舞弊事件特征来看, 舞弊动机和压力往往来源于公司利益最大化或者管理层利益最大化。所以我们从这两个角度设置舞弊动机和压力指标。

表 3-3 舞弊动机和压力指标

指标符号	指标名称	指标定义	指标类型
PLF	舞弊前是否亏损	舞弊前一年的净利润如果小于 0, 则取值为 1; 否则取值为 0	离散型
POP	舞弊前盈利能力	舞弊前 1 年的净资产收益率	连续型
GROW	舞弊前公司的成长性	舞弊前 1 年的营业收入增长率	连续型
SOC	舞弊前公司偿债能力	舞弊前 1 年的速动比率	连续型
TMS	股份薪酬激励	舞弊当年高管人员持股数	连续型
SALA	高管薪酬占比	当年高管前三名薪酬总额/应付职工薪酬	连续型
CHAN	高管更迭	如果舞弊当年发生了董事长或总经理的变更, 则取值为 1; 否则取值为 0	离散型

指标说明:

公司为了获取在资本市场上投资者的青睐或者是为了获取债权人的信任, 有可能会选择通过包装公司财务报表达成目标。我们通过舞弊前是否亏损 (PLF)、盈利能力 (POP)、成长性 (GROW)、偿债能力 (SOC) 指标来识别这一舞弊动机。

股份薪酬激励(CHAN)。管理层持股比例高, 秦江萍 (2015) 认为当管理层

持股且追求高风险的时候容易产生贪婪心理，通过舞弊获得更高的非法收益。陈婷(2007)发现管理层财务舞弊的原因大多是为了提高持股比例和提升薪酬待遇。

高管薪酬占比(SALA)。王静(2007)以2009-2013年上市公司作为研究对象，发现高管薪酬和公司财务舞弊显著负相关。

高管更迭(TMS)。管理层为了获取董事会的信任，可能在继任之后对公司变革，渴望达成目标，甚至不惜选择舞弊，预测高管更迭之后更加容易产生舞弊动机。

(3) 舞弊态度和借口指标

舞弊规模较大且受到中国证监会违规处罚的上市公司发生的财务舞弊行为通常是由公司高层管理者主导。舞弊态度和借口变量指标主要从舞弊者的行为依据、道德水平、诚信意识等来考虑，而这些因素往往很难量化判断，所以我们通过舞弊者的特征归纳来选取指标。

表 3-4 舞弊态度和借口指标

指标符号	指标名称	指标定义	指标类型
ALEV1	舞弊前应计水平	净利润-企业自由现金流	连续型
ALEV2	舞弊前应计方向	哑变量，当舞弊前应计为正时，取值为1；否则取值为0	离散型
AGES	管理层平均年龄	高级管理人员平均年龄	连续型
EDU	管理层平均学历	管理层本科以下学历取值1；本科学历取值2；研究生学历取值3；博士取值4；博士以上取值5；取平均值	离散型
GEND	管理层性别比例	高级管理人员中男性所占比例	连续型
NSOP	舞弊之前三年曾被出具非标意见	哑变量，如果公司在舞弊前三年曾被出具非标审计意见，则取值1；否则取值为0	离散型

指标说明：

舞弊前应计水平、舞弊前应计方向(ALEV)：净利润与现金净流量的差额体现了管理层对现金的利用程度，在一定程度上反映了管理层对于使用会计政策的

激进度。

管理层平均年龄（AGES）、平均学历（EDU）、性别比例（GEND）。杨梦恬（2017）发现管理层年龄、学历水平、男性管理层比例都与内部控制质量显著关联。卢馨（2015）通过研究财务舞弊和公司管理层特征的关系再次得出，财务舞弊与管理层学历相关联，受过良好教育的管理层对公司治理的重视程度和对法律法规的敬畏程度都较高，更不容易发生舞弊行为。并且管理层年龄较大、女性占比比较高的，内部控制质量较高，更不容易发生舞弊。

舞弊之前三年曾被出具非标意见（NSOP）：如果曾被出具非标意见，说明之前经第三方独立审计机构审计就曾有舞弊迹象或者信息质量不高，进而说明管理层的诚信意识、规则意识值得怀疑。

3.3.2 财务异常指标构建

财务异常指标主要以财务报表为依据，根据财务指标的异常变动来识别舞弊行为。本章分类参照国泰安数据库中公司研究的分类方法，同时加入识别财务舞弊的特有分类，从公司会计信息质量、营运能力、盈利能力、偿债能力、变现能力、财务风险水平和其他财务指标 7 个方面构建财务指标。

表 3-5 财务异常指标

衡量维度	指标符号	指标定义
会计信息质量	TAOI	主营业务税金及附加/营业利润
	TATP	所得税/利润总额
	OITP	营业利润/利润总额
营运能力	FATR	固定资产周转率
	TATR	总资产周转率
	ARTR	应收账款周转率
	STR	存货周转率
	NYTR	应付账款周转率
盈利能力	RONA	净资产收益率
	CEM	营业毛利率

	GPMS	销售费用率
	NPRS	成本费用利润率
偿债能力	CURA	流动比率
	QURA	速动比率
	NCU	净现金比率
	ALRA	资产负债率
	DETA	有形净值债务率
	STRE	经营活动产生的现金流量净额 / 带息债务
	CARE	现金满足投资比率
	DFL	财务杠杆系数
财务风险水平	DOL	经营杠杆系数
	TL	综合杠杆系数
	TIN	是否微利公司。以净资产收益率为标准，当净资产收益率<1%时，为微利公司，取值为1；否则取值为0。
发展能力	NL	固定资产增长率
	RL	营业收入增长率
	CL	营业总成本增长率
	IL	可持续增长率
	AL	净资产收益率增长率

指标说明：

会计信息质量指标。舞弊公司虚增利润，但是并不按照舞弊后的利润缴纳税金，会导致税金与利润的比率与实际税率无法匹配，所以我们在这里设置税金与利润的比值作为会计信息质量指标。同时上市公司还会通过非业务收入，例如营业外收入、投资收益等进行舞弊，所以会导致营业利润与利润总额之间差额较大。我们设置营业利润/利润总额这一指标来进行识别。

营运能力指标。衡量各类资产的使用效率，各资产周转速度越快，资产管理效率越高，企业从资产中获得的收益越高，对于资金的需求越少，从市场获得资

金的需求降低可能会减少舞弊行为的发生。

盈利能力指标，是对公司利润率的深层分析，反映在投入成本、资产、初始资本能够创造的利润额。企业获取利润能力较差那么舞弊的可能性就会加大。

偿债能力指标。其中流动比率、速动比率和净现金比率是衡量企业短期偿债能力的指标，都是正向指标，三个比率越来越严格的衡量短期偿债能力。资产负债比率和有形净值债务率是衡量企业长期偿债能力的存量指标，其中有形净值债务率是对产权比率更为严格的计算方法，将无形资产刨除在外，衡量破产清算时对债权人的保护程度。利息保障倍数是衡量企业长期偿债能力的流量指标，利息保障倍数越大，企业偿还利息的缓冲效果越多，对外部资金的需求越少。

财务风险水平指标。曾小青（2021）通过将近十年的舞弊公司和非舞弊公司共计 7014 条样本研究，构建了基于逻辑回归、神经网络、支持向量机和决策树四种模型并进行优化，指出财务舞弊特征中应当特别关注综合杠杆、财务杠杆财务指标。Person（2006）研究发现高财务杠杆的公司更容易发生舞弊。财务杠杆系数反映财务风险，较高的财务杠杆说明财务风险较大，债务融资的可能性性较小。综合杠杆系数是两者的乘积，既可以反映企业每股获利能力，也可以反映经营杠杆与财务杠杆的平衡。

其他财务指标。国内外学者发现，当企业处于利润边界点和重度亏损的情况下，发生财务舞弊十分普遍，企业有强烈的舞弊动机维持公司形象。所以我们设置 TIN 指标，以净资产收益率为判断依据，当净资产收益率小于 1%时认定为微利企业。

发展能力指标。主要衡量公司发展的潜力，主要指标有固定资产增长率、营业收入增长率、营业总成本增长率、可持续增长率、净资产收益率增长率等。

3.4 模型介绍

基于人工的传统识别方法覆盖指标较少并且参数规则设置随意，较难发掘指标之间的深层次关系，而机器学习可以解决这一问题，并且可以对数据规模化处理，实现自动化。随着数据量的增多，仅依靠人工识别财务舞弊的时间和成本都会大大增加，人工智能、机器学习等现代化金融科技的介入，为财务舞弊的识别提供了重要的技术工具。

为了对上市公司的财务舞弊行为进行预测与识别，并进而分析在预测过程中较为重要的公司特征，本文选取了 Logistic 回归模型、SVM 支持向量机、决策树和 LightGBM 作为最终财务舞弊识别模型的候选。选择上述模型作为舞弊识别模型的原因有：第一，它们具有较强的可解释性，可以通过训练得到的模型来判断较为重要的输入变量。第二，上述模型的模型结构较为简单，在本文研究问题样本数量较少的情况下，不适合采用复杂度较高的模型，容易出现过拟合的问题。第三，这些模型既可以处理连续值也可以处理离散值。本节主要对以上模型逐一介绍。

3.4.1 Logistic 回归模型

Logistic 回归是统计学经典算法之一。它相当于通过 *Sigmoid* 函数对特征的线性组合 $\theta^T x$ 进行了放缩，将目标值压缩至 $[0,1]$ 内，如下图所示。

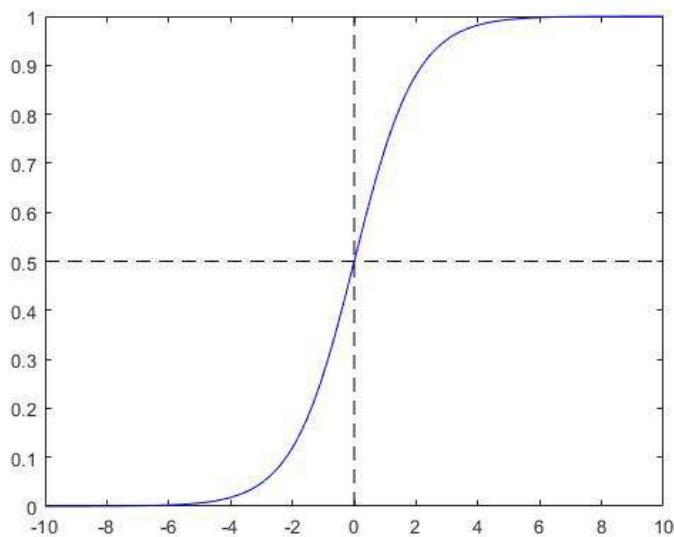


图 3-1 sigmoid 函数

Logistic 回归模型的预测形式如下所示，经过 *Sigmoid* 函数的压缩之后， $h_\theta(x)$ 就代表样本分类为 1 的概率。一般来说，逻辑回归的分类阈值为 0.5，若模型的输出值 $h_\theta(x) > 0.5$ ，模型将会把样本分类为类别 1；若模型的输出值 $h_\theta(x) < 0.5$ ，模型将会把样本分类为类别 0。

$$h_\theta(x) = g(\theta^T x) = \frac{1}{1 + e^{-\theta^T x}} \quad (1)$$

Logistic 回归模型的损失函数应用了交叉熵损失函数，具体可参考公式（2）的内容，常规来讲，会应用梯度下降法求解参数。Logistic 回归模型中特征的重

要程度可以根据特征所对应的的 t 统计量进行比较。

$$\begin{aligned} J(\theta) &= \frac{1}{m} \sum_{i=1}^m \text{cost}(h_\theta(x_i), y_i) \\ &= -\frac{1}{m} \left[\sum_{i=1}^m (y_i \log h_\theta(x_i) + (1 - y_i) \log(1 - h_\theta(x_i))) \right] \end{aligned} \quad (2)$$

逻辑回归模型适合二分类问题，并且可以看到不同指标对最后预测结果的影响，并且模型的可解释性较好，所以我们选取逻辑回归模型作为舞弊识别模型的候选。

3.4.2 支持向量机

支持向量机是传统机器学习研究领域中最重要的成果之一，它具有完备的数学理论与数学证明，它的结构十分优雅，在深度学习尚未普及的时期在工业界应用广泛。

识别传统财务舞弊一般会利用分类计数建立预测模型，评分的问题一般不会被考虑，使得识别成效不佳，当数据挖掘技术全面优化之后，智能技术兼具评分和分类的功能，应用比较广泛，支持向量机和神经网络最常应用。支持向量机的核心思路是在高维特征空间中找到超平面，使其可以以最大信度划分不同样本。在实际应用中，一个由特征线性组合表示的超平面能够将所有样本分隔开来的情况较少，支持向量机能够将核函数和软间隔方法引入，使得线性不可分样本的问题得到妥善处理。其中软间隔是模型允许某些样本分类错误，但会在目标函数中增加惩罚项以限制误分类的程度。核函数通过定义不同形式的内积 $K(x, x')$ ，把特征从低维空间映射到高维空间，这样就能够建立非线性分类器，使得线性不可分样本的问题得到妥善处理。

支持向量机的决策函数为：

$$y(w^T x + b) \geq 1$$

而 $w^T x + b = 1$ 即上文所提到的超平面。而离该超平面最近的样本距离超平面的距离为：

$$d = \frac{y(w^T x + b)}{\|w\|}$$

最大化这个距离，就可以得到支持向量机的参数 w 、 b 。

因此支持向量机可以转换为：

$$\begin{aligned} \min \frac{1}{2} \|w\|^2 \\ \text{s.t. } y_i(w^T x_i + b) \geq 1 \end{aligned}$$

同时可以通过引入松弛变量、拉格朗日函数以及对偶算法进行求解。

支持向量机适合解决二分类问题，抓取关键样本的能力强，并且能够将冗余样本剔除，可以处理高维数据，所以我们选取支持向量机模型作为舞弊识别模型的候选。

3.4.3 CART 决策树

决策树是由一系列 if-else 规则组成的二叉树预测模型，在决策树中每个节点都会通过一种 if-else 规则对样本集进行分裂生成新的节点，并由最终停止分裂的叶子节点对样本的目标值进行预测。CART 决策树即可以完成分类任务，也可以完成回归任务。如下图所示，其中包括条件的节点是决策树中的节点，具体有 c_1, c_2, c_3, c_4 预测值的节点，它们是叶子节点，代表一种预测结果。

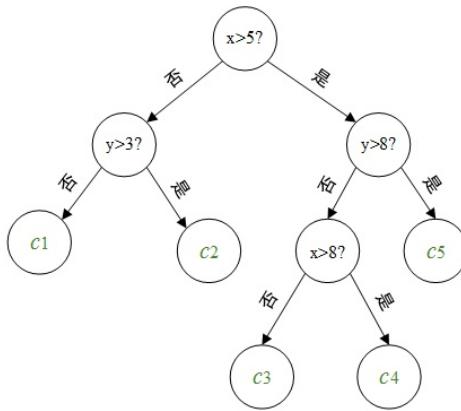


图 3-2 CART 决策树

决策树模型的训练主要有分裂和剪枝两个部分。决策树从最初的根节点开始分裂，在每个非叶子节点上，通过寻找最优的分裂特征及其阈值对样本进行划分。重复这个分裂的过程直到模型达到如最大树深度、最大叶子节点个数等限制条件。决策树在进行回归任务时，在根节点进行分裂的依据是使分裂后两个子节点的样本目标值的平方误差和最小化，如以下公式所示。

$$\min_{j,s} \left[\min_{c_1} \sum_{x_i \in R_1(j,s)} (y_i - c_1)^2 + \min_{c_2} \sum_{x_i \in R_2(j,s)} (y_i - c_2)^2 \right]$$

其中 $\hat{c}_1 = \frac{1}{N_1} \sum_{x_i \in R_1(j,s)} y_i$, $\hat{c}_2 = \frac{1}{N_2} \sum_{x_i \in R_2(j,s)} y_i$, R_1 为左叶子节点的样本集, R_2 为右叶子节点的样本集。

在完成决策树的分裂以后, 为了简化决策树, 避免出现过拟合的问题, 需要剪枝已经产生的决策树, 也就是说, 把一些不关键的叶结点或子树剪掉, 使得模型整体更为简单。

决策树对于异常点的容错能力很强, 健壮力很高, 可解释性强。对于本文离散型和连续型指标都可以很好的处理。所以我们选取决策树模型作为舞弊识别模型的候选。

3.4.4 LightGBM

LightGBM 属于在决策树基础上产生的学习方法, 具备集成特点, 其延续了 Boosting 方法的思路。Boosting 的思路为串联各种弱学习器, 从而建立强分类器, 每一个弱学习器所拟合的内容是上一个学习器的预测误差。Boosting 方法在回归任务上的预测即所有弱学习器的预测值总和。在 LightGBM 中, 弱学习器即决策树模型。在决策树的生长中, LightGBM 进行了一系列的优化手段, 使得模型更为高效。首先, LightGBM 对连续特征进行分箱, 而非遍历所有可能的分裂点。虽然会损失一定的精度, 但会大大提升运行效率。同时这样的分箱方式还可以增强模型整体的泛化性能。其次, LightGBM 对稀疏特征做了“互斥特征捆绑”的优化功能, 将许多互斥的特征自动组合为一个新的特征, 以达到降维的目的。最后, LightGBM 中决策树的生长方式由广度优先搜索转化为了深度优先搜索, 大大提高了算法的效率。

但由于 Boosting 算法的特点, 每个学习器都在减少其在训练集上的预测残差, 很容易引起过拟合风险。因此, 在模型训练的过程中可以适当提高 L1、L2 惩罚度, 以简化模型, 提高泛化性能。

3.4.5 非线性主成分

针对传统的逻辑回归无法处理特征之间的非线性问题，本文引入了非线性主成分方法为逻辑回归引入非线性特征。首先，通过主成分分析来消除财务特征数据中的冗余数据和噪音。主成分分析认为数据中方差越大表明其中包含的信息量就越多，因此主成分分析的思想就是在使方差最大化的同时降低特征空间的维度，将初始的样本通过线性变换从高维特征空间投影到低维特征空间，以达到特征降维的效果。主成分个数的选取可以通过主成分的累计方差贡献率来判断，一般来说，可以选取累计方差贡献率达到 80% 的主成分个数。其次，通过以上方法构建的主成分仍然是线性的，为了引入非线性关系，必须构建非线性特征。非线性特征的构建方法一般有特征分箱、特征 Encoding、特征卷积、构建交互项等方式。在逻辑回归中，可以通过构建特征交互项和高次方项来构建非线性特征。

为了控制模型的复杂度，降低过拟合风险，本文通过最高次方为 2 的多项式来引入非线性特征。假设有主成分 f_1 、 f_2 ，则构建二次多项式回归方程，如下式所示：

$$h(\hat{\theta}) = \theta_0 + \theta_1 z_1 + \theta_2 z_2 + \theta_3 z_3 + \theta_4 z_4 + \theta_5 z_5$$

其中 $z_1 = f_1$, $z_2 = f_2$, $z_3 = f_1^2$, $z_4 = f_2^2$, $z_5 = f_1 f_2$, 而 z_3 、 z_4 和 z_5 即非线性特征。当原始特征数量较多时，通过以上方式构建的非线性特征的数量也会很多，容易造成特征冗余、过拟合的问题。因此本文通过后逐步回归的 Wald 法来对上述构造的非线性特征进行筛选，将有用的特征提取出来，以增加财务舞弊预警模型的识别能力。

第四章 实证分析

在本章中，通过对指标多重共线性的检验，使用非线性主成分分析法对数据降维，然后基于机器学习构建财务报告可信度模型和财务异常模型，分别对连续舞弊样本集和首次舞弊样本集进行训练和测试，训练和测试完后收集样本实验结果，通过实现构造的评价指标对模型识别能力比较，最后构架综合指标模型，观察是否可以通过指标的综合提高识别舞弊风险的能力。

4.1 评价指标的构建

对学习器泛化性能的评估需要有效可行的评价标准，这就需要用到模型性能度量指标。不同度量指标可能会产生不同的评价结果，模型的“好坏”是相对的，取决于任务需求。为了评价本章模型的识别能力，本文主要采用总准确率指标、查准率指标、查全率指标及 F 度量指标，以下是对各个指标的简要介绍。

4.1.1 总精度指标（accuracy）

总精度指标是正确分类的样本数与总样本数之比，是被模型正确识别的概率。对于样例集 D，分类可以定义为：

$$\begin{aligned} acc(f; D) &= \frac{1}{m} \sum_{i=1}^m \prod(f(x_i) = y_i) \\ &= 1 - E(f; D) \end{aligned}$$

在二分类混淆矩阵中，它的定义为：

$$accuracy = \frac{TP + TN}{ALL}$$

在样本数据均衡，并且犯两类错误的代价差额较小时，总精度指标可以很好地衡量分类器的分类效果。当存在上述两种问题时，样本的识别率必然会向样本占比较高的类型识别率倾斜，不能够准确全面的反映识别情况，因此我们还需构建其他指标共同评价学习器的识别效果，科学评估模型的识别效率。

4.1.2 查准率指标 (precision)

在二分类问题中，可以用二阶矩阵将样本真实情况和预测结果表示出来，行代表真实情况，列代表预测结果，可以将结果划分为真正、假反、假正和真反这四种情形，令 TP、FN、FP、TN 分别代表样本数，则有 $TP+FN+FP+TN=$ 总样本个数，二分类结果的矩阵如表 4-1 所示。

表 4-1 分类结果混淆矩阵

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

将查准率指标 (precision) 定义为：

$$precision = \frac{TP}{TP + FP}$$

查准率指标是指在预测为财务舞弊的样本中，真正为舞弊样本的比率，直接体现了模型识别结果的正确率。

4.1.3 查全率指标 (recall)

将查全率指标 (recall) 定义为：

$$recall = \frac{TP}{TP + FN}$$

查全率指标是指在所有财务舞弊样本中，被成功识别出的数量。查全率指标从样本识别的完全度如何这一角度反映了模型的识别能力。

4.1.4 F 度量指标 (F-score)

查准率指标 (precision) 和查全率指标 (recall) 是一对矛盾的评价指标，通常情况下，两者此消彼长，只有在一些简单的机器学习任务中，才能使查全率和

查准率都很高或者都很低。因此使用两个指标共同评价分类器的分类效果更好，由此引入 F 度量指标，取两个指标的调和平均数。

$$F-score = \frac{2 \cdot precision \cdot recall}{precision + recall}$$

在本文的实证中，舞弊样本遗漏的代价和错判的代价不分伯仲，所以我们赋予查准率和查全率的权重相同，由此带入模型计算 F-score 指标，从而评价模型识别效果。

4.2 特征多重共线性检验

在上一章，通过对以往实证文献的梳理选取了财务报告可信度指标和财务异常指标。由于过多同类型的指标存在，会不可避免地出现共线性的问题，但是，仅仅一味削减指标数量，会导致很多信息受损，识别舞弊的效率也会因此大大降低。所以，本章主要研究了如何将原始信息有效保存下来，促进识别效率的全面提升。

在使用机器学习模型在数据集上训练之前我们需要检验各变量彼此的多重共线性。若存在多重共线性，计算自变量的偏回归系数时矩阵不可逆。本文在检验各变量彼此的多重共线性时应用了方差膨胀因子 VIF 方法，它表示回归系数估计量的方差与假设自变量间不线性相关时方差相比的比值。它的计算公式如下：

$$VIF = \frac{1}{1 - R_i^2}$$

常规而言，随着方差膨胀因子 VIF 的值增大，表示变量彼此有更为严重的共线性。通常以 10 作为判断边界，如果 VIF 超过 10 的话，表示特征彼此具有很严重的多重共线性；当 $VIF < 10$ ，不存在多重共线性；当 $10 \leq VIF < 100$ ，存在较强的多重共线性；当 $VIF \geq 100$ ，存在严重多重共线性。我们将分别检验财务报告可信度指标、财务异常指标和综合指标的多重共线性。

4.2.1 财务报告可信度指标

通过表 4-2 能够了解到，本文对财务报告可信度模型中的所有指标进行 VIF 检验，发现有部分指标的 VIF 因子超过了 10，包括 POCS、SMNE、OCRH、S、

REL、SUG、PUB、PUD、PUV、AGES、EDU、GEND，共计 12 个指标，占所有指标的 50%。这说明在数据集中可能存在较为严重的多重共线性，可能会使模型对特征值的变动十分敏感，进而导致模型在样本外的预测性能变差。因此，在进行进一步的建模之前我们必须对数据集的多重共线性进行处理。本文将在之后的内容应用主成分分析 PCA 降维数据集，从而使得多重共线性问题得到妥善处理。

表 4-2 财务报告可信度指标方差扩大因子

指标	VIF Factor	指标	VIF Factor
POCS	28.04724	PLF	1.375737
SMNE	20.69961	POP	1.058386
OCRH	11.14707	GROW	1.028919
S	10.2514	SOC	1.439807
Z	1.557061	TMS	1.21231
REL	22.07344	SALA	1.026045
SUG	12.66049	CHAN	1.517933
PUB	50.78659	ALEV1	1.072859
PUD	37.89744	ALEV2	2.484945
PUV	202.0405	AGES	206.14
AUDI	2.765329	EDU	41.8122
DOUB	1.127671	GEND	55.33491

4.2.2 财务异常指标

如表 4-3 所示，本文对财务异常模型中的所有指标进行 VIF 检验，发现有 4 个特征的 VIF 因子超过了 10，即 CURA、QURA、DFL、TL。这说明在数据集中可能会存在一定的多重共线性，因此，在进行进一步的建模之前我们需要对数据集的多重共线性进行处理。本文在之后利用主成分分析 PCA 降维数据，从而使得多重共线性问题得到妥善处理。

表 4-3 财务异常指标方差扩大因子

指标	VIF Factor	指标	VIF Factor
TAOI	1.244544	NCU	7.466451
TATP	1.884818	ALRA	1.662628
OITP	1.027642	DETA	1.046269
FATR	1.029079	STRE	1.001812
TATR	1.456198	CARE	1.003487
ARTR	1.019275	DFL	21.90558
STR	1.0375	DOL	2.275821
NYTR	1.006759	TL	21.74995
RONA	1.097046	TIN	1.560881
CEM	2.416153	NL	1.017661
GPMS	1.224446	RL	1.755968
NPRS	1.700015	CL	1.803194
CURA	47.41121	IL	1.062675
QURA	55.61051	AL	1.046567

4.2.3 综合指标

在综合模型中，本文结合了财务报告可信度模型中的指标与财务异常模型中的指标，试图构建一种精确率更高的财务舞弊行为识别模型。由于指标数量的增加，需要更加注意数据中是否存在相关性过高的情况。如表 4-9 所示，本文对综合模型中的所有特征进行 VIF 检验，发现有大量特征的 VIF 因子超过了 10，共计 20 个指标。这说明在数据集中可能存在较为严重的多重共线性，而这对本文之后的建模是不利的，可能会使模型对特征值的变动十分敏感，进而导致模型在样本外的预测性能变差。因此，在进行进一步的建模之前我们必须对数据集的多重共线性进行处理。本文会在之后的内容中利用主成分分析 PCA 来降维数据，使得多重共线性问题得到妥善处理。

表 4-4 综合指标方差扩大因子

指标	VIF Factor	指标	VIF Factor
TAOI	1.264436	NBMD	8.013863
TATP	1.911022	NOBS	13.92594
OITP	1.035066	NBSD	16.18284
FATR	1.056371	POID	30.66113
TATR	2.05354	POCS	29.67981
ARTR	1.028072	SMNE	21.42353
STR	1.058249	OCRH	11.64563
NYTR	1.019885	S	10.57353
RONA	1.131249	Z	1.624312
CEM	4.282927	REL	22.47354
GPMS	1.267628	SUG	15.00039
NPRS	1.74662	PUB	58.52689
CURA	51.03675	PUD	42.25079
QURA	59.43933	PUV	210.3877
NCU	7.77517	AUDI	2.850933
ALRA	1.764962	DOUB	1.155625
CARE	1.025015	GROW	1.056955
DFL	26.77442	SOC	2.883218
DOL	2.782261	TMS	1.259425
TL	26.07646	SALA	1.045171
TIN	1.928261	CHAN	1.625685
NL	1.034969	ALEV1	1.100231
RL	1.801647	ALEV2	2.812743
CL	1.864636	AGES	222.6028
IL	1.148796	EDU	45.10568
AL	1.066923	GEND	58.75034
NOBM	30.34219	NSOP	1.594961

我们可以发现，在财务报告可信度指标和综合指标中存在严重的多重共线性，在财务异常指标中也存在一定程度的多重共线性，因此，在进行进一步的建模之前我们必须对数据集的多重共线性进行处理。下一节内容中会利用主成分分析PCA来降维数据，使得多重共线性问题得到妥善处理。

4.3 主成分分析及非线性指标的构建

轻微的共线性可以容忍，但是当样本量较少时，共线性问题会被放大，极易导致模型不稳定，因此对处理共线性问题很有必要。本文使用主成分分析这一种在数据科学实验中主流的特征降维技术，通过矩阵分解的方法，把高纬空间的数据投影到低维空间，在数据集中方差被有效保留的基础上，使降维后数据集与原数据之间的偏差最小化。

如表 4-5、表 4-6 和表 4-7 所示，展示了财务报告可信度指标数据集、财务异常指标数据集和综合数据集的特征贡献百分比，我们在确定主成分时，设置了 80% 的累计贡献率，避免遗漏太多信息，分别分别取前 20 个、前 17 个和前 35 个符合条件的变量分别作为财务报告可信度指标数据集、财务异常指标数据集和综合数据集的主成分。

表 4-5 财务报告可信度指标方差贡献率表

主成分	方差贡献率	累计方差贡献率	主成分	方差贡献率	累计方差贡献率
1	9.920%	9.920%	16	2.943%	71.363%
2	7.622%	17.541%	17	2.906%	74.268%
3	6.470%	24.011%	18	2.838%	77.106%
4	4.701%	28.712%	19	2.594%	79.700%
5	4.408%	33.120%	20	2.585%	82.285%
6	4.123%	37.243%	21	2.480%	84.766%
7	3.966%	41.210%	22	2.462%	87.228%
8	3.802%	45.012%	23	2.358%	89.586%
9	3.524%	48.536%	24	2.180%	91.767%

10	3.459%	51.995%	25	2.089%	93.855%
11	3.424%	55.419%	26	1.957%	95.813%
12	3.352%	58.771%	27	1.823%	97.635%
13	3.277%	62.048%	28	1.582%	99.217%
14	3.231%	65.279%	29	0.540%	99.757%
15	3.141%	68.420%	30	0.243%	100.000%

表 4-6 财务异常指标方差贡献率表

主成分	方差贡献率	累计方差贡献率	主成分	方差贡献率	累计方差贡献率
1	10.894%	10.894%	15	3.501%	74.983%
2	8.801%	19.695%	16	3.489%	78.473%
3	6.669%	26.364%	17	3.352%	81.824%
4	5.899%	32.264%	18	3.266%	85.090%
5	4.913%	37.176%	19	3.131%	88.221%
6	4.524%	41.700%	20	2.847%	91.069%
7	4.194%	45.894%	21	2.552%	93.621%
8	3.805%	49.700%	22	1.949%	95.570%
9	3.744%	53.444%	23	1.368%	96.938%
10	3.709%	57.152%	24	1.330%	98.268%
11	3.634%	60.786%	25	1.219%	99.487%
12	3.593%	64.379%	26	0.391%	99.877%
13	3.560%	67.939%	27	0.077%	99.954%
14	3.543%	71.483%	28	0.046%	100.000%

表 4-7 综合指标方差贡献率表

主成分	方差贡献率	累计方差贡献率	主成分	方差贡献率	累计方差贡献率
1	6.778%	6.778%	30	1.601%	73.775%
2	5.417%	12.195%	31	1.564%	75.339%
3	4.393%	16.588%	32	1.555%	76.893%
4	4.213%	20.801%	33	1.505%	78.398%
5	3.798%	24.599%	34	1.466%	79.865%
6	2.958%	27.557%	35	1.439%	81.303%
7	2.632%	30.189%	36	1.414%	82.717%
8	2.530%	32.719%	37	1.321%	84.037%
9	2.435%	35.154%	38	1.315%	85.352%
10	2.254%	37.408%	39	1.251%	86.603%
11	2.180%	39.587%	40	1.242%	87.845%
12	2.130%	41.717%	41	1.195%	89.039%
13	2.020%	43.737%	42	1.147%	90.187%
14	1.991%	45.728%	43	1.097%	91.284%
15	1.946%	47.674%	44	1.075%	92.359%
16	1.890%	49.564%	45	1.037%	93.395%
17	1.882%	51.445%	46	0.949%	94.345%
18	1.864%	53.309%	47	0.912%	95.257%
19	1.810%	55.120%	48	0.848%	96.105%
20	1.788%	56.908%	49	0.765%	96.870%
21	1.766%	58.673%	50	0.679%	97.549%
22	1.744%	60.418%	51	0.638%	98.186%
23	1.739%	62.157%	52	0.618%	98.804%
24	1.725%	63.882%	53	0.563%	99.367%
25	1.708%	65.589%	54	0.275%	99.642%
26	1.691%	67.280%	55	0.179%	99.821%
27	1.663%	68.943%	56	0.122%	99.943%
28	1.622%	70.565%	57	0.035%	99.978%
29	1.609%	72.174%	58	0.022%	100.000%

以财务报告可信度数据集的第一个主成分为例，其计算公式如下所示，是原始特征集的线性组合：

$$\begin{aligned}
 f0 = & -0.03 * NOBM + 0.10 * NBMD + 0.03 * NOBS + 0.14 \\
 & * NBSD + 0.19 * POID + 0.37 * POCS + 0.10 * SMNE \\
 & + 0.39 * OCRH + 0.14 * S + 0.19 * Z + 0.08 * REL + 0.45 \\
 & * SUG - 0.34 * PUB - 0.08 * PUD + 0.01 * PUV + 0.20 \\
 & * AUDI + 0.05 * DOUB - 0.08 * PLF + 0.01 * POP - 0.02 \\
 & * GROW - 0.03 * SOC - 0.01 * TMS + 0.02 * SALA - 0.05 \\
 & * CHAN + 0.08 * ALEV1 + 0.12 * ALEV2 + 0.33 * AGES \\
 & + 0.19 * EDU + 0.01 * GEND - 0.18 * NSOP
 \end{aligned}$$

为了使模型具有对变量之间非线性关系的拟合能力，本文参考 Taylor 的非线性延展思维，建立了非线性特征：

$$F_i F_j$$

其中 $i, j \in (0, 12)$ 代表 n 个主成分的交互项，共 n^2 个非线性特征。为了减少冗余特征，本文采用向后逐步回归的 Wald 法和向机器学习模型的损失函数中加入 L1、L2 惩罚项的方法，以处理在模型中重要程度较低的特征，以避免过拟合风险。

4.4 模型实证结果

为了对比非线性-主成分逻辑回归模型、支持向量机、LightGBM、决策树模型在应用于财务报告可信度模型、财务异常模型和综合指标模型时对财务舞弊公司的识别效果。本文将 2000–2017 年的连续舞弊样本数据和首次舞弊样本数据分别作为训练集，2017–2019 年三年的连续舞弊样本数据和首次舞弊样本数据分别作为测试集，将主成分及其衍生的交互项特征筛选后的特征作为模型的输入变量，上述模型在训练集上进行训练之后得到对应的分类器，将分类器应用于测试集即可得到最终的预测结果。各个模型对财务舞弊公司的预测效果将在下面一一展示。

4.4.1 基于首次舞弊样本

本文选取的首次舞弊样本和控制样本数量相当，舞弊样本数量 75 个，非舞弊样本数量 73 个，不存在样本量失衡的情况，所以在基于首次舞弊的样本集对模型测评中我们使用总精度指标（accuracy）来评价模型的识别效果。基于首次舞弊样本的研究可以让我们看到模型预警财务舞弊的精度。

(1) 财务报告可信度模型的识别效率

表 4-8 财务报告可信度模型效果对比—基于首次舞弊样本集

		非线性-主成分			
		Logistic 回归		支持向量机	
		训练集	测试集	训练集	测试集
舞弊	0.9152	0.4615		0.6271	0.7692
非舞弊	0.9666	0.5833		0.7000	0.8333
合计	0.9411	0.5200		0.6638	0.8500
		LightGBM		决策树	
		训练集	测试集	训练集	测试集
舞弊	0.7627	0.6923		0.8983	0.6153
非舞弊	0.6923	1.0000		0.9159	0.7500
合计	0.7647	0.8400		0.9333	0.6800

从表 4-8 结果可知，非线性-主成分 Logistic 回归的表现最差，虽然它在训练集上表现良好，但在测试集上的表现非常差，这说明模型严重过拟合。在四种模型中，综合表现最好的是支持向量机模型，在测试集上对舞弊公司的识别准确率到了 0.7692，综合识别率达到 0.85。虽然 LightGBM 模型综合识别率与支持向量机模型基本相当，但它对舞弊公司的识别准确率远低于支持向量机模型。在财务舞弊的预警当中，我们更看重模型对舞弊公司的识别能力。因此，在基于首次舞弊样本的财务可信度模型中，支持向量机是预测能力更强的模型，对舞弊的识别率达到 77%。

(2) 财务异常模型识别效率

表 4-9 财务异常模型效果对比—基于首次舞弊样本集

非线性-主成分				
	Logistic 回归		支持向量机	
	训练集	测试集	训练集	测试集
舞弊	0.6440	0.5384	0.6440	0.6194
非舞弊	0.8333	0.9166	0.8500	1.0000
合计	0.7394	0.7200	0.7478	0.6000
LightGBM				
	LightGBM		决策树	
	训练集	测试集	训练集	测试集
舞弊	0.8474	0.6185	0.9322	0.6185
非舞弊	0.6833	0.7500	0.9333	0.5833
合计	0.7647	0.6800	0.9327	0.8000

从表 4-9 结果可知，非线性-主成分 Logistic 回归达到了较为理想的预测效果，但仍是预测能力不足。决策树模型在测试集上对舞弊公司的识别准确率到了 0.6185，综合识别率达到 0.8000。虽然 LightGBM 的对舞弊公司的识别准确率为 0.6185 与决策树相同；而支持向量机对舞弊公司的识别准确率为 0.6194，略高于决策树，但由于它们对所有样本的综合识别准确率远低于决策树。因此，在基于首次舞弊样本的财务异常指标建模中，我们认为决策树的预测性能为最优，达到 61.85%。

(3) 综合识别模型的识别效率

表 4-10 财务舞弊综合识别模型效果对比-基于首次舞弊样本

非线性-主成分				
	Logistic 回归		支持向量机	
	训练集	测试集	训练集	测试集
舞弊	0.9152	0.3846	0.7966	0.6932
非舞弊	0.9666	0.7500	0.8333	0.8333
合计	0.9411	0.5600	0.8151	0.7600
LightGBM				
	训练集	测试集	训练集	测试集
舞弊	0.8135	0.4615	0.9830	0.6153
非舞弊	0.8166	0.8333	0.9333	0.5833
合计	0.8151	0.6400	0.9579	0.6000

从表 4-10 结果可知, 逻辑回归对于财务舞弊公司的识别表现很差, 仅为 0.3846, 甚至低于随机判断。在四种模型中, 综合表现最好是支持向量机, 它在测试集上对舞弊公司的识别准确率达到了 0.6923, 综合识别率达到 0.76。其他模型的综合识别准确率与舞弊公司识别准确率都远低于支持向量机。逻辑回归模型和决策树模型对于识别首次舞弊行为虽然在训练集表现良好, 但是泛化能力弱。显然, 在综合预警模型中, 支持向量机是预测能力最强的模型, 识别率 69.3%。

所以, 基于首次舞弊样本我们可以得出, 财务报告可信度指标-支持向量机模型对首次舞弊行为识别精度最高, 达到 76.92%; 综合模型对于首次舞弊行为欠佳, 识别率为 69.32%, 可能是由于首次舞弊行为财务异常波动小, 财务报告可信度指标对首次舞弊行为解释力强, 加入财务异常指标反而影响了识别效果。逻辑回归模型和决策树模型对于识别首次舞弊行为虽然在训练集表现良好, 但是泛化能力弱。

4.4.2 基于连续舞弊样本

本文选取的连续舞弊样本和控制样本数量存在差距，舞弊样本数量 625 个，非舞弊样本数量 1273 个，存在样本量失衡的情况，所以在基于连续舞弊的样本集对模型测评中我们综合使用总精度指标和 F-score 指标来评价模型的识别效果。基于连续舞弊样本的研究可以让我们看到模型侦别财务舞弊的精度。

(1) 财务报告可信度模型

表 4-11 财务报告可信度模型效果对比—基于连续舞弊样本集

非线性主成分				
	Logistic 回归		支持向量机	
	训练集	测试集	训练集	测试集
Accuracy	0.79	0.80	0.80	0.78
Precision	0.78	0.73	0.79	0.71
Recall	0.76	0.72	0.78	0.71
F-score	0.76	0.72	0.78	0.71
合计	0.775	0.76	0.79	0.745
LightGBM				
	训练集	测试集	训练集	测试集
Accuracy	0.82	0.80	0.81	0.75
Precision	0.87	0.77	0.80	0.66
Recall	0.76	0.66	0.79	0.65
F-score	0.78	0.68	0.79	0.66
合计	0.80	0.74	0.80	0.705

注：合计项为总精度指标和 F-score 指标的简单平均

结果显示，从总精度指标、查准率指标和查全率指标评价效果来看，非线性主成分 Logistic 回归和 LightGBM 的识别率都位列前两名，但是在综合评价指标 F-score 方面，非线性主成分 Logistic 回归要优于 LightGBM。根据本文的实证结果得出，财务报告可信度模型基于连续舞弊样本数据集，非线性主成分 Logistic 回归最优，达到 0.76。

总之，财务报告可信度模型对于首次舞弊和连续舞弊的上市公司识别效果都很好。对于首次舞弊公司的识别效果略优，支持向量机是预测能力最强的模型，识别率达到 77%，与我们初期预测基本一致。

(2) 财务异常模型识别效率

表 4-12 财务异常模型效果对比—基于连续舞弊样本集

		非线性主成分	
		Logistic 回归	支持向量机
		训练集	测试集
Accuracy	0.72	0.78	0.74
Precision	0.72	0.70	0.73
Recall	0.67	0.67	0.69
F-score	0.67	0.68	0.7
合计	0.695	0.73	0.735
		LightGBM	
		Cart 决策树	
		训练集	测试集
Accuracy	0.78	0.77	0.77
Precision	0.85	0.70	0.76
Recall	0.71	0.60	0.75
F-score	0.73	0.61	0.75
合计	0.755	0.69	0.76

注：合计项为总精度指标和 F-score 指标的简单平均

结果显示，财务异常模型对连续财务舞弊的识别效果均低于财务报告可信度模型。从总精度指标、查准率指标和查全率指标评价效果来看，非线性主成分 Logistic 回归和支持向量机的识别率都位列前两名，但是在综合评价指标 F-score 方面，支持向量机要优于非线性主成分 Logistic 回归。根据本文的实证结果得出，财务异常模型基于连续舞弊样本数据集，支持向量机最优，达到 0.735%。

总之，财务异常模型对于连续舞弊的上市公司识别效果很好，可以达到

73.5%，但是对于首次舞弊的上市公司识别效果欠佳，可能是由于首次舞弊公司的财务异常波动不显著。

(3) 综合识别模型的识别效率

表 4-13 财务舞弊综合识别模型效果对比-基于连续舞弊样本

非线性主成分				
	Logistic 回归		支持向量机	
	训练集	测试集	训练集	测试集
Accuracy	0.86	0.79	0.75	0.83
Precision	0.85	0.72	0.74	0.79
Recall	0.84	0.72	0.72	0.76
F-score	0.84	0.72	0.72	0.77
合计	0.85	0.755	0.735	0.80
LightGBM				
	训练集	测试集	训练集	测试集
Accuracy	0.81	0.82	0.8	0.78
Precision	0.83	0.77	0.79	0.71
Recall	0.76	0.72	0.78	0.72
F-score	0.78	0.74	0.78	0.71
合计	0.795	0.78	0.79	0.745
Cart 决策树				
	训练集	测试集	训练集	测试集

注：合计项为总精度指标和 F-score 指标的简单平均

结果显示，综合模型对连续舞弊行为识别效果较好，比单独使用财务报告可信度模型和财务异常模型识别效果更好。从总精度指标、查准率指标和查全率指标和 F-score 指标评价效果来看，支持向量机识别效果都是最优。根据本文的实证结果得出，综合模型基于连续舞弊样本数据集，支持向量机最优，达到 80%。

所以，基于首次舞弊样本可以得出，将财务报告可信度指标和财务异常指标同时纳入模型会提高连续舞弊识别效果，针对连续舞弊行为，使用综合指标模型预测效果更佳。这四个学习器在连续舞弊样本集中泛化能力都表现良好。

4.4.3 模型中的重要指标

根据上述建模分析，我们已知支持向量机在综合模型中表现为最佳。本文不仅希望构建机器学习模型对上市公司舞弊行为进行预测，还要识别在财务舞弊行为中具有显著影响和贡献的特征。由于支持向量机的特性，它是通过在特征空间上构建一个超平面从而对样本进行分类。而本文在将数据输入模型前对数据做了标准化处理，因此，模型系数绝对值越大的特征对该超平面的影响越大，也就表明它对整个模型的重要性越大。

(1) 首次舞弊样本

如表 4-14 所示，是在以连续舞弊样本为基础构建的综合模型中，识别效果最优的支持向量机模型系数绝对值排名前 10 的特征，根据表 4-14 的结果，我们发现 SOC、CEM 的模型系数显著为正，而 TIN、PLF、ALRA、SMNE、ALEV、NSOP、OCRH、CHAN 模型系数显著为负。这说明 SOC、CEM 指标数值越高的机构，越容易产生财务舞弊行为，而 TIN、PLF、ALRA、SMNE、ALEV、NSOP、OCRH、CHAN 指标数值越高的机构，越不容易产生财务舞弊行为。

表 4-14 top10 重要指标-首次舞弊

重要指标	指标名称	SVM 模型系数
SOC	舞弊前公司偿债能力	0.38
CEM	营业毛利率	0.23
TIN	是否微利公司	-0.13
PLF	舞弊前是否亏损	-0.12
ALRA	资产负债率	-0.11
SMNE	董事长兼任总经理	-0.11
ALEV	舞弊前应计方向	-0.11
NSOP	舞弊前三年曾被标出 非标意见	-0.10 -0.10
OCRH	股权集中度	-0.10
CHAN	高管更迭	-0.10

如表 4-14 所示，识别首次舞弊行为时模型系数绝对值排名前 10 的特征有 7 个属于财务报告可信度指标，只有 3 个是财务异常指标。这也从侧面说明了财务报告可信度指标对首次舞弊行为的识别效果更好。

其中董事长兼任总经理、股权集中度属于舞弊机会变量，兼任会产生治理结构四方制衡失衡，过高或过低的股权集中度都不利于管理。舞弊前是否亏损、舞弊前公司偿债能力、高管更迭属于舞弊动机变量，公司为了获取在资本市场上投资者的青睐或者是为了获取债权人的信任，有可能会选择通过包装公司财务报表达成目标，通过对这些指标监测可以帮助识别公司的舞弊动机。舞弊前应计方向、舞弊前三年曾被标出非标意见属于舞弊态度和借口变量，管理层对会计政策的激进程度和对管理层诚信规则意识的度量可以帮助识别管理层舞弊态度和借口。

是否微利公司、资产负债率、营业毛利率分别代表了企业处于盈亏临界点的指标、偿债能力指标和盈利能力指标。当企业面临债务压力、处于盈亏临界点边缘和盈利能力较差时，企业会有强烈的舞弊动机维持公司形象。

(2) 连续舞弊样本

如表 4-15 所示，展示了在支持向量机模型中，模型系数绝对值都排名在前 10 的特征，我们发现 GROW、CURA、STR、ALRA 和 NSOP 的模型系数显著为正，而 GPMS、DETA、ALEV、TATP 和 QURA 的模型系数显著为负。这说明 GROW、CURA、STR、ALRA 和 NSOP 财务指标数值越高的机构，越容易产生财务舞弊行为，而 GPMS、DETA、ALEV、TATP 和 QURA 财务指标数值越高的机构，越不容易产生财务舞弊行为。

表 4-15 top10 重要指标-连续舞弊

重要指标	指标名称	SVM 模型系数
GROW	舞弊前公司成长性	1.541683
CURA	流动比率	0.871653
STR	存货周转率	0.439457
ALRA	资产负债率	0.402746
NSOP	舞弊前三年曾被 标出非标意见	0.40144

GPMS	销售毛利率	-0.40138
DETA	有形净值债务率	-0.42755
ALEV	舞弊前应计方向	-0.42794
TATP	所得税/利润总额	-0.50307
QURA	速动比率	-0.6687

GROW 是公司舞弊前一年的销售增长率，属于公司舞弊动机变量。销售增长率高的公司说明成长性较好，说明公司正在进行大规模扩张，这种扩张和维持经营需要大量资金支持，而资本市场是一个巨大的资金池，是企业资金来源的重要途径和方式，所以产生了管理层舞弊的动机。NSOP 和 ALEV 指标代表舞弊借口因素，分别表示公司在舞弊前三年是否曾被出具非标审计意见以及舞弊前应计方向。由于舞弊前就曾存在迹象，而虚构利润等舞弊行为都是延续为之，是管理层自欺欺人，使其行为合理化。可见，舞弊动机变量和舞弊借口变量的加入可以提高财务舞弊预测模型的识别效果。

CURA、QURA、DETA 和 ALRA 指标代表了公司的债务管理能力，几乎包含了本文中所有代表债务管理能力的指标，由此可见公司财务杠杆比例、债务结构与财务舞弊之间存在很大的关系，不良债务结构会增加财务舞弊的可能性。

STR、GPMS 和 TATP 指标分别是存货周转率、舞弊当期销售毛利率和所得税比率，分别代表了资产管理能力、经营管理能力和会计信息质量。存货周转率与财务舞弊行为正相关，可能是因为舞弊公司在虚假增加销售收入的基础，为了不让销售成本与销售收入比值过于悬殊，也使得虚假销售成本增加。虚增利润公司可能由于虚增收入导致利润虚高而导致当期销售毛利率较高，也有可能由于当期毛利率较低而导致舞弊行为，在本文实证分析中得出的结论是当期毛利率越低越容易发生舞弊行为。所得税率指标越低，发生舞弊的可能性越大，是因为舞弊公司的应税所得可能显著小于税前会计利润。

在识别公司财务舞弊行为时，可以重点关注舞弊前应计方向、舞弊前三年曾被标出非标意见、资产负债率这三个指标，对于连续舞弊行为和首次舞弊行为的识别都具有较好的效果。

4.5 模型实证结果分析

表 4-16 财务舞弊识别模型效果对比

	财务报告可信度指标		财务异常		综合指标	
	最优模型	识别效果	最优模型	识别效果	最优模型	识别效果
连续舞弊样本	逻辑回归	76%	支持向量机	73. 50%	支持向量机	80%
首次舞弊样本	支持向量机	77%	决策树	61. 85%	支持向量机	69. 30%

汇总实证分析结果，将各部分识别效果最优的结果放入表 4-16 中，通过分析实证结果，我们可以发现：

- (1) 将财务报告可信度指标和财务异常指标同时纳入模型会提高预测连续舞弊识别效果，针对连续舞弊行为，使用综合指标模型预测效果更佳。
- (2) 财务异常指标的加入反而影响了首次舞弊的预测效果，针对首次舞弊行为，对财务报告可信度的预测，即对舞弊发生可能性的评估，对舞弊行为的发生及时甄别精度更高。
- (3) 基于传统计量模型的逻辑回归模型已不再适用于目前错综复杂的金融市场信息，即使我们认为地对其加入非线性信息，仍旧无法有效提高其样本外的预测能力。从结果中我们也可以看到，支持向量机对财务舞弊各样本和各类型指标均表现良好。这也启示我们财务舞弊领域对于机器学习的应用应该向更广、更深、更细的方向发展。
- (4) 财务报告可信度指标主要是非财务指标信息，根据实证结果显示，这些非财务指标信息的预测精度高于财务指标信息，这也启示我们在构建财务舞弊识别模型时，要综合考虑财务指标和非财务指标，并且重点关注非财务指标的选取。

第五章 研究结论及展望

5.1 研究结论

5.1.1 机器学习模型识别能力

根据本文实证分析的结果，在舞弊公司识别中不同类型的指标会对不同模型的识别效率产生影响。具体来说，无论对于哪类样本集，支持向量机综合识别模型中对于舞弊公司的识别精确率最高，这可能是由于支持向量机模型以最大置信度的情况将样本隔离开来以完成分类预测，其复杂度只依赖于支持向量而非模型参数，因此模型结构十分简洁，不易过拟合，因此在小样本的情况下表现优异。逻辑回归模型和决策树模型泛化能力较弱。逻辑回归在财务报告可信度模型中表现良好，对于连续舞弊样本的识别效果可以达到 76%，但由于非线性 Logistic 模型的交互项特征数量过多，所以可能在其他模型上表现不佳。决策树在财务异常指标中表现较好，但总体仍表现不佳，这可能是由于决策树对于数据的划分不依赖于缩放，当二元特征与连续特征同时出现时，决策效果会更好。

这启示我们，基于传统计量模型的逻辑回归模型已不再适用于目前错综复杂的金融市场信息，即使我们认为地对其加入非线性信息，仍旧无法有效提高其样本外的预测能力。非线性特征工程需要消耗大量的人力和经验，效果却不增反降。仅仅依靠专家和注册会计师经验判断应对财务舞弊风险的时代在过去，借助机器学习模型作为辅助工具可以帮助识别舞弊风险。

5.1.2 财务报告可信度模型、财务异常模型和综合模型对比

根据本文实证分析的结果，在舞弊公司识别中财务报告可信度指标和财务异常指标对舞弊识别效率不同。

对于连续舞弊样本集，综合模型识别效果最好，说明财务异常指标和财务报告可信度指标中所选的公司特征对公司连续舞弊行为都具有良好的解释力。因此，在构建连续舞弊识别模型时综合考虑财务异常指标和财务报告可信度指标会提高模型预测效果。

对于首次舞弊样本集，加入财务异常指标后的综合模型反而识别效果下降了，这说明本文在财务异常模型中所选的公司特征可能与公司发生首次舞弊行为没有明显的联系，加入这些特征使得整个模型出现了特征冗余的情况，反而使得模型得预测降低。

因此，本文所构建得财务报告可信度模型是预警上市公司首次财务造假行为的更优选择，财务报告可信度模型中所选公司特征对企业首次舞弊行为的解释效力更高。构建得综合模型是甄别上市公司连续财务舞弊行为的更优选择，财务报告可信度模型和财务异常模型中所选公司特征对企业连续舞弊行为的解释效力都很好。

5.1.3 财务舞弊预警模型中的特征重要性

在预警企业首次舞弊行为时，可以重点关注舞弊前公司偿债能力、营业毛利率、是否微利公司、舞弊前是否亏损、资产负债率等指标。

在甄别企业连续舞弊行为时，可以重点关注舞弊前公司成长性、流动比率、存货周转率、资产负债率、舞弊前三年曾被标出非标意见等指标

在识别公司财务舞弊行为时，可以重点关注舞弊前应计方向、舞弊前三年曾被标出非标意见、资产负债率这三个指标，对于连续舞弊行为和首次舞弊行为的识别都具有较好的效果。

5.2 展望

根据本文对财务舞弊识别模型的研究，提出以下研究展望：

第一，利用网络数据挖掘进一步研究。本文的财务指标和非财务指标构建大多来自于公司财务报表和公司公告，具有一定的滞后性，并且数据量小，无法完全发挥机器学习的优势。在下一阶段的研究中可以将数据挖掘结合文本挖掘，从新闻报道、舆论大众、专家经验等多个维度考量，发掘更多的财务舞弊特征。

第二，分行业研究财务舞弊模型。本文虽然在样本选取中考虑了行业分布情况，但是缺乏对细分行业的研究。不同行业公司舞弊行为存在差异，精准定位行业可比公司，对不同行业舞弊企业画像，下沉到行业构建行业图谱是下一阶段的研究重点。

第三，完善机器学习综合模型。在实证分析中发现对于财务报告可信度指标和财务异常指标采用不同的机器学习模型识别效率更好，但是简单粗暴地将两个指标混合在一起的综合模型有些生硬，在后续研究中，如何将财务报告可信度模型和财务异常模型深度融合，提升综合模型识别效率是下一阶段研究重点。

参考文献

- [1] Albrecht, W.S and M.B Romney. Red flagging Management Fraud:A Validation[J].Advances in Accounting, 1986 (3):323-333.
- [2] Anderson, Roald C;Mansi, Sattar A;Reeb, David M.2004, Board Characteristics, Accounting Report Integrity, and the Cost of Debt[J]. Journal of Accounting & Economics, 37(3):315-342.
- [3] Bahram Soltani. The Anatomy of Corporate Fraud: A Comparative Analysis of High Profile American and European Corporate Scandals [J]. Journal of Business Ethics, 2014(2):251-274.
- [4] Beasley, Carcello, Hermanson and Lapidus.Fraudulent Financial Reporting: Consideration of Industry Traits and Corporate Governance Mechanisms. Accounting Horizons.2000.
- [5] Beaver W H.Financial ratios as predictors of failure [J].Journal of Accounting Research, 1966, (Supplement) :71-111.
- [6] Bell T, Carcello J. A Decision Aid for Assessing the Likelihood of Fraudulent Financial Reporting[J]. Auditing A Journal of Practice & Theory , 2000 , 19(1):169-184.
- [7] Bell Timothy B and Carcello Joseph V. A Decision Aid for Assessing the Likelihood of Fraudulent Financial Reporting, Auditing, 2000 (Spring) .
- [8] 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):271-309.
- [9] Beneish, M. D. Incentives and penalties related to earnings overstatements that violate GAAP, Accountine Review.1999(10): 425-457.
- [10]Beneish, M.D. Detecting GAAP violation: implications for assessing earnings management among firms with extreme financial performance , Journal of Accounting and Public Policy, 1997 (16): 271~309.
- [11]Bologna J. The one minute fraud auditor[J]. Computers & Security, 1989 , 8(1):29-31.
- [12]Bruno Frey.Yes, Managers Should Be Paid Like Bureaucrats[J].Economics and Organization, 2013, (8).

- [13] Cohen J, Ding Y, Cédric Lesage, et al. Corporate Fraud and Managers' Behavior: Evidence from the Press[J]. *Journal of Business Ethics*, 2010, 95(2 Supplement):271-315.
- [14] Dechow P M, Ge W, Larson C R, et al. Predicting Material Accounting Misstatements[J]. *Contemporary accounting research*, 2011, 28(1): 17-82.
- [15] Fanning K M, Cogger K O. Neural Network Detection of Management Fraud Using Published Financial Data. *International Journal of Intelligent Systems in Accounting, Finance & Management*, 1998, 10(7):21-41.
- [16] Fraud in Organizations[J]. *Journal of Business Ethics*, 2011, 101(4):601-618.
- [17] Glancy F H, Yadav S B. A computational model for financial reporting fraud detection[J]. *Decision support System*, 2011, 50(3):595-601.
- [18] Goel S, Uzuner O. Do Sentiments Matter in Fraud Detection? Estimating Semantic Orientation of Annual Reports[J]. *Intelligent Systems in Accounting Finance & Management*, 2016, 23(3):215-239.
- [19] Green B. P and J. H. Choi. Assessing the Risk of Management Fraud through Neural Network Technology. *Auditing*. 1997, Spring (16): 14~28.
- [20] Lee, Ingram and Howard. The Difference between Earnings and Operating Cash Flow as an Indicator of Financial Reporting Fraud[J]. *Contemporary Accounting Research*, 1999(4): 749-846.
- [21] McCarthy, John. Using Altman Z-score and Beneish M-score Models to Detect Financial Fraud and Corporate Failure: A Case Study of Enron Corporation[J]. *International Journal of Finance and Accounting*, 2017, 6(6): 159-166.
- [22] Mark S. Beasley. An Empirical Analysis of the Relation Between the Board of Director Composition and Financial Statement Fraud[J]. *The Accounting Review*, 1996(71):443-465.
- [23] Murphy M, Dacin T. Psychological Pathways to Fraud: Understanding and Preventing
- [24] P.Ravisankar, V. Ravi, G.RaghavaRao, Detection of financial statement fraud and feature selection using data mining techniques [J]. *Decision Support Systems*, 2011(50):491-500.
- [25] Persons. O. S. 2006. Corporate Governance and Non-Financial Reporting Fraud[J]. *Journal of Business& Economics Studies*, 12(1):27-38.
- [26] Persons. Using Financial Statement Data to Identify Factors with Fraudulent

- Financing Reporting[J].Journal of Applied Business Research, 1995(11):38-46.
- [27] Tarjo, Herawati N. The Comparison of Two Data Mining Method to Detect Financial Fraud in Indonesia[J]. Social Science Electronic Publishing, 2017, (2):1-8.
- [28] Wilks T, Zimbelman M. Decomposition of Fraud Risk Assessments and Auditors' Sensitivity to Fraud Cues[J]. Social Science Electronic Publishing, 2004, 21(3):719–745.
- [29] 陈国欣, 吕占甲, 何峰.财务报告舞弊识别的实证研究——基于中国上市公司经验数据[J]. 审计研究, 2007, 03:88-93.
- [30] 陈庆杰.基于经理人特征的财务报告舞弊识别模型的改进研究---来自中国上市公司的实证检验[J].经济问题, 2012 (8) :118-122.
- [31] 程安林, 梁芬莲, 季洁.基于会计舞弊的内部控制形式有效性研究[J]. 南京审计学院学报, 2013, 02:77-87.
- [32] 崔东颖, 胡明霞.“雅百特”财务舞弊案例研究——基于舞弊三角理论的视角 [J]. 财会通讯, 2019(04):6-9.
- [33] 戴丹苗, 刘锡良. 中概股公司财务舞弊的文献综述[J]. 金融发展研究, 2017(1):11-19.
- [34] 邓庆山, 梅国平. 基于 BP 神经网络的虚假财务报告识别[J]. 系统工程, 2009, 10:70-75.
- [35] 邓庆山, 梅国平. 基于支持向量机的虚假财务报告识别研究[J]. 当代财经, 2009, 07:105-108.
- [36] 方军雄.我国上市公司财务欺诈鉴别的实证研究[J].上市公司, 2003(9):23-25.
- [37] 韩丹, 闵亮, 陈婷. 管理层股权激励与上市公司会计造假相关性的实证检[J]? 统计与决策, 2001(18):69-72.
- [38] 洪芸, 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.
- [39] 洪芸, 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, 08:8.
- [40] 皇甫冬雪.基于 Lib-SVM 的损益调整类财务报告舞弊识别模型研究——来 自中国证券市场的证据[J]. 会计之友, 2011, 9(1):75-78.
- [41] 阚宝奎, 刘志新, 宋晓东, 杨众.改进支持向量机在虚假财务报告识别中的应 用[J]. 管理评论, 2012, 05:144-153.
- [42] 李清, 任朝阳.基于非线性—主成分 Logistic 回归的会计舞弊识别研究[J].统 计与信息论坛, 2016, (3) : 75—80.

- [43] 娄权.财务报告舞弊四因子假说及其实证检验[D]. 厦门大学, 2004.
- [44] 卢馨, 李慧敏, 陈炼辉.高管背景特征与财务舞弊行为的研究——基于中国上市公司的经验数据[J]. 审计与经济研究, 2015, 06:58-68.
- [45] 蒙肖莲, 李金林, 杨毓. 基于概率神经网络的欺诈性财务报告的识别研究[J]. 数理统计与管理, 2009, 28(1):36-45.
- [46] 秦江萍. 企业会计舞弊及其控制:基于博弈理论的分析[J]. 财经论丛(浙江财经大学学报), 2005(4):54-58.
- [47] 饶斌. 基于 GONE 理论视角的上市公司财务报告舞弊研究[J]. 江西财经大学学报, 2011(4):34-40.
- [48] 施金龙, 韩玉萍.基于 GONE 理论的上市公司财务舞弊分析 [J]. 会计之友, 2013(23):98-100.
- [49] 王泽霞, 谢冰.审计质量替代指标谁更有效:来自被查处上市公司的经验证据 [J].中国注册会计师, 2010(7):23-29.
- [50] 王泽霞, 谢冰. 审计质量替代指标谁更有效: 来自被查处上市公司的经验证据[J]? 中国注册会计师, 2010 (7): 23-29.
- [51] 吴金菊.会计舞弊的不确定性视角分析[J]. 商业会计, 2012, 08:104-105.
- [52] 夏明, 李海林, 吴立源. 基于神经网络组合模型的会计舞弊识别[J].统计与决策, 2015, 16:49-52.
- [53] 谢小莹, 孙燕东.基于舞弊三角形理论的会计舞弊影响因素研究[J]. 财会通讯, 2014, 27:3-7.
- [54] 谢小莹, 孙燕东.基于舞弊三角形理论的会计舞弊影响因素研究[J].财会通讯, 2014(27):3-7.
- [55] 熊方军, 张龙平.上市公司财务舞弊的风险识别与证据收集[J]. 经济与管理研究, 2016, 10:138-144.
- [56] 许存兴.基于现金流量指标的企业财务舞弊分析 [J].中国注册会计师, 2013(2):107-114.
- [57] 杨薇, 姚涛. 公司治理与财务舞弊的关系——来自中国上市公司的经验证据 [J].重庆大学学报 (社会科学版), 2006, 12 (5): 24-30.

致谢

来到南京读书已经 7 年，转眼间研究生生涯也要结束，马上就要告别学生生涯，心中无限感慨。这篇文章是我以前从未涉及的全新领域，虽然仍不能令我满意，但也实属不易。论文的完成包含了导师对我的敦敦教导，包含了同学对我的帮助，也包含了父母对我的支持。此时此刻，有太多人想要感谢。

在研究生涯中，我最想感谢的是我的导师。研究生期间，我有幸成为李心丹教授的学生，李老师对学术的严谨，对工作的热情投入永远都是我学习的榜样。在研究生一年级李老师鼓励我们参加华泰证券的量化实习，让我对金融领域有了全新的认识，研究生二年级让我参与了科创板评估课题，研究能力和文字撰写能力有了很大提升。研究生阶段，在李老师身上学到的优秀品质，将会成为我人生宝贵的财富。

我也想感谢李心丹老师团队的所有老师，在每次论文研讨会上，都给我们提出很多宝贵意见。尤其想感谢李昊骅师兄，深夜还在帮我们批改论文，帮我们理清论文框架。还想感谢工程管理学院所有的老师，在研一研二期间，由于你们的悉心教导，我学到了扎实的金融学知识和专业技能，你们为人师表的风范令我敬仰，你们严谨治学的态度令我敬佩。

我还想感谢关心和帮助我的同学们，三年来，我们朝夕相处，共同进步，一起上课，一起去牛津大学交流学习，是你们让我这一路走来倍感温暖。也要感谢我的舍友季娟宇和江畅，陪我度过了快乐的研究生时光。我还要感谢我的父母，悉心爱护培养我这么多年，无论在什么时候都是默默的支持我陪伴我。

最后，我想感谢南京大学，“诚朴雄伟，励学敦行”的校训深深刻在我的脑海，在以后的生活工作中，我会时刻谨记，不忘初心。

想感谢的人太多，限于篇幅无法一一列举，在此表示由衷的感谢！

《学位论文出版授权书》

本人完全同意《中国优秀博硕士学位论文全文数据库出版章程》(以下简称“章程”),愿意将本人的学位论文提交“中国学术期刊(光盘版)电子杂志社”在《中国博士学位论文全文数据库》、《中国优秀硕士学位论文全文数据库》中全文发表。《中国博士学位论文全文数据库》、《中国优秀硕士学位论文全文数据库》可以以电子、网络及其他数字媒体形式公开出版,并同意编入《中国知识资源总库》,在《中国博硕士学位论文评价数据库》中使用和在互联网上传播,同意按“章程”规定享受相关权益。

作者签名: 冯嘉金
2021 年 5 月 20 日

论文题名	<u>基于机器学习的财务舞弊识别模型研究</u>				
研究生学号	<u>MF1815012</u>	所在院系	<u>工程管理学院</u>	学位年度	<u>2021年</u>
论文级别	<input type="checkbox"/> 学术学位硕士 <input checked="" type="checkbox"/> 专业学位硕士 <input type="checkbox"/> 学术学位博士 <input type="checkbox"/> 专业学位博士 (请在方框内画钩)				
作者 Email					
导师姓名	<u>李心丹</u>				

论文涉密情况:

不保密

保密, 保密期(____年____月____日至____年____月____日)