分类号 F222　　密级
UDC 311　　编号

中南财经政法大学

硕士专业学位论文

上市公司财务舞弊的识别研究

研究生姓名：刘　小　霞
校内导师姓名、职称：李　占　风　教　授
校外导师姓名：熊　东　平
专业学位类别：应用统计硕士专业学位
专业名称：应　用　统　计
研究方向：金融统计方向
入学时间：二〇二〇年九月

二〇二二年五月十五日

Research on the Identification of Financial Fraud of Listed Companies

Xiaoxia Liu

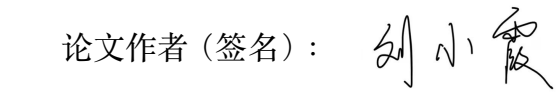
2022.05.15

中南财经政法大学学位论文独创性声明和使用授权声明

学位论文独创性声明

本人所呈交的学位论文，是在导师的指导下，独立进行研究所取得的成果。除文中已经注明引用的内容外，本论文不含任何其他个人或集体已经发表或撰写的作品对本文的研究做出重要贡献的个人和集体，均已在文中标明。

本声明的法律后果由本人承担。

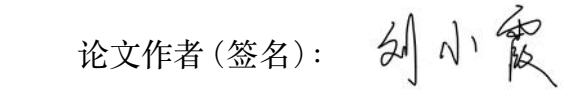
论文作者（签名）： 

2022 年 5 月 15 日

学位论文使用授权书

本论文作者完全了解学校关于保存、使用学位论文的管理办法及规定，即学校有权保留并向国家有关部门或机构送交论文的复印件和电子版，允许论文被查阅和借阅。本人授权中南财经政法大学将本学位论文的全部或部分内容编入有关数据库，也可以采用影印、缩印或扫描等复制手段保存或汇编本学位论文。

注：保密学位论文，在解密后适用于本授权书。

论文作者（签名）： 

2022 年 5 月 15 日

摘要

目前，我国经济正处于快速发展阶段，证券市场欣欣向荣，上市公司层出不穷，但相关审查机制和监管法规相对不够完善，导致上市公司财务舞弊行为频发，虽然目前采取的监管措施在一定程度上减少了财务舞弊的发生，但财务舞弊依然是近几年监管部门的心头大患。企业为了谋取自身利益从而通过某些不正当的手段进行财务舞弊不仅影响上市公司自身的发展，让公司时刻面临着停市的风险，还影响公司内外各利益相关者的权益，阻碍了我国资本市场经济的稳步发展。所以如何有效过滤和抵制上市公司财务舞弊对现在的中国证券市场来说具有重要意义。

本文将数据挖掘技术应用于上市公司财务舞弊识别的研究中，在对现有文献进行梳理之后，从 CSMAR 数据库中选取 2011-2020 年因财务舞弊受到惩罚和披露的上市公司的违规信息，并且为了保证样本质量避免重复样本，最终确定了两千多个舞弊样本，根据配比原则选择同年、同规模和同行业的四千多个非舞弊企业的数据作为对照；在变量选取上，本文以风险因子理论为基础，从财务指标和非财务指标层次对公司整体情况进行刻画，初步选出了 42 个初始指标，然后分别用树算法、RFE 搜索算法、lasso 回归和随机森林进行特征选择；将四种特征选择算法选择的指标与未经指标筛选的原始指标相结合，形成的数据集分别在逻辑回归、支持向量机、随机森林、K 近邻算法和神经网络这五种单一识别模型下进行检测。为了得到更好的模型识别效果，我们基于之前的单一识别模型分别用平均法、投票法和集成学习的方法构建了三种不同的综合识别模型，最后再对各模型的识别效果进行比较分析。

最终得到本文的结论如下：第一，舞弊类型多样化并且行业分布存在明显差异。舞弊类型最多的前三项分别是违规买卖股票，推迟披露和重大遗漏，归根结底是公司为了获取某种不正当的利益；从行业类别上看制造行业舞弊数量最多，所以审计人员应对制造业的上市公司进行重点监督。第二，在财务舞弊识别指标的变量选取方面，舞弊识别的关键变量主要集中在财务指标当中，说明这些指标对财务舞弊的识别更加显著。第三，在单一数据挖掘模型的财务舞弊识别效果方面，随机森林在不同的条件下都具有较好的识别效果，不管是针对初始指标还是经过特征选择后的指标，其准确率均处于较高水平。第四，在财务舞弊识别模型方面，通过混淆矩阵的相关评估指标和 ROC 综合曲线比较了各模型的识别效果，最终发现综合识别模型的识别效果普遍高于基础的单一模型，其中效果最好的是基于 stacking 集成学习法下的综合识别模型，说明 Stacking 集成算法集成了各个分类算法的特点达到了博采众长的目的，对于识别上市公司是否财务舞弊的问题上更加高效和可靠。

关键词：舞弊识别；特征选择；数据挖掘算法；Stacking 集成学习

Abstract

At present, China's economy is in the stage of rapid development, the securities market is booming, and listed companies emerge one after another. However, the relevant review mechanism and regulatory regulations are relatively imperfect, resulting in frequent financial fraud of listed companies. Although the current regulatory measures have reduced the occurrence of financial fraud to a certain extent, financial fraud is still a major problem for the regulatory authorities in recent years. In order to seek their own interests, enterprises carry out financial fraud through some improper means, which not only affects the development of listed companies and makes the company face the risk of market suspension at all times, but also affects the rights and interests of stakeholders inside and outside the company and hinders the steady development of China's capital market economy. Therefore, how to effectively filter and resist financial fraud of listed companies is of great significance to China's securities market.

This paper applies data mining technology to the research of financial fraud identification of listed companies. After combing the existing literature, this paper selects the violation information of listed companies punished and disclosed for financial fraud from 2011 to 2020 from CSMAR database. In order to ensure the sample quality and avoid duplicate samples, more than 2000 fraud samples are finally determined, and the same year and The data of more than 4000 non fraudulent enterprises of the same scale and industry are used as a comparison; In terms of variable selection, based on the risk factor theory, this paper describes the overall situation of the company from the level of financial indicators and non-financial indicators, preliminarily selects 42 initial indicators, and then selects the characteristics with tree algorithm, RFE search algorithm, Lasso regression and random forest respectively; Combining the indexes selected by the four feature selection algorithms with the original indexes without index screening, the data sets are tested under five single recognition models: logistic regression, support vector machine, random forest, k-nearest neighbor algorithm and neural network. In order to get better model recognition effect, based on the previous single recognition model, we constructed three different comprehensive recognition models with average method, voting method and integrated learning method respectively. Finally, we compared and analyzed the recognition effect of each model.

Finally, the conclusions of this paper are as follows: first, the types of fraud are diverse and there are obvious differences in industry distribution. The top three types of fraud are illegal trading of stocks, delayed disclosure and major omissions. In the final analysis, the company is in order to obtain some improper interests; From the perspective of industry category, the manufacturing industry has the largest number of fraud, so auditors should focus on the supervision of Listed Companies in the manufacturing industry. Second, in the variable selection of financial fraud identification indicators, the

key variables of fraud identification are mainly concentrated in financial indicators, indicating that these indicators are more significant in the identification of financial fraud. Thirdly, in terms of the financial fraud identification effect of a single data mining model, the random forest has a good identification effect under different conditions. Whether for the initial indicators or the indicators after feature selection, its accuracy is at a high level. Fourthly, in the aspect of financial fraud identification model, the identification effect of each model is compared through the relevant evaluation indicators of the confusion matrix and the ROC comprehensive curve. Finally, it is found that the identification effect of the comprehensive identification model is generally higher than that of the basic single model, and the comprehensive identification model based on the stacking integrated learning method is the best. It shows that the stacking integrated algorithm integrates the characteristics of each classification algorithm to achieve the purpose of drawing on the strengths of others, It is more efficient and reliable for identifying whether listed companies are financial fraud.

Key Words: Fraud identification; Data mining algorithm; Machine learning algorithm; Stacking integrated learning

目录

绪论

一、选题背景与研究意义

（一）选题背景

随着中国经济的飞速发展,越来越多的公司争相上市，截至 2020 年底其数量已达 4224 家，然而随着数量的增加上市公司财务舞弊的现象也在不断增加。ACFE[1]发布的统计数据显示，截至 2020 年，因财务舞弊给上市公司自身造成的金额损失高达 70 亿美元，其数量较 2016 年相比增加了 14.7%。发生财务舞弊的主要原因是公司内部人员失信、国家监管体系不完善以及外部审计机构失责等。近年来，瑞幸咖啡、索菱股份、东方金钰等重大财务舞弊事件在社会上引起了轩然大波。经证监会的审查发现，瑞幸咖啡从第二季度到第四季度的虚报总额为 22 亿元，占总金额的 42%，瑞幸股价暴跌 85%，面临 6 次停牌，市值损失约 350 亿元；康美药业财务舞弊金额高达 300 亿，造假规模之大、手段之恶劣，导致投资者损失总金额达 24.59 亿元。因此证监会不仅对康美药业采用了停市的处罚，并且连带董事长在内的 6 人被禁止进入证券市场[2]。由此可知，企业财务舞弊不仅影响公司自身的发展，让公司时刻面临着停市的风险，还阻碍了证券市场的稳定发展。为了防范财务舞弊行为，国家证券监督管理委员会在新实施的证券法中强调，对涉及财务舞弊的公司和相关负责人要加大处罚力度和加强惩戒措施。

目前，各研究学者对财务舞弊的研究主要停留在理论阶段，比如舞弊的手段、成因、行为等方面。这种方式对财务舞弊的识别主要是通过探究财务报表中各指标间的相关关系是否异常。显然，这种方法成效甚微，并且需要丰富的从业经验和过硬的知识储备。随着企业舞弊方式和手段的多样化以及企业的财务报表的难以获得，这种人工识别方法已逐渐不能满足需要。于是识别财务舞弊的统计模型应运而生，这种方法相较于财务报表确实有了很大的改进，但依旧不够高效和简便。为了进一步识别财务舞弊，各研究学者试图用各种方式对财务舞弊进行识别，再后来人们便逐渐发现了数据挖掘技术，从传统的数据挖掘模式到高度智能的数据挖掘模式，都显示出了优秀的技术特性。在对财务舞弊的识别上运用数据挖掘技术具有很多优点，它不需要丰富的审计经验，也不要求传统统计模型中复杂的条件假定，并且由于计算机系统的强大，他可以快速高效的处理分析海量数据。因此，本文在国内外学者已有研究的基础上，基于数据挖掘技术，通过对各模型实证研究的对比和分析，以期寻找一个最佳的舞弊识别模型，以此来帮助投资者做出正确的决策，为公司内部和外部的审计人员提供识别财务舞弊的一种方式，维持我国证券市场的健康稳定发展。

[1]ACFE 指的是注册舞弊审查师协会，是全球公认的反舞弊培训调查机构，它是一家专门从事反欺诈风险管理的全球权威组织。

[2]本数据来源于证监会关于财务舞弊调查处置工作情况的通报。

（二）研究意义

在上市公司舞弊手段逐渐多样化的背景下，寻求一种最优的财务舞弊识别方法对我国证券市场有着重大且深远的意义，接下来将分别从理论上和现实上进行阐述：

1.理论意义

本文主要通过研究公司的财务指标和非财务指标，并结合挖掘方式构建的模型识别财务舞弊现象。从理论上，由于非财务指标不能简单的被篡改或者隐藏，因此可以比较客观真实的反映公司的实际运营与管理状况，从而可以比较合理地发现公司内部日益隐蔽的财务舞弊现象。但是公司利用财务舞弊手段所构成的虚假数据很可能会在统计构造上具有内在缺陷，而这些内在缺陷通常无法借助肉眼加以辨认，不过通过数据挖掘和统计分析等方式就可以找到这些显著性差异。而近年来，随着数据挖掘技术的越来越完善，在企业经营管理等领域中应用广泛。数据挖掘方法可以利用海量的财务报表数据和非财务数据中的信息，采用自我学习的方式，发现财务数据信息之间可能的关联，为财务人员舞弊识别提供了技术基础，从而有效防止虚假的会计信息，提高会计信息的质量。

2.现实意义

本文的研究有利于维护利益相关者的决策，公司的利益相关者不但包含了投资方和债务人，而且还有供货商和消费者等。这些利益相关者相对于上市公司本身而言，却常常得不到准确和及时的消息，而同时又面临着欺瞒公众的心理动机，从而无法确定自身权益在公司中的稳定性和收益性。所以寻找一种有效的识别财务舞弊的工具对这种内外利益者进行合理决策具有重要而深远的意义。它有助于审计人员发现财务舞弊的苗头，也有助于证券监管部门的监督管理工作。中国证券市场还处在进一步探索与改革的过程中，财务舞弊可能会误导政策法规与制度，歪曲市场资源分配，从而减低市场资源配置效果。面对日益复杂的监管环境和舞弊类型，怎样建立健全管理制度的同时革新监管方式，从而形成一个合理的财务舞弊识别方法，对于监管部门来说也有着重要的意义。在数字化时代的背景下，数据挖掘技术发展是一个集机器学习[3]、人工智能和数据科学等众多应用领域内容为一体的强大技术手段，非常适合用来识别上市公司财务舞弊的问题。对财务舞弊运用数据挖掘技术进行识别具有很多强大的功能，它不需要具备非常充分的审计经验，并且由于计算机系统的强大，他可以快速高效的处理分析庞大的数据。除此之外，本文还基于舞弊风险因子理论选取了财务指标和非财务指标，以上市公司的违规数据为主要调查样本，来建立财务舞弊识别模型，有利于帮助投资者进行决策并且给监管部门和公司内外审计提供一定的技术手段。

[3]机器学习是一门强大的信息交叉学科，它以概率论和统计学的知识作为其理论基础，运用计算机来模拟人的学习行为，被广泛应用于分类、识别和回归等领域。

二、国内外文献综述

由上文分析我们可以知道，财务舞弊的理论研究已经相当丰富，而在财务舞弊识别模型方面相对较少，目前已有的文献大致上可分成二类，一类是基于基础统计模型构建的，如 Probit 回归[4]；另一类是建立在数据挖掘的基础之上的，以随机森林、支持向量机等居多，并且数据挖掘技术还在不断发展。本文将根据已有的财务舞弊识别技术，分为国外财务舞弊模型的研究和国内财务舞弊模型的研究加以整理，为本文中财务舞弊识别模型奠定理论基础。

(一)国外关于数据挖掘技术下财务舞弊模型的研究

Calderon 和 Green(1997)从审计的角度，通过公司的财务报表数据，探索其关联性和矛盾点，以此来探究财务舞弊的动机和可能。Fanning 与 Cogger(1998)发现随机森林可以用于财务舞弊识别的研究，因此把随机森林引入了财务舞弊识别模型，并根据重要性进行了排序，最终选择了现金资产比率、权益乘数和资本累计率这几个指标为最重要的三个识别指标[5]。自此之后人们便发现了数据挖掘模型对于财务舞弊识别上的适用性和有效性，并开始争相效仿。Gottlieb、Shek 和 Salisbury 等(2006)选取了 200 家公司为舞弊样本并配比了 200 家非舞弊公司，采用逻辑回归的方法对样本进行检验，结果表明基于公司发展水平的指标具有显著性差异，实验得到的准确率最高[6]。Koumanakos、Kotsiantis 与 Tzelepis 等(2006）以运输业三百二十家企业为对象，把逻辑回归、K 近邻算法、决策树和神经网络等四个技术手段纳入了研究范围，结果表明这四种方法均可以对财务舞弊进行识别并且神经网络的效果表现最好[7]。Kirkos、Manolopoulos 和 Spathis(2007)总结了前人的经验，在比较财务舞弊的识别效果时不仅仅只看它的识别准确率，还首次把混淆矩阵的概念引入了财务舞弊识别结果的比较范畴，有利于对各模型的效果进行综合判别分析。Pediredla、Rao 和 Ravi 等(2011)采用了六种数据挖掘技术，分别为朴素贝叶斯、决策树、神经网络、k 近邻算法、逻辑回归和随机森林。结果表明，神经网络相较于其他五种方法财务舞弊的识别效果最好。Chiu、Lin 和 Huang 等(2015)采用 BP 神经网络技术对财务舞弊进行了识别研究，通过不断调节神经网络的层数，神经元的个数以及其复杂程度，最终取得了良好的研究效果，为该领域的识别做出了巨大贡献。

(二)国内基于数据挖掘技术下财务舞弊识别模式的研究

相比于国外对财务舞弊识别模型的研究，国内相对较少，但其对于中国财务舞弊的识别具有重大而深远的意义。秦江萍(2005)选择了近五年的重大财务舞弊事件为样本，再选取了二十七个关联指数，并通过支持向量机创建了财务舞弊识别模型。陈俊、王明(2005)选取了 2004 年年末上市公司财务舞弊数据为样本，通过与正常样本的各

[4]Probit 模型是一种服从正态分布的线性模型，常用于离散数据的分类问题，其模型结构与 Logit 模型类似。

[5]FANNING K M,COGGER K O. Neural network detection of management fraud using published financial data[J]. International Journal of Intelligent Systems In Accounting,Finance and Management. 1998,7(5):21-24.

[6]GOTTLIEB O,SALISBURY C,SHEK H,et al. Detecting Corporate Fraud:An Application of Machine Learning[C]. CS 229 Machine Learning Final Projects,Autumn 2006.

[7]KOTSIANTIS S,KOUMANAKOS E,TZELEPIS D,et al. Forecasting fraudulent financial statements using data mining[J]. International Journal of Computational Intelligence,2006,11(2): 104-110.

项指标的对比，选择了差异性最大的 8 个指标代入支持向量机模型对财务舞弊进行识别。夏明、李海林和吴立源(2015)通过对现有文献的研究，发现虽然目前已经有了很多识别财务舞弊的方法，但其识别准确率都不够高，所以他们选择对舞弊模型进行了改良，通过不断的实验改变了某些参数，结果表明，经过改良后的神经网络模型的识别效果最好。邹译萱(2018)以不同行业的财务指标为样本，分别构建了 BP 神经网络和随机森林模型，发现对于不同的行业模型的效果略有不同，但其整体识别效果均表现出良好的特性。曹德芳和刘柏池(2019)根据各企业的年度报告，选取了 24 个能综合反映公司整体经营状况的财务指标代入向量机模型并取得了较好的实验效果。崔东颖和胡明霞(2019)把近十年的数据以年为单位分成了十个样本并代入随机森林和决策树模型，结果表明，对于不同的数据随机森林的识别效果都优于决策树。

(三)文献述评

通过以上国内外学者的研究，我们发现财务舞弊的识别已经经历了充足而漫长的研究过程，而且是广泛而深入的。并且国外对于财务舞弊的研究早于国内学者，无论是在理论上还是实践上都相对成熟，所以我国很多研究学者都是在国外对财务舞弊识别的研究基础上进行的。数据挖掘技术在证券市场上的应用已经非常广泛，比如能够更好地对企业财务报表中的各项指标数据进行分类，对于企业改善财务管理舞弊识别的效率有着积极而深远影响。在相关研究中，构建财务舞弊模型使用较多的数据挖掘技术有逻辑回归、支持向量机以及 SVM[8]等。因此，通过对多个数据挖掘方法的对比分析，采用了五种最适合本文数据的数据挖掘技术分别建立财务舞弊识别模型，并根据三种不同的结合方式建立了综合识别模型。在选择纳入模型的指标时，中国国内研究者的方式也是比较多样的，但其大都集中在财务指标上，本文为了使研究更加全面，结果更加准确，在财务指标的基础上加入了非财务指标，使模型的结果更具整体性。

三、研究内容与方法

（一）研究目标

本文的研究目标是在国内外文献研究的基础上，总结财务舞弊行为发生的诱因、特征及现状分析，并据此寻找最适合目前证券市场上市公司财务舞弊的识别模型，为上市公司自身、内外审计及投资者识别财务舞弊提供一定的帮助。在分析财务舞弊的诱因时，借鉴了目前运用最广泛的四种诱因理论，具体详见下文介绍。然后根据最成熟的舞弊风险因子理论对财务舞弊的行为、特征及现状进行分析，最后根据各种数据挖掘技术的特点以及财务舞弊的特征，选取最能区分财务舞弊和非财务舞弊的相关指标和能识别该指标的数据挖掘模型，为各利益相关者提供若干建议。

（二）研究内容

本文在明确研究目标和研究意义的基础上，根据财务舞弊和数据挖掘的相关理论，总结财务舞弊的特点，以寻找最适合当前证券市场中的财务舞弊识别模型。研究的主

[8]SVM(Support Vector Machine)是一种以监督学习的方式对数据进行二分类的线性分类器，详见后文理论介绍。

要结构如下：

首先为绪论。这部分包括本篇论文的目的、背景和意义，然后在这些内容的支撑下，阐述了本文的研究方法和研究的创新点。

第一章为理论分析。首先介绍了财务舞弊的概念及相关界定、财务舞弊的各种诱因理论；其次，通过财务舞弊的特征及现状引出用于识别它的数据挖掘理论，并介绍了数据挖掘的概念、常用的分类方法和实施过程。

第二章为舞弊识别的指标体系及特征提取。首先基于国泰安数据库选取了部分舞弊样本，并根据配比原则选取其余的非舞弊样本，其次从财务指标和非财务指标的角度选取了用于构建模型的 42 个变量，并对变量进行缺失值和标准化处理。然后把这些指标分别通过曼惠特尼显著性检验和皮尔森相关性分析后再进行过滤式、包裹式和嵌入式特征选择，选出最适合代入模型的变量。

第三章为财务舞弊识别方法的选择及实证分析。首先分别将原始指标和经过特征选择后的指标代入五种单一数据挖掘模型，比较各个模型的识别效果，选出识别效果最好的指标和模型，然后基于集成学习的方法构建综合识别模型，如基于 stacking 集成学习法、投票法和平均法的财务舞弊综合识别模型，依据基于混淆矩阵的各项指标和 ROC 曲线对模型识别的效果进行比较，求出模型识别效果最优的模型。

最后是研究结论与展望。此章节通过以下几个部分来展开，一是研究结论，对本文要达到的目的以及全文的实证部分进行总结，二是研究建议，通过前面的结论，有针对性的对证券市场中扮演的各个角色提出建议，三是研究展望，针对本文的不足之处提出的展望。

（三）研究方法

本文在财务舞弊指标的建立方面通过文献研究法，对财务舞弊识别模型的各种优缺点和实践结果进行总结，从而确定本文的研究方向和研究思路。通过定性分析方法与定量分析方法相结合选择舞弊样本和非舞弊样本，为后面舞弊识别模型的建立做准备。具体内容介绍如下：

1.文献研究法。在财务舞弊指标的建立方面通过文献研究法，从不同视角全方位总结现有研究成果。区分国内外学术界的研究成果并加以分析与总结，为本文的指标体系建立提供借鉴。

2.理论知识和实验研究相结合。在财务舞弊的识别模型建立部分，首先对支持向量机、随机森林、逻辑回归、K 近邻算法以及神经网络等数据挖掘方法从理论上加以介绍，随后开展了大量实证的模型构建和验证。

3.定性分析方法与定量分析方法结合。本文首先定性分析企业中最容易直接影响财务舞弊的各种因素，并总结和建立了能够有效识别财务舞弊的评估指标体系，然后再结合定量分析选取主要财务指标和非财务指标，从而最终形成了综合财务舞弊识别指标，在此基础上进行舞弊模型的构建。

4.对比分析法。在特征选择的基础上评价与比较各模型的有效性，并将结果进行对比分析，进一步阐述模型改进的必要性，以增强模型识别的准确率。

四、本文特色与创新点

本文的特色与创新点主要体现在指标选取和识别方法两个方面，具体包括：

1.在初始指标的选取上，选取的指标更加全面。本文在总结国内外文献综述的基础上根据数据的实际情况综合选择了反映公司偿债能力、盈利能力、发展能力、经营能力、比率结构、现金流分析、风险水平的财务指标和反映公司治理结构和股权结构的非财务指标。这些指标相较于已有的大多数研究引入了非财务指标，能够比较全面的反映公司的发展水平、经营情况和公司内部的结构，使识别结果更加准确可靠。

2.在财务舞弊识别的方法上进行创新，在构建了上市公司财务舞弊的单一识别模型之后，根据三种结合策略构建了综合识别模型，运用更加复杂的模型结构对公司财务报表的相关指标进行更加全面的识别和应用。最后分别对各舞弊识别模型的识别性能进行比较，选出最适合目前证券市场的财务舞弊识别模型。

第一章　舞弊相关概念及舞弊识别理论

第一节　财务舞弊的概念

一、财务舞弊的概念及相关界定

财务舞弊一般是指财务报告的舞弊，是指企业有预谋地对财务报告中所列示的数值，或者说财务报表附注部分内容予以误报和忽视，从而达到欺诈财务报告的用户的目的。财务舞弊的行为主要包括：操纵财务报告据以编造财务记录或凭证文件；错误提供或故意忽视有关财务报告的交易、事件或其他关键信息；有意错用与数量、类别、信息提供方法以及揭示方式相关的审计准则。财务舞弊的行为损害了市场参与者的信心，降低了企业的内部控制结构，同时更是严重影响了整个证券市场的稳定性。

从上述对财务舞弊概念的认识我们可以得到财务舞弊主要的界定，可以从以下几个方面来概括：首先财务舞弊的对象是目前证券市场中的上市企业，然后上市公司进行舞弊的行为是一种故意的行为，其舞弊的表现形式主要集中在上市公司的月报、季报和年报，最后舞弊行为其实是一种触犯规章制度的违法行为。根据上述对财务舞弊理论的总结，我们可以这样定义财务舞弊：上市公司为了实现自己的利益，采用各种不同的方式和手段，发布了具有错误信息的财务报表，从而误导公司内外的利益相关者，从而实现自身利益的违法行为。

二、财务舞弊的诱因理论

1.冰山理论

罗伯特与杰克波罗格纳共同提出了舞弊冰山理论，也就是我们通常所说的二因素论。也就是说冰山理论把形成财务舞弊的诱因分成两个部分，分别是行为上的原因和结构上的原因。该理论认为如果把财务舞弊看作一个冰山，构成它的因素即是冰山暴露海面的一角，这部分构成是普遍存在并易于被人类直接观测到的；而舞弊行为上的原因即是由埋藏于海面之下的冰山的部分所构造，这一类内容通常带有一定的主观色彩，虽然它也是客观存在，但是经常被人们忽视。舞弊冰山理论所强调的是一家企业是否具有舞弊行为，除要重视结构因素之外，还应重视企业实际操纵者的行为方面的因素，并剖析企业操纵者的人性方面的舞弊危险性。

2.舞弊三角理论

上世纪时代的劳伦斯索耶通过对目前文献的总结，发现财务舞弊形成的原因包括以下三个方面：异常行为、机会以及合乎情理。劳伦斯在一定程度上说明了财务舞弊的形成，主要来自于欲望，也就是说人一旦有了欲望就会采取各种方式来达成自己的欲望，不管这种方法是否是违法犯罪的。美国心理学家史蒂夫艾伯伦奇特在劳伦斯的理论基础上更进一步发展了舞弊理论，并对异常行为进一步分析转化解释为压力，使合乎情理的更加书面化为借口，并重新提出了舞弊形成的三个原因：压力，借口和机

遇。其中机遇是公司为了实现自身的发展，从而创造一些可以吸引投资的机会，借口原因的形成主要来自于前面所说的欲望的形成，而压力显而易见来自于公司内外的各种压力，比如同行业对手之间的竞争压力，公司上市所面临的压力等等。

3.GONE 理论

GONE 理论由舞弊三角理论发展而来，该理论一经发表就引起了各大学者的广泛共鸣。GONE 理论也称四因素理论，也就是说舞弊的形成包括四个因素：机遇、贪婪、需要和显露。其中机遇和显露主要是指外部因素形成的财务舞弊，机遇和前面一样是为了吸引投资，模糊各大利益相关者的决策，而显露是指公司要定期披露各大财务报表，公司为了掩饰自身的实际财务状况，从而进行舞弊；贪婪和需要则是指由于公司自身的原因形成的财务舞弊，是指公司为了满足自身的欲望，取得不切实际的发展而采取一些不正当的手段进行财务舞弊。

4.舞弊风险因子理论

随着舞弊的形式逐渐多样化，布洛格等人发现 GONE 理论已经不能满足目前财务舞弊的市场需求了，所以他们在 GONE 理论的基础上发展了舞弊风险因子理论。该理论认为可以将舞弊形成的原因分成一般风险因子和个别风险因子。其中个别风险因子是指全部由公司自身所形成的原因，但并非企业可以掌控的所有各种因素，还包括个人的素质等；一般性风险因子通常泛指能够被企业和机构所触达的各种因素，包括外部环境因素和公司所面临的各种压力等。该理论认为当一个公司同时存在这两种风险因子时，舞弊就会发生。

三、财务舞弊的特征及现状分析

目前，越来越多的公司进入中国证券市场，被处罚的上市公司也越来越多。近年来，瑞幸咖啡、索菱股份、东方金钰等重大财务舞弊事件在社会上引起了掀然大波。企业财务舞弊不仅影响公司自身的发展，让公司时刻面临着停市的风险，还影响公司内外各利益相关者的权益，严重影响了社会经济，阻碍了证券市场的稳定发展。有的公司为了谋取相关利益从而通过某些不正当的手段进行财务舞弊，所以如何有效过滤和抵制上市公司财务舞弊对现在的中国证券市场来说具有重要意义。发生财务舞弊的主要原因为公司内部人员失信、国家监管体系不完善以及外部审计机构失责等。所以提高审计质量、减少舞弊行为是目前证券市场主要的目标。为维持证券市场的正常秩序，国家对上市公司的行为制定了严格的规范措施。但仍然有部分上市公司为了保护自己的利益，欺瞒群众，提交虚构过的财务报告。如今违反上市公司财务制度的方式逐渐多样化，形式也日益隐蔽。本文认为，进行财务舞弊是上市公司为了达到某种目的，满足投资者的期望和自身利益而违反财务制度，具体特征如下：高估资产，低估负债，操作关联性交易，提前确认收入、成本和支出重估等。

目前，各研究学者对财务舞弊的研究主要停留在理论阶段，对于舞弊的手段、行为等方面的分析已相当透彻。在识别财务舞弊方面，虽然起源很早，但早期的研究还

不成熟。目前舞弊的识别总的来说可以包括二种，一种是主要依靠经验的方式，即分析性复核方式，其目的主要是通过知识和经验来发现虚拟的财务报表与实际会计报表之间的差异来鉴别，但是通常这个方式对鉴别者的要求比较高，并且必须掌握较为丰富的会计专业知识，另一类则是通过建立的识别模型来做出判别，目前应用最多的是基础统计建模和数据挖掘模型。通过对前面文献的总结发现，采用数据挖掘模型对财务舞弊进行识别更具普遍性，并且具有许多优点，可以快速高效的处理分析海量数据。

第二节　财务舞弊识别的相关理论

近年来，随着数据仓库的发展，这种集成了数据存储、数据清理并实现联机分析处理的技术的功能已经足够强大，然而对于海量数据的分类和聚类分析仍需要其他数据分析工具。于是数据挖掘应运而生，它的诞生引起了各学术界不小的轰动，因为它不仅可以处理海量的数据而且比之前的方法更加简洁高效。随着数据挖掘的发展，如今已广泛应用于医学、教育、金融等各大领域。本文用于财务舞弊识别的相关理论和方法就是基于数据挖掘的方法和理论的基础上对财务舞弊进行识别，所以下文在数据挖掘的基础上进行展开。

一、数据挖掘的概念

数据挖掘是 1995 年各研究学者在一场技术交流年会上提出的概念。数据挖掘信息技术开展以来，不仅基础理论研究工作十分活跃，同时多种信息挖掘产品与应用体系也相继诞生，而且取得了显著性成效，因此受到了电子信息产业界的高度重视。数据挖掘技术由于其兼容性较强，功能强大所以在不同领域中都具有广泛应用。在金融行业中，数据挖掘可以通过对不同客户消费产生的数据进行分类，分析不同客户的消费习惯，并由此推荐相应的产品；在医学中，可以利用大数据分析技术分析检查结果中的各项指标，以此为基准来判断患者的病情种类和严重程度并推荐相关治疗手段；在证券市场中，内外审计和投资者也可以利用数据挖掘技术对相关财务数据进行识别，以此来判断本公司发生财务舞弊的风险和可能性。由于不同专家和研究学者之间对于大数据分析有着不同的认识，并且数据挖掘的应用领域也不同，大家对数据挖掘概念的辨析也不同。但是，中国国内外许多学者都提出，所谓数据挖掘就是从大量的数据中发掘潜在的有用的信息的过程。

二、数据挖掘的过程

数据挖掘技术发展到今天已经非常成熟，所以其挖掘过程也形成了一套非常完备的体系。通常，数据挖掘的过程分为以下五个流程：

1.设立研究的目标。在进行数据挖掘之前我们首先要明确的就是研究的目标。有了目标才有前进的方向，才能有目的性地进行研究，才能明白我们做数据挖掘研究想要达到的效果是什么。所以设立研究目标是整个数据挖掘过程的研究之本。

2.数据准备。数据准备也就是数据的预处理，首先是对数据进行采集，根据研究的目标选择要下载的数据；然后就是对数据进行清洗，以获得符合本文实验要求的数据；对数据进行缺失值处理，对一些缺失信息较多的数据直接进行删除，其他的可以采用均值、众数等方法进行填充；最后是对数据进行标准化处理，以消除因计量单位不同对结果造成的影响。

3.数据挖掘。这个部分是整个数据挖掘的关键所在，它是根据之前下载并处理的数据选择合适的算法模型，然后把清洗过的数据代入到数据挖掘算法的模型当中，通过不断的训练和学习，最终找到最适合该数据的挖掘模型。

4.模型效果的评估。找到了合适的数据挖掘的模型之后我们要根据本文研究的目的寻找一套合适的评估标准，进而才能准确的对模型的效果进行评价。

5.模型推广。在整个过程最后我们要将模型进行推广，不仅要将数据挖掘得到的结论运用到相关决策当中，还要判断该模型是否能运用到更加广泛的领域当中。

三、数据挖掘的方法

数据挖掘技术虽起源较晚，但其发展十分迅速，目前已经形成了非常多成熟的数据挖掘方法。最常用的数据挖掘算法主要有：决策树、随机森林、支持向量机、K 近邻算法、朴素贝叶斯、回归分析、神经网络等。下面将对这些方法分别进行介绍。

（一）决策树

决策树(Decision Tree，DT)在数据挖掘领域起源非常早，其本质是利用一系列规则对数据进行划分的过程。决策树通过不断地向下选择、不断地分支对样本进行分类，决策树包括三个组成部分：状态节点、叶节点和决策结点。决定树中最顶端的结点叫做根结点，每分枝都是一种新的决策节点，每个节点分支都代表了一个种选择或者一种类别，在各个结点上都会出现一类试验结果，对各个结点上测试的结论形成了各种各样的分枝，最终得出分类结果最多的叶子，这便是使用决策树完成分类的流程。

优点：分类的结果通俗易懂，便于后续对结果进行分析；该方法运算速度快，运算量也相对较小，对多种类型的数据都适用；决策树的产出也包括了属性的排序方式，对各个样本进行分类，并统计各个特征所占的比重，在解决分类问题时把特征所占的比重纳入考虑范围。

缺点：不适用于分类类型较多的情况，因为决策树技术中采用了"贪心"算法[9]，总是选择最好的那个结果，导致分类结果非常片面；该方法要是用于连续型数据的分类也非常困难。

（二）随机森林

随机森林(Random Forest，RF)是最常用的数据挖掘算法之一，因为其在数据挖掘领域具有普遍适用性，所以被广泛应用于各大研究领域。随机森林就是利用集成思维，将多颗决策树进行综合运算。但事实上从直觉思维视角来理解，每个决策树都是一种

[9]是指在对问题求解时总是选择最优解，而没有从整体上对问题进行考虑，从而可能会导致结果不准确。

分类器(假定现在面临的是一类问题)，所以对同一种输入的样本，N 颗树就会有 N 个解析结论。而随机森林则汇集了全部种类与选择结果，把选择次数最高的种类确定为最后的输出结果，这也是一个最朴素的 Bagging 思想[10]。个别树选择最多的那个结果则为分类的最终结果。在回归问题中，回归结果就是通过决策树的加权平均得到的。

优点：它可以计算很高层次的数值，而且不必进行特征选取(因为特征子集是随意选取的)；在练习完成之后，就可以得出这些特征哪些较为关键；在构建随机森林的时期，由于采取的方法是无偏估计方法，所以建模泛化能力较强；练习速率快，也易于做成并行化方式(练习时树与树中间是互相独立性的)。

缺点：在处理回归的问题时，效果并没有它在处理分类问题中呈现出的效果那么好；对许多大数据构建者而言，随机森林给人的感受就像是个黑盒子，内部没法控制。

（三）K 近邻算法

K 近邻算法(K Nearest Neighbors，KNN)，利用该方法处理分类问题时，先存储一个由已知类型的训练样本所构成的实例集，在判断某个新实例的类型时，与其最接近的 k 个例子中多数属于什么类型，新例子就将被划分到哪一种类型。因此，需要给出一些计算方法以反映实例间的相似程度。

优点：快速训练并且很容易完成，因为实际上没有复杂的操作过程；当数据的特征数量比较多时也同样适用。

缺点：把所有的样本信息均储存到一起，那么分类器会变得非常笨重并且反应缓慢，最理想的解决方法是保存搜集到所有信息的原型样品，或采用最高效的搜索技术；但若具有许多的不相关属性，则会直接影响分类结果的准确性。

（四）支持向量机

支持向量机(Support Vector Machine，SVM)在解决不确定性问题和高维度模式的辨识流程中具有诸多的独特优点。近年来，SVM 方式在图像辨识、数据信息加工以及基因组图像辨识等领域技术方面都取得了成功的应用。而支持向量机主要是采用分类器进行分类的，通过提高分类器的识别效率来提高分类的准确率。它的基本思路是：由训练样本寻找分类的最优超平面，同时将训练集中的焦点离分析平面尽量远离，即搜索每一个类型平面使其两边的空白范围(margin)都最大；而在非线性情形下，则先寻找一个具有合适超平面的高维空间，这样就可以把非线性问题转化为线性问题。

优点：模块构造简便，参数变化较小；支持向量是 SVM 方法的主要技术训练成果，可以决定最终的分类结果；当数据维度比较高时也同样适用，因为有多个核函数可使用，能够处理各种形式的数据问题。

缺点：对于大量训练样本无法实现，当数据量较多时就很难找到可以划分不同类别的超平面；处理多种类型问题时也十分困难，因为一般支持向量机解决的都是二分类问题，对于多分类问题非常复杂。

[10]是指通过结合几个模型降低泛化误差的技术。

（五）朴素贝叶斯

朴素贝叶斯(Naive Bayesian Classifier，NB)方式则是在标准贝叶斯算法的基础上作出了一定的精简，即在假设给出目标值与属性间的交互要求独立。也就是说没有什么特征变量对于决策结果来说占据着极大的比例，也没有什么特征变量对决策结果占据着相对较小的比例。朴素贝叶斯的运算原理是概率统计中的贝叶斯定理，他需要一定的假定条件才能运用，因此在实际生活中应用比较局限。

优点：由于朴素贝叶斯运算假设了各种数据组属性间是彼此相互独立的，所以运算的逻辑性极为简洁，而且由于运算流程相对稳定，在数据出现了多种多样的特点时，朴素贝叶斯的分类也不会有太大的区别。

缺点：不适应与数据量大特征繁多的分类当中；需要很多前提假设，比如数据之间必须具有独立性，而且由于数据集中的特点属性间通常都存在着交错关联，因此一旦在数据分类过程中就存在了这些问题，会使得数据分类的效率大为下降。

（六）回归分析

回归分析主要是为了探寻两变量之间存在的某种关系，在相关关系研究领域中应用广泛。而回归统计分析的方式根据求解变量和被求解变量间的关联类型，可能进行线性和非线性回归分析；根据解释变量的大小，可能进行一元回归分析和多元回归分析。逻辑回归（Logistics Regression，LR）属于一种非线性回归的方式，适应于对比调查、跟踪调查和横断面调查，其结论中出现的变量取值一定是二分的，或多分类。

优点：比较简单和便捷；回归式分析方法不仅可以判别变量之间的相关关系还可以衡量相关关系的大小；回归分析不仅适用于二分类模型，还适用于多分类模型；适用性高，回归分析被广泛应用于分类、回归领域。

缺点：回归分析对于两变量之间的关系只是一种推测，结果不够准确；模型中的变量个数过多时，计算起来就会非常复杂。

（七）神经网络

神经网络(Artificial Neural Network，ANN)是一种新型智能学习算法，从它的起源至今已经经过了一百年的发展，所以神经网络技术的研究目前已经十分透彻，它也被广泛应用于各个领域当中。神经网络顾名思义就是指的计算机模拟人的行为，其内部机制像人的大脑一样交错复杂，通过神经元来传达指令协调全身的肢体协动。神经网络可以解决很多信息冗杂数据量大的问题。

优点：处理信息冗杂和数据口径不一致的数据有很大优势；神经网络广泛应用于不同的研究领域，针对不同的需求可以通过不断的自我学习取得非常不错的效果；可以通过自我学习的方式优化模型效果。

缺点：神经网络的运作模式是不透明的，所以没办法对得到的结果进行详细的解释；神经网络运作时会把信息处理成数据传入系统，所在把信息转化为数据的过程中可能会导致信息的缺失。

第二章　舞弊识别的指标体系及特征提取

第一节　样本选取

一、数据来源

本文的数据主要来自 CSMAR 金融研究数据库[11]，包括公司研究系列中的财务指标数据、违规数据[12]、治理结构和内部控制中的数据。根据资料的可获得情况和本文的研究目的选取了 2011 年到 2020 年年末的财务报表数据作为本文的数据来源，其中舞弊样本数据主要来自于违规信息中的数据和相应的证监会处罚公告；非舞弊样本主要来自于国泰安数据库中的财务指标数据和上市公司自身发布的年度报告。

二、舞弊样本的选择

结合 CSMAR 数据库中的资料和证监会的处罚公告，本文选取了近 10 年来的舞弊样本数据，并对存在重大缺失数据的样本进行删除，将剩下的数据进行归纳整理，得到以下分类：重大遗漏、一般会计处理不当、内幕交易、披露不实、推迟公开、违法买卖证券、虚假记录和虚构利润。上市公司财务舞弊行为种类的分布情况如表 2-1 所示：

表 2-1 上市公司财务舞弊类型汇总表

	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	总计
内幕交易	11	11	4	3	5	6	0	2	0	0	42
一般会计处理不当	42	35	18	8	7	5	9	4	3	4	135
披露不实	12	15	9	15	18	11	1	5	1	2	89
虚假记载	43	54	21	14	20	32	37	30	26	21	298
违规买卖股票	43	39	63	37	85	78	81	82	67	44	619
推迟披露	60	49	44	39	68	93	69	75	64	33	594
重大遗漏	75	53	33	30	41	42	34	32	17	22	379
虚构利润	145	128	80	62	87	111	103	112	73	63	964
总计	431	384	272	208	331	378	334	342	251	189	3120

为了保证样本的质量，避免重复样本，更精准地分析财务舞弊的特征，本实证研究选取的样本为首次发生财务舞弊的数据，由于一些公司未注明违规时间和有些公司缺失值过多，经过一系列筛选处理后，得到样本数据共 2022 条。由表 2-1 中可以看出，我国上市公司财务舞弊原因最多的三项是违规买卖股票、推迟披露和重大遗漏，其中占比最大的就是违规买卖股票，高达 619 次，其次推迟披露也达到了 594 次。按照规定，中国上市公司有义务披露重要事项和财务信息，但推迟披露的原因不外乎是披露了对公司不利的信息，中国市场对信息非常敏感。为了自身利益而保护、拖延或

[11]全称为 CSMAR 经济金融研究数据库，是国泰安从学术研究的需求出发，借鉴芝加哥大学 CRSP、标准普尔 Compustat、纽约交易所等知名数据库的开发模型，再结合我国实际情况开发的金融型数据库。
[12]该数据来源于证监会每年对上市公司财务舞弊的公开处罚结果，并对其进行整理汇总得到。

不公布上市公司相关信息，这是由投资者鼓吹并严重损害股东利益的。违规买卖股票是指采取一些非正当的手段从中谋取差价，从而获利。但上述做法通常不是很隐蔽，因此被监管部门披露的可能性很大，舞弊行为也就更多。

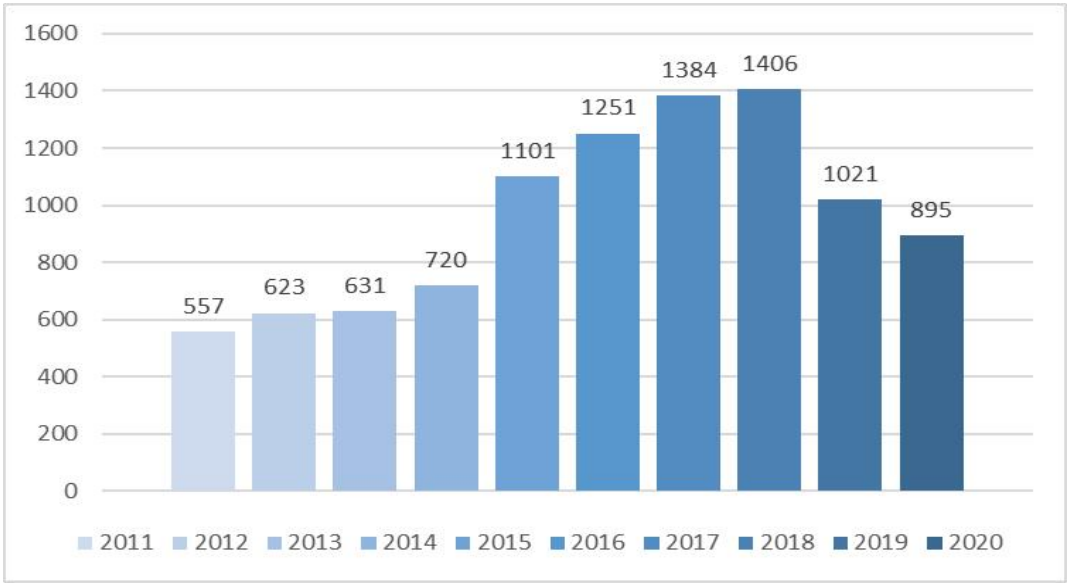


图 2-1 近年来发生财务舞弊的公司数量

从图 2-1 可以发现，近年来上市公司财务舞弊的数量总体上呈现出先上升后下降的趋势，在 2018 年达到顶峰，达到了 1406 家，然后 2019 年和 2020 年略微减少，说明上市公司自身或者国家很可能采取了某种政策导致了舞弊公司数量的减少，但不可否认的是上市公司舞弊的数量依然居高不下。

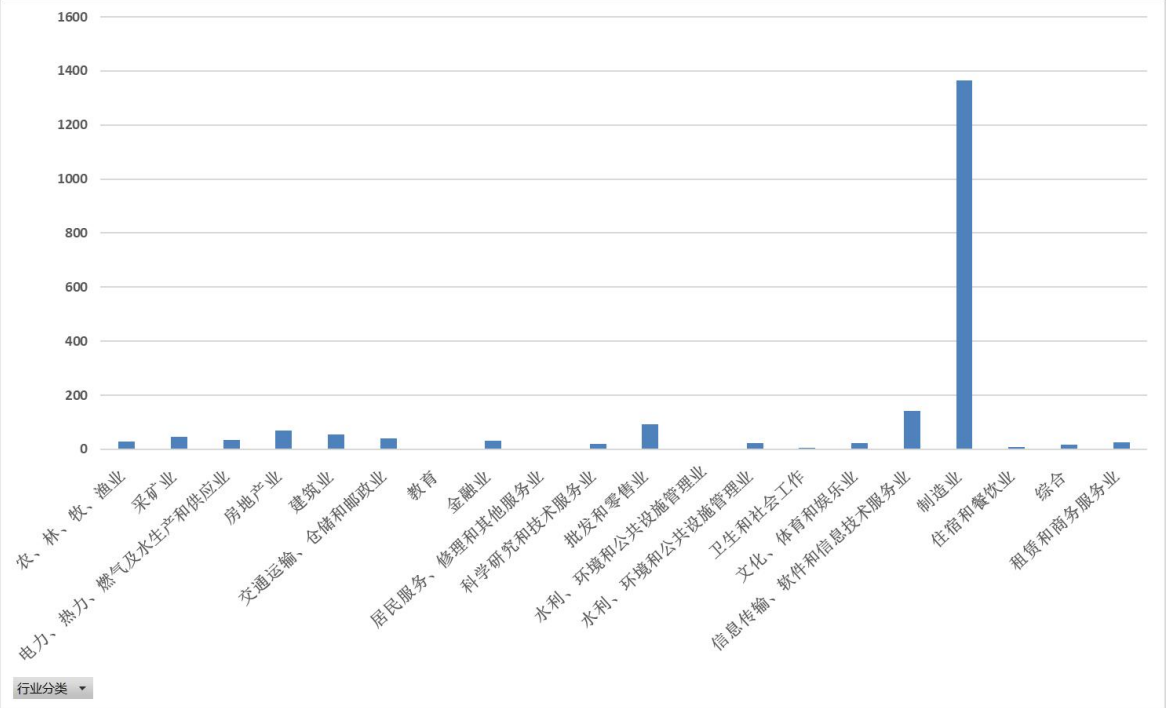


图 2-2 2020 年上市公司财务舞弊行业分布图

从图 2-2 可知，上市公司的行业分类[13]中几乎都存在财务舞弊现象，但制造行业舞弊数量最多，占了 67.46%，其次是软件、传输和信息技术服务业。其原因可能是由于制造业的在我国起源较早，累积到今天数量繁多，并且制造业的公司操作流程非常庞大，从采购到原料的生产再到制造和销售，牵涉的会计核算科目相对较多，为粉饰财务数据提供了许多可能性。所以审计部门在之后的审查中对制造行业的财务指标的审查应该提高警惕。

三、非舞弊样本的选择

虽然我国上市公司的数量众多，产生舞弊的上市公司数量也在逐年上升，但相对于上市公司的总量，发生舞弊的公司数量相对较少，每年发生财务舞弊的公司所占上市公司的比例不到 10%。显而易见，这样的样本将不利于我们对财务舞弊的特征进行分析，也可能会对后文选取财务舞弊的识别模型造成影响。所以针对该不平衡数据集，我们采用过采样和欠采样相结合的方式，适当增加舞弊样本的比例，以提高模型的识别效果。根据本文的研究目的，最终选择了 1：2 的配对比例选择非舞弊样本的数量。

在明确配对样本的选取比例后，参照贝叶斯的配对原则[14]并考虑到财务数据的特征，保证配对样本的质量，应按照以下原则进行配对：

第一，舞弊样本和匹配样本的一致性。两样本要来自同一个行业且公司的规模相似、业务类型相似、关联交易相似，这样才能保证整体结构的相似性，以提高模型的识别效率。

第二，舞弊样本和匹配样本年度相同。为了避免因外部环境变化造成的不同年份数据的不同，所以匹配样本数据应和舞弊样本来自于同一个年度。

第三，匹配样本应该是未经过证券交易所等监管机构处罚的，也就是没有发生财务舞弊行为的，财务报告的审计意见为标准的无保留意见并且公司规模相对稳定，业务比较成熟，还有最关键的一点是近几年内发生舞弊的可能性较低。

基于以上选取原则，最终通过整理归纳了 2022 条舞弊数据和 4044 条非舞弊数据纳入舞弊识别模型。

第二节　变量选取及数据的预处理

本文以舞弊风险因素理论为基础，从财务舞弊动机的角度分析了财务舞弊的成因和手段。将公司治理等非财务指标纳入财务违规识别变量的范围，对财务舞弊的认识有一定的作用。因此，通过对上市公司财务和非财务的两方面分析，确定了本文的可量化指标。由于指标中变量选择的原则侧重于指标的客观价值、操作意义和代表意义，避免使用信息交叉指标和强调信息的完整性。所以本文采用的指标是一个可变的指标，

[13]这里的行业分类按照证监会行业分类 2012 年版进行分类。
[14]在配对时尽量除解释变量外其他条件保持一致，保证两个样本预测时结果更加准确。

能反映整个上市公司的运作的。而且结合以往的风险因素理论，本文还提出了一种新的指标选择方法：利用财务指标确定影响企业经营压力的因素，特别是盈利能力、经营能力和偿债能力等。还有能够体现公司的具体资产结构的比率结构也纳入我们财务指标的构建中。为了更好地考虑到公司的经营状况和财务风险，也选取了股权结构和公司治理两方面来衡量。

一、财务指标的构建

1.偿债能力

偿债能力反映了公司对各种债务的偿还能力。根据财务舞弊动因理论我们知道财务舞舞的形成很大一部分来自“压力”，如果企业外部给到的压力过大，导致企业无力偿还，则很有可能发生财务舞弊。

2.发展能力

发展能力反映了公司在原有的基础上不断壮大自身的能力，能够持续健康发展的能力。公司的资金来源之一是外部投资，而外部投资者对公司的判别标准就是看它的发展能力，所以公司自身很可能为了吸引投资进行财务舞弊。

3.比率结构

比率结构反映了公司各种资产的具体结构的比率。一般而言，正常公司的内部结构资产、负债和所有者权益会达到一种动态平衡，但如果公司的财务状况很差的话要达到这种平衡就很困难，从而会采取某些手段粉饰其财务报表。

4.盈利能力

盈利能力也称获利能力，反映了公司获取利润的能力。他和发展能力一样，是外部投资者非常看重的一项指标，所以很多公司为了掩饰其真实的获利能力吸引投资而进行财务舞弊。

5.经营能力

反映选定上市公司营运能力的关键指标。在证券交易所上市的周转公司利用虚假收入来降低成本和费用，这就违反了财务相关的条例。同时，应收账款、流动资产经常增加，通过增加这些资产，公司的财务状况以及这些资产中舞弊公司的经营指标都高于非舞弊公司。

4.现金流分析

现金流主要反映了公司内部资产的流动情况。在现行会计制度下，大量上市公司利用会计制度操纵自身权益，现金流量平衡是以收支制度为基础的，因此通过对现金流量指标的分析，可以降低人为因素。

5.风险水平

风险水平反映了企业对于外部环境变化的应变能力。只有企业的风险承受能力较高时，企业才能健康稳定的发展下去。

二、非财务指标的构建

1.治理机制

公司治理结构的缺陷将更有可能增加公司财务舞弊发生风险。本文选择董事会规模、独立董事占比、以及股东大会召开次数来体现公司财务违规的指标。

2.股权结构

股权结构可以清晰地划分公司股权持有者的情况，集中控制公司。一个恰当的股权结构可以克服公司治理的缺陷，提高公司治理效率。初步选择了两权分离度[15]等指标来表示股权结构。具体指标如表 2-2 所示：

表 2-2 初始指标构建

指标名称	一级指标	二级指标	变量	一级指标	二级指标	变量
财务指标	偿债能力	流动比率	X_1	盈利能力	资产报酬率	X_{18}
		速动比率	X_2		总资产净利润率（ROA）	X_{19}
		现金比率	X_3		流动资产净利润率	X_{20}
		固定支出偿付倍数	X_4		净资产收益率	X_{21}
		权益乘数	X_5		投入资本回报率	X_{22}
	发展能力	固定资产增长率	X_6		营业毛利率	X_{23}
		总资产增长率	X_7		成本费用利润率	X_{24}
		净利润增长率	X_8	经营能力	应收账款与收入比	X_{25}
		资本保值增值率	X_9		应收账款周转率	X_{26}
		资本积累率	X_{10}		存货周转率	X_{27}
		可持续增长率	X_{11}		流动资产周转率	X_{28}
	比率结构	流动资产比率	X_{12}	现金流分析	营业收入现金净含量	X_{29}
		现金资产比率	X_{13}		营运指数	X_{30}
		营运资金比率	X_{14}		资本支出与折旧摊销比	X_{31}
		固定资产比率	X_{15}	风险水平	财务杠杆	X_{32}
		所有者权益比率	X_{16}		经营杠杆	X_{33}
		流动负债比率	X_{17}		综合杠杆	X_{34}
非财务指标	治理结构	独立董事占比	X_{35}	股权结构	直接控股股东持股比例(%)	X_{39}
		董事会规模	X_{36}		实际控制人拥有公司所有权比例	X_{40}
		董事会会议次数	X_{37}		实际控制人拥有公司控制权比例	X_{41}
		股东大会召开次数	X_{38}		两权分离度(%)	X_{42}

三、数据的预处理

1.缺失值处理

由于下载的数据有缺失值，在进一步分析处理之前，数据要进行处理。在实际数据丢失超过 50%的指标后，就将该指标删除。但如果将所有只要有缺失值的指标都删除的话，样本大小会大大减小。如果数据量不足，许多重要的隐藏信息就会丢失，数据的特征和分布也会受到影响。若手动填充是对误差值的主观估计，且不一定是真实的数值，错误的填充会增加噪声。因此为了保证数据的可靠性，本文利用 python 中

[15]本文所指的两权分离度中的两权不是所有权和经营权，而是指公司治理中的控制权与所有权的分离。

的填充工具[16]进行填充，填充值为每列数据的平均值。

2.标准化处理

本文选取的指标量纲不同，为了避免量纲差异造成结果的不准确，所以要对样本数据进行标准化处理。指标对评价方案的影响也不同，指标的直接相加不能准确反映其总体效果，这就必须要改变数据的类型：即数据同化，确定数据值之间水平差异的影响指数，并对数据进行无量纲处理。对于原始数据来说，改善数据关系也是非常重要的，这样可以保证模型的收敛速度和精度及其结果的可靠性。最后我们为每个样本添加标签类型：舞弊样本为 1，非舞弊样本为 0。

第三节　指标的特征选择

一、指标的显著性检验

上文选出的指标可能对识别财务舞弊的行为没有显著的区别能力，如果这些指标仍保留在确定的指标体系中，就不能用来确定财务是否违规，而且会因为这些指标有更高的维度导致误差较大和数据冗余问题，所以我们需要对它们分别进行显著性检验[17]。在两个样品的分类检验中，最常用的独立检测方法是曼惠特尼 U 检验。该检验是把收集到的样本信息进行混合，然后求出各样本信息的秩，对两组不同样本的秩分别求出其平均值，并据此判断两相对独立的样本的平均数是否具有显著性差异。本文的曼惠特尼 U 检验结果是通过 spss 软件得到的，通过检验最终得到了这 42 个指标的显著性检验结果，由于表格较大，所以我们只列出了未通过显著性检验的相关指标，如表 2-3 所示：

表 2-3 曼惠特尼检验不显著指标

指标名称	显著性	指标名称	显著性
固定资产增长率	0.208	独立董事占比	0.244
流动负债比率	0.646	董事会规模	0.828
应收账款周转率	0.094	股东大会召开次数	0.600
资本支出与摊销折旧比	0.243	两权分离度	0.926

由输出结果可知，有 34 个初始指标通过了曼惠特尼 U 检验，它们的 P 值均小于 0.05，即认为这 34 个指标对财务舞弊和未财务舞弊的样本有显著性差异；而固定资产增长率，流动负债比率等指标的 P 值均大于 0.05，即可以认为这 8 个指标对识别上市公司财务舞弊的行为没有显著性差异，故应剔除这 8 个指标。

二、指标的相关性检验

除此之外，我们还得判断变量与变量之间是否存在相关性，其强弱如何，所以本文对通过显著性检验的 34 个指标进行 person 相关性检验[18]，以避免变量之间的关联

[16]这里使用的是 fillna 方法对缺失值进行批量填充。

[17]显著性检验就是首先对要检验的样本做一个假设，然后利用样本信息来判断这个假设是否合理，即判断总体的真实情况是否与原假设存在显著性差异。

[18]该检验是用来衡量两个变量间的线性关系的。

性对结果造成的误差。由相关性系数的概念可知，两变量的关联性越强，皮尔森相关性系数越接近 1 或-1，关联性越弱，person 相关性系数越接近 0。对于多变量我们通常使用的是 person 相关系数矩阵，但由于本文变量较多，所以结合热力图的呈现效果，绘制了 person 相关系数热力图，具体如图 2-3 所示：

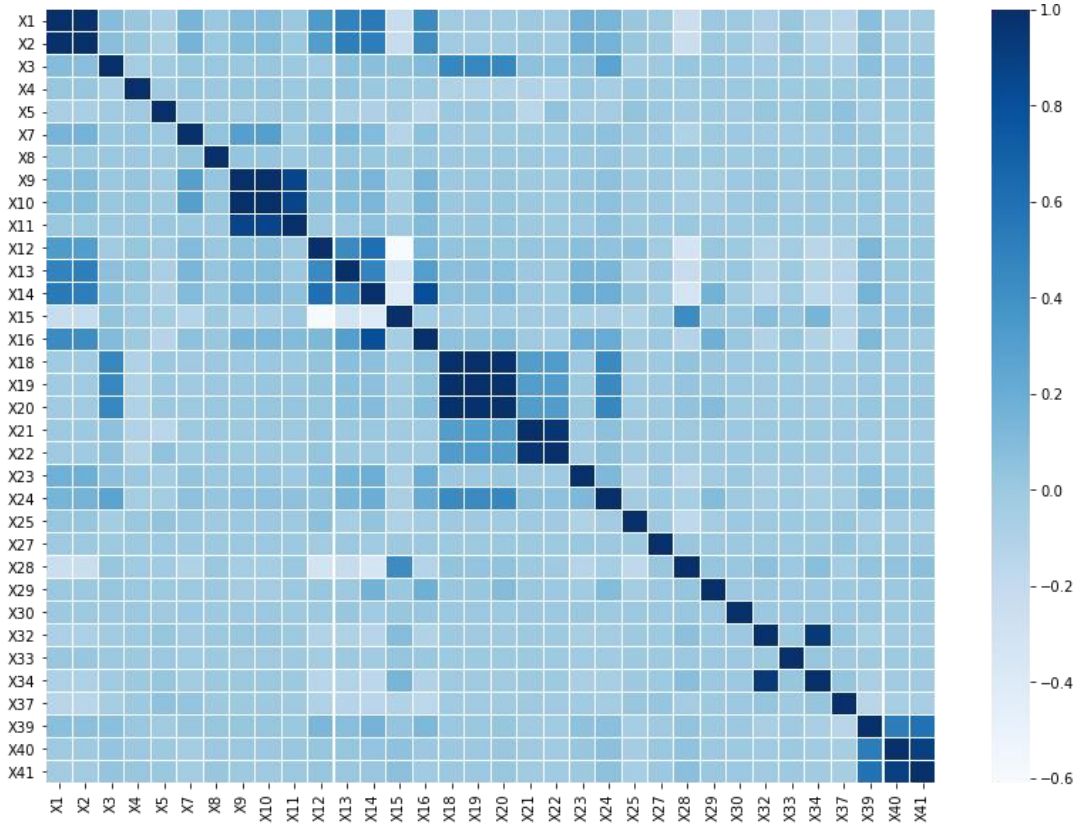


图 2-3 person 相关性热图

由图 2-3 可知，有 6 组指标之间存在强相关性，具体分析情况如下：

1.流动比率（X_1）和速动比率(X_2)这组指标中的相关系数为 0.988，指标之间存在强相关性。流动比率用来反映公司资产在部分负债资产结束之前，能够变成现金资产用以清偿资产的力量，而速动比率是企业速动资产与流动资产的比例，只是对流动比率的一种补充，而且也能够通过流动资产计算得到，所以本文考虑舍去速动比率，而把流动比率纳入模型的构建当中。

2.资本保值增值率（X_9），资本累计率（X_{10}）和可持续增长率（X_{11}）这组指标中的相关系数分别为 0.999，0.869，0.870，指标之间存在强相关性。这三个指标都是通过一些营运指标来反映公司的可持续发展状况的，但可持续增长率相对比较全面，与其他变量的相关系数也较低，所以选择在这里选择可持续增长率。

3.资产报酬率（X_{18}），总资产净利润率（X_{19}）和流动资产净利润率（X_{20}）这组指标中的两两相关系数分别为 0.999，0.986，0.986，指标之间存在极强的相关性。首先总资产净利润率和流动资产净利润率的区别在于对分母的维度处理不同，前者是以平均资产作为分母，后者是以流动资产作为分母，为了更加全面的涵盖公司信息，选

择了平均资产总额，而分子也选择更加综合的资产报酬率纳入指标建设中。

4.净资产收益率（X_{21}）和投入资本回报率（X_{22}）这组指标中的相关系数为 0.966，指标之间存在强相关性。净资产收益率可反映出企业对主要股东直接投入资金的效率，它补充了每股税后利润指数的不足之处。投入资本回报率是指公司投入的成本和收回的利润之间的比值，用以反映投出资金的使用效果。我们判定财务舞弊的行为通常会更加关注给股东或公司带来的效益，所以在这里选择的是净资产收益率。

5.财务杠杆（X_{32}）和综合杠杆（X_{34}）这组指标中的相关系数为 0.943，指标之间存在强相关性。其中综合杠杆是包含财务杠杆和经营杠杆的，是它们共同作用的结果，是用于衡量销售量的变动对普通股每股收益变动的影响程度的。所以本文考虑舍去财务杠杆，而把综合杠杆纳入模型的构建当中。

6.实际控制人拥有所有权的比例（X_{40}）和实际控制人拥有控制权的比例(X_{41}）这组指标中的相关系数为 0.902，指标之间存在强相关性。这两个指标的区别就在于所有权和控制权的不同，而所有权和控制权区别就是实权和虚权的区别，实际控制人拥有公司的所有权并不能参与公司决策，所以对于识别财务舞弊的行为来说，我们更应该关注的是后者，故将后者纳入模型的构建当中。

三、基于特征选择算法的指标提取

为了使模型达到最好的识别效果，本文综合采用过滤式[19]、包裹式[20]、嵌入式[21]多种方法对模型的特征进行选择。具体为 Scikit-Learn[22]中的树算法、RFE 搜索算法、Lasso 回归和随机森林四种方式。先把原始数据代入到各个模型当中，然后对各个特征进行排序，再进行拟合投票表决。其中 Scikit-Learn 树算法内部实现是采用了调优过的 CART 树计算，既能够做回归，又能够做分类；RFE 搜索算法是通过对特征子集的有效搜寻，以便找到最佳的子集，也就是说该模型在这些子集上有最好效果，递归消除的核心思想是通过反复迭代，剔除没有预测意义的特征，与向后逐步回归非常相似，属于纯技术性的变量选择；Lasso 回归是在线性回归模型的代价函数后面加上 L1 范数的约束项模型，通过控制参数 Lambda 进行变量筛选和复杂度调整；随机森林则可以直接对特征重要性进行排序，确定要剔除的比例，然后依据特征重要性剔除相应比例的特征。我们在 python 中导入 sklearn 库，把通过显著性检验和相关性检验的原始指标分别代入四种特征选择的方法中，并且按照各特征的频次进行筛选和排序，选取了前十二个特征，构造出新的特征子集，具体见表 2-4 所示：

[19]过滤式就是先特征选择后训练的方法，通过设定不同的阈值进行特征选择。
[20]包裹式是通过不断训练学习器来选择最佳子集的一种方法。
[21]嵌入式是指在选择特征子集的过程中自由进行选取。
[22]Scikit-learn 是针对 Python 编程语言的免费软件机器学习库，它具有各种分类，回归和聚类算法。

表 2-4 特征筛选表

特征	Tree	RFE	Lasso	RF	Total
总资产增长率	True	True	True	True	4
现金比率	True	True	True	True	4
成本费用利润率	True	True	True	True	4
综合杠杆	True	True	True	True	4
营运指数	True	True	True	True	4
可持续增长率	True	True	True	True	4
资产报酬率	True	True	True	True	4
净资产收益率	True	True	True	True	4
流动资产比率	True	True	False	True	3
现金资产比率	True	False	True	True	3
经营杠杆	True	True	False	True	3
净利润增长率	True	True	True	False	3

第三章　财务舞弊识别方法的选择及实证分析

第一节　舞弊识别模型的选择及模型评估指标的介绍

一、舞弊识别模型选择

本文主要是为了研究上市公司财务舞弊的识别方法，在国内外文献研究的基础上，我们发现数据挖掘技术对财务舞弊的识别具有良好的表现。然后我们收集了数据挖掘算法的适用条件和优缺点，通过比较各种数据挖掘方法的特点，并结合本文样本数据，发现数据挖掘技术可以用于财务舞弊的分类问题。通过构造不同的分类器，可以实现上市公司财务舞弊的识别。数据挖掘包括线性判别分析、决策树、支持向量机、朴素贝叶斯分类、神经网络等算法，可以有效地对数据量庞大，特征繁多的数据进行分类。并且为了达到更好的识别效果，我们引入了组合分类器的概念。组合分类器顾名思义是对各基础分类器的预测结果进行组合，它综合了所有单个分类器的优点，使得识别结果更加全面和高效。组合分类器通常被认为是弱学习算法的补充。并且对于综合识别模型来说还有很多种组合形式，不同的组合形式效果也不同，这点我们将在后面的研究当中进行验证。

基于以上对分类器的研究和对各种识别模型优缺点的比较，根据本文数据量的大小及其特点，选择了适合本文数据量的五种基础分类器来构建上市公司舞弊识别模型，再根据平均法、投票法和学习法构建了综合识别模型，最后对各种单一模型和根据不同组合方式的综合识别模型的效果进行比较，选出最优的财务舞弊识别模型。

二、模型评估指标的确定

（一）准确率、精确率、召回率及 F_1 值

评估分类模型的方法有很多，但最常用的方法是计算混淆矩阵中的各项指标。而这些指标也是基于混淆矩阵定义的基础上提出的，简单来讲混淆矩阵定义就是先依次计算分类模型的预测结果哪些是正确的，哪些是错误的，之后再将结果放到同一个表里中得出，而这种表示方式便是混淆矩阵[23]。具体如表 3-1 所示：

表 3-1 混淆矩阵

		真实值	
		1(True)	0(False)
预测值	1(Positive)	TP	FP
	0(Negative)	FN	TN

其中：TP 表示实际为正预测也为正、FP 表示实际为负但预测为正、TN 表示实际为负预测也为负、FN 表示实际为正但预测为负。

通过混淆矩阵，可以给出以下各指标的值：

准确率(Accuracy)：反映识别模型对实验样本的预测水平，也就是能将正的预测为正，负的预测为负的能力，计算公式：

[23]混淆矩阵是预测分类模型结果的分析表，数据集根据分类模型的实际类别和预测类别进行聚合。

$$Accurary = \frac{TP + TN}{TP + TN + FP + FN} \tag{3-1}$$

精准率(Precision)：被预测为真中真正为真的个数，即预计对的正例数占真的正例数量的比例：

$$\Pr ecision = \frac{TP}{TP + FP} \tag{3-2}$$

召回率(Recall)：所有实际是正的样本中有几个被真正预计为正样本，也就是实际正样本与真正正样本的比值：

$$\mathrm{Re}call = \frac{TP}{TP + FN} \tag{3-3}$$

F_1 值：也可以说是负样本的精准率，或者可以说是负样本中的数据有哪些是被真正预测为真的，其计算公式：

$$F1Score = \frac{2PR}{P + R} \tag{3-4}$$

（二）ROC 曲线和 AUC 值

ROC 曲线也称接受者操作特性曲线，得此名的主要原因就是曲线的各点都表现同样的感受性，它们均是人类对于同一种信号干扰程度的反映，只不过是处于一种不同的判断准则下所产生的结果而已。ROC 曲线主要是以虚惊概率为交错轴，基于击中概率的操纵结果而生成的，对于特殊干扰条件而言，由于采用了不同的判别标准所以画出来的曲线也是各不相同。AUC 值是在 ROC 曲线下方围成的总面积，为图中阴影部分的总面积。很明显，AUC 值越大，分类器的分类效率就越好。如果 AUC=1，那么作为完美分类器，在应用这个预测模式时，无所谓选取哪个阈值点都能够判断得非常完美，所以他并不常见；当 AUC>0.5 时，称为优于随机猜测，在这种情况下，其应用价值主要看研究的目标和研究数据的特点；当 AUC=0.5 时，就如同随机猜想，猜对和猜错的概率都为 1/2；当 AUC<0.5 时，便毫无应用价值可言，这里便不做具体的阐述了。具体如图 3-1 所示[24]。

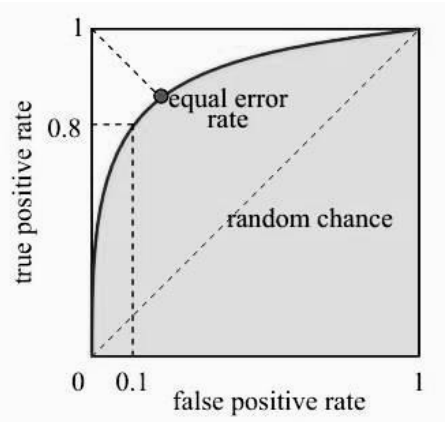


图 3-1 ROC 曲线示意图

[24]图片来源于 ImaginationTech 的机器学习之分类器性能指标之 ROC 曲线、AUC 曲线。

第二节　财务舞弊单一识别模型的构建及识别效果分析

一、基础舞弊识别模型的构建

（一）支持向量机

支持向量机在解决分类问题上具有非常广泛的运用，它的关键在于其核心函数的不同，因此它可以解决线性或非线性的问题。而在支持向量机中核心函数却是十分关键的，也正是支持向量机的灵魂所在，他的参数能够直接影响支持向量机的复杂程度。经过对不同核心函数的特性的综合研究，选择了高斯函数作为支持向量机的核心函数。

本文采用的核函数为高斯函数[25]，其函数形式为：

$$K(x,x_i)=\exp(-\frac{|x-x_i|^2}{\sigma^2}) \tag{3-5}$$

得到的基分类器表达式为：

$$f(x)=\mathrm{sgn}(\sum_{i=1}^{n}a_iK(x,x_i)+b) \tag{3-6}$$

支持向量机模型不仅可以解决线性问题，还可以解决很多非线性的问题。它是解决很多复杂的分类问题的有力手段，可以直接用 python 中的支持向量内置函数对上市公司样本数据进行分类。

（二）K 近邻算法

K 近邻算法又称 KNN，是计算机教学与数据挖掘应用领域中常见的学习类型，更是计算机教学中最简洁的一个模型。KNN 的应用范围相当广，在样本数足够大的前提下它的精度也相当高。其计算的核心思想就是某个样本类型和统计集的 k 个样本都很接近，假设这 k 个样本中的数据大多都靠近某一个类型，那么该样本也就属于这种类型。如图 3-2 所示，本文通过该算法处理新数据时，首先计算处理数据的点和距离，为了更好的确定 k 值，本文运用的是欧几里德距离函数。然后选择训练数据的下一个点来匹配距离的相似度计算，通过少数服从多数的原则进行分类。

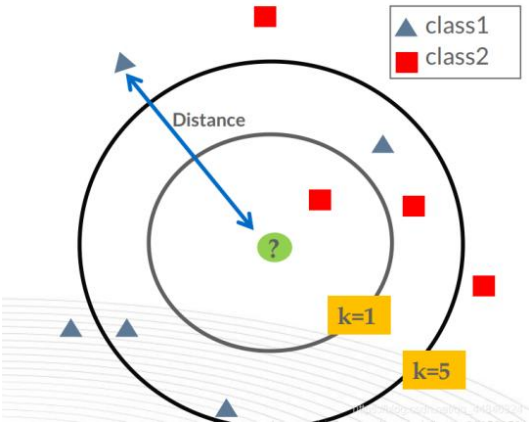


图 3-2 K 近邻算法示意图

[25]高斯函数呈特征对称的钟形形态，其应用非常广泛，如模糊、正态分布统计和卷积网络。

（三）逻辑回归模型

Logistic 回归模型的主要功绩就是它能够消除统计学假定中条件上的约束，它不需要过多的假设条件。Logistic 模型不仅能够高效地处理回归问题，还可以用于分类问题当中，因此被广泛用于财务舞弊识别的研究当中。设本文的自变量用 x_1，x_2，.....，x_k 表示，被解释变量用 P 表示。P 也可以理解为企业发生财务舞弊的概率，1 表示发生财务舞弊，0 表示正常公司的样本。

本文采用的 Logistic 回归模型为：

$$\ln\frac{p}{1-p}=a_0+\sum_{i=1}^{k}a_ix_k \tag{3-7}$$

将然后把上文得到的指标代入到该模型当中，如果 P 小于 0.5 则为正常公司的样本，如果 P 大于 0.5 则为舞弊公司样本。Logistic 回归模型不要求样本数据满足各种苛刻的前提假设，总体来说模型的计算简单，适用性强。

（四）随机森林

随机森林用通俗的话来说就是一棵棵决策树综合起来的结果。因为他是由不同决策树[26]累计起来的，就像一片森林是由一棵棵树木构成的，这也是其名字的由来，其分类的结果则是由各个决策树综合决定的。在论文中，随机森林算法主要包括了以下的四个过程：

(1)把 6066 个舞弊样本和非舞弊样本按照 70%的比例抽取训练集，再对其进行有放回的随机抽样，把抽取出来的样本作为一个新的训练集；

(2)再把过程一中的新训练集按照有放回的随机抽样的原则，抽出 k 个特征构成子集去训练决策树；

(3)然后不断重复前两个步骤，直到把样本训练完成。

(4)根据各分类结果投票决定最优分类。如图 3-3 所示。

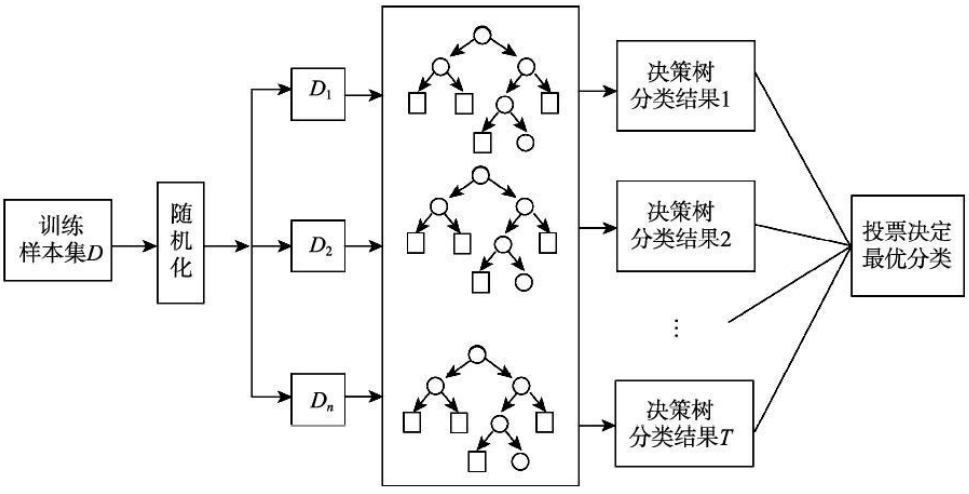


图 3-3 随机森林示意图

[26]该方法是通过不断的向下选择、不断地进行分支来对样本进行分类的一种方式。

（五）BP 神经网络

BP 神经网络系统是目前应用最多的一类人工神经网络形态。BP(Back Propagation)神经网络系统是人工神经网络的一类，它采用反方向传递的方式来调节系统中的各项参数，采用梯度下降方式，实现神经网络误差函数最低的目标，从而实现信息的获取。其构造分为输入层、隐含层和输出层。基于前面描述，我们可以选择三层的神经网络系统构成，也就是一组输入层、一组隐含层、一组输出层。图 3-4 是 BP 神经网络系统的基本内部结构示意图。输入层的神经元总量取决于样本中的各指标的数量，隐含层的神经元数量通过不断地测试和自我学习，最终选择了效果最好的层数 7，而输入输出层的神经元总量则选定为 1。我们以 1 表示发生财务舞弊的样本，0 表示正常样本。初始连接权值必须位于(-1，1)之间，并使其随机地生成。然后刚开始的步长我们设置为 0.1，然后不断开始叠加。如果样本输出值>0.5，则该样本很可能是发生财务舞弊的公司；如果样本输出值<0.5，则该样本为正常公司的数据。

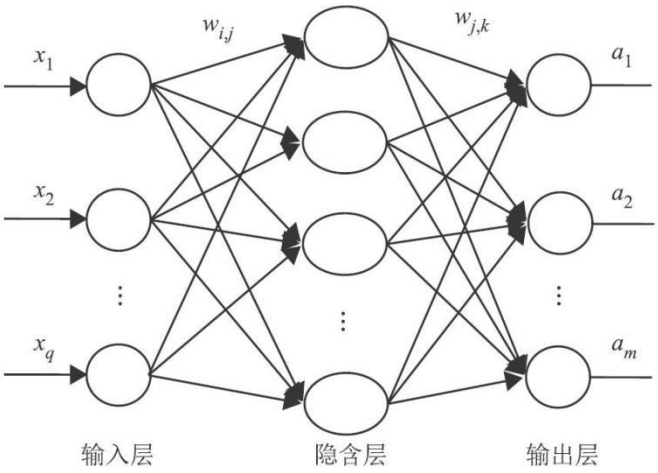


图 3-4 神经网络示意图

二、基于初始指标的基础分类模型的评估

根据上文的样本选择，目前已经得到了 6066 个样本，其中财务舞弊和非财务舞弊的样本数为 1：2，在纳入基础分类模型之前，需要将这 6066 个样本分为训练集和测试集。为了保证精确度，我们按照随机抽样的原则抽取总样本的 70%作为训练集，剩余 30%的样本用作测试集[27]。为了使不同公司的各项指标之间能够进行运算和比较，并且具有良好的识别效果，我们选用通过显著性检验和相关性检验的标准化指标作为初始指标带入模型。然后我们把训练样本中的指标数据分别用支持向量机、神经网络、逻辑回归、随机森林和 K-近邻算法建立模型，并且用测试集对训练集产生的模型加以验证，最终得到如图 3-5 所示的实验结果：

[27]划分时用 random_state 设置随机数种子，用 test_set 设置测试集比例。

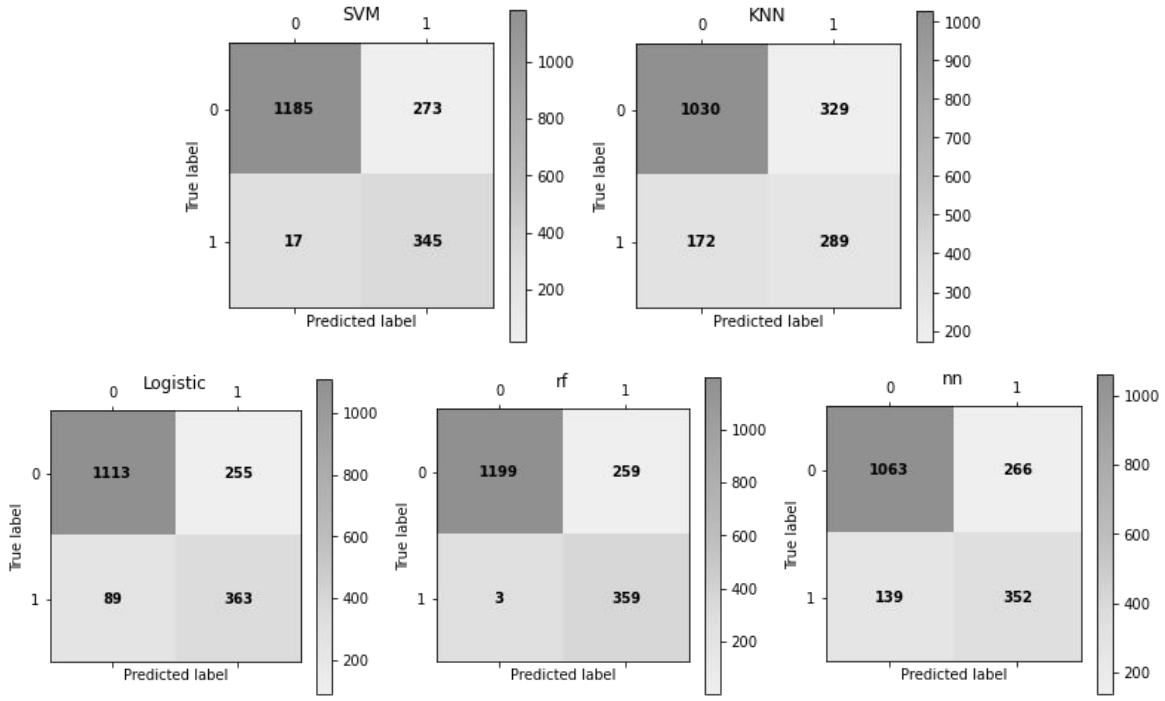


图 3-5 各分类模型的混淆矩阵

总的来说，构建模型的目的是为了识别财务舞弊的企业，所以本文中模型对财务舞弊的识别效果决定了模型最终的实用价值。从上述矩阵中我们可以看出支持向量机模型测试集中有 1185 个公司为非财务舞弊公司，识别出来也是非财务舞弊公司，345 个样本被准确识别为财务舞弊的公司，正确识别总数为 1530 个样本，占测试集总数的 84.07%；K-近邻算法模型中有 1030 个公司为非财务舞弊公司，识别出来也是非财务舞弊公司，289 个样本被准确识别为财务舞弊的公司，正确识别总数为 1319 个样本，占测试集总数的 72.47%；逻辑回归模型中有 1113 个公司为非财务舞弊公司，识别出来也是非财务舞弊公司，363 个样本被准确识别为财务舞弊的公司，正确识别总数为 1476 个样本，占测试集总数的 81.10%；随机森林模型测试集中有 1199 个公司为非财务舞弊公司，识别出来也是非财务舞弊公司，359 个样本被准确识别为财务舞弊的公司，正确识别总数为 1558 个样本，占测试集总数的 86.21%；神经网络模型测试集中有 1063 个公司为非财务舞弊公司，识别出来也是非财务舞弊公司，352 个样本被准确识别为财务舞弊的公司，正确识别总数为 1415 个样本，占测试集总数的 77.75%，五种模型都分别取得了不错的识别效果。

由上文分析可以看出混淆矩阵得到的模型训练结果不够直观清晰，所以基于混淆矩阵我们得到了如表 3-2 所示的各种衍生指标的评估效果表。

表 3-2 初始指标的基础模型评估效果

基础模型	准确率（%）	精确率（%）	召回率（%）	F_1 值（%）
SVM	84.0659	85.3038	77.2054	79.7529
KNN	72.4725	62.6898	66.2271	67.0027
LG	81.0989	80.3097	75.6667	77.2326
RF	86.2087	88.6737	79.8891	82.6421
ANN	77.7472	71.6904	72.6969	73.7395

由表 3-2 可以看出，支持向量机和随机森林的准确率相对较高，分别达到了 84.07%和 86.21%，而且他们的精确率、召回率和 F1 值也是五种模型中效果最好的。但支持向量机的准确率只有 72.47%，其他各项指标率也很低，则该识别模型对原始指标的财务舞弊识别率很差，其原因可能是因为支持向量机更适合小样本数据，所以识别的准确率没有其他模型高。

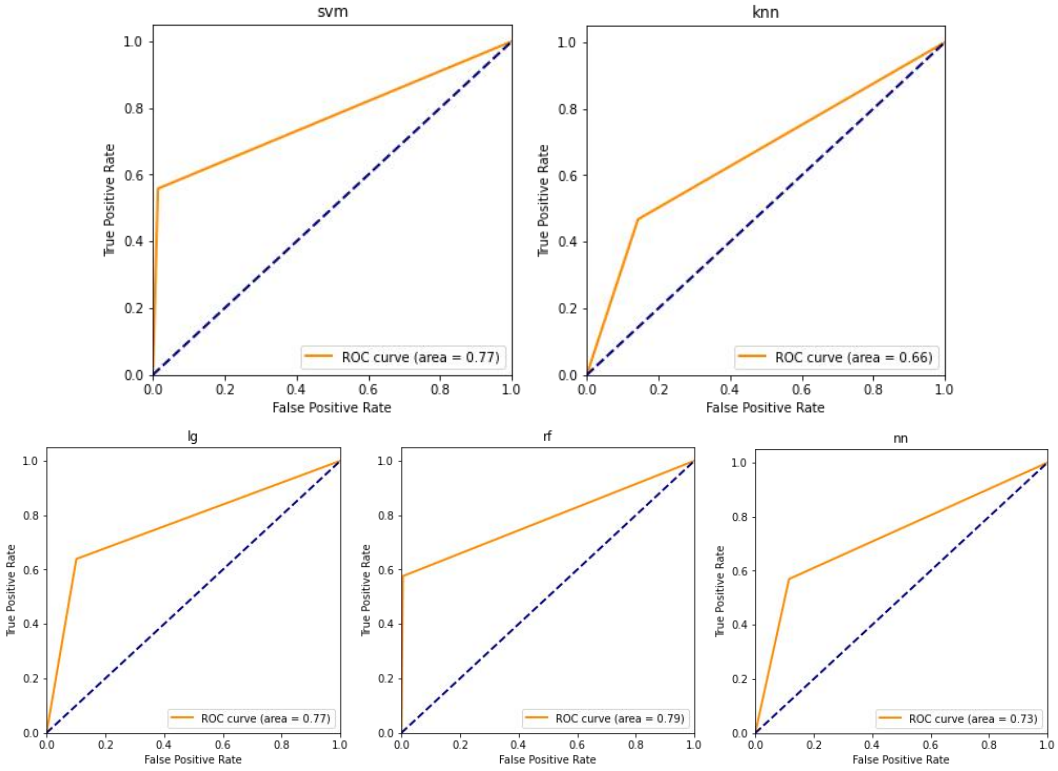


图 3-6 各分类模型的 ROC 曲线

由上述定义可知，ROC 曲线以下不大于 1 的曲线面积指的就是 AUC 的值。一般来说 AUC 值位于 0.5 到 1 之间说明该模型是有预测价值的，并且越靠近 1，预测价值越大。本次预测的各分类模型的效果基本上与混淆矩阵的分析结果一致，由图 3-6 可知，五种模型 AUC 值都在 0.6 以上，说明各模型都取得了较好的预测效果，具有一定的预测价值。

三、基于特征选择的基础分类模型的评估

我们由上述初始指标得到的模型效果可知，各分类模型对识别财务舞弊有一定的效果，但效果并不够显著，这可能是由于指标冗余造成的，所以我们对各指标进行了特征选择。经过上文提到的特征选择算法筛选后，剩下的指标分别是现金比率（X_3），总资产增长率（X_7），净利润增长率（X_8），可持续增长率（X_{11}），流动资产比率（X_{12}），现金资产比率（X_{13}），资产报酬率（X_{18}），净资产收益率（X_{21}），成本费用利润率（X_{24}），营运指数（X_{30}），经营杠杆（X_{33}），综合杠杆（X_{34}）。我们可以发现被选出的指标基本上都是财务指标，这就说明非财务指标在识别财务舞弊方面没有显著效果，故在以后的舞弊识别中也可以只考虑财务指标。把上述分类指标代

入各分类模型，结果如表 3-3 所示：

表 3-3 特征筛选后各个模型的评估效果

基础模型	准确率（%）	精确率（%）	召回率（%）	F_1 值（%）
SVM	91.3186	94.4015	88.3568	89.8908
KNN	89.2307	88.0866	86.7367	87.6674
LG	88.0769	88.7814	84.7233	86.1690
RF	92.1978	98.9712	88.7079	90.7691
ANN	92.1428	89.9159	90.7889	91.1594

从上表可以看出，经过特征选择后的准确率都比初始指标下的各个模型要高，尤其是神经网络模型，准确率从之前的 77.75%，提升到了现在的 92.14%，说明经过特征选择算法筛选后，筛选掉了一些不显著的指标，各模型的效果才有了显著性的提高。从准确率来看，发现随机森林和神经网络的识别已经取得了较好的效果，分别达到了 92.20%和 92.14%，而逻辑回归的效果最差，只有 88.08%，对于它来说，对模型进行特征选择后的模型效果没有太大的提升。总体而言，使用特征选择算法使得模型的评估效果达到了更好的水平，说明初始指标中确实存在冗余现象，在之后的研究中可以把冗余的指标剔除。

第三节　财务舞弊综合识别模型的构建及识别效果分析

上一节我们采用五种基础模型分别构建了财务舞弊的单一识别模型，并取得了较好的识别效果。下面为了进行更深层次的研究，进一步取得更好的识别效果，我们还将引入综合识别模型对财务舞弊进行识别。因此本文尝试将五种单一模型进行综合，以期获得一个具有更好的财务舞弊识别效果的综合识别模型。

一、财务舞弊综合识别模型理论介绍

在传统的数据挖掘有监督学习算法中，我们期望得到一个各方面都表现良好的算法模型，但往往现实中的模型都各有其优缺点。这时候就急需一个比较全面而稳定的算法，于是集成学习法应运而生。集成化教学就是组合这里的几个弱监测模式以求获得一种识别效果更好更全方位的强监测模式，集成化教学潜在的思维是即使有一种方法的预测结果不够准确，也能通过其他的方法进行纠正。集成中包括了各种类型的“个体学习器”和组建成的各种“组建学习器”。

经过以上阐述，我们已经初步认识了集成学习，简要而言，集成化计算只是先培训了一堆单一学习器，而后再利用某些方法将他们的训练结果加以综合。接下来，我们就来认识一下最常用的结合策略：

（一）平均法

最常见的方法为平均法，它是对于数值型数据的输出 $h_i(x) \in R$ 最常用的方法。

--简单平均法：

$$H(x)=\frac{1}{T}\sum_{i=1}^{T}h_i(x) \tag{3-8}$$

--加权平均法[28]：

$$H(x)=\sum_{i=1}^{T}\omega_i h_i(x) \tag{3-9}$$

其中ω_i为权重，通常ω_i要求：$\omega_i \geq 0$，$\sum_{i=1}^{T}\omega_i=1$

一般而言，在单一学习器之间相差不多时简单平均法比较合适，当单一学习器之间相差很大时加权平均法则比较合适。

（二）投票法

投票法也就是从个别学习器的预测结果中取其众数所在的类别，这也是运用的最多的组合方法。

--绝对多数投票法：

$$H(x)=\begin{cases} c_j, & if\ \sum_{i=1}^{T}h_i^j(x)>0.5\sum_{k=1}^{N}\sum_{i=1}^{T}h_j^k(x); \\ reject, & otherwise \end{cases} \tag{3-10}$$

即属于某种类型获得的票数最多，则预测结果就为该类型。

--相对多数投票法：

即选择投票数最多的那个类别，若几个类别同时出现多个票数最多，则任选其一。

$$H(x)=c_{\arg_j \max\sum_{i=1}^{T}h_i^j(x)} \tag{3-11}$$

---加权投票法：

$$H(x)=c_{\arg_j \max\sum_{i=1}^{T}\omega_i h_i^j(x)} \tag{3-12}$$

与加权平均法类似。

（三）学习法

当数据过多时，上面介绍的几种方法使用起来便显得非常繁琐，一种更加科学高效的方法是学习法。其经典代表形式是stacking学习法，在stacking中我们把个体学习器称为初级学习器，用于结合的学习器称为次学习器或者元学习器，它的主要思想为：先从初始数据集中培训出初次教学器，而后产生一种新的数据集中用来培训次级学习器。产生的新数据集中，初级学习器的输出被当成样例输入特征，而原始数据中的特

[28]加权平均法是指给不同的变量设定相应的权重，并通过权重和变量相结合的方式进行组合。

征会随着初级学习器的输出进入下一个层级的学习器。具体如图3-7所示[29]。

输入： 训练集 $D=\{(\boldsymbol{x}_1,y_1),(\boldsymbol{x}_2,y_2),\ldots,(\boldsymbol{x}_m,y_m)\}$;
初级学习算法 $\mathfrak{L}_1,\mathfrak{L}_2,\ldots,\mathfrak{L}_T$;
次级学习算法 $\mathfrak{L}$.

过程：

```
for t = 1,2,...,T do
    h_t = L_t(D);
end for
D' = ∅;
for i = 1,2,...,m do
    for t = 1,2,...,T do
        z_it = h_t(x_i);
    end for
    D' = D' ∪ ((z_i1, z_i2, ..., z_iT), y_i);
end for
h' = L(D');
```

输出： $H(\boldsymbol{x})=h'(h_1(\boldsymbol{x}),h_2(\boldsymbol{x}),\ldots,h_T(\boldsymbol{x}))$

图 3-7 stacking 集成学习流程图

二、综合识别模型的评估效果

上文共讲述了三种关于综合识别模型的结合策略，下面我们把经过特征选择后的指标代入三种结合策略中得到如表 3-4 所示结果。

表 3-4 各种综合识别模型的评估效果

综合模型	准确率（%）	精确率（%）	召回率（%）	F_1 值（%）
平均法	90.5934	92.0313	87.8627	89.1311
投票法	91.3187	79.1262	94.4015	86.0915
Stacking 集成法	98.1428	92.9159	97.7889	95.1594

其中平均法由于个体学习器的性能比较接近，所以我们这里直接选用简单平均法，直接将五种模型的结果进行加总后除以模型个数。

投票法选用的是加权投票法，即不仅要考虑每种单一识别模型的识别结果，还要根据他们的识别准确率来设定相应的比重，这样经过投票法得到的结果会更加可靠。

学习法选用的是基于 stacking 的集成学习法，首先把全部舞弊样本和非舞弊样本按照 7：3 的比例划分成训练集和测试集，分配完成后把用于训练的数据分成五份，把每个部分的数据都带入初级学习器，这里我们经过不断测试选择的初级学习器为神经网络、支持向量机、随机森林和 K 近邻算法，然后将初级分类器的预测概率值作为次级分类器的输入，这里的次级分类器选择的是 Logistic 回归，即逻辑回归得到的结果为最终的预测结果。最后用原始数据的测试集代入模型来对该 Stacking 集成学习法进行测试评估，具体如图 3-8 所示。

[29]本图来自于周志华老师的《机器学习》中对 stacking 集成学习的介绍。

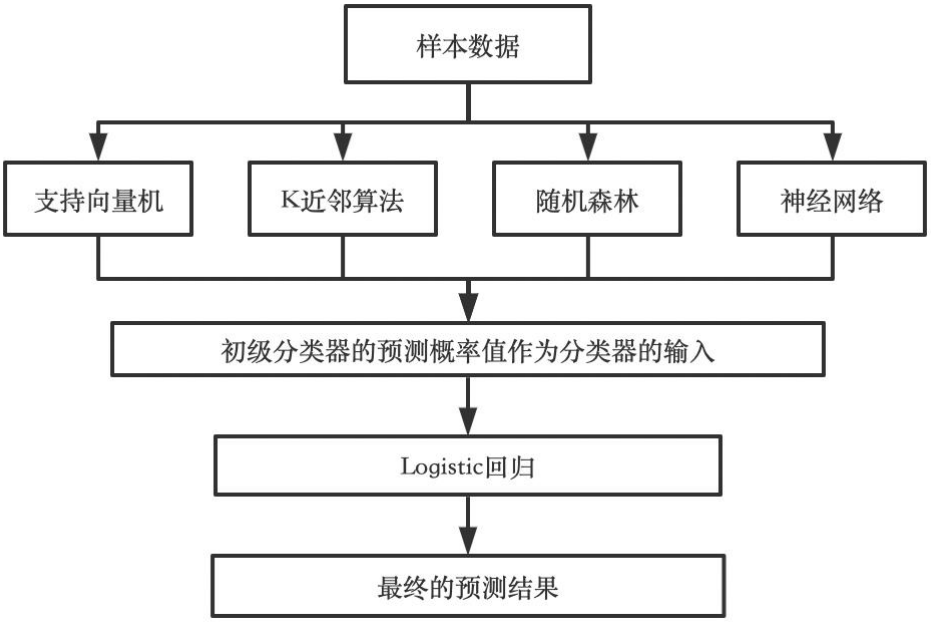


图 3-8 stacking 集成流程图

从表 3-4 可以发现，基于平均法和投票法的组合方式与单一模型的效果类似，并没有得到很好的提升，但是它们比单一模型的识别效果更加综合和全面，不容易出错。而 stacking 集成后的模型效果要比之前只采用单一的模型准确率要高，不论是从准确率、精确率还是 F_1 值，都可以说明集成效果比较好，识别效果也有了显著性的提高，他的识别准确率达到了 98.14%，召回率也达到了 97.79%。

第四节　模型识别效果的对比分析

一、基于混淆矩阵的指标对比分析

由前所述，我们可以得到经过特征选择后的指标具有更好的识别效果，并且已经得到了识别效果比较好的单一识别模型和综合识别模型，为了进一步对模型进行比较，下面我们将经过特征选择后的指标代入单一识别模型和综合识别模型的结果汇总到一起进行综合对比分析，如下表 3-5 和图 3-9 所示：

表 3-5 各种模型的评估效果

识别模型	准确率（%）	精确率（%）	召回率（%）	F_1 值（%）
SVM	91.3186	94.4015	88.3568	89.8908
KNN	89.2307	88.0866	86.7367	87.6674
LG	88.0769	88.7814	84.7233	86.1690
RF	92.1978	98.9712	88.7079	90.7691
ANN	92.1428	89.9159	90.7889	91.1594
平均法	90.5934	92.0313	87.8627	89.1311
投票法	91.3187	79.1262	94.4015	86.0915
Stacking 集成法	98.1525	92.5459	97.7739	95.1784

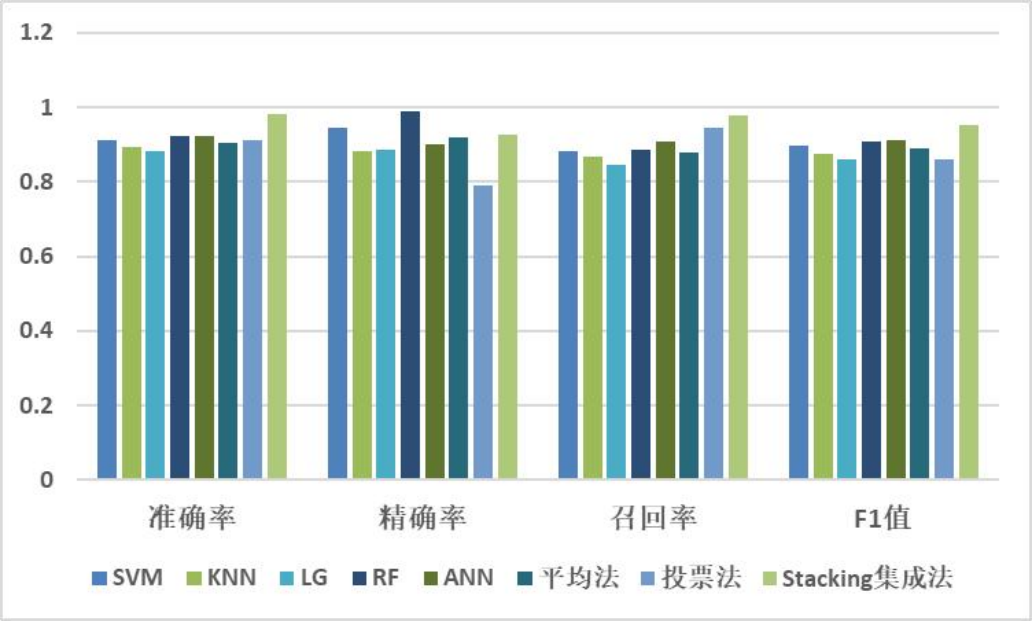


图 3-9 单一分类器和综合分类器的性能对比图

由表 3-5 和图 3-9 可以看出，基于 stacking 集成学习的综合识别模型的各项指标均高于单一识别模型和其他两种综合识别模型。在单一识别模型中，相较于其他四个模型随机森林的各项指标都达到了较高水平，分别为 92.19%、98.97%、88.70%和 90.77%。与综合模型对比来看，综合模型在准确率指标中，比最高的随机森林的模型高出 6.0%，具体为 98.15%；在精确率方面，比最高的基于 BP 神经网络的模型高出 7.6%，具体为 97.77%；在 F_1 分值方面，综合模型比基于 BP 神经网络的模型分别高出 4.0%，具体为 95.18%。财务舞弊综合识别模型是各种基础模型综合判断的成果，克服了财务舞弊单个判断模式的内在缺点，其可靠性相比于单个判断模式更高，有利于提高应用价值。因此，本文构建的基于集成学习下的财务舞弊综合识别模型达到了较好的财务舞弊识别效果。

二、基于综合 ROC 曲线的对比分析

由上文分析可知，评判模型的识别可以从两个方面来分析，一方面是通过混淆矩阵得出的各项指标，另一方面则是通过 ROC 曲线和 AUC 值进行判别。为了更加清晰直观的看出财务舞弊的识别效果，我们利用 python 的画图软件将多个模型的 ROC 曲线绘制在图 3-10 所示的一张图中：

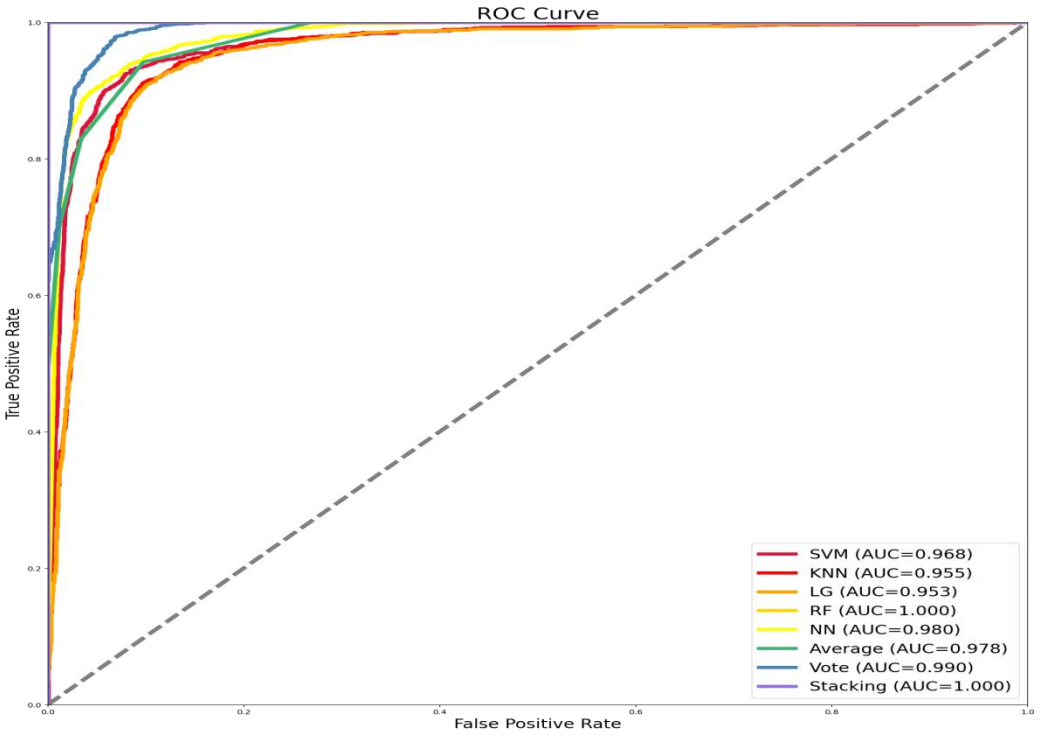


图 3-10 ROC 曲线的综合对比图

由图 3-10 可以知道，最靠近 ROC 曲线左上角的依次是基于 stacking 集成学习的综合模型，随机森林模型，基于投票法的综合模型，神经网络模型，基于平均法的综合模型，逻辑回归，支持向量机，和 K 近邻算法，说明 stacking 集成学习的综合模型和随机森林的识别效果最好，逻辑回归模型对于财务舞弊行为的识别效果最差。

其次可以看 AUC 曲线下的面积，即 AUC 值。我们可以发现基于单一模型的识别效果中随机森林的识别效果是最好的，它的 AUC 值达到了 1，而在综合模型的识别效果中，基于 stacking 集成学习得到效果是最显著的，进一步可以看出基于 ROC 曲线和 AUC 值的判别效果与基于混淆矩阵的判别效果是一致的。因此可以认为基于 stacking 集成学习的综合模型达到了最好的财务舞弊的识别效果，可以将此模型用于目前证券市场中财务舞弊行为的识别。

第五节　基于 stacking 集成学习综合识别模型的预测

一、预测样本的选取和数据预处理

通过上一节对单一识别模型和综合识别模型效果的对比分析，我们可以得到基于 stacking 集成学习的综合识别模型达到了最好的舞弊识别效果。为了进一步验证该方法是否普遍适用于目前上市公司的财务数据，也就是说对于不同的财务相关数据该方法能否取得同样好的实验结果。如果研究所得的准确率、精准率等指标仍处于较高水平，那么我们有理由认为该方法可以用于目前证券市场上市公司的财务舞弊的识别。

因此我们从 2021 年年末的上市公司研究的相关系列表中选取了 3153 家上市公司的财务指标数据，根据前文对指标进行特征提取的结果，我们选择了总资产增长率，

现金比率，净资产收益率，成本费用利润率，营运指数，可持续增长率，经营杠杆，现金资产比率，资产报酬率，净利润增长率，流动资产比率和综合杠杆这几个指标，以避免信息冗杂对就过造成影响。然后分别对数据进行缺失值处理和标准化处理，再把处理过的数据直接带入 stacking 集成学习综合识别模型。

二、预测结果的分析

通过前面对预测样本的选取和数据的预处理，我们把 3153 条上市公司的财务数据按照 7：3 的比例分为训练集和测试集，然后把训练集直接代入到基于 stacking 集成学习的综合识别模型，再对测试集进行预测，得到如表 3-6 所示结果（由于数据量太大，所以此表只展示前 20 条数据）。

表 3-6 基于 stacking 集成学习模型对财务舞弊的预测

证券代码	公司简称	预测类别	实际类别	预测结果
000005	ST 星源[30]	0	0	True
000010	美丽生态	1	1	True
000036	华联控股	1	1	True
000046	泛海控股	1	1	True
000056	皇庭国际	1	1	True
000070	特发信息	1	1	True
000078	海王生物	1	0	False
000150	宜华健康	1	1	True
000403	派林生物	1	1	True
000408	*ST 藏格	0	0	True
000410	ST 沈机	0	0	True
000413	东旭光电	1	1	True
000416	民生控股	1	1	True
000426	兴业矿业	1	1	True
000502	绿景控股	0	0	True
000504	南华生物	1	1	True
000506	中润资源	1	1	True
000510	新金路	0	0	True
000518	四环生物	1	1	True
000521	长虹美菱	1	1	True

由实验结果可知，参与本文训练的 946 条数据中有 889 个样本为非财务舞弊公司的样本，有 857 条识别出来为非财务舞弊公司，所以对于非财务舞弊公司的识别准确率为 96.40%，总共发生舞弊的企业有 57 家，被识别正确的有 51 家，所以对于财务舞弊公司的识别准确率为 89.47%，总的正确识别总数为 908 个样本，占测试集总数的 95.98%；它的准确率、精确率、召回率和 F1 值分别为 95.98%、96.40%、99.30%和 97.83%。由此可知，把基于 stacking 集成学习的综合识别方法用于财务舞弊的识别取得了较好的效果，说明该方法可以普遍适用于目前证券市场上财务舞弊的识别。

[30]ST 指的是财务状况或其他指标异常的股票，*ST 则表示有退市的风险。

结论

一、研究结论

本文选取了 2011-2020 年因财务舞弊受到惩罚和披露的上市公司的违规信息，并且为了保证样本的质量避免重复样本，最终确定了 2022 个舞弊样本。为了实验的完整性，根据配比原则选择同年、同规模和同行业的 4044 个非舞弊企业的数据作为对照；其次，按照全面性、可行性和客观性从财务指标和非财务指标层次对公司整体情况进行刻画，初步选出了 42 个初始指标，然后进行显著性检验和相关性检验后筛选出了 26 个指标。将这 26 个指标分别用树算法、RFE 搜索算法、lasso 回归和随机森林进行特征选择；然后，将四种特征选择算法选择的指标与未经指标筛选的原始指标相结合，形成的数据集分别在支持向量机、逻辑回归、随机森林、神经网络和 K 近邻算法这五种单一识别模型下进行检测。为了得到更好的模型识别效果，我们基于之前的单一识别模型用集成学习的方法构建了三种不同的综合识别模型。通过前面几个章节的分析，我们可以得到以下结论：

（一）舞弊类型多样化并且行业分布存在明显差异。把国泰安数据库中的舞弊样本数据经过整理分析，可以得到目前上市公司舞弊的类型和手段多种多样，但数量最多的前三项分别是违规买卖股票，推迟披露和重大遗漏，归根结底是为了获取某种不正当的利益。从行业类别上看制造行业舞弊数量最多，占了 67.46%，远超其他行业。所以审计人员应对制造业的上市公司进行重点监督。

（二）在财务舞弊识别指标的变量选取方面，违规识别的关键变量主要集中在财务指标当中。按照全面性、可行性和客观性从财务指标和非财务指标层次对公司整体情况进行刻画，初步选出了 42 个初始指标。然后通过指标的显著性检验，相关性检验以及 sklearn 中的树算法、RFE 搜索算法、lasso 回归和随机森林特征选择算法筛选后的指标分别是总资产增长率，现金比率，净资产收益率，成本费用利润率，营运指数，可持续增长率，经营杠杆，现金资产比率，资产报酬率，净利润增长率，流动资产比率和综合杠杆，说明这些指标对财务舞弊的识别更为重要。

（三）在单一数据挖掘模型的财务舞弊识别效果方面，随机森林在不同的条件下都具有较好的识别效果。总的来说，基于原始指标的各模型均取得了一定的识别成效，但效果不是很好。并且通过混淆矩阵和 ROC 曲线对五个模型进行比较发现，支持向量机和随机森林的准确率相对较高，分别达到了 84.07%和 86.21%，而支持向量机的准确率只有 72.47%。然后把经过特征选择后的数据代入各分类模型，发现随机森林和神经网络的识别准确率已经取得了较好的效果，分别达到了 92.20%和 92.14%。

（四）在财务舞弊识别模型方面，综合识别模型的识别效果普遍高于基础的单一模型，其中最好的是基于 stacking 集成学习法下的综合识别模型。并且通过混淆矩阵的相关评估指标和 ROC 综合曲线对模型的识别效果进行了分析。最后通过对比发现，

通过特征选择后的财务指标代入 Stacking 集成识别模型效果最好，在评估指标和适应性等方面都有不错的表现。说明 Stacking 集成算法集成了各个分类算法的特点达到了博采众长的目的，对于识别上市公司是否财务舞弊的问题更加高效和可靠。

二、建议

（一）对上市公司自身来说，无论其进行舞弊的原因是什么，以怎样的类型和手段进行舞弊，它终将成为一种隐患将公司推向深渊。它不仅会损害公司的形象和公司内部人员的利益，还会对投资者和整个社会公众造成严重的影响。因此，财务舞弊的遏制必须由上市公司自己来进行，比如优化公司内部结构，设立内部审查制度等。

（二）对审计人员来说，可以设定风险指标。某些重要指标如可持续增长率，净利润增长率，资产报酬率等可以直接影响财务舞弊的识别效果，并且与公司资金流动情况息息相关。因此，为了从源头上进行控制，审计人员可以设定风险指标体系，在进行审计的时候着重关注这些指标。

（三）对监管部门来说，不能只单纯关注公司的财务报表或者只是用最原始的统计方法，更加科学高效的方法是可以通过数据挖掘的相关算法来识别财务舞弊的行为。并且通过本文的研究发现基于 Stacking 集成识别模型效果最好，在评估指标和适应性等方面都有不错的表现，其结果的可信度也更高，因此监管部门可以将此模型用于以后的证券市场中财务舞弊行为的识别。

三、研究展望

虽然本文针对上市公司数据挖掘技术下的财务舞弊识别模型做出了许多研究，但仍然存在很多不足之处，所以本文提出了以下展望：

第一，本文对财务舞弊类型的划分只有两种：舞弊和非舞弊，将舞弊的样本计为 1，非舞弊的计为 0。这样的划分比较简单直接，不够具体，所以之后的研究可以根据舞弊严重程度、受处罚的金额等进行分类，从而可以将舞弊样本研究的更为深入。

第二，本文选用的指标不管是财务指标还是非财务指标都是数量指标，但是每年的年度财务报表和证监会的处罚公告中都包含了很多文本信息，如果能将这些文本信息和我们选择的数量指标结合起来分析，将更有利于识别财务舞弊，这也为数据挖掘领域和文本分析领域作出巨大贡献。

第三，本文构建的综合识别模型虽然已经取得了较高的识别效果，但是综合识别模型的识别是建立在基础识别模型的基础之上的。如果各基础模型的识别效果不好，那综合识别模型也不能取得好的效果。所以我们还得对基础模型进行深入的研究，不断对基础模型进行优化，这样才能不断提高综合识别模型的识别效果。

参考文献

[1]陈国亭. 我国上市公司财务报告舞弊因素的实证分析[J]. 审计研究, 2007, 56(5): 91-96 .
[2]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, 23(3): 88-103.
[3]陈俊, 王明. 上市公司会计欺诈预警模型的应用研究[J]. 财会通讯(学术版), 2005, 13(4): 50-54.
[4]成雪娇. 基于数据挖掘的中国上市公司财务舞弊识别研究[D]. 重庆:重庆理工大学, 2018.
[5]曹德芳, 刘柏池. SVM 财务欺诈识别模型[J]. 东北大学学报(自然科学版), 2019, 40(2): 295-299.
[6]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, 2(4): 6-9.
[7]邓庆山, 梅国平. 基于 B P 神经网络的虚假财务报告识别[J]. 财务研究, 2009, 11(10) :70-75.
[8]房琳琳. 财务困境上市公司财务报告舞弊预警模型研究[J]. 经济与管理研究, 2013, 19(30): 116-121.
[9]龚青青. 我国上市公司财务报告舞弊识别实证研究[D]. 江西:江西理工大学, 2016.
[10]郭月. 基于数据挖掘的财务舞弊识别研究[D]. 山西:山西财经大学, 2017.
[11]顾宁生. 基于 LVQ 神经网络的财务舞弊识别模型实证研究[J]. 价值工程, 2009, 10(1): 11-13.
[12]韩建光, 惠晓峰, 孙洁. 遗传算法选择性集成多分类器的企业财务困境预测[J]. 系统工程, 2010, 08(3): 9-15.
[13]贺颖. 基于偏最小二乘法--支持向量机的上市公司财务舞弊识别模型研究[D]. 新疆:石河子大学, 2010.
[14]黄世忠. 从 SAS99 看财务报表舞弊风险因素有效性分析[J]. 中国注册会计师, 2006, 10(6): 71-74.
[15]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, 14(10): 24-28.
[16]黄世忠. 上市公司财务造假的八因八策[J]. 财务与会计, 2019, 12(16): 4-11.
[17]刘立国, 杜莹. 公司治理与会计信息质量关系的实证研究[J]. 会计研究, 2003, 21(2): 28-36.
[18]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, 2006, 3(1): 102-105.
[19]鹿小楠, 傅浩. 中国上市公司财务造假问题研究[J]. 上海证券交易所研究中心,

2005, 2(12): 2-8.
[20]梁杰, 王璇, 李进中. 现代公司治理结构与会计舞弊关系的实证研究[J]. 南开管理评论, 2004, 3(7): 47-51.
[21]李康. 制造业上市公司财务报告舞弊识别混合模型研究[D]. 甘肃:兰州大学, 2011.
[22]李清, 任朝阳. 上市公司会计舞弊风险指数构建及预警研究[J]. 西安交通大学学报(社会科学版), 2016, 36(1): 36-44.
[23]李燕. 财务舞弊的若干特征分析[J]. 会计之友, 2006, 4(3): 64-65.
[24]李臣臣. 基于数据挖掘的上市公司财务舞弊的关联规则研究[D]. 吉林:吉林大学, 2011.
[25]李新朋. 数据挖掘技术在财务舞弊审计中的应用研究[D].上海:上海工程技术大学,2020.
[26]吕峻. 基于不同指标类型的公司财务危机征兆和预测比较研究[J]. 山西财经大学学报, 2014, 36(1): 103-113.
[27]年靖宇. 上市公司财务舞弊识别研究[D]. 安徽:安徽大学, 2019.
[28]潘梦雪. 基于随机森林的上市公司舞弊风险识别模型研究[D]. 杭州:杭州电子科技大学, 2019.
[29]钱苹, 罗玫. 中国上市公司财务造假预测模型[J]. 会计研究, 2015, 7(2): 18-25.
[30]秦江萍. 上市公司会计舞弊:国外相关研究综述与启示[J]. 会计研究, 2005, 6(11): 69-74.
[31]施华. 上市公司财务舞弊问题研究及监管对策--以近年经典案件为例[J]. 商业会计, 2019, 17(14): 8-11.
[32]施东晖. 上市公司十大管理舞弊案分析及侦查研究[J]. 审计研究, 2001, 12(15): 45-53.
[33]孙青霞, 贾瑞敏, 韩传模. 基于舞弊风险因子理论的会计舞弊识别研究[C]. 天津:天津财经大学出版社, 2010.
[34]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数据[J]. 中国注册会计师, 2010, 2(7): 23-29.
[35]吴勇,何长添,方君,张超.基于大数据挖掘分析的财务报表舞弊审计[J].财会月刊,2021(03):90-98.
[36]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(2): 49-52.
[37]许文静, 王君彩, 梁静. 博元投资会计舞弊行为及根源探究[J]. 中国注册会计师, 2017, 10(1): 117-120.
[38]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, 37(10): 138-144.
[39]杨贵军, 周亚梦, 孙玲莉, 等. 基于 Benford 律的 Logistic 模型及其在财务舞弊识

别中的应用[J]. 统计与信息论坛, 2019, 23(11): 1-7.
[40]杨敏. 对粉饰财务报表的识别[J]. 甘肃科技, 2006 , 10(22) : 194-197.
[41]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:浙江大学, 2021.
[42]张新民, 吴革. 财务报告舞弊的特征与识别研究[J]. 财贸经济, 2008, 12(12):30-40.
[43]张曾莲, 高雅. 财务舞弊识别模型构建及实证检验[J]. 统计与决策, 2017, 21(9): 172-175.
[44]郑伟宏, 李晓, 张婷, 等. 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, 11(22): 19-25.
[45]张雅宁,杜昀昊,王瑞,张晓萍,崔维康.基于数据挖掘的京津冀上市公司财务舞弊识别研究[J].江西电力职业技术学院学报,2020,33(02):126-127.
[46]赵英林, 陈素华. 基于因子分析的会计舞弊识别模型[J]. 山东财政学院学报, 2007, 12(6): 57-60.
[47]邹萱. 数据挖掘在上市公司财务造假识别中的应用研究[D]. 山东:山东大学, 2018.
[48]ALDEN M E, BRYAN D M, LESSLEY B J, et al. Detection of Financial Statement Fraud Using Evolutionary Algorithms[J]. Journal of Emerging Technologies in Accounting, 2012, 9(1):71-94.
[49]CALDERON T G, GREEN B P. Signaling Fraud by Using Analytical Procedures[J]. The Ohio CPA Journal, 1994, 53(4): 27-38.
[50]FANNING K M, COGGER K O. Neural network detection of management fraud using published financial data[J]. International Journal of Intelligent Systems In Accounting, Finance and Management. 1998, 7(5):21-24.
[51]GOTTLIEB O, SALISBURY C, SHEK H, et al. Detecting Corporate Fraud:An Application of Machine Learning[C]. CS 229 Machine Learning Final Projects, Autumn 2006.
[52]KOTSIANTIS S, KOUMANAKOS E, TZELEPIS D, et al. Forecasting fraudulent financial statements using data mining[J]. International Journal of Computational Intelligence, 2006, 11(2): 104-110.
[53]WANG Z, CHEN M H, CHIN C L, et al. Managerial Ability, Political Connections, and Fraudulent Financial Reporting in China[J]. Journal of Accounting and Public Policy, 2017(2): 141-162.
[54]ACHAKZAI M, ATIF K,JUAN P,et al. Using machine learning Meta-Classifiers to detect financial frauds[J]. Finance Research Letters,2021(2),48.

致谢

随着论文的即将结束，我在中南财经政法大学的学习时光也即将告一段落，首先我很感谢中南财经政法大学这两年给予我的教育和帮助，它给我们提供了学习氛围浓厚的校园环境和丰富教学资源，它开拓了我的视野，锻炼了各方面的能力，使自己成为了一个更优秀的人。

其次，我要感谢在论文的写作过程中帮助过我的每一位老师，我发现每一个老师都非常的负责，而且很有耐心，就算你的论文漏洞百出，他也会不厌其烦的指导你进行修正。在我论文撰写的各个阶段都离不开各位老师的辛勤付出。尤其我要感谢的是我的导师，我的论文题目就是和我的指导老师商量后确定的，从前期的数据收集整理，到中期的开题报告，再到后来正式论文的撰写，都是经过反复考虑确定的，并且他给我提出了很多珍贵的建议，使我整个论文撰写过程中都受益匪浅。

我的论文能有今天的成就，也来自于很多人的帮助。我要感谢我的父母，因为他们我才能在如此健康稳定的环境下成长，并且给了我上学的机会，是他们从小的悉心培养才有了我的今天；然后我要感谢我的室友和同门，是他们在我遇到问题时及时给予我帮助，特别是在一些深奥的领域，大家一起学习，一起讨论，共同进步，是他们让我感受到了学习的乐趣和友情的可贵。

至此，我的论文就正式结束了，我希望我的文章能为广大上市公司提供有价值的参考，我想再次感谢一下在写论文过程中帮助过我的人，我会在今后的学习生活中更加努力，以不辜负你们对我的期盼。