硕士学位论文

（学术学位论文）

基于机器学习的中国上市公司财务舞弊识别方法研究

RESEARCH ON FINANCIAL FRAUD IDENTIFICATION METHOD OF CHINESE LISTED COMPANIES BASED ON MACHINE LEARNING

周明昊

哈尔滨工业大学

2022 年 6 月

国内图书分类号：C931.6　　学校代码：10213
国际图书分类号：004.62　　密级：公开

硕士学位论文

基于机器学习的中国上市公司财务舞弊识别方法研究

硕士研究生　周明昊
导　　师　芦鹏宇副教授
申请学位　管理学硕士
学　　科　管理科学与工程
所在单位　经济与管理学院
答辩日期　2022 年 6 月
授予学位单位　哈尔滨工业大学

Classified Index: C931.6
U.D.C: 004.62

Dissertation for the Doctoral Degree

RESEARCH ON FINANCIAL FRAUD IDENTIFICATION METHOD OF CHINESE LISTED COMPANIES BASED ON MACHINE LEARNING

Candidate： Zhou Minghao
Supervisor： Associate Prof.Lu Pengyu
Academic Degree Applied for： Master of Management
Speciality： Management Science & Engineering
Affiliation： School of Economics Management
Date of Defence： June, 2022
Degree-Conferring-Institution： Harbin Institute of Technology

摘要

上市公司的财务舞弊问题时常发生，财务舞弊的行为，严重损害了相关投资人的利益，干扰了市场的正常秩序。大多数投资人无法对上市公司的财务报表进行有效的甄别，监管部门的资源精力有限，无法对所有公司进行细致的检查。因此研究财务舞弊的特征、构建有效的财务舞弊识别模型具有重要意义。

本文首先回顾了国内外的研究文献，明晰了财务舞弊的相关概念，为研究的进行提供了理论依据和方向。本文以 2007-2019 年的中国上市公司(非金融行业)为研究对象，选取期间的上市公司年度财报数据样本，以相关监管部门的处罚公告为标准，在明确财务舞弊定义后对舞弊公司记录进行了标记。对于公司舞弊数据，只保留了连续舞弊记录的第一次舞弊数据，并在文章后续对比证明了对连续舞弊样本处理的必要性。

其次，为讨论在基于机器学习的中国上市公司财务舞弊识别中，原始会计指标体系是否会更优于财务比率指标体系，本文依据财务舞弊三角理论选择了财务指标，并找到对应的原始会计数据指标。在删除缺失值较多的指标后，分别构建了二套指标体系，27 个财务比率指标作为财务比率指标体系、27 个原始会计指标作为原始会计指标体系。实证建模结果表明，在中国上市公司财务舞弊识别中，原始会计数据指标体系并不比财务比率指标体系更优。

考虑到财务数据在时间维度上的连续性，对各指标进行时间维度上扩充，使得每条样本记录包含其前三年在内的数据指标及方差，通过对财务比率指标体系、原始会计指标体系及合并的指标体系分别进行时间维度的扩充，实验证明指标时间维度扩充是有效的。本文使用随机森林、支持向量机和自适应增强三种主流的机器学习算法，构建并分析了多个的财务舞弊识别模型，实验表明，随机森林算法表现最好。

最后，将本文提出的最优的财务舞弊识别模型与基准模型(Benchmark Model)做对比，依据基准模型使用的美国会计指标，选取了 32 个中国的原始会计数据指标，删除缺失值过多的指标后，获得 25 个原始会计数据指标。将基准模型在中国上市公司数据进行建立后与本文提出的模型对比，结果表明，本文提出的中国上市公司财务舞弊识别模型的 AUC 十次均值为 0.697，NDCG@k 均值为 0.200，明显高于基准模型的 0.626 与 0.134，有较好的识别效果和实际应用价值。

关键词：财务舞弊识别；机器学习；原始会计数据；连续舞弊；时间维度

Abstract

The financial fraud of listed companies often occurs, which seriously damages the interests of relevant investors and interferes with the normal order of the market. Most investors are unable to effectively identify the financial statements of listed companies, and the regulatory authorities have limited resources and energy, so they are unable to carefully inspect all companies. Therefore, it is of great significance to study the characteristics of financial fraud and build an effective identification model of financial fraud.

Firstly, this paper reviews the research literature at home and abroad, clarifies the related concepts of financial fraud, and provides a theoretical basis and direction for the research. This paper takes the Chinese listed companies (non-financial industry) from 2007 to 2019 as the research object, selects the annual financial report data samples of listed companies during the period, and marks the records of fraudulent companies after clarifying the definition of financial fraud based on the punishment announcements of relevant regulatory authorities. For the company's fraud data, only the first fraud data of continuous fraud records is retained, and the subsequent comparison in the article proves the necessity of processing continuous fraud samples.

Secondly, in order to discuss whether the original accounting indicator system will be better than the financial ratio indicator system in the identification of financial fraud of Chinese Listed Companies Based on machine learning, this paper selects the financial indicators according to the financial fraud triangle theory and finds the corresponding original accounting data indicators. After deleting the indicators with many missing values, two sets of indicator systems are constructed, 27 financial ratio indicators as the financial ratio indicator system and 27 original accounting indicators as the original accounting indicator system. The empirical modeling results show that the original accounting data index system is not better than the financial ratio index system in the identification of financial fraud of Chinese listed companies.

Considering the continuity of financial data in the time dimension, each indicator is expanded in the time dimension, so that each sample record includes the data indicators and variances of the previous three years. Through the time dimension expansion of the financial ratio indicator system, the original accounting indicator system and the consolidated indicator system, the experiment shows that the index time dimension expansion is effective. In this paper, three mainstream machine learning algorithms, random forest, support vector machine and adaptive enhancement, are used to build and analyze several financial fraud identification models. The experiments show that the random forest algorithm performs best.

Finally, the optimal financial fraud identification model proposed in this paper is compared with the benchmark model. According to the American Accounting indicators used in the benchmark model, 32 Chinese original accounting data

indicators are selected. After deleting the indicators with too many missing values, 25 original accounting data indicators are obtained. After the benchmark model is established in the data of Chinese listed companies, it is compared with the model proposed in this paper. The results show that the AUC ten times mean value of the financial fraud identification model of Chinese listed companies proposed in this paper is 0.697, NDCG@k The average value is 0.200, which is significantly higher than 0.626 and 0.134 of the benchmark model. It has good recognition effect and practical application value.

Keywords： Identification of financial fraud, Machine learning, Original accounting data, Continuous fraud, Time dimension

目录

第 1 章 绪论

1.1 研究背景及研究意义

1.1.1 研究背景

著名的财务舞弊案件始于 1720 年的英国南海公司，自那以后，美国的世通、日本的东芝、奥林巴斯舞弊案等纷纷涌现。1929 年美国经济危机期间，许多上市公司为了掩盖商业危机，在市场上散布虚假的财务信息，这也是大萧条期间许多投资者做出错误决策的原因之一。从 21 世纪开始，世界上最大的能源交易商安然公司被爆涉嫌伪造虚假利润 5.86 亿美元，该丑闻直接导致这个曾经的巨头破产。而安然在审计中的蒙蔽行为也致使安达信事务所，这一排名世界前五的会计师事务所宣布破产。

20 世纪 90 年代以来，我国证券市场逐步形成和发展，但财务舞弊案件也屡见不鲜。中水国际、长城机电、深圳原野这三大造假案严重破坏了投资者们的信心，紧接着的蓝田和银广厦舞弊事件更是全国闻名。相关监管部门对 2013 年发生的万福生科财务造假案中上市证券的管理人员和相关人员实施了相当严厉的制裁，但在紧接着的 2014 年，南纺股份、新中基就接连被爆出财务舞弊。2016 年金亚科技因涉嫌包括建立多个账套、虚增利润、虚增银行存款、虚增预付工程款等在内的多项财务造假行为被证监会处罚，后退市。2020 年 4 月，瑞幸咖啡财务造假事件震惊中外。该公司成功进入美国上市后，创造了一个资本神话，后却被浑水公司曝光虚增 22 亿交易，股价下跌了 85%，被迫退市。

图 1-1 是 2007 年——2019 年间因虚构利润、虚列资产、重大遗漏、虚假记载这 4 个问题而被证监会、交易所等监管机构公开处罚的上市公司数量情况(同一家公司的连续舞弊记作多次)。

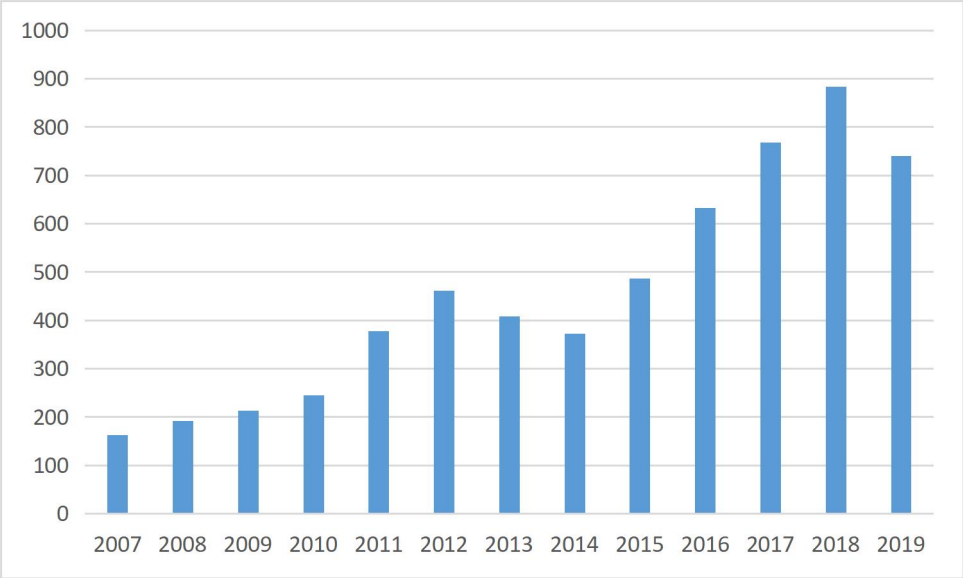

图 1-1 2007 年到 2019 年中国上市公司财务舞弊情况

由上图 1-1 可以看到，因 2007 年——2019 年间进行财务舞弊而被相关监管机构公开处罚的上市公司数量整体呈现增加趋势，由于与非上市公司相比，一般来说上市公司的规模更大，若发生财务舞弊则造成的影响也更加恶劣。

1.1.2 研究意义

如前所述，由于各种复杂利益的驱使，财务舞弊成为资本市场难以根治的问题，会对财会信息的质量造成严重且直接的影响。财务舞弊会对市场产生严重的干扰行为，严重破坏投资者信心，形成不良氛围，不利于国家市场经济的正常发展，而财务舞弊识别模型可以在一定程度上帮助解决这类问题。各审计单位和职能部门的检查和制度使财务舞弊问题得到了一定的遏制，但如今社会上对财务报表粉饰的方法极多，财务舞弊的手段也越来越复杂和隐蔽，单纯的人工审计越发的困难，且调查、检查的成本较高。不过无论造假的手段多么高超，虽能使各报表达到平衡，但会计上的勾稽关系都难免出现异常，甚至矛盾。鉴于财务数据之间存在不可分割的联系性，机器学习方法在财务舞弊识别问题上提供了一种新的解决思路的同时，仍有很多问题有待讨论。具体而言，本文的研究意义及价值在于：

（1）有助于外部监管机构聚焦监察重心，提高监管效率

由于审计程序本身的局限性和审计合谋的存在，审计师很难完美得完成审计工作。同时证监会、证券交易所、财政部等外部监管机构受监管资源有限，只能重点审查部分公司，而财务舞弊识别模型可以帮助监管机构将更多注意力

放在高舞弊风险的公司上，在一定程度上提高监管效率。

（2）有助于投资者及其他相关信息使用者规避投资风险

由于很多投资者精力及专业能力的有限，外加信息的匮乏，对上市公司的财务舞弊行为不具有识别能力。有效的财务舞弊识别模型可以在一定程度上为投资者提供建议信息，使得投资者在风险控制和规避上多了一个手段，从而降低因财务舞弊导致的投资风险。

（3）丰富我国现有上市公司财务舞弊识别方法研究理论

上市公司每年都会对外披露大量数据，包括财务数据及非财务数据，这为机器学习提供了较为良好的应用环境，从而带来传统审计方法无法实现的巨大优势。本文从财务舞弊理论和现有研究出发，阐述分析了财务舞弊的发生原因，为财务舞弊的识别提供了相关理论基础。同时针对机器学习在该领域的具体问题(如连续舞弊问题、指标体系构建问题等)进行深入讨论，丰富了我国现有上市公司财务舞弊识别方法研究的理论。

1.2 国内外研究现状及分析

1.2.1 财务舞弊概念辨析

在对上市公司财务欺诈识别模型展开深入研究之前，我们需要对财务舞弊这一概念进行明确界定。

美国《审计准则公告第 16 号》将“financial fraud”定义为：蓄意编制虚假财报。张佳佳(2021)认为，美国定义的“financial fraud”应与中文里的财务舞弊相对应。中国对“舞弊”的定义和美国对“financial fraud”的定义高度相似。而“财务舞弊”与“财务欺诈”，是“financial fraud”的两种译法，前者是意译，后者为直译，两个词语的概念一致[1]。

国内外审计机构、注册会计师协会对财务欺诈的定义略有不同。COSO 委员会 1999 年报告中有关财务报告舞弊的论述，认为财务报告舞弊是在财务报表或财务相关信息披露中存在的有意错报，或从事其他对财务报表或财务披露有重大直接影响的非法行为。美国注册会计师协会(AICPA)在其 2002 年发布的第 99 号《审计准则公告》(SAS No. 99)对会计舞弊的定义同样关注了舞弊的主观故意性和违法违规性，认为会计舞弊是为了欺骗财务报表使用者而对财务报告或相关财务信息进行有意识地错报或漏报[2]。中国《独立审计具体准则第 8 号》则将财务舞弊定义为致使会计报表产生虚假反映的有意行为。《中国注册会计师审计准则第 1141 号》指出：“舞弊是指被审计单位的管理层、治理

层、员工或第三方使用欺骗手段获取不当或非法利益的故意行为[3]。”事实上需要指出，“财务舞弊”与“财务报告舞弊”的概念不完全相同，“财务舞弊”在含义拓展性上应大于“财务报告舞弊”，但在实际应用时，常常混用。

1.2.2 财务舞弊的特点和主要手段

财务舞弊的手段多种多样，国内外学者在此方面已有较为丰富的研究。1999 年发布的 COSO 报告总结得出了美国公司最常采用的造假手段为虚增收入、虚增资产的结论。Beasley(2000)针对不同行业进行了研究，发现不同行业的财务舞弊方式存在不同，金融相关公司常常滥用资产，而科技型公司多使用虚增收入的方法[4]。Dechow(2011)认为舞弊公司进行盈余操作是一个重要的的欺诈手段，并进行了相关研究[5]。Zager(2016)的研究表明，虚增资产是财务舞弊的常用方式[6]。

在国内研究方面，阎达五等(2001)研究发现，我国上市公司往往通过关联交易来提高投资收益或操纵利润。同时，他们倾向于增加赊销或确认不应提前记录的销售收入，或直接利用虚假销售增加营业利润[7]。韩文明(2005)发现，虚构交易是上市公司舞弊的常用手段[8]。吴晓迪(2011)财务造假的手段随着政策的发展而演变，往往变得更加复杂，调节收支、操纵利润和隐瞒重大事件信息是常见的财务欺诈手段[9]。曾汝林(2011)认为公司在业务上的造假以“假账真做”和“账假做真”的方式存在。“假账真做”本公司无实际业务，但按照适当的会计政策进行变更的业务，例如虚构销售业务，虚构业务招待费、广告费、办公费等支出，虚列职工人数等；“真账假做”指拥有真实业务但在会计上“粉饰”业务，包括使用不当的确认和计量方法对会计要素进行确认、计量和利用关联交易操纵利润[10]；对在对许多财务舞弊公司进行调查后，王淑玲(2012)认为不披露关联方交易是一些公司的舞弊手段[11]。

刘永(2013)对绿大地的财务造假事件分析，发现虚增资产、虚列收入等是其造假手段[12]。相似的是，曹媛等(2015)通过对新大地舞弊事件进行研究，指出虚增利润、资产和收入以及隐瞒关联方交易是新大地财务舞弊最主要的手段[13]。刘元，林爱梅等(2015)根据证监会 2008 年至 2013 年的刑事诉讼统计，发现上市公司一般采用多种方式进行舞弊，包括调整资产负债、以各种方式操纵利润、对公司报告进行虚假陈述[14]。刘石球(2016)的研究发现，财务舞弊主要有三种方法：伪造客户或供应商、虚构资金流、消化“虚构毛利占用的资金”的手法[15]。黄进敏(2017)通过实证研究发现企业通常采取多种舞弊手段，包括虚增收入、多增成本、虚列费用等[16]。万朝辉(2019)研究发现，财务舞弊的主

要手段包括操纵收入成本、滥用会计政策、资产重组等方法[17]。张彤(2019)总结了近年来财务舞弊的主要表现形式，如改变会计核算方法、修改折旧规则、利润水平调整、收支错报、关联交易等多种方式[18]。郑伟宏等(2019)对中国证监会处罚报告的分析表明，上市公司通常同时使用不同的手段进行财务欺诈，其中最常见的是错误披露财务信息和操纵损益表[19]。

1.2.3 财务舞弊的识别特征

在有关财务舞弊特征的研究中，逐渐由定性研究转为定量研究。国外的研究发展较早，国内相对较晚。其中“红旗”研究是较早的财务舞弊识别的研究，Albrecht(1986)通过使用调查问卷的方式，表明一些变量可作为“红旗”指标，用来作为财务报告舞弊的特征，对财务造假行为进行了分类研究，实验总结并证明了数十个指标具有显著性预兆的作用[20]，但在实际应用中仍有较大的局限性。Loebbecke 等(1988)提出从定性角度把企业财务舞弊模型的特征变量分成三类，分别是舞弊条件、舞弊态度和舞弊动机，通过这三个方面选取指标[21]。Dechow 等(1996)发现在公司治理结构方面，舞弊公司的审计委员会较少，内部董事比例过高；而虚列利润是舞弊的常用手段，同时总资产中的经营活动产生的净现金流量相对较少[22]。Beneish(1997)和(1999)的两次研究表明，舞弊公司具有更高的财务杠杆指数、应计利润率和应收款项[23-24]。Abbott(2004)认为公司内部的管理制度对于财务合规的监管起到重要作用，例如审计委员会的合理设立等可以减少舞弊发生的概率[25]。Farber(2005)研究了 87 家被查处的财务舞弊公司，结果表明，这些公司的公司治理水平明显更低，具体表现为：四大审计机构所占比例较小、外部董事数量不足等[26]。Ficha 等(2007)通过对财务舞弊相关纠纷案件的抽样调查，发现了公司管理水平的提高会增加造假者的离职率[27]。Erickson 等(2010)比较了 1996 年至 2003 年间被美国证券交易委员会(SEC)指控会计欺诈的公司的高管股权激励与两个未被指控欺诈的公司样本，研究表明没有证据表明高管激励对公司财务舞弊存在影响[28]。Yang 等(2020)验证了美国上市公司下的财务原始数据与集成学习模型结合的有效性[29]，但未考虑非财务数据。

国内在此方面的研究虽然起步较晚，但也已有较为丰富的研究。方军雄(2003)实证研究表明，资产负债率和应收账款比率与企业财务舞弊呈一定的正相关，而速动比率和应收账款周转率则与企业财务舞弊呈负相关[30]。洪荭(2012)研究发现，公司内控失效、治理不完善或会导致财务舞弊[31]。许存兴(2013)通过对 2011 年非标准审计意见的公司为财务舞弊公司进行研究，发现现金流

量指标与企业财务舞弊的可能性有关[32]。程鑫(2015)得出纳税申报信息造假主要与权益资本比率、投资收益比率、资产净利润率、留存收益比率、流动比率、主营业务成本率、应交税金/主营业务收入等因素有关[33]。但没有将非财务指标考量在内。张苏彤(2016)证明了奔福德定律在舞弊侦测方面具有有效性[34]。洪荭等(2017)研究后得出盈余管理与财务舞弊之间的关系，其认为无论是过去是本年度，企业的盈余管理均与财务舞弊之间表现出一定的相关性，从而得出了随着时间的推移,盈余管理向财务舞弊逐步演变的结论[35]。李清等(2017)通过对 2010-2014 年的 A 股市场进行研究，发现财务舞弊发生的概率与董事会会议次数等变量呈现正相关，与董事会规模适度性、股权集中度、监事会会议次数、董监高平均年龄和平均受教育程度等指标呈现负相关[36]。向晖(2018)认为，审计人员若想及时识别企业财务舞弊，应着重关注财务报表波动较大科目、报表中是否存在矛盾的事实等方面[37]。黄世忠等(2019)通过对证监会处罚决定书进行分析发现上市公司财务造假与地区经济发达程度、公司规模呈反比关系[38]。众多学者指出了财务舞弊的特点，在选取指标体系时，对财务指标给予关注的同时，可将非财务指标也纳入其中，对关注指标波动较大的情况，应进行分析。总的来说，国内外研究均表明，财务指标和非财务指标与公司财务舞弊之间存在显著相关性，这为建立模型所需的指标体系选取提供了一定依据。

1.2.4 财务舞弊的识别模型

早年的研究多使用的是 Logistic 回归算法，这种算法适合较小的样本，且具有简单和相对可靠的优点。

Persons(1995)选用了 10 个财务相关指标，通过 Logistic 逐步回归算法可以发现哪些变量对识别舞弊具有指示作用，实证表明，财务杠杆与资产周转率等指标在模型中表现重要[39]。Leeet(1999)构建了 Logistic 逐步回归模型，对 1978-1991 年间的 56 个舞弊公司及其配对非舞弊公司进行了考察。研究发现，盈余现金流量差对财务舞弊识别具有重要作用，与非舞弊公司相比，舞弊公司舞弊前一年的盈余现金流量差更高[40]。Bell 和 Carcello (2000)选取了几十家舞弊公司及上百家未舞弊公司采用 Logistic 回归算法构建了财务舞弊识别模型[41]，但该研究所使用的样本量较少。

随着人工智能的发展，更多的机器学习算法也被不断研究和应用，包括支持向量机、决策树、随机森林、神经网络、自适应增强等算法，而这些算法可以更好得解决舞弊识别问题中存在的一些非线性问题。Kotsiantis 等(2006)系统分析了各算法的性能，其选取了数十家欺诈公司及一百余家未舞弊公司数据

样本，比较了 C4.5 决策树、KNN 投票 、Logistic 回归等几种单分类器算法的效果，结果表明 C4.5 决策树性能最好[42]。Feroz 等(2000)对 42 家财务舞弊公司及其配对样本进行研究，对比了 Logistic 回归方法和人工神经网络两种方法，实验结果表明人工神经网络方法的实验结果在一定情况下要好于 Logistic 回归[43]。Cecchini (2010)使用支持向量机的方法，并针对财务舞弊识别问题创建了一个新的核函数，financial kernel。此研究将不同时间段区分开来，使用 1991-2000 年数据作为训练集，将 2001-2003 年的数据作为测试集，取得了较好的效果[44]。Ozdagoglu 等(2017)使用决策树、Logistic 回归和人工神经网络的算法对土耳其危机后的特定时期(2009-2013)的上市公司财务舞弊情况进行了预测，得出三种算法效果均不错，但人工神经网络的效果最好的结论[45]。

国内在此方面的研究也已较为丰富。刘君等(2006)运用 RBPNN 构建出财务造假识别模型，对 36 家上市公司进行进行训练和学习。研究结果显示准确率可达到 86.7%[46]，然而该研究使用的样本过小，易于过拟合，且很难具有普适性。陈国欣等(2007)使用 Logistic 回归算法，建立了一个基于少量变量的回归模型[47]。曾月明等(2008)提出了智能识别的概念，并使用神经网络和支持向量机方法进行判别，效果好于传统方法[48]。蒙肖莲等(2009)发现概率神经网络模型在构建欺诈性财务报告识别方面具有很高的预测力，并发现该模型的性能优于人工神经网络模型以及 Logistic 回归模型[49]，但所使用的样本量较为有限。阚宝奎等(2012)对 SVM 算法进行了模糊化的改进，引入非财务指标，构建了一种模糊的支持向量机算法，在模型训练时对真实样本和舞弊样本进行不同处理[50]。钱苹等对 1994-2011 年的部分样本作为数据，重新建立了适合中国财务舞弊识别的 M-score 模型[51]。夏明等(2015)提出了一种 RBF-BP 组合神经网络模型，使模型能够较基于 RBF 或 BP 神经网络算法中的任意单独一种模型更有效地识别财务造假[52]，但在指标的选择上十分单一，缺少考虑其它相关信息。吴杰等(2016)采用 FCM 算法识别舞弊的财务报表，分析舞弊财会报表的数据的特定特征，使用模糊 C 均值聚类方法把属性特征差异性放大[53]，但识别效果仍可加强。冯炳纯(2019)讨论了不同算法的财务舞弊识别效果，得出随机森林与 Relief 特征选择算法结合效果最好的结论[54]。

1.2.5 文献评述

通过回顾和总结国内外研究财务舞弊的文献资料，可以发现目前该领域已有不少的研究成果。国外对财务舞弊的研究相对较早，在财务舞弊成因方面，学者们提出了舞弊双因素理论、舞弊三角理论、GONE 理论和风险因子等理论，

这些理论模型多是由国外学者提出；在财务舞弊的特点及手段方面，国内外均进行了丰富的研究，主要的手段包括虚列资产、操纵利润、虚构业务等，不同行业的舞弊手法被证实存在较为明显的差异，这些研究为财务舞弊识别提供了切入角度；从财务舞弊识别特征来看，多使用财务比率指标，非财务指标也已被证明在识别中起到重要作用；从财务舞弊识别模型方面的研究情况来看，早期的研究多在国外进行，使用定性的研究，后逐渐流行为定量研究。最早使用Logistic 回归居多，后来机器学习方法的有效性得到证实，并被不断应用于财务舞弊识别领域，我国在该方面的研究多为依靠相关的机器学习技术进行实证研究。目前来看，我国在基于机器学习的财务舞弊识别研究中仍存在一些可以讨论的空间：

第一，在指标体系选取方面，我国大部分研究使用财务比率指标，这些比率通常由专家根据理论制成，但原始会计指标作为财务比率指标最基本的组成部分，十分容易从财报中获取，却很少被应用于识别模型中。基于原始财务数据的财务舞弊识别模型是否优于基于财务比率的模型是未知的，尽管 Yang 等（2020）证明了基于原始会计指标体系的机器学习方法可能更有优势，但这尚未在中国数据下被讨论证实。

第二，财务舞弊识别模体系中，多使用的为当年的财务指标，而未对财务数据在时间维度上的相互关联做出过多的考虑，公司经营业务具有连续性，其财务数据在时间维度上可能反应出一些有用信息，这在具有让数据“自言自语”的自动学习能力的机器学习方法中可能会发挥出作用。

第三，连续舞弊行为作为财务舞弊问题中特有的常见行为，应被给予更多考量。传统的研究中对连续舞弊问题大多未进行过多讨论或专门化处理，在实际问题中，企业常常连续进行财务舞弊，但却在数年后才被发现并通告，因此若不考虑连续舞弊的情况，将会使机器学习模型的训练集中存在先验信息，即各条数据被单独看待，从而导致连续舞弊中的部分数据被划分至训练集并被使用，这可能会造成模型的高估，且不符合实际应用场景。

本文将对这些问题进行进一步的讨论，并通过真实的中国上市公司数据进行实验证明。

1.3 主要研究内容

1.3.1 研究内容

本文主要研究基于机器学习的中国上市公司财务舞弊识别模型，包含从样

本选取到指标体系构建，再到算法选择等多个流程的实验研究和讨论。共分为四章，各章主要内容如下：

第一章绪论。介绍论文的研究背景、研究意义，以介绍国内外财务舞弊的研究现状和分析，主要研究内容及论文技术路线；

第二章指标体系构建与数据预处理。从财务舞弊三角理论入手，构建财务比率指标体系、原始会计指标体系、各指标时间维度扩充后的指标体系等多个指标体系，选择研究对象，进行含样本标记、缺失值处理、连续舞弊情况考虑和数据集划分等步骤在内的数据预处理工作，最后对用来比较的基准模型使用的指标体系进行构建；

第三章模型构建与评价方法。介绍本文所选择的评价指标以及随机森林、支持向量机和自适应增强算法的方法原理与基于三种算法的模型构建流程；

第四章模型效果比较与分析。对比和分析基于不同指标体系和算法的模型效果，论证指标体系的有效性。论证对连续舞弊情况处理的必要性，最后与基准模型相比较，证明本文提出模型的有效性。

1.3.2 技术路线

本文将数据集进行处理后，划分出十份数据集，使用三种算法，在指标体系方面进行讨论，分别对基于财务比率指标体系、原始会计指标体系的模型效果进行比较，讨论原始会计指标体系下的模型是否优于财务比率指标体系下的模型；判别合并的指标体系模型效果；再对三种指标体系进行时间维度的扩展，并比较证明时间维度扩充的有效性。在样本选取方面，将基于连续舞弊处理与未处理的样本模型的差异比较，论证连续舞弊情况考虑的必要性。最后，与选取的基准模型进行对比，证明本文提出模型的有效性。

本文的技术路线可以概括为图 1-2。

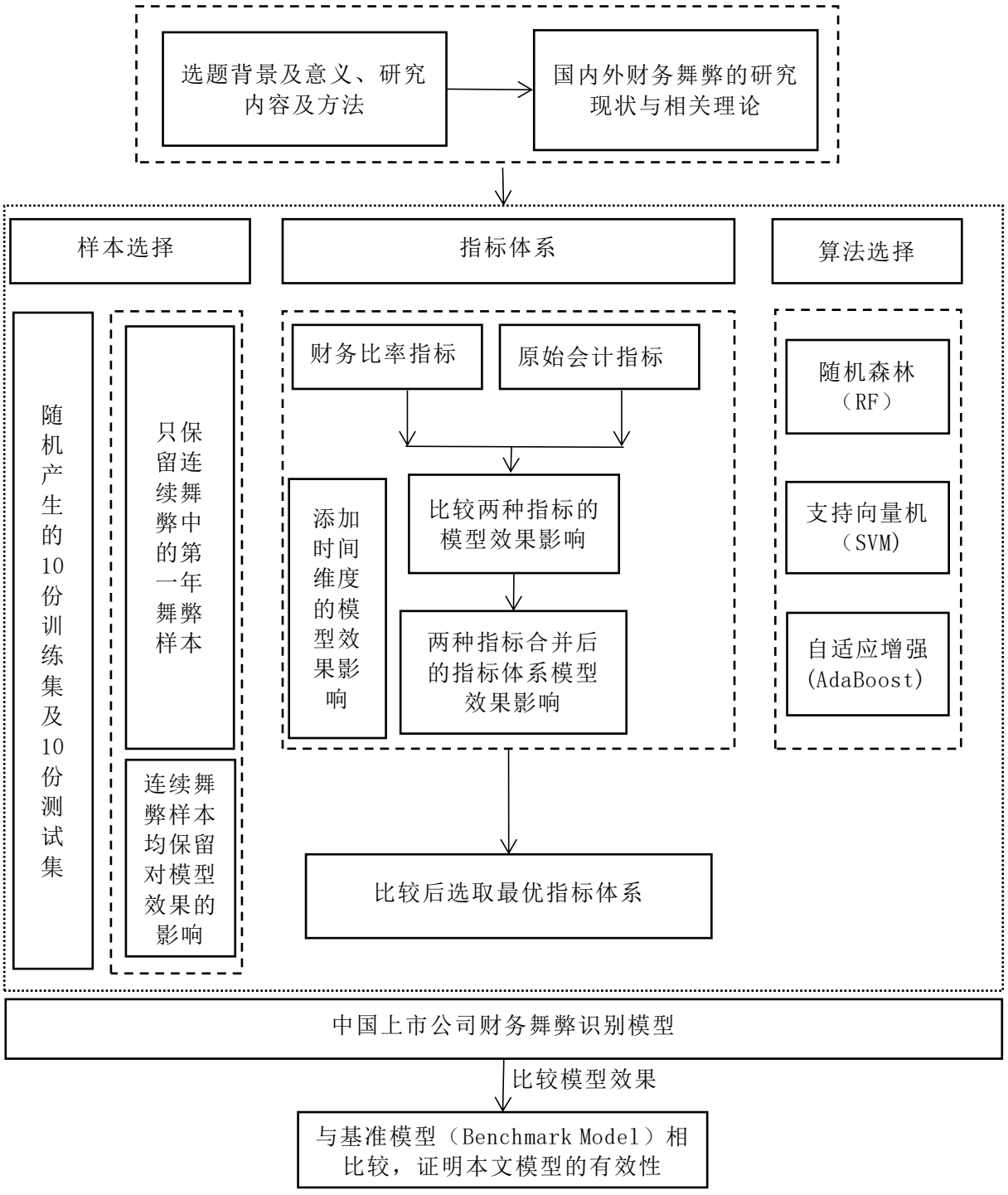

图 1-2 本文技术路线

第 2 章 指标体系构建与数据预处理

2.1 财务舞弊成因理论

财务舞弊成因的相关理论可以帮助构建财务比率指标体系，使得指标体系的构建更加合理，本小节介绍主流的财务舞弊成因理论。

2.1.1 冰山理论

冰山理论，也称为双因素理论，将财务舞弊比作大海中的冰山，我们所看到的财务舞弊只是冰山在海面上的部分。该理论认为结构与行为两个方面是财务舞弊的不同层面的体现，结构指的的公司内部的管理问题，包含等级制度、财务资源、组织效率等；而行为更多关注个人，这一海平面以下的部分同样需要被关注，财务舞弊由组织层面的结构与管理者个人层面的行为共同导致。

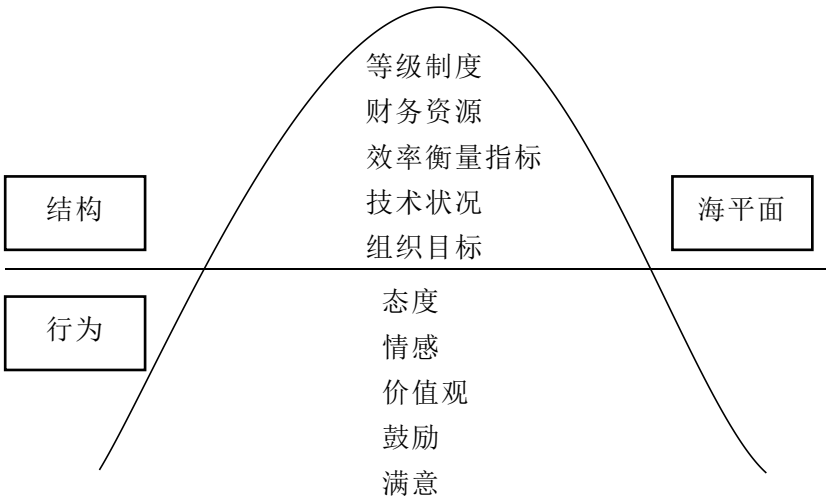

图 2-1 财务舞弊冰山理论模型

2.1.2 舞弊三角理论

1953 年，Donald R .Cressey 通过访谈方法，得到了基于金融犯罪理论的三角理论[55]。Albrech 于 1995 年对此进一步研究，将其从犯罪学研究引入到会计研究领域，认为压力是财务舞弊产生的直接动机，机会是舞弊的条件，而借口则为舞弊者对舞弊行为的解释与合理化。表明预付财务舞弊行为要加强内

部控制和外部监管的同时，还要进行减小压力、消除借口，这三个方面都应被重视[56]。

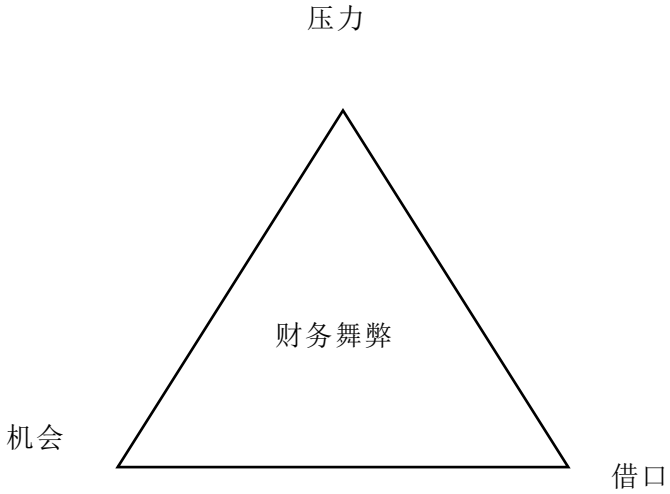

图 2-2 财务舞弊三角理论模型

2.1.3 GONE 理论

Gone 理论，又称四因素理论，由 Bologna 等提出，是著名的公司舞弊理论之一。Gone 指的是需求、贪婪、机会和暴露，这四个方面的问题共同导致了财务舞弊行为。与三因素理论相比，该理论将舞弊动机说明为需要、道德价值取向说明为贪婪，并增加了一个“暴露”因素。它认为，舞弊行为暴露的可能性以及暴露后的惩处将影响欺诈者是否进行财务舞弊。其中，前两个因素为舞弊人员层面的个人因素；后两个因素则组织层面的，这四个因素的共同作用导致了财务舞弊行为。

2.1.4 风险因子说理论

风险因子理论是 Bologna 等人在 Gone 理论的基础上提出的，该理论较 Gone 理论更为完善。风险因素理论将风险因子划分为两个层次，分别是个人及组织层面，称做个别风险因子与一般风险因子。与冰山理论中的海平面下的部分相似，个别风险因子多指的是管理层个人层面；一般风险则包括了公司内外部的诸多因素，如舞弊的机会、被揭露的可能性与制裁等。在这一理论中，个人与组织都应被给予关注[57]。

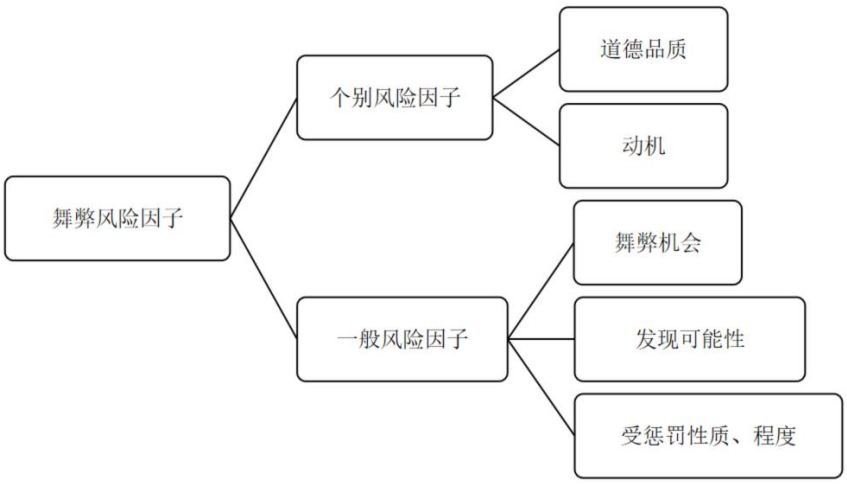

图 2-3 GONE 理论模型

2.1.5 其他理论

随着财务舞弊研究的逐渐深入，越来越多的舞弊理论被提出，具有代表性的是 Wolfe 等提出的舞弊菱形理论，这一理论是在舞弊三角理论上发展出的，此理论增加了舞弊能力因素[58]。还有一些其他新发展的理论，如 Kassem 在 2012 年基于原有理论提出的新舞弊三角形理论[59]、同年 Tugas 拓展出的舞弊五角理论[60]等。我国学者娄权等人(2004)提出了“四因素假说”，其认为如果社会缺乏诚信，管理层有一定的财务舞弊意愿，并获得了良好的机会，同时根据理性判断，如果进行造假的当前收益高于当前造假的成本，就有可能会引发财务舞弊行为[61]。

2.2 指标体系选择

2.2.1 财务比率指标体系选择

2.2.1.1 指标体系选取依据

本文以经典的财务舞弊三角理论为依据，通过上述财务舞弊成因理论的介绍，可以发现，财务舞弊三角理论并不过多关注舞弊者个人的情况，这便于财务比率指标的选取，在这里进一步对财务舞弊三角理论进行介绍分析，本文主要选取压力方面的财务指标。

（1）压力

一般认为，压力是舞弊三角中的顶点，也是财务舞弊的直接原因，张竹怡

（2019）将压力分为财务状况、外部压力和管理层财务状况三个方面，并依据这三个方面选取了相关的财务指标[62]。本文也将压力指标分为此三方面进行选取。

（a）财务状况　　如果公司整体的财务状况下降，压力便会产生，管理层为了增强利益相关方的信心或为了符合一些资质的条件要求，便有进行财务舞弊的动机。公司的财务状况反映了公司的经营情况，经营情况的下滑是企业压力的内部原因。

（b）外部压力　　在企业的经营受阻之后，利益相关者的信心下降，将会对上市公司的股价和进一步筹资能力等造成不良影响，此类影响将对公司管理层形成较大的外部压力。公司为进一步筹资，可能通过财务舞弊提高公司偿债能力。来自外部的压力可能会增大公司财务舞弊的概率。

（c）管理层财务状况　　管理层控股对其自身会产生激励作用，但当企业市值下降时，其自身利益也会受到影响，因此公司的股价等会影响管理层自身的财务情况。

（2）机会

一般来说，股权的集中程度等会造成话语权的集中，股权若是较为集中，则相对便缺少更广泛的监督，发生财务舞弊的可能性便越高。

（3） 借口

在面临压力、拥有机会后，公司进行舞弊还需要一个要素——借口(自我合理化)，也就是说，无论这种解释是否真的合理，公司舞弊者都需要找到一些理由，使公司欺诈行为符合自己的道德概念和行为准则。舞弊人员会为自己的舞弊寻找理由，使自己的行为合理化，如大家都是这么做的、我也是为公司股东着想。而这与管理层和会计从业人员的认知水平、所处行业、审计结论有着密不可分的关系。

其中的机会和借口两个方面，涉及的指标为非财务指标，不在本文研究范围内，因此不予讨论。选取指标类数据后，找出与之对应的原始会计数据指标，共同纳入财务舞弊识别指标体系。

2.2.1.2 财务比率指标体系确定

（1）财务状况

上市公司的经营能力、发展能力、盈利能力等均会影响投资者的信心，这些财务指标的下降可能会对企业管理层产生压力。参考以往学者的研究，多从这几个方面进行，因此将以下指标纳入指标体系。

纳入指标：应收账款与收入比、存货与收入比、存货周转率、应付账款周转率、流动资产周转率、固定资产与收入比、固定资产周转率、总资产周转率、股东权益周转率、可持续增长率、资产报酬率、总资产净利润率、长期资本收益率、营业毛利率、息税前营业利润率、投资收益率。

（2）外部压力

上市公司的偿债能力和风险水平影响进一步筹资的能力;对外部融资造成压力，因此考虑将以下指标纳入指标体系。

纳入指标：资产负债率、流动比率、速动比率、利息保障倍数、财务杠杆、经营杠杆、综合杠杆、经营活动产生的现金流量净额 / 流动负债。

（3）管理层财务状况

公司股利分配情况会影响管理层自身的财务状况，若股利分配情况较差，则可能对管理者的自身财务造成压力。因此考虑将以下指标纳入指标体系。

纳入指标：每股收益、每股税前现金股利、股利分配率、收益留存率、基本每股收益。

本文共选取了 28 个财务比率指标，对其中缺失值过多的指标进行删除处理，将缺失值超过 30%的指标删除后，得到了 27 个财务比率指标，如表 2-1 所示，其中最后一列标明了该指标经缺失值处理后是否最终入选。

表 2-1 财务比率指标与对应的原始会计数据指标体系

		指标	解释	最终入选
压力	财务状况	应付账款周转率	营业成本 / 应付账款平均占用额	是
		流动资产周转率	营业收入 / 流动资产平均占用额	是
		固定资产与收入比	固定资产 / 营业收入	是
		固定资产周转率	营业收入 / 固定资产平均净额	是
		总资产周转率	营业收入 / 平均资产总额	是
		资产报酬率	(利润总额+财务费用)/平均资产总额	是
		总资产净利润率(ROA)	(净利润) / 总资产平均余额	是
		长期资本收益率	(净利润+所得税费用+财务费用) /（所有者权益+长期负债）	是

表 2-1 （续表）

		指标	解释	最终入选
压力	财务状况	营业毛利率	(营业收入—营业成本) / (营业收入)	是
		息税前营业利润率	(净利润+所得税费用+财务费用)/(营业收入)	是
		投资收益率	本期投资收益/(长期股权投资本期期末值+持有至到期投资本期期末值+交易性金融资产本期期末值+可供出售金融资产本期期末值+衍生金融资产本期期末值)	否
		股东权益周转率	营业收入 / 平均股东权益	是
		可持续增长率	(净利润/所有者权益合计期末余额)*[1-每股派息税前/(净利润本期值/实收资本本期期末值)]/(1-分子)	是
		应收账款与收入比	应收账款/营业收入	是
		存货与收入比	存货/营业收入	是
		存货周转率	营业成本 / 存货平均占用额	是
	外部压力	经营活动产生的现金流量净额 / 流动负债	经营活动产生的现金流量净额 / 流动负债合计分母为流动负债合计	是
		财务杠杆	(净利润+所得税费用+财务费用)/(净利润+所得税费用)	是
		经营杠杆	(净利润+所得税费用+财务费用+固定资产折旧、油气资产折耗、生产性生物资产折旧+无形资产摊销+长期待摊费用摊销)/(净利润+所得税费用+财务费用)	是
		速动比率	(流动资产－存货) / 流动负债	是
		综合杠杆	(净利润+所得税费用+财务费用+固定资产折旧、油气资产折耗、生产性生物资产折旧+无形资产摊销+长期待摊费用摊销)/(净利润+所得税费用)	是
		资产负债率	负债合计 / 资产总计	是

表 2-1 （续表）

		指标	解释	最终入选
压力	外部压力	流动比率	流动资产 / 流动负债	是
		利息保障倍数	(净利润+财务费用) / 财务费用	是
	管理层财务状况	每股税前现金股利	每股派息税前	是
		每股收益	净利润本期值 / 最新股本	是
		股利分配率	每股派息税前/(净利润本期值/实收资本本期期末值)	是
		收益留存率	1—(每股派息税前)/(净利润本期值/实收资本本期期末值)	是

2.2.2 原始会计指标体系选择

在以往的国内外财务舞弊识别研究中，多使用专家提出的财务比率作为指标，这种比率指标基于一些财务分析知识体系，便于人们对财务情况的分析把握。但财务比率指标作为人工进行财务分析时期的重要工具，未必会在机器学习过程中表现的更好，原因之一可能是这些基于知识的财务比率指标，在使财务数据更易被直观发现问题的同时，在一定程度上损失部分信息。另一种可能是财务比率指标所依赖的财务理论知识本身并不完美，仍存在一定进化空间。

直接利用财务原始数据进行机器学习，在国内外的研究较为少见。Yang等(2020)曾利用美国数据进行研究，证明了在财务原始数据与集成学习在美国上市公司数据的情况下是有效的，效果超过了财务比率指标与集成学习模型的搭配使用。但财务原始数据由于没有利用专家知识，未必会有更优的效果，且中国的上市制度与美国不同，企业经营环境及文化也有巨大差别，财务原始数据的有效性尚未在国内得到验证。

本文的原始会计指标体系来自于所选用的财务比率指标，这些原始会计指标可以直接从财报及相关披露获取，并可通过这 30 个原始会计指标经计算生成前述的财务比例指标体系，对缺失值过多的原始会计指标进行剔除后，保留了 27 个原始会计指标，具体指标可见表 2-2。

表 2-2 原始会计数据指标体系

所属财务报表	原始会计指标	最终入选
资产负债类	应收账款净额	是
	存货净额	是
	应付账款	是
	固定资产净额	是
	资产总计	是
	所有者权益合计	是
	实收资本(或股本)	是
	长期股权投资净额	是
	持有至到期投资净额	否
	交易性金融资产	是
	可供出售金融资产净额	是
	衍生金融资产	否
	负债合计	是
	流动资产合计	是
	流动负债合计	是
	长期负债合计	是
损益类	营业收入	是
	营业成本	是
	净利润	是
	所得税费用	是
	基本每股收益	是
	利润总额	是
	投资收益	是
	财务费用	是

表 2-2 （续表）

所属财务报表	原始会计指标	最终入选
现金流量类（直接法）	经营活动产生的现金流量净额	是
现金流量类（间接法）	长期待摊费用摊销	是
	无形资产摊销	是
	固定资产折旧、油气资产折耗、生产性生物资产折旧	是
其他披露类	每股派息税前	是

2.2.3 时间维度指标扩充

在考虑上市公司财务舞弊识别的指标体系时，需考虑到该类识别不同于传统的判别识别问题，事实上，财务舞弊行为会导致财务相关数据发生变化，这就使得财务及非财务相关数据具有前后的时间相关性，因此在做财务舞弊识别分析时，将公司近年的相关数据纳入其中是有十分必要的。本文将上市公司的上述相关指标进行时间维度的扩充，使得第 n 年的样本记录同时将 n-1、n-2、n-3 及这 4 年数据的方差纳入指标体系，对指标体系进行时间维度的扩充，以便考察时间上的变化性。

这使得指标数量大大增加，其中财务比率指标及原始会计指标均由 27 个扩充至 135 个，同时，这将使得 2007-2009 年三年的样本不被纳入识别中，对样本造成少量但可接受的损失。通过纳入考察年份的近三年数据及变化作为指标体系，具有一定领域应用价值。

2.3 研究对象及来源

本文将财务舞弊定义为：上市公司违背会计准则或会计制度及相关规范对信息质量的要求，在财务报表或财务披露中存在的蓄意错报，通过虚构利润、虚列资产、重大遗漏、虚假记载这 4 个手段，对企业经营成果进行粉饰的行为。具体认定以相关机构审查处罚为准。在研究中提及的“财务舞弊”、“财务造假”、“管理欺诈”等视为同一概念。

基于财务舞弊的定义、参考以往研究，舞弊样本选取虚构利润、虚列资产、

重大遗漏、虚假记载 4 个方面的非金融行业违规公司，其中，同一企业连续违规时，仅保留连续舞弊中第一次的违规数据，未舞弊的样本则选择相同时间段的国内非金融上市公司。

本文所使用的数据集来自 CSMAR (国泰安)数据库。CSMAR 数据库中包含有丰富的国内上市公司的财务情况数据及公司违规情况的数据，对研究中国上市公司财务舞弊十分有帮助。

CSMAR 迄今共公开从 1990 至 2021 年间的公司情况数据，考虑到中国在 2005 年开始了股权分置改革，股权分置改革对中国的上市公司有着较大的影响，因此选择数据应避免这种影响，一般认为，在 2006 年底，股权分置改革已基本完成。黄世忠等(2021)对我国上市公司财务舞弊的研究表明，平均财务舞弊被曝光需要 3 年[63]。且考虑到 2007 年开始实行了新的会计准则，会导致财务报表发生一定变化，故选择 2007-2019 年共 12 个年份的非金融行业的中国上市公司作为研究对象。剔除金融公司是主要考虑到金融行业的财务报表与非金融行业的财务报表有着很大差别。

2.4 数据预处理

2.4.1 样本标记

在选择指标后，对各表进行连接等操作，生成数据集，如图 2-4 所示。其中 Stkcd 为股票代码，Accper 为所在年度，对公司是否舞弊进行标注，如式 2-1.

$$\text{isviolation} = \begin{cases} 0, \text{该上市公司本年度未财务舞弊} \\ 1, \text{该上市公司本年度发生财务舞弊} \end{cases} \tag{2-1}$$

将所有特征变量考量入内，数据集共有 27 个财务比率指标、27 个原始会计指标，及大量的时间扩充指标，其中有 3921 条违规记录，31558 条未违规记录，共有 35479 条记录。

Stkcd	Accper	D000103000	D000104000	D000105000	D000109000	C001000000	A002100000	B001302000	B001000000
000002	2007	1.082281e+08	NA	2410643.03	359500074.40	-1.043772e+10	4.877398e+10	208030696.10	7.641606e+09
000002	2008	9.416901e+07	NA	11649080.27	657253346.42	-3.415183e+07	6.455372e+10	209411393.50	6.322286e+09
000002	2009	9.565321e+07	9.390604e+06	NA	573680423.04	9.253351e+09	6.805828e+10	924076829.10	8.617428e+09
000002	2010	9.598956e+07	NA	NA	504227742.57	2.237255e+09	1.296508e+11	777931240.02	1.194075e+10
000002	2011	1.134478e+08	NA	NA	509812978.62	3.389425e+09	2.007242e+11	699715008.48	1.580588e+10
000002	2012	1.545651e+08	2.805590e+07	0.00	764757191.68	3.725958e+09	2.598336e+11	928687953.69	2.107019e+10
000002	2013	1.534443e+08	2.579125e+07	0.00	891715053.49	1.923869e+09	3.289218e+11	1005187804.32	2.429101e+10
000002	2014	3.760286e+08	1.728757e+08	0.00	640839545.38	4.172482e+10	3.456540e+11	4159261963.52	2.525236e+10
000002	2015	4.901612e+08	2.256451e+08	0.00	477735809.60	1.604602e+10	4.200618e+11	3561908083.68	3.380262e+10
000002	2016	7.082103e+08	4.404218e+08	NA	1592067967.14	3.956613e+10	5.799985e+11	5013835862.38	3.925361e+10
000002	2017	1.101115e+09	8.279227e+08	NA	2075256781.28	8.232283e+10	8.473554e+11	6244561688.39	5.114195e+10
000002	2018	1.697226e+09	1.224710e+09	NA	5998574652.64	3.361818e+10	1.121914e+12	6787934513.16	6.746020e+10

图 2-4 数据集样示

2.4.2 缺失值处理

缺失值是指现有数据集中某个或某些属性的值是不完全的，在本文中对缺失值处理采用删除缺失个案和缺失值填充两种处理手段相结合的方式。对于同一条记录中缺失值超过阈值的个案进行删除，在此处采用列数的 25%作为阈值。违规数据集删除 41 条记录，未违规数据集删除 1777 条记录。

对于缺失值填充，需要考虑财务数据的特殊性，对比了部分财报，发现国泰安数据库中原始会计数据的缺失值，多是由于对应公司的财报未披露相关数据导致的，这可能是有多种原因，如会计准则要求披露内容的新增调整等，其中一条重要原因是该公司在该项科目上未发生相关业务，而对这部分数据，国泰安数据库部分以 0 处理、部分以 NA 缺失值处理，因此以 0 对原始会计数据的缺失值进行填充是合理的。

如表 2-3 所示，国农科技(代码 000004)在 2016 年的年报中，未披露“发行债券收到的现金”科目，而对应国泰安数据库中的该值为 NA。可认为该上市公司在该年度未发生此项业务活动，可以 0 填充缺失值。

而对于财务比率指标，数据库中的缺失值大量是其计算公式中的分母为 0 或缺失所导致的，因此本文将这些缺失的财务比率指标数据统一由 0 填充。

表 2-3 国农科技 2016 年财报（部分）

项目	本期发生额	上期发生额
三、筹资活动产生的现金流量		
吸收投资收到的现金		
其中：子公司吸收少数股东投资收到的现金		
取得借款收到的现金		10,000,000.00
发行债券收到的现金		
收到其他与筹资活动有关的现金		
筹资活动现金流入小计		10,000,000.00
偿还债务支付的现金	40,000,000.00	
分配股利、利润或偿付利息支付的现金	7,506,650.84	5,225,801.22
其中：子公司支付给少数股东的股利、利润		5,225,801.22
支付其他与筹资活动有关的现金		
筹资活动现金流出小计	47,506,650.84	5,225,801.22
筹资活动产生的现金流量净额	-47,506,650.84	4,774,198.78

2.4.3 连续舞弊情况的考虑

在考虑样本的选取时，考虑连续舞弊问题。连续舞弊是上市公司舞弊事件中的经常性行为，指的是若该公司进行舞弊后未被发现，则在接下来的数年内仍有舞弊的动机和倾向，直到被发现为止。因此，监管机构对上市公司财务舞弊的公告及处罚，常常是一次公告该公司在过去多个年度的舞弊行为。在传统的研究中，大多未考虑连续舞弊情况，这不符合模型实际的应用场景，且会使得识别模型的准确率可能被高估(训练集利用了连续舞弊样本的先验信息)。因此在舞弊样本选取时，若发现存在连续舞弊行为，本文则只选取第一次舞弊样本，后续连续舞弊样本被删除。不考虑连续舞弊情况的识别模型是否虚高了模型的判别能力将在后续章节被讨论。

在删除连续舞弊样本和时间维度扩充后，数据集基本情况如下：数据集共有 1233 条违规记录，20733 条未违规记录，共有 21966 条记录，与未考虑连

续舞弊的样本选取对比，连续舞弊行为约占了舞弊行为的 2/3。

2.4.4 数据集划分

2.4.4.1 样本不平衡处理方法

（1）欠采样法

该方法常用于对类别种类有限的情况进行处理，欠采样方法会减少样本较多的类别（大类）的观测数对数据集进行样本平衡。这一方法在样本数目较小的类别（小类）的观测数目不低的时候较为合适，该方法的另一优点是通过降低训练集的整体规模使得计算时间和存储开销降低，但是由于部分样本被剔除，欠采样方法会使得数据样本集产生信息损失。欠采样法共有两类：随机欠采样和依据信息的欠采样。

随机欠采样法会随机剔除一定大类的样本使得大类的样本数与小类的样本数相平衡，依据信息的欠采样法则会依照一定的准则删除样本。

（2）过采样法

又名升采样，该方法与欠采样方法相反，主要是通过对小类进行重复处理的方式来解决样本不平衡问题。过采样方法也可以分成两大类，按照过采样过程中是否使用信息，可以划分成随机过采样和依据信息的过采样。依据信息的过采样会依照一定的准则生成小类样本，而随机过采样通过使小类样本不断随机重复直至与大类样本数量相平衡。过采样不会产生信息损失，但由于重复了小类样本的观测，易使得训练集上的拟合效果较好,但在测试集的泛化能力较差。

（3）数据生成法

数据生成法通过模拟仿真出一些小类样本的数据来解决小类样本不足的问题，本质上是一类过采样技术，之所以将其单独列出，是因为该方法被广泛使用。

在数据生成法中, SMOTE 方法是一种常被使用的方法。该方法使用“样本间距”来比较不同观测值之间的相似程度，并基于现有数据随机生成一些仿真数据，自助法(Bootstrap)和最近邻法(KNN)在这一过程中被使用，算法步骤为:

(a)寻找一个采样点，并基于特征空间找到一个最近邻点。

(b)以两点之间的所有特征空间内的点为集合，从集合中随机抽取一点。

(c)步骤 b 中生成的点即为人工模拟出的新样本，加入小类集合中。

（4）代价敏感学习

代价敏感学习通过测量错误分类观测的代价来解决不平衡问题；该方法不

产生平衡数据集，而是产生一个成本矩阵来解决不平衡问题。成本矩阵是一种工具，用于描述在特定情况下由于错误分类的观测结果而造成的损失。这种方法可以找到一个最小化总成本的分类器。

2.4.4.2 欠采样划分

财务舞弊问题是一种典型的样本不平衡问题，本文采用随机的欠采样方法进行数据集划分，主要出于以下 3 点考虑：

（1）财务舞弊以是否舞弊为划分，可将其分为二大类，而欠采样适合二类划分的样本进行处理。

（2）舞弊样本数有 1233 个，这个数量较大，这使得欠采样虽然损失了部分未舞弊样本的信息，但所生成的训练集样本数量可以满足模型训练需要。

（3）基准模型采用了随机欠采样的方式，本文采取方欠采样方式更便与其对比。

所选取数据集共有 1233 条违规记录，20733 条未违规记录。选取随机数，进行不放回的随机抽样，分别对违规样本和未违规样本进行抽样，选取 70%的样本作为训练集，剩余 30%作为测试集；由于未违规记录较多，考虑到样本平衡，在训练集中随机取其中与舞弊样本匹配数量的未违规样本用作训练。为了使实验结论具有稳定性和可信性，对数据集进行十次随机抽样划分，形成十份训练集和与之对应的十份测试集。

综上，共有十份训练及测试集，每个训练集有 15376 个观测样本，其中 863 个舞弊样本和 863 条随机选取的未舞弊样本用作训练；每个测试集有 6590 个观测样本。

2.5 对比的基准模型指标体系

Yang Bao and Bin Ke 对美国上市公司的财务舞弊识别问题进行研究后，参考了 Cecchini(2010)和 Dechow(2011)的研究，选择这二份研究所对应的原始会计数据，将 40 余个指标进行挑选，剔除缺失值超过 25%的指标，最终选出了 28 个原始会计数据指标，并证明了原始会计数据与 RUSBoost 算法结合的有效性。该指标体系应用于美国数据，未必适用于中国上市公司，本研究对 Yang 所选的 28 个原始会计数据指标进行讨论，按照与 Yang(2020)同样的方法，剔除缺失值超过 25%的指标，最终得到 25 个变量，如表 2-4 所示。

表 2-4 基准模型所采用的原始会计指标及对应中国原始会计指标

英文指标	中文释义	拆分成的中文指标	最终入选
资产负债类			
Cash and short-term investments	现金和短期投资	货币资金	是
		短期投资净额	否
Receivables, total	应收账款	应收账款净额	是
Inventories, total	存货	存货净额	是
Short-term investments, total	短期投资	短期投资净额	否
Current assets, total	流动资产	流动资产合计	是
Property, plant and equipment, total	不动产、厂房和设备	固定资产净额	是
Investment and advances, other	投资和预付款，其他	长期投资	否
		持有至到期投资	否
		交易性金融资产	是
		可供出售金融资产	是
Investment and advances, other	投资和预付款，其他	衍生金融资产	否
		预付款项净额	是
Assets, total	总资产	资产总计	是
Accounts payable, trade	应付账款、贸易	应付账款	是
Debt in current liabilities, total	流动负债中的债务总额	应付票据	是
		一年内到期的非流动资产	否
Income taxes payable	应交所得税	应交所得税	否
Current liabilities, total	流动负债总额	流动负债合计	是
Long-term debt, total	长期债务总额	非流动负债合计	是
Liabilities, total	负债总额	负债合计	是

表 2-4 （续表）

英文指标	中文释义	拆分成的中文指标	最终入选
资产负债类			
Common/ordinary equity, total	普通股/普通股，总计	实收资本(或股本)	是
		资本公积	是
Preferred/preference stock (capital), total	优先股/优先股(资本)，总计	优先股	否
Retained earnings	留存收益	盈余公积	是
		未分配利润	是
损益表			
Sales/turnover (net)	销售额/营业额(净额)	营业收入	是
Cost of goods sold	销货成本	营业成本	是
Depreciation and amortization	折旧和摊销	折旧摊销	是
Interest and related expense, total	利息及相关费用合计	利息支出	是
Income taxes, total	所得税总额	所得税费用	是
Income before extraordinary items	非经常项目前的收入	营业收入	是
Net income (loss)	净收入(亏损)	营业成本	是
现金流量表			
Long-term debt issuance	长期债务发行	发行债券收到的现金	否
Sale of common and preferred stock	普通股和优先股的出售	吸收权益性投资收到的现金	是
市场价值项目			
Price close, annual, fifiscal	结算价	/	否
Common shares outstanding	已发行普通股	实收资本(或股本)	是

由于中美的会计准则存在一定差异，美国学者采用的会计原始数据指标需要转化成一致或相似的中国会计原始数据指标，在此对部分指标对应进行阐明。

“Investment and advances, other”指标中的投资，由于中国财报不直接披露此项，因此转化为“长期投资”、“持有至到期投资”、“交易性金融资产本”、“可供出售金融资产”、“衍生金融资产”共计 5 个原始会计指标；“Debt in current liabilities, total”在中文中直译为“流动负债中的债务总额”，根据 U.S. and Canadian GAAP 对该科目的定义：

This item represents the total amount of short-term notes and the current portion of long-term debt (debt due in one year). This item is the sum of:

(1)Long-Term Debt Due in One Year (DD1)

(2)Notes Payable (Short-Term Borrowings) (NP)

可将该科目拆分为“应付票据”、“一年内到期的非流动资产”两个原始会计指标；“Retained earnings”直译为“留存收益”，可拆分为“盈余公积”、“未分配利润”两个原始会计指标；而现金流量表中的“Sale of common and preferred stock”直译应为“普通股和优先股的出售”，对应国内的“吸收权益性投资收到的现金”。

2.6 本章小结

本章构建了中国上市公司财务舞弊识别的指标体系，依据财务舞弊三角理论，选取了财务比率指标，为了讨论原始会计指标是否会由于财务比率指标，本文生成了财务比率指标对应的原始会计指标，形成了两套指标体系，用来对比效果。考虑到公司经营及财务数据的连续性，对指标体系进行时间维度的扩展，用以讨论指标时间维度扩展的有效性。同时，为与基准模型进行比较，对基准模型采用的美国会计指标体系进行翻译和转化，选取了对应的、合适的中国会计指标体系，用以在中国上市公司财务数据上进行实验比较。

本章也对样本数据进行了处理，在对研究的对象来源及选取进行了说明后，对数据进行违规标记、缺失值处理、连续舞弊样本的处理和数据集划分。数据的预处理工作为后续的模型建立提供了基础。

第 3 章 基于机器学习的模型构建与评价方法

3.1 机器学习简介

机器学习是一类算法的总称，在广义上讲，机器学习可以指“使用计算的方式对已有经验进行学习，从而改善和提高自身系统的性能”的方法。机器学习往往需要较多的数据，且这些数据含有有价值的信息，系统利用数据内在的信息及逻辑进行学习，从而利用内部逻辑生成模型。在某种程度上也可以说，机器学习方法可以在设定规则的基础上，自动利用数据发现规律，从而形成具有预测、识别能力的模型。

由于现实生活中存在大量可以被应用于机器学习的数据，目前机器学习方法已经得到广泛的应用，在自然语言处理、图像识别、DNA 测序等诸多领域均已得到良好的使用和发展，其在财务舞弊识别领域也已有较丰富的研究。

机器学习算法按照不同的角度可以进行不同的归类。按照学习方式可以分为：有监督学习、无监督学习和强化学习。其中，有监督学习，指的是在学习训练时已知所需学习数据的类别，无监督学习则与之相反，针对的是未知所属类别的数据集，通过学习训练，发掘其中存在的内在逻辑联系，如聚类算法、主成分分析等；强化学习与前二者差别较大，一般设置一个奖励函数，通过系统不断尝试的方式，自动生成大量模拟数据，判断何种情况下可以使得奖励达到最大，使得系统可以在不同情况下产生不同的行为，多用于游戏、电商方向。

本文主要应用有监督的学习方式，包括随机森林、支持向量机、自适应增强等算法。

3.2 基于随机森林算法的财务舞弊识别模型构建

3.2.1 随机森林概述

决策树是随机森林的基础，是一种具有可解释性的传统的分类方法。决策树的思想很朴素，即通过对特征属性的划分与切割，将样本集划分成多个子集，而每个子集对应着各自所属的类。

具体来说，可以将待训练的样本集整体视为决策树的根节点，以此开始，寻找一个特征变量，将该特征变量以合适的方式进行划分，生成两个子树节点，该划分方式会使得两个节点对应的正负样本子集被尽可能合理的划分，基尼（Gini）系数是一种常用的评价标准，对基于每一个特征变量的基尼系数进行

计算后选出最高的，是常见的节点确定策略。

Gini 系数的计算方式如式(3-1)所示：

$$G(X_i) = \sum_{j=1}^{J} P_r\left(X_i = L_j\right)\left(1 - P_r\left(X_i = L_j\right)\right) = 1 - \sum_{j=1}^{J} P_r\left(X_i = L_j\right)^2 \quad (3\text{-}1)$$

其中X_i为某一分割变量，L_1, ..., L_j表示相应的级数。

通常来讲，决策树需要剪枝操作，否则会造成过拟合。这是由于决策树在分完全部数据集后，其树的层数过多，较深的树结构十分冗杂，虽然对训练集会有很好的划分作用，但泛化能力很差。剪枝操作可以减去多余的子树，使得决策树结构更加健康。

随机森林(Random Forest,RF)是一种基于决策树的集成学习方式，由大量的决策树组成。不同的是，这些决策树多为弱分类器，是通过随机特征变量和样本集构建的仅比随机挑选略优的弱分类器，这些弱分类器不需要剪枝操作。随机森林通过 bootstrap 取样的方式，使得各决策树之间减少相关性，一般认为，随机森林中的众多决策树之间是独立的，这便可由数学推导证明其生成的强分类器具有远超组成它的众多弱分类器的判别能力。

随机森林在许多分类问题上都具有良好的表现，同时拥有很多优点：可以处理数量较多的变量特征而不用担心维度灾难、对异常点的宽容性较高、并行的树结构使得其计算速度高于许多其他机器学习方法、建立过程简单等。这些优点使得随机森林算法变得十分流行。

3.2.2 随机森林模型构建流程

如图 3-1，基于 RF 的财务舞弊识别模型利用以下算法构造每一棵决策树：

（1）选择出 N 条财务训练用例(样本)，特征总数目用 M 表示。

（2）输入每次选择的特征数目 m，其中 m 小于 M，用于每次决策树使用抽取特征变量。

（3）对 N 个训练样本进行反复有放回的抽样，取样 N 次，生成一个训练集(即 bootstrap 取样)，并用未被抽取的样本作为袋外预测，以评估误差。

（4）对于每一个节点，随机选择 m 个特征，并根据这些特征在决策树中确定每个点的决策，计算这些节点上的基尼系数，依据此选择其最佳的分叉方式。

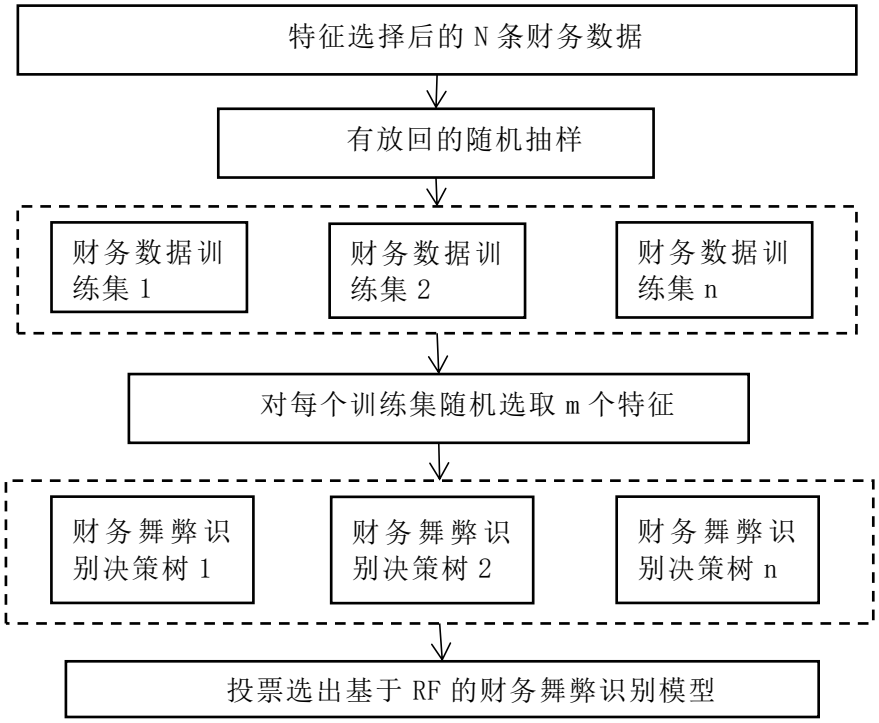

图 3-1 基于 RF 的财务舞弊识别模型构建

3.3 基于支持向量机算法的财务舞弊识别模型构建

3.3.1 支持向量机概述

支持向量机(Support Vector Machine, SVM)是一种流行的机器学习算法，它因基于精巧的数学理论与良好的性能而受到欢迎。支持向量机使用最大边距超平面作为决策分类平面，其分类边界寻求结构风险最小，一般来说用来解决二分类问题，是一种基于有监督学习模式的广义线性分类器。SVM 模型原理较为复杂，构建一个有效的分类模型较为困难，但一旦建成，则会有着较好的泛化能力。SVM 模型对小样本和高维度分类问题有着较好的性能，在财务舞弊识别中，财务舞弊的公司数据量有限、维度较高，在此类问题上有着一定优势，因此选择 SVM 模型。

在线性分类的问题上，支持向量机的分类平面是有效的，但若问题不是线性可分的，则需要引入核函数将特征变量映射到更高维度，从而达到线性可分的目的，否则将无法满足约束(3-2)

$$y_i(x_i w^T + b) \geq 1 \tag{3-2}$$

此时需加入松弛变量ξ_i，使得约束条件变为(3-3)

$$y_i(x_i w^T + b) \geq 1 - \xi_i \ , \ 0 < \xi_i < 1 \tag{3-3}$$

若满足该约束，则数据分类正确；当$\xi_i > 1$时，数据则会分类错误。通过在目标函数加入惩罚项$c\sum_i^n \xi_i$对该问题进行解决，加入惩罚因子后的目标函数为式(3-4)：

$$\begin{gathered} \min\left(\frac{1}{2}\sum_i^n w_i^2 + c\sum_i^n \xi_i\right) \\ \text{subject to} \\ y_i(x_i w^T + b) \geq 1 - \xi_i, \ i = 1,2,\ldots,n \end{gathered} \tag{3-4}$$

目标函数是一个二次凸优化问题，如果分类问题是线性的，则对偶问题求解起来更容易。利用拉格朗日乘子法与 KKT 条件对 w, b 进行求解，如式(3-5)：

$$\begin{gathered} \max\ W(\alpha) = L(w,b,\alpha) = \sum_i^n \alpha_i \ - \frac{1}{2}\sum_{i,j=1}^n \alpha_i\alpha_j y_i y_j x_i^T x_j \\ subject\ to \\ \alpha_i \geq 0, \ i = 1,2,\ldots,n \\ \sum_i^n \alpha_i y_i \ = 0 \\ \alpha_i\left(1 - y_i\left(\sum_i^n \alpha_i y_i \ \langle x_j, x_i\rangle + b\right)\right) = 0, \ i = 1,2,\ldots, \ n \end{gathered} \tag{3-5}$$

其中$\langle x_j, x_i\rangle$是$x_j x_i$的内积，相当于$x_j^T x_i$。

3.3.2 支持向量机模型构建流程

在输入训练集后，写出目标函数，对偶化后选择高斯核函数，调整参数，求解目标函数得到分类器，构建流程图如图 3-2 所示：

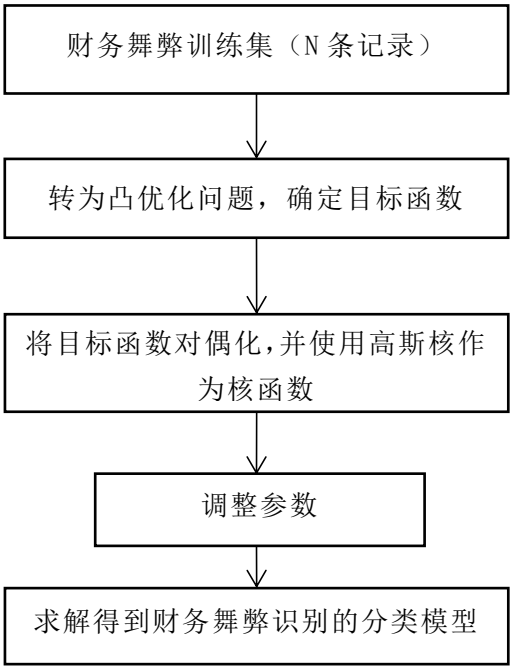

图 3-2 基于支持向量机的财务舞弊识别模型的构建流程图

3.4 基于自适应增强算法的财务舞弊识别模型构建

3.4.1 自适应增强概述

自适应增强（Adaptive Boosting,AdaBoost）算法是一种具有自适应学习能力的分类算法，其前身是 Boosting 算法。Boosting 是一种包含多个弱分类器的集成学习模型，每个弱分类器之间存在关联，即后一个弱分类器是基于前一个弱分类器生成的，通常通过赋予被错误判断样本更大权重的办法，对于前一个弱分类器识别错误的样本进行着重的学习训练。除此之外，Boosting 在每个弱分类器训练时，对样本集的抽样采用无放回方式，这与 Bagging 方法存在差异。

AdaBoost 算法中，将弱分类器的迭代次数设为 T，并将第 k 个弱分类器中样本训练集的输出权重定义为 $D(K) = (w_{k1}, w_{k2}, \cdots, w_{kn})$，则样本的初始化权重表示为 $D(K) = (w_{11}, w_{12}, \cdots, w_{1n})$，且 $w_{11} = w_{12} = \cdots = w_{1n} = \frac{1}{n}$，即每个样本在最开始具有相同的权重。

算法的数学过程为：

（1）对于 $k = 1,2, \cdots T$，使用权重为 D(k)的样本集进行数据的训练，获得

第 k 个弱分类器$F_k(x)$。

（2）确定样本分类问题的错误率，训练集的第 k 个弱分类器$F_k(x)$的加权错误率e_k表示为式 3-6：

$$e_k = P(F_k(x) \neq y_i) = \sum_{i=1}^{n} w_{ki} I(F_k(x) \neq y_i) \tag{3-6}$$

其中，$I(F_k(x) \neq y_i)$反映误差样本点，一般为 0 或 1，正确样本分类取值为 0，错误分类取值为 1。因此，对每个弱分类器错误分类的样本进行权重求和即可得到该分类器的错误率。

（3）求出弱分类器的权重系数，设置弱分类器的权重α_k为：

$$\alpha_k = \frac{1}{2}\ln\frac{1-e_k}{e_k} \tag{3-7}$$

根据式 3-7 可知，错误率e_k越大，则该弱分类器所拥有的权重α_k越小。

（4）求出更新后的样本权重，设第 k 个弱分类器$F_k(x)$所对应的各样本的集合的权重是 $D(K) = (w_{k1}, w_{k2}, \cdots, w_{kn})$，则可求出下一次在抽样训练时其对应的样本集合的权重$w_{k+1,i}$如式 3-8：

$$w_{k+1,i} = w_{ki} e^{\{-\alpha_k y_i F_k(x_i)\}/Z_k} \tag{3-8}$$

其中Z_k为归一化因子，表示为$Z_k = 2[e_k(1-e_k)]^{\frac{1}{2}}$。由式 3-8 可知，若在第 k 个分类器中的第 i 个样本被错误分类，则$y_i F_k(x_i)$值为负，使得样本在下一个分类器中的权重提高。

（5）经过弱分类器组合可得到强分类器，最终分类器为式 3-9：

$$f(x) = \mathrm{sign}\left(\sum_{t=1}^{T} \alpha_k F_k(x)\right) \tag{3-9}$$

3.4.2 自适应增强模型构建流程

基于 AdaBoost 的财务舞弊识别模型构建过程如下：

（1）对于训练样本集中的每一个样本，均赋值相同的初始化权重；

（2）对各子步骤进行重复，使误差小于阈值或者达到迭代次数限制：

(a) 对训练样本依据权重进行抽样训练，得到一个弱分类器；

(b) 用 a 步骤产生的弱分类器对样本集计算，得到判别值；

(c) 对每个样本的识别情况进行记录，获得总误差率；

(d) 将被错误识别的样本权重提高；

(e) 重复 a 至 d 步骤；

（3）依据各分类器性能进行加权投票，合并生成强分类器。

构建流程图如图 3-3 所示：

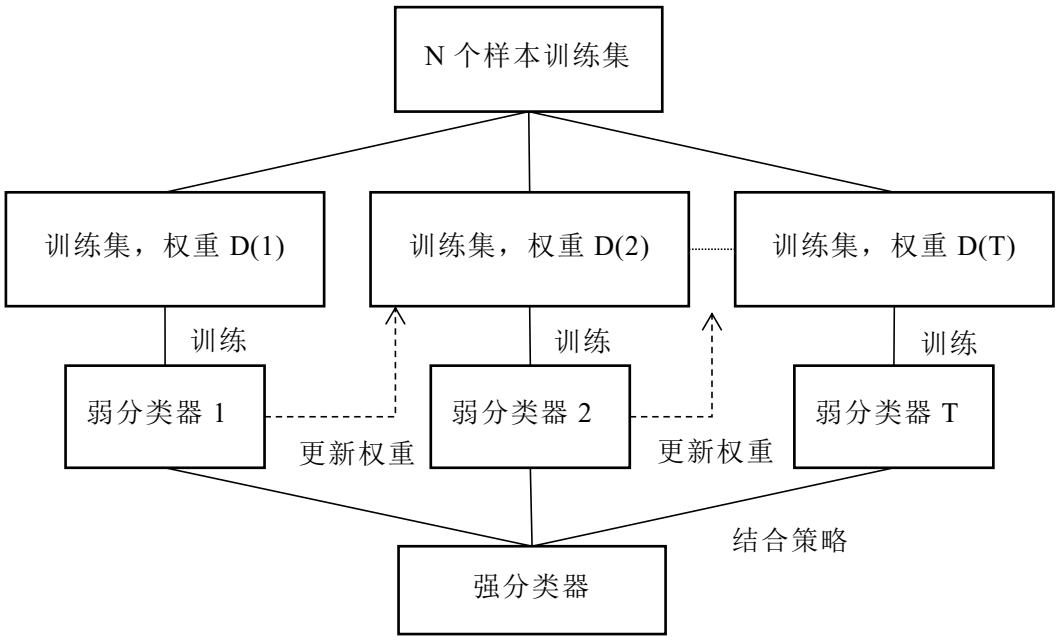

图 3-3 AdaBoost 模型构建流程

3.5 评价指标体系

3.5.1 混淆矩阵与准确度

混淆矩阵是用来直观评价模型效果的一种工具，其使用交叉表的形式对样本的判别结果与真实标签结果进行表示。通过混淆矩阵可以直观的看到数据集在模型上的识别或预测效果。对于二分类问题，混淆矩阵通常是一个 2 乘 2 的表格。2 个横行代表样本的真实值，阳性样本（又称正例，通常用 1 表示）或阴性样本（又称负例，通常用 0 表示），纵列代表模型的判别情况，交叉形成了样本真实值与模型预测值的直观效果。混淆矩阵的形式如表 3-1 所示：

表 3-1 混淆矩阵

		预测类别	
		1	0
实际类别	1	TP	FN
	0	FP	TN

敏感性与特异性分别是用来评价正负样本判断情况的重要指标，可依据混淆均值算出，其计算公式分别如式(3-11)及(3-12)所示：

$$敏感性 = \frac{TP}{(FN + TP)} \tag{3-11}$$

$$特异性 = \frac{TN}{(FP + TN)} \tag{3-12}$$

从混合矩阵可以计算出许多指标来评价模型的判别效果。其中，准确率(Accuracy)是一个常用的评价指标，准确率可以整体的表达模型的识别能力，即整体上有多少样本被正确识别，其计算方法如式(3-13)所示：

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{3-13}$$

3.5.2 ROC 曲线与 AUC 值

准确率虽然可以对模型整体效果进行评价，但在样本不平衡问题中却一般不被使用，这主要是因为当大量假阳性样本存在时，准确率仍然会很高，但却失真。因此，在对模型评估中，常使用 ROC 曲线及 AUC 值对分类器效果进行判别。

ROC(Receiver Operating Characteristic)曲线，又名接受者操作特征曲线，，该曲线常用于评价模型的判别能力，ROC 曲线的横纵坐标是基于特异性和敏感性画出的。

在财务舞弊识别模型的分类性能评价中，希望尽可能地将舞弊记录识别出来，但另一方面，要尽可能避免对未舞弊企业的误判。所以需要对敏感性及特异性指标进行平衡，而这二者在实际应用中常常是处于对立状态，即一者随着另一指标的增高而降低，而 AUC 值是对这两个指标的一个平衡计算。

AUC 是曲线下面积(Area under the Curve)，在比较同一问题的不同的分类模型时，通常会绘制出每个模型的 ROC 曲线，并用曲线下的面积作为模型性能的一个重要指标。需要表明 AUC 的几点重要性质：

（1）ROC 曲线横纵坐标的取值范围是 0 到 1，因而 AUC 值也在 0~1 之间；

（2）一般认为大于分类阈值的样本为阳性，小于分类阈值的为阴性；

（3）如果随机选取一个正样本和负样本，则分类器判断正样本值大于负样本值的概率为 AUC 值；

（4）简单说：在其他指标不变的情况下，分类器的 AUC 越大，则其性能越好。

3.5.3 NDCG@k 值

目前的已有财务舞弊识别研究中，大多数用的是传统的 AUC 的评价指标，用来评价模型的识别效果。由于相对于非舞弊的公司，上市公司财务舞弊是小概率事件，即使是表现最好的欺诈预测模型也会导致大量误报，远远超过测试期间的真阳性数。显然，鉴于打击此类欺诈的可用资源有限，监管机构或公司监管者调查所有预测的欺诈案例是不切实际的。即使有人希望调查所有的欺诈预测，直接和间接的成本将是巨大的，而收益将是小的(因为大多数预测欺诈观察是假阳性)。因此，监管机构和其他监管者自然会寻求调查数量最少、预测欺诈可能性最高的观察结果。

显然，鉴于打击此类欺诈的可用资源有限，监管机构或公司监管机构调查所有预测的舞弊案件是不切实际的。Yang 等(2020)提出的将排序指标 NDCG@k 应用于财务舞弊识别问题上，是更合适的。

NDCG@k 又称归一化折损累计增益，常被用于评价排序推荐问题的效果，是一种加权排序，即各结果的预测准确性与其排序位置进行加权。NDCG@k 是 DCG@k 指标的标准化，DCG@k 的计算公式如式(3-14)所示：

$$\mathrm{DCG@K} = \sum_{i}^{K} \frac{r(i)}{log_2(i+1)} \tag{3-14}$$

其中 i 代表该结果所排序的位置，$r(i)$ 在判别正确时为 1，判别错误时为 0，K 在本内容里取值为测试实例数量的 2%，选定 2%这个比例是因为依据本文定义，我国的上市公司财务舞弊的比例大约为 2%。

$$\mathrm{NDCG@K} = \frac{DCG@K}{ideal\ DCG@K} \tag{3-15}$$

其中 $ideal\ DCG@K$ 为 $DCG@K$ 的理想值，即所判别的公司中，前 K 个均判别正确时的 $DCG@K$ 值。

由式(3-15)可知，NDCG@K 的值在 0 到 1 之间，其值越大越好。

3.6 本章小结

本章主要说了模型构建的方法和模型效果评价的指标。介绍了随机森林、支持向量机和自适应增强这三种机器学习方法原理，并说明了对应的模型构建流程，基于不同算法的模型构建为基于不同指标体系的有效性论证提供了充分

的比较基础。在评价指标方面，采用传统的 AUC 值和更适合财务舞弊识别问题的 NDCG@K 值，并对两种评价指标的概念及算法进行了详细的论述，对采用 NDCG@K 值的原因进行了说明，NDCG@K 值在财务舞弊识别问题中具有较好的实际应用的价值和管理意义上的价值，即可避免由于样本不平衡导致的大量假阳性样本和十分有限的监管资源导致模型的实际不可用。评价指标体系的选取将便于讨论不同模型之间的有效性，使得不同模型之间存在可比性。

第 4 章 模型效果比较与分析

4.1 基于财务比率指标与原始会计指标的模型效果比较

目前国内的财务舞弊识别模型中多使用财务比例指标，而鲜有使用原始会计数据的。在机器学习模型中，原始会计数据是否较财务比率指标更优是值得探讨的。一方面原始会计数据保留了更多的信息，同时机器学习会对数据自动学习，其效果可能更好；另一方面由于原始会计数据没有利用财务比率指标的财务理论，可能效果并不好。Yang Bao(2020)证明了在财务原始数据与集成学习在美国上市公司数据的情况下是有效的，效果超过了财务比率指标与集成学习模型的搭配使用。中国企业经营环境和财报要求与美国存在较大差异，原始会计数据在帮助识别中国上市企业财务舞弊问题上是否有效仍需探讨。

为了表现识别模型效果的可靠性，本文对训练集和测试集进行了 10 次的随机划分，形成 10 套训练测试集。为比较各识别模型在统计意义上，其判别效果是否存在差异，本文使用 t 检验。t 检验是用 t 分布理论来推论差异发生的概率，从而比较两个平均数的差异是否显著，若 p 值小于 0.05（在表格中标记**），则被认为二者在统计上存在差异。

4.1.1 基于随机森林的模型

针对随机抽取的 10 次训练样本集进行基于不同指标体系的 RF 模型构建，如图 4-1 所示，在 0-200 棵树木期间，随着决策树数量的增加，模型的错误率整体逐渐下降，而后逐渐平稳。

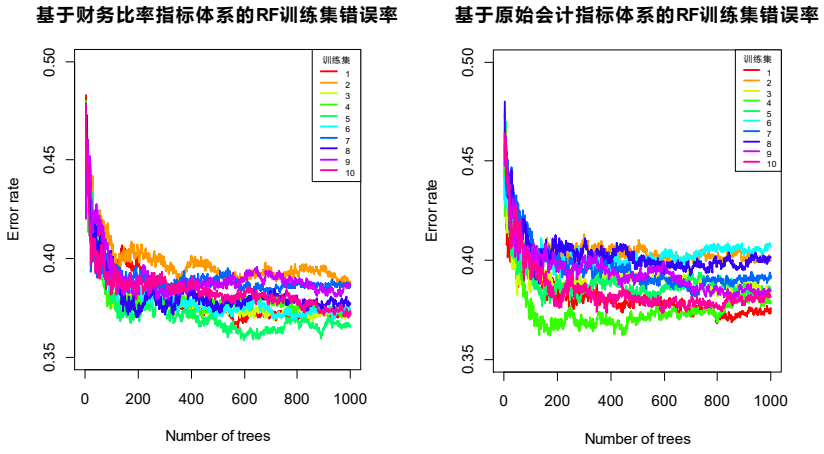

图 4-1 基于两种指标体系的随机森林模型错误率随着树数量的增加变化情况

基于财务比率指标指标体系和原始会计指标体系的随机森林模型在测试集上的 ROC 曲线如图 4-2 所示，可以看到两种指标体系下的 RF 模型判别效果相近，而在同一指标体系下，各测试集之间的判别效果存在一定差异，这种由于随机扰动产生的误差有限，本文进行的多次实验可以使得随机扰动的误差被减少。

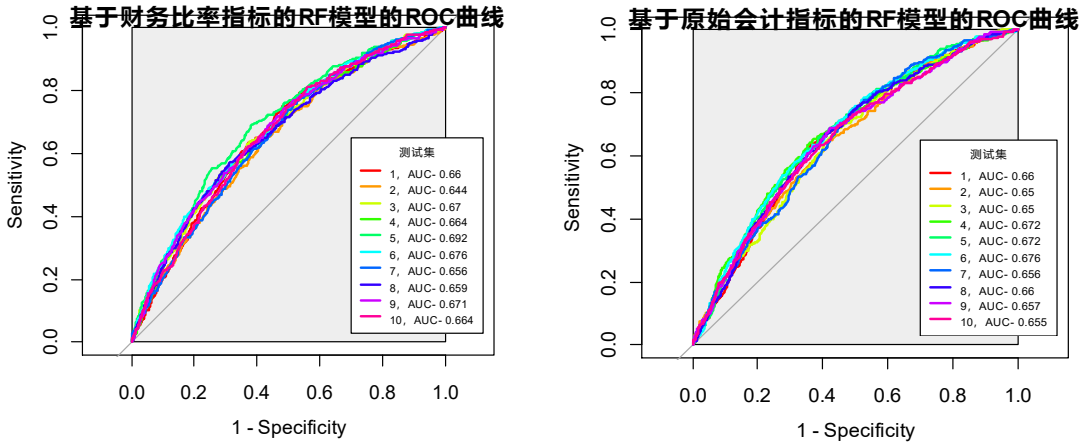

图 4-2 基于不同指标体系的 RF 模型的 ROC 曲线

为验证十次实验的 AUC 值与 NDCG@K 值是否符合正态分布，进行正态分布的 Q-Q 图检验，由图 4-3 可知，数据基本符合正态分布，可以用于独立样本的 t 检验，本文所使用的 t 检验均使用了 welsh 修正自由度，因此可以不考虑方差齐次性的前提条件。

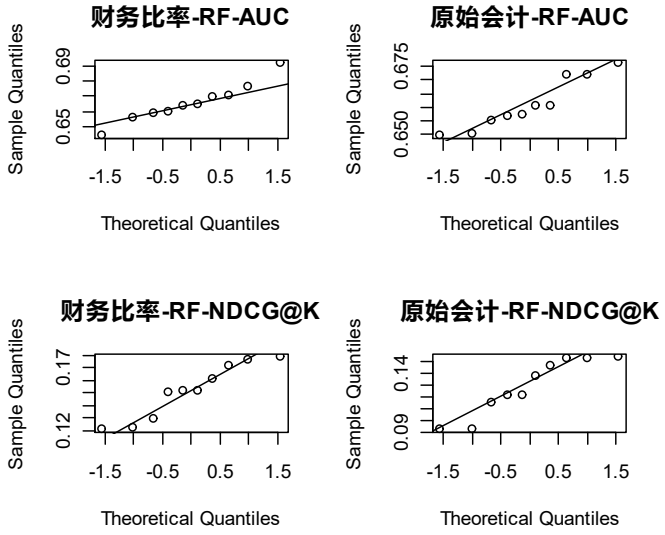

图 4-3 基于两种指标体系的 RF 模型多次实验结果的 Q-Q 图

表 4-1 展现了基于财务比率指标体系和原始会计指标体系的随机森林模型在不同测试集上运行十次的均值结果：基于财务比率指标体系的 RF 模型测试集的十次运行的 AUC 均值为 0.665，平均 NDCG@K 值为 0.151；而基于原始会计指标体系的 RF 模型测试集的 AUC 均值为 0.661，NDCG@K 均值为 0.128。对二种模型的 AUC 进行 t 检验，p 值为 0.370，这表明接受原假设，即可以认为两组 AUC 的均值没有差别。同理，NDCG@K 值的 t 检验得出的 p 值为 0.034，可认为基于财务比率指标的 RF 模型在评价指标 NDCG@K 值上，表现更好。

表 4-1 基于财务比率指标和原始会计指标体系的 RF 模型测试集结果

	财务比率	原始会计	t 检验-p 值
AUC 均值	0.665	0.661	0.370
NDCG@k 均值	0.151	0.128	0.034**

4.1.2 基于支持向量机的模型

图 4-4 展示了基于两种指标体系的 SVM 模型在 10 组测试数据集下的 AUC 和 NDCG@K 表现情况，可以看到，在同一数据集下，基于财务比率指标体系与基于原始会计指标体系的模型效果总体相近。

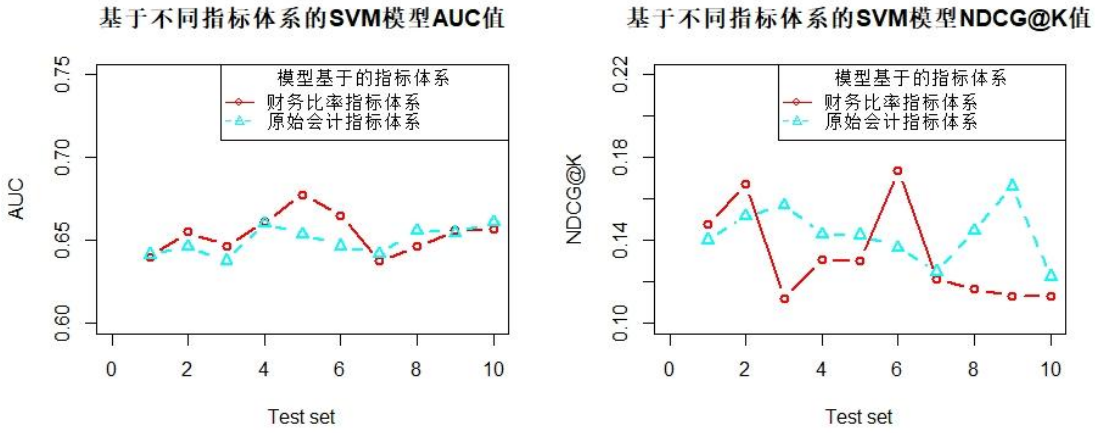

图 4-4 不同指标体系下的 SVM 模型测试集判别情况

为验证基于不同指标体系的 SVM 模型是否在统计上存在均值差异，先对测试集判别的评价指标进行正态 Q-Q 图的观测，如图 4-5 所示，可以认为各 AUC 和 NDCG@K 值大体上符合正态分布，符合 t 检验条件。

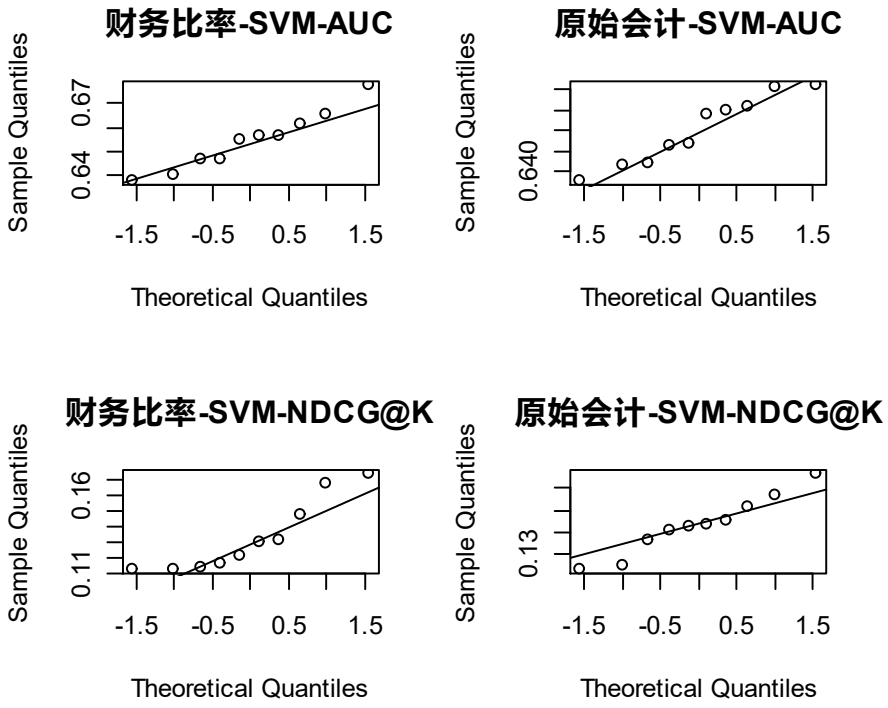

图 4-5 基于两种指标体系的 SVM 模型多次实验结果的 Q-Q 图

表 4-2 展现了基于财务比率指标体系和原始会计指标体系的支持向量机模型测试集结果：基于财务比率指标体系的 SVM 测试集的十次 AUC 均值为 0.654，NDCG@K 均值为 0.133；基于原始会计指标体系的 SVM 模型测试集的 AUC 均值为 0.650，NDCG@K 值为 0.143。两个评价指标的 t 检验均不显著，这表明财务比率指标体系和原始会计指标体系对 SVM 算法的 AUC 和 NDCG@K 影响不大。

表 4-2 基于财务比率指标和原始会计指标体系的 SVM 模型测试集结果

	财务比率	原始会计	t 检验-p 值
AUC	0.654	0.650	0.367
NDCG@k	0.133	0.143	0.234

4.1.3 基于自适应增强的模型

图 4-6 展示了基于两种指标体系的 AdaBoost 模型在 10 组测试数据集下的 AUC 和 NDCG@K 表现情况，可以看到，在同一数据集下，基于财务比率指

标体系与基于原始会计指标体系的模型效果总体相近。

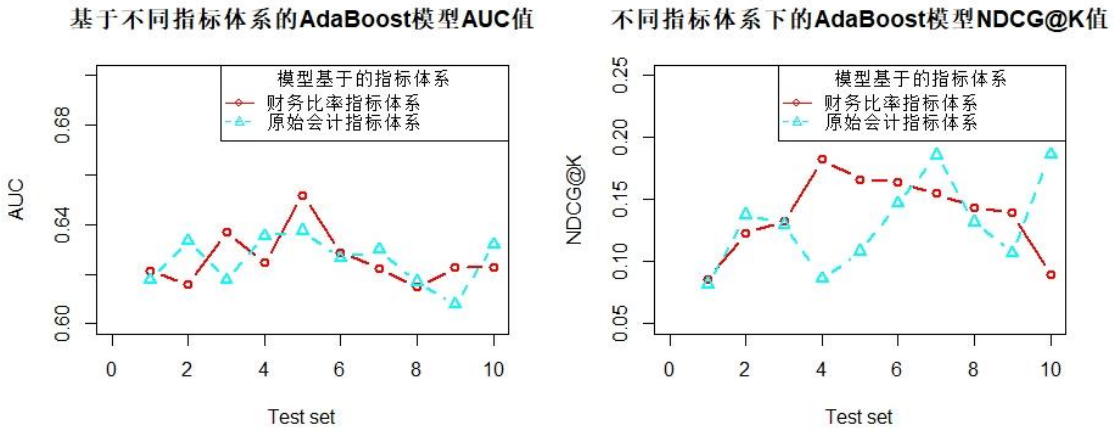

图 4-6 不同指标体系下的 AdaBoost 模型测试集判别情况

为验证基于不同指标体系的 AdaBoost 模型是否在统计上存在均值差异，先对测试集判别的评价指标进行正态 Q-Q 图的观测，如图 4-7 所示，可以认为各 AUC 和 NDCG@K 值大体上符合正态分布，符合 t 检验条件。

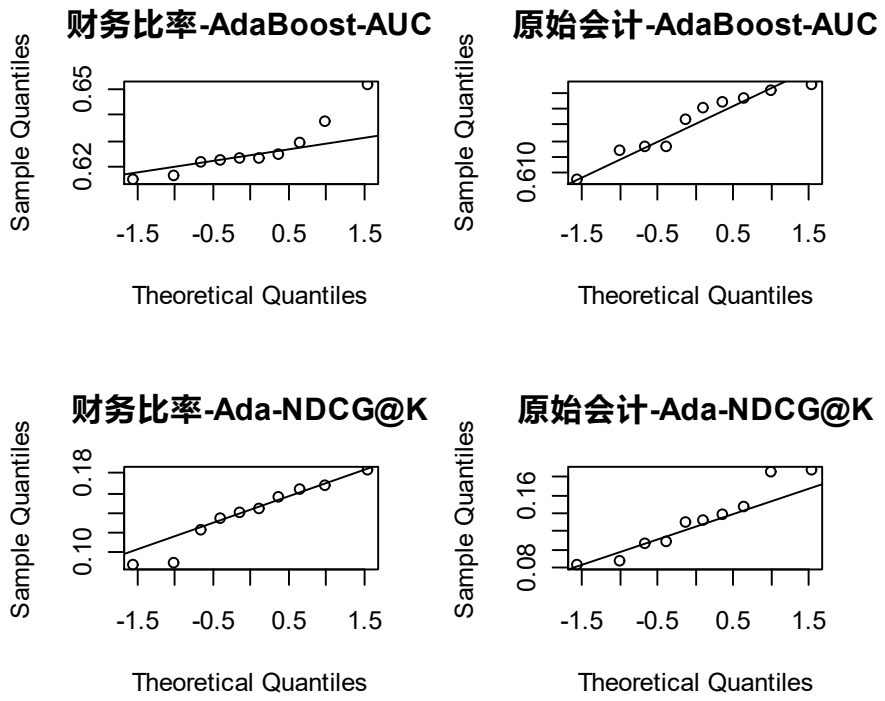

图 4-7 基于两种指标体系的 SVM 模型多次实验结果的 Q-Q 图

基于自适应增强算法的财务比率指标体系和原始会计指标体系的两个模型的测试集结果如表 4-3 所示：基于财务比率指标体系的 AdaBoost 模型测试集的 AUC 均值为 0.626，NDCG@K 均值为 0.138；基于原始会计指标体系的 AdaBoost 模型测试集表现情况为，十次 AUC 均值为 0.625，NDCG@K 均值为 0.130。AUC 值的 t 检验的 p 值为 0.881，NDCG@K 的 t 检验 p 值为 0.612，二者均不显著，表明基于财务比率指标体系和原始会计指标体系的 AdaBoost 算法效果没有差异。

表 4-3 基于财务比率指标和原始会计指标体系的 AdaBoost 模型测试集结果

	财务比率	原始会计	t 检验-p 值
AUC	0.626	0.625	0.881
NDCG@k	0.138	0.130	0.612

综上所述，为讨论在国内上市公司的数据条件下，原始会计指标是否会比财务比率指标更有效，本文构建了基于财务比率指标体系的多个模型和基于原始会计数据指标体系的多个模型，包括随机森林、支持向量机、自适应增强在内的机器学习算法，所使用的财务比率指标为 27 个，原始会计数据指标为 27 个，三种算法在两种指标体系及不同数据集进行了 10 次运行，结果表明，暂时没有证据表明中国上市公司数据下的原始会计数据指标体系较财务比率指标体系更优。

4.2 基于合并指标体系的模型效果讨论

虽然没有证据表明，中国上市公司数据下的原始会计数据指标体系较财务比率指标体系更优，但原始会计指标本身可能含有一些财务比率指标失去的信息，将两个指标体系合并成一个指标体系，有可能会使得模型效果提高。

为对这一问题进行讨论，构建了基于合并指标体系的财务舞弊识别模型，结果如表 4-4 所示，其中随机森林和自适应增强算法的模型识别效果有所提高，二者 AUC 均有一定提高。支持向量机算法在合并指标体系上的模型效果并无明显提高，这主要是由 SVM 算法和所采用指标体系的特点决定的，SVM 算法使用核函数（本文使用的为高斯核函数）使得指标特征在原有的基础上通过组合变形映射到了更高维的空间，而原始会计指标体系又可以通过计算组合成财务比率指标体系，这使得合并指标体系对模型的效果影响不大。

总的来讲，合并指标体系对基于 RF 和 AdaBoost 算法的识别模型效果有一定的提升。

表 4-4 基于三种指标体系的识别模型测试集结果比较

	随机森林		支持向量机		自适应增强	
	AUC	NDCG@k	AUC	NDCG@k	AUC	NDCG@k
合并指标体系	0.672	0.156	0.656	0.138	0.635	0.138
财务比率指标	0.665	0.151	0.654	0.133	0.626	0.138
原始会计指标	0.661	0.128	0.650	0.143	0.625	0.130

4.3 基于时间维度扩充指标体系的模型有效性讨论

如表 4-5 所示，无论是基于财务比率指标的时间扩展的指标体系，还是基于原始会计指标的时间扩展的指标体系，或是基于合并指标的时间扩展的指标体系，随机森林与自适应增强算法均表现出了比未进行时间拓展的指标体系更好的判别效果，而支持向量机则没有明显差别。如基于 RF 的算法中，对财务比率指标体系下的模型的 AUC 均值和 NDCG@k 均值由 0.665、0.151 分别提升至 0.690、0.203，模型的判别效果明显提升；基于 AdaBoost 算法的三种指标体系的模型在 AUC 值上也均有明显的提升。

实验结果表明，对指标体系进行时间维度的拓展，可以在一定程度上提高基于 RF 或 Adaboost 算法的模型判别能力。

表 4-5 基于三种指标时间扩展指标体系的识别模型测试集结果

	随机森林		支持向量机		自适应增强	
	AUC	NDCG@k	AUC	NDCG@k	AUC	NDCG@k
对合并指标体系的时间扩展	0.697	0.200	0.651	0.159	0.664	0.146
合并指标体系	0.672	0.156	0.656	0.138	0.635	0.138
t 检验-p 值	0.000**	0.022**	0.245	0.210	0.000**	0.672

表 4-5 （续表）

	随机森林		支持向量机		自适应增强	
	AUC	NDCG@k	AUC	NDCG@k	AUC	NDCG@k
对财务比率指标的时间扩展	0.690	0.203	0.652	0.146	0.655	0.145
务比率指标体系	0.665	0.151	0.654	0.133	0.626	0.138
t 检验-p 值	0.000**	0.012**	0.592	0.393	0.000**	0.699
对原始会计指标的时间扩展	0.690	0.170	0.658	0.147	0.656	0.140
原始会计指标体系	0.661	0.128	0.650	0.143	0.625	0.130
t 检验-p 值	0.000**	0.006**	0.038**	0.746	0.000**	0.508

4.4 连续舞弊样本处理的必要性讨论

由于连续财务舞弊是上市公司财务舞弊中的常见行为，而传统研究中大多未针对此问题进行考虑或讨论，因此在研究中应进行考量。

本文使用表现较好的随机森林算法进行对比实验，以原始会计指标体系为例，比较考虑与不考虑连续财务舞弊情况时，模型在测试集上的差异变化情况。在采用相同算法的情况下，对未剔除连续财务舞弊的样本，划分出相同的测试集数据，其余用来训练。

表 4-6 基于不同样本集的随机森林财务舞弊识别模型效果

表现较好的随机森林算法	训练集	连续舞弊样本处理后的测试集		连续舞弊样本未处理的测试集
	AUC 均值	AUC 均值	t 检验-p 值	AUC 均值
剔除了连续舞弊样本中的重复同一公司样本	0.659	0.665	0.221	/
未剔除连续舞弊样本中的重复同一公司样本	0.716	0.682	0.000**	0.719

如表 4-6 所示，未剔除连续财务舞弊样本时，其 AUC 均值在训练集中为 0.716，而连续舞弊样本处理后的测试集 AUC 均值为 0.682，且 t 检验显著，而对连续舞弊样本进行处理了的识别模型则 t 检验不显著，同时，未剔除连续舞弊的模型在未剔除连续舞弊的测试集上的 AUC 均值未 0.719，明显高于真实值，这表明模型效果被高估了。同时，未进行剔除处理的模型，不符合实际应用场景，这主要是由于训练时利用了连续舞弊的先验信息，而这些先验信息在实际应用中是无法获得的。

综上所述，在财务舞弊识别问题中，对连续舞弊这一关键问题给予考量是必要的，仅保留连续舞弊公司记录的第一条舞弊样本是有效可行的，若不处理连续舞弊问题，将会造成识别模型的结果被高估及实际不可用。

4.5 与基准模型效果对比

为讨论模型的有效性，以 Yang 等(2020)的最优模型为基准模型，进行比较和分析。基准模型的复现采用表 2-3 中的指标体系，这是 Yang 等(2020)对美国上市公司的财务舞弊识别问题进行研究后，参考了 Cecchini(2010)和 Dechow(2011)的研究，选择这二份研究所对应的 40 余个原始会计数据，剔除缺失值超过 25%的指标后最终选出的 28 个原始会计数据指标。其证明了美国上市公司原始会计数据与 RUSBoost 算法结合的有效性。

本研究对 Yang 所选的 28 个原始会计数据指标进行讨论，比较了中美会计准则及财报的差异，翻译形成了 32 个原始会计指标，按照与 Yang 等(2020)同样的方法，剔除缺失值超过 25%的指标，最终得到 24 个变量。

RUSBoost 算法是 AdaBoost 算法的一种变体，同样是一种基于树的

Boosting 集成学习方法。RUSBoost 算法的不同之处是其针对样本不平衡问题进行了额外一步的处理。采用随机欠采样的方式，使得财务舞弊样本与非财舞弊务样本进行平衡，在实现过程中，我们采取与 Yang 等(2020)同样的随机欠采样方式，使得舞弊样本与未舞弊样本数量完全一致，即舞弊与未舞弊各 863 个样本。采用 500 个树模型作为基模型进行投票，如表 4-7 所示，基准模型在测试集上的 AUC 均值为 0.626，NDCG@K 均值为 0.134。

表 4-7 中国上市公司财务舞弊识别模型与基准模型效果比较

	AUC	NDCG@k
中国上市公司财务舞弊识别模型	0.697	0.200
基于中国上市公司数据的 Benchmark Model	0.626	0.134
t 检验-p 值	0.000**	0.001**

在本文所提出的模型中，采用效果较好的基于时间维度扩充的合并指标体系的随机森林模型与基准模型进行对比，AUC 值及 NDCG@K 值如表 4-7 所示，本文所提出的中国上市公司财务舞弊识别模型在性能上更优，其 AUC 为 0.697，高于基准模型的 0.626，NDCG@K 为 0.200，远高于基准模型的 0.134，两个评价指标均明显高于基准模型，且均通过了 t 检验，因此本文提出的中国上市公司财务舞弊识别模型具有一定的学术价值及管理应用价值。

4.6 本章小结

本章比较并讨论了多种模型的有效性，对基于财务比率指标体系和原始会计指标体系的财务舞弊识别模型进行比较，证明了在中国上市公司财务舞弊识别模型中，原始会计指标体系并不优于财务比率指标体系，合并指标体系可以在一定程度上提高基于随机森林和自适应增强算法的模型效果。对财务比率指标体系、原始会计体系、合并指标体系这三种指标体系进行时间维度扩展实验，结果表明对指标进行时间维度的扩展可以有效提高基于随机森林和自适应增强算法的模型判别能力。

本章对连续舞弊这一常见问题的处理进行了讨论，证明了财务舞弊识别问题中，连续舞弊样本处理的必要性。实验结果表明，不进行连续舞弊样本处理的样本集会高估模型的识别能力，同时模型不符合实际应用场景。

最后，本章将本文所提出的基于时间维度扩充的合并指标体系的随机森林模型与基准模型相对比，实验结果表明，本文所提模型具有一定的有效性。

结论

本文以2007-2019年的真实中国上市公司数据为样本，依据财务舞弊理论和参考文献，在数据预处理和特征选择方面做了专门化处理，建立了一套指标体系，并以随机森林、支持向量机、自适应增强为算法基础，构建了中国上市公司财务舞弊识别模型。本文通过对各模型的比较和分析，取得的研究成果如下：

（1）**构建了中国上市公司财务舞弊识别指标体系**　本文比较了多套指标体系，并证明有效性。依据财务舞弊理论最终选取了财务比率指标体系和原始会计指标体系；将二者指标进行合并生成合并指标体系。实验结果表明，暂无证据表明中国财务舞弊识别问题中，原始会计指标体系优于财务比率指标体系。实验还表明，通过从时间维度对指标体系进行扩充，可明显提高基于RF或AdaBoost算法的模型效果。

（2）**论证连续舞弊样本处理的必要性**　在样本选择方面，选取了20000多条真实记录进行研究，涵盖了大部分的中国上市公司，使得模型在实际应用中存在一定的意义。特别论证了连续舞弊样本的处理问题，由于连续舞弊是上市公司舞弊行为中的常见行为，因此应给予重视。不做特殊处理的样本集会高估模型的识别能力，且不符合实际应用场景。

（3）**构建了中国上市公司财务舞弊识别模型**　本文讨论了基于RF、SVM和AdaBoost算法的财务舞弊识别模型。三种算法中，RF与SVM算法具有较好的识别效果，其中RF在基于时间扩充或合并指标体系下表现更强，与基准模型的判别效果比较，本文提出的基于时间扩充后的合并指标体系的随机森林财务舞弊识别模型的AUC为0.697，NDCG@k为0.200，明显高于基准模型的0.626与0.134。

同时，本文提出的方法也存在一定的不足。时间维度扩充后的指标对模型判别有较高的价值，但本文未对进行更深入的探索。在后续研究中，应进一步加强这方面的研究。

参考文献

[1] 张佳佳. 基于数据挖掘的上市公司财务报告舞弊识别模型研究[D].浙江大学,2021.DOI:10.27461/d.cnki.gzjdx.2021.001360.

[2] American Institute of Certified Public Accountants(AICPA). 2002. Consideration of Fraud in a Financial Statement Audit. Statement on Auditing Standards No. 99. New York, NY: AICPA.

[3] 中国注册会计师协会.注册会计师审计法律与准则.北京.中国法制出版社，2006：113.

[4] Beasley M S, Carcello JV ,Hermanson D R, et al.(2000). Fraudulent financial reporting: Consideration of industry traits and corporate governance mechanisms[J]. Accounting Horizons, 14(4): 441-454

[5] DECHOW, P. M., GE, W., LARSON, C. R. and SLOAN, R. G. (2011), Predicting Material Accounting Misstatements.Contemporary Accounting Research, 28: 17–82. doi: 10.1111/j.1911-3846.2010.01041.x

[6] Zager L,Malis S S,Novak A.The Role and Responsibility of Auditors in Prevention and Detection of Fraudulent Financial Reporting[J].Procedia Economics & Finance.2016,39:693-700.

[7] 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):21-25.

[8] 韩文明.中国上市公司会计造假行为的统计特征分析[J].审计与经济研究,2005(05):56-60.

[9] 吴晓迪.财务造假的手段剖析及防范措施[J].现代商业,2011(21):251-251.

[10] 曾汝林. 浅析企业财务欺诈的主要手段及其防范[J]. 中国商论,2011(9X):2.

[11] 王淑玲.财务舞弊的手段与治理方法分析[J].现代商业,2012(10):242-243.

[12] 刘永.上市公司绿大地造假案例分析[J].中国市场,2013(45).

[13] 曹媛.上市公司财务舞弊问题研究[D];财政部财政科学研究所,2015.

[14] 刘元,林爱梅,单雅迪.我国上市公司财务报告舞弊的特征和手段——基于2008-2013 年证监会处罚公告[J].财会月刊,2015(28):16-19.

[15] 刘石球.虚构经济业务型财务造假手法剖析及识别[J].中国注册会计师,2016(12):73-77+2.

[16] 黄进敏.浅谈企业财务报表舞弊及审计对策[J].会计师,2017(11):53-54.

[17] 万朝辉.识别虚假财务信息[J].中国乡镇企业会计,2019(02):75-76.

[18] 张彤.中国上市公司财务报告舞弊手段、成因及对策研究[J].财会学习,2019(12).
[19] 郑伟宏,李晓,张婷,黄敬龄.上市公司财务报告舞弊与审计揭示——基于证监会行政处罚决定书的分析[J].财会通讯,2019(22):19-25.
[20] Albrecht,W.S. andRomney.Red-flagging management fraud:a validation.[J] Advances in Accounting,1986,3,323-333
[21] Loebbecke J, J Willingham. Review of SEC accounting and auditing enforcement releases: [dissertation]. Salt Lake City: University of Utah, 1988,1-5
[22] Dechow P M, Sloan R G, Sweeney A P. Causes and Consequences of Earnings Manipulation: An Analysis of Firms Subject to Enforcement Actions by the SEC. Contemporary Accounting Research,1996, 26(13): 1-36
[23] 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(16):271-309.
[24] Beneish M.D. Incentives and Penalties Related to Earnings Overstatements That Violate GAAP[J].The Accounting Review，1999，74 (4): 425-457.
[25] Abbott L J, Parker S, Peters G F. Audit committee characteristics and restatements. Auditing: A Journal of Practice and Theory, 2004, 24(23): 69-87
[26] David B.Farber,2005.Restoring trust after fraud Does corporate governance matter?The Accounting Review,Vol.90,No.2(Apr.,2005),pp.539-561.
[27] Eliezer M.Ficha and Anil Shivdasanib, Financial fraud,director reputation and shareholder wealth.Journal of Financial Economics 86(2007) 306-336
[28] Erickson M , Maydew H E L . Is There a Link between Executive Equity Incentives and Accounting Fraud?[J]. Journal of Accounting Research, 2010, 44(1):113-143.
[29] Yang Bao,Bin Ke,Bin Li,Detecting Accounting Fraud in Publicly Traded U.S. Firms Using a Machine Learning Approach [J]. Journal of Accounting Research,2020,58(1).
[30] 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
[31] 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(08):84-90+97.
[32] 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(02):107-114.
[33] 程鑫.上市公司纳税申报信息的真实性识别研析——基于代价敏感支持

向量机模型[J].税务研究,2015(07):81-84.
[34] 张苏彤.奔福德定律与舞弊审计——基于“人为造假”与随机数样本的实证测试[J].会计之友,2016(12):7-15.
[35] 洪荭,胡华夏,王晶.盈余管理与财务舞弊关系的演变与动态拓展[J].会计与经济研究,2017,31(03):32-55.
[36] 李清，任朝阳.上市公司会计舞弊风险指数影响因素研究[J].当代经济科学，2017,39(5):67-75
[37] 向晖.企业财务舞弊常见手段及审计要点[J].中国国际财经(中英文),2018(08):110.
[38] 黄世忠,叶钦华,徐珊.上市公司财务舞弊特征分析——基于 2007 年至 2018 年 6 月期间的财务舞弊样本[J].财务与会计,2019(10):24-28.
[39] PERSONS O S. 1995. Using Financial Statement Data to Identify Factors Associated with Fraudulent Financial Reporting[J]. Journal of Applied Business Research, 11(3): 38-46.
[40] LEE T A, INGRAM R W, HOWARD T P. 1999. The Difference between Earnings and Operating Cash Flow as an Indicator of Financial Reporting Fraud[J]. Contemporary Accounting Research, 16(4): 749-786.
[41] BELL T B, CARCELLO J V. 2000. A Decision Aid for Assessing the Likelihood of Fraudulent Financial Reporting[J]. Auditing: A Journal of Practice & Theory, 19(1): 169-184.
[42] Kotsiantis S , Koumanakos E , Tzelepis D , et al. Forecasting fraudulent financial statements using data mining[J]. Enformatika, 2006, 3(2):104-110.
[43] Feroz, E.H., Kwon, T. M., Pastena, V., & Park, K.J. The Efficacy of Red Flags in Predicting the SEC’s Targets: An Artificial Neural Networks Approach[J].International Journal of Intelligent Systems in Accounting,Finance &Management.2000,9:145-157.
[44] Cecchini M , Aytug H , Koehler G J , et al. Detecting Management Fraud in Public Companies[J]. Management Science, 2010, 56(7):1146-1160.
[45] Ozdagoglu G, Ozdagoglu A, Gumus Y, et al. The application of data mining techniques in manipulated financial statement classification: The case of turkey[J]. Journal of AI and Data Mining, 2017, 5(1): 67-77.
[46] 刘君, 王理平. 基于概率神经网络的财务舞弊识别模型[J]. 哈尔滨商业大学学报：社会科学版, 2006(3):4.
[47] 陈国欣，吕占甲，何峰.财务报告舞弊识别的实证研究[J].审计研究，

2007(3)：88-93.
[48] 曾月明，宋新平，葛文雷财务报表舞弊可能性的智能识别方法[J]财经论坛，2008,(2):115-117
[49] 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(01):36-45.
[50] 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,2012,24(05):144-153.
[51] 钱苹，罗玫.中国上市公司财务造假预测模型[J]. 会计研究，2015：18-25,96.
[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] 吴杰,耿新青.虚假财务报表的 FCM 算法识别[J].辽宁科技大学学报,2016,39(04):307-310.
[54] 冯炳纯.基于数据挖掘技术的财务舞弊识别模型构建[J].财会通讯,2019(05):93-97.
[55] Cressey,D.R. Other People's Money-a Study in the Social Psychology of Embezzlement［M］. Glencoe, IL, Free Press. 1953.
[56] Albrecht,W.S.,Wernz,G.W.& Williams,T.L.(1995).Fraud:Bring the Light to the Dark Side of Business[J].New York Irwin Inc,22(3):15-52.
[57] Bologna G. J., R.. S. Lindquist, and J. T. Wells. The Accountant’s Handbook of Fraud and Commercial. New York; John Wiley ＆ Sons.1993.
[58] Wolfe D T, Hermanson D R. 2004. The Fraud Diamond: Considering the Four Elements of Fraud［J］. The CPA Journal, 74(12): 38-42.
[59] Kassem R, Higson A. 2012. The New Fraud Triangle Model[J]. Journal of Emerging Trends in Economics and Management Sciences (JETEMS), 3(3): 191-195.
[60] Tugas F C (2012). Exploring a New Element of Fraud: A Study on Selected Financial Accounting Fraud Cases in the World[J]. Am Int J Contemp Res.
[61] 娄权.财务报告舞弊的四因子假说[J].财会通讯, 2004(13):63-63.
[62] 张竹怡.基于舞弊三角形理论的财务报表舞弊识别的实证研究[J].纳税,2019,13(30):41-43.
[63] 黄世忠,叶钦华,徐珊,叶凡.2010～2019 年中国上市公司财务舞弊分析[J].财会月刊,2020(14):153-160.

哈尔滨工业大学学位论文原创性声明和使用权限

学位论文原创性声明

本人郑重声明：此处所提交的学位论文《基于机器学习的中国上市公司财务舞弊识别方法研究》，是本人在导师指导下，在哈尔滨工业大学攻读学位期间独立进行研究工作所取得的成果，且学位论文中除已标注引用文献的部分外不包含他人完成或已发表的研究成果。对本学位论文的研究工作做出重要贡献的个人和集体，均已在文中以明确方式注明。

作者签名：[signature]　　日期：2022年 6月 19日

学位论文使用权限

学位论文是研究生在哈尔滨工业大学攻读学位期间完成的成果，知识产权归属哈尔滨工业大学。学位论文的使用权限如下：

（1）学校可以采用影印、缩印或其他复制手段保存研究生上交的学位论文，并向国家图书馆报送学位论文；（2）学校可以将学位论文部分或全部内容编入有关数据库进行检索和提供相应阅览服务；（3）研究生毕业后发表与此学位论文研究成果相关的学术论文和其他成果时，应征得导师同意，且第一署名单位为哈尔滨工业大学。

保密论文在保密期内遵守有关保密规定，解密后适用于此使用权限规定。

本人知悉学位论文的使用权限，并将遵守有关规定。

作者签名：[signature]　　日期：2022年 6月 19日

导师签名：芦鹏宇　　日期：2022年 6月 19日

致　谢

时光飞逝，转眼间硕士生涯就要结束，回首过去两年，时间虽不长，但我却有很多成长和收获。尽管已经在哈工大学习、生活了六年，但硕士这两年比我预想的还要短暂和珍贵，在这两年中，我收获了珍贵的友情、爱情和师生情，在这里，请允许我向帮助过我的所有人表达最诚挚的感谢！

首先感谢我的指导教师芦鹏宇，没有她的帮助和指导，我将无法论文如此完善得完成。芦老师不仅时常督促我论文的进展，还耐心地指出我论文存在的问题，芦老师在工作上兢兢业业的态度无时无刻不在感染着我，让我在每次懈怠和懒惰时又重振精神，榜样就在身边，可以说芦老师就是我工作上的榜样。除了论文方面，芦老师在生活中也对给予了很多关心，经芦老师的帮助，我得以在硕士期间赴英国参加短期交流项目，这极大的开阔了我的视野，对我的人生观产生了积极影响。

其次感谢我的朋友们，他们对我的帮助和包容让我更加享受生活，在我孤单失意时，给予我支持和鼓励，在我得意高兴时，分享我的快乐。特别感谢我的室友，提供给了我安静的睡眠环境，这我保持了相对充沛的精力和能量。

感谢我的父母，他们不但给了我经济上的支持，使我能够不为金钱而烦恼，还给了我足够的关心和爱，让我随着年龄的增长，越发的明白什么是心灵的港湾，那是世界上最宝贵安全的避风港，置身其中，便可体会到内心的宁静与安然。

感谢我实习期间的同事们，他们让我学到了书本上难以学到的知识，这开阔了我的视野、加快了我的成长；感谢我的女朋友，陪伴了我成长。

最后感谢学校和学院的培养，感谢所有的老师的辛勤付出。学校学习的知识为我们打开了多个方向的多扇窗，未来我们将沿着窗外的风景，走得更远、看得更远！衷心的祝福母校越来越好，祝学院发展越来越好！