天津财经大学

专 业 学 位 硕 士 学 位 论 文

基于机器学习的上市公司财务舞弊识别模型研究

所属学院：统计学院

专业名称：应用统计

指导教师：尚翔

论文作者：姜涵

学　　号：2019670082

二〇二一年三月

分类号：
密　级：

硕士学位论文（硕士专业学位论文）

基于机器学习的上市公司财务舞弊识别模型研究

Research on Financial Fraud Identification Models of Listed Companies Based on Machine Learning

所 属 学 院：统计学院

年　　级：2019 级

学　　号：2019670082

论 文 作 者：姜涵

内容摘要

利益相关者了解公司运营状况的一个有效依据为财务报告，其反映了企业在一定时期内的财务状况和经营成果。然而，一些上市公司或是出于躲避监管的目的，或是出于公司管理层个人利益需要，通过多手段进行财务舞弊。财务舞弊不仅会误导投资者，使投资者信心降低，而且会扰乱市场秩序，对资本市场的有序发展产生阻碍。

在对财务舞弊识别相关文献进行梳理总结后，本文选取 2007 年 1 月 1 日至 2019 年 12 月 31 日 62 家虚列资产与虚构利润的舞弊公司，并依据配对样本与舞弊样本细分行业相同等标准构建配对样本。本文选取偿债能力、发展能力等六类财务指标以及治理结构、股权性质、内部控制三类非财务指标构建指标体系，利用决策树等进行指标筛选。此外，本文利用 SMOTE 过采样方法进行样本重建，利用 Bagging、GBDT、支持向量机三种方法分别构建模型，并依据召回率等指标评价模型效果。通过研究本文得到以下结论：

在模型识别效果方面，支持向量机、Bagging、GBDT 各模型识别准确率均在 70%以上。其中，径向基核函数的支持向量机模型与 GBDT 模型识别效果较好。在基于相关性进行指标筛选后，径向基核函数的支持向量机模型效果最好。在基于决策树进行指标筛选后，GBDT 模型效果最好。因财务报告舞弊的特殊性，模型应旨在尽可能地识别出舞弊样本，故召回率为本文的一个主要评价指标。在基于较少指标的情况下，GBDT 模型的召回率较好。因而在综合考虑数据获取成本等因素后，本文建议利用 GBDT 模型识别财务舞弊。而当数据易得时，则可考虑使用精度更高的支持向量机模型。在指标重要性方面，非财务指标中可重点关注前十大股东持股比例及未领取薪酬监事人数。而在财务指标中，资产报酬率、流动资产净利润率等 7 个指标对于财务舞弊的识别较为重要。此外，每股折旧和摊销、财务费用率以及长期借款与总资产比对于财务舞弊的识别更为重要。

关键词：财务舞弊　机器学习　识别模型

Abstract

An effective basis for stakeholders to understand the operating status of a company is the financial report, which reflects the financial status and operating results of an enterprise in a certain period of time. However, some listed companies either for the purpose of avoiding supervision, or for the personal interests of the management of the company, through multiple means to commit financial fraud. Financial fraud will not only mislead investors and reduce investor confidence, but also disrupt the market order and hinder the orderly development of the capital market.

After sorting out and summarizing relevant literatures on the identification of financial fraud, this paper selects 62 fraudulent companies with false assets and fictitious profits as of January 1, 2007 and December 31, 2019, and constructs a paired sample according to the standards of the paired sample and the same subdivision industry of the fraud sample. In this paper, six types of financial indicators such as debt paying ability and development ability and three types of non-financial indicators such as governance structure, equity nature and internal control are selected to construct an index system, and the decision tree is used to screen the indicators. In addition, SMOTE oversampling method is used for sample reconstruction in this paper, and three methods of Bagging, GBDT and support vector machine are used to build the model respectively, and the effect of the model is evaluated according to the recall rate and other indicators. Through the study of this paper, the following conclusions are drawn:

In terms of model recognition effect, support vector machine, Bagging and GBDT all have a recognition accuracy of more than 70%. Among them, the support vector machine model of radial basis kernel function and GBDT model have better recognition effect. After the index screening based on correlation, the support vector machine model with radial basis kernel function has the best performance. After index screening based on decision tree, GBDT model has the best effect. Due to the particularity of fraud in financial reporting, the model should be aimed at identifying fraud samples as much as possible, so recall rate is a major evaluation index in this paper. The recall rate of GBDT model is better when it is based on fewer indicators. Therefore, after comprehensively considering the data acquisition cost and other factors, this paper suggests using GBDT model to identify financial fraud. When data is readily available, a more accurate support vector machine model can be considered. In terms of the importance of indicators, non-financial indicators can focus on the shareholding ratio of the top ten shareholders and the number of supervisors who are not paid. In the financial indicators, the rate of return on assets, net profit rate of current assets and other seven indicators are important for the identification of financial fraud. In addition, depreciation and amortization per share, financial expense ratios, and long-term borrowings versus total assets are more important for the identification of financial fraud.

Key words: Financial fraud;Machine learning;Identification model

目录

第 1 章　问题提出

1.1 研究背景

我国股票交易市场自诞生以来越来越受到人们的关注。在选择股票时，投资者了解公司运营状况的一个有效依据为财务报告，其反映了企业在一定时期内的财务状况和经营成果。然而，财务舞弊问题在国内外资本市场屡禁不止。财务舞弊是指故意虚假陈述、遗漏重要事实或数据，以误导财务报表使用者并使其决策错误的行为，是一种带有欺骗性、预谋性的故意行为。财务舞弊行为存在的根本原因在于利益驱动，舞弊收益大于舞弊成本使企业选择进行财务舞弊。为了躲避监管或出于公司管理层个人利益的需要，一些上市公司采取虚假增加利润、资产等手段进行财务舞弊。此外，因企业财务舞弊手段不断更新且隐蔽性增强，财务舞弊隐蔽时长最长达七年，最短为一年以内，大部分财务舞弊公司的隐蔽时长为两年，财务舞弊披露年度相较于财务舞弊年度会有所推迟。财务舞弊不仅会误导投资者，使投资者信心降低，而且会扰乱市场秩序，对资本市场的有序发展产生阻碍。

美国安然公司和日本东芝公司等知名企业的财务舞弊事件在社会上引起了轩然大波。安然公司 2000 年的销售额已超过 1000 亿美元，拥有两万多名员工，其曾是全球最大的能源型公司之一，然而，正是这样一家发展前景大好的公司却因其财务舞弊问题在一个月内无奈破产。据美国监管机构调查结果显示，安然公司连续多年通过与关联方进行复杂交易虚构利润并隐瞒高额负债。这一事件导致安然公司股价暴跌并破产，相关人员也受到监禁等严重处罚。日本东芝公司自 1904 年成立以来发展迅速，20 世纪 90 年代更是完成了由家电行业向 IT 行业的转变，成为全球著名的电子产业领头企业，其公司治理也曾作为典型成功案例供人研究。然而，2015 年知情人向日本证监会举报该公司存在财务舞弊问题。其后，经日本证监会调查，东芝公司连续多年虚构利润。最终，东芝公司被处以巨额罚款，这一事件也导致东芝公司的股价急剧下跌。

在我国，财务舞弊问题同样常见。2020 年上半年，我国的上市公司中已有数十家因此问题受到了处罚。2015 年至 2017 年龙昕科技利用虚构交易等方式致使账面上存在相当数额的虚假应收账款，虚假增加大量收入，2017 年被康尼机电收购。2020 年 5 月，康尼机电受到 30 万元罚款等惩罚。2020 年 7 月，因惠而浦虚增其 2015 年及 2016 年利润，安徽证监局对惠而浦实施 40 万元罚款等惩罚措施。

国内外上市公司的财务舞弊事件在被多次禁止后仍继续存在且规模巨大。2019 年 7 月，

一度被认为A股市场白马股的康得新被曝财务舞弊，其在2015年至2018年的年度报告中通过虚假记载等方式造假规模达到百亿元。此消息引起社会一片哗然，高达119亿元的造假金额也是刷新了公众对于财务舞弊规模的认知。2020年4月，瑞幸咖啡自曝存在财务舞弊行为，舞弊金额达到22亿。瑞幸咖啡作为中国新兴连锁咖啡品牌，其新零售模式深受消费者的喜爱。该公司自2017年成立以来发展迅速，门店数目高速增长，2019年在美国上市。此次事件使得瑞幸咖啡股价暴跌，市值大幅下降。财务舞弊已成为全球性问题，高效、准确地识别企业财务舞弊行为，对投资者正确决策等具有十分重要的意义。

1.2 研究目的及意义

1.2.1 研究目的

舞弊手段专业性、隐蔽性的逐渐增强使得信息之间可能存在隐蔽信息，此时利用传统方法识别企业的财务舞弊行为会存在一定困难。机器学习通过设计模型与方法，从已知数据集中学习及寻找数据间的联系，并基于此预测未知数据的规律。其可将数据宝贵的隐藏信息提取出来，现已广泛应用于医疗、金融等多个领域。本文在现有文献的基础上，以财务舞弊相关理论为支撑，基于机器学习对上市公司财务舞弊识别模型进行比较研究，为利用机器学习识别财务舞弊这一领域做出一定贡献，同时为投资者等进行判断与决策提供一定的参考。

1.2.2 研究意义

（1）本文以我国2007年至2019年上市公司数据为研究对象，从财务与非财务两方面选取财务舞弊识别指标。考虑到在现实情况中出现财务舞弊问题的上市公司相对于未出现财务舞弊问题的上市公司数量较少的情况，本文利用过采样方法进行样本重建。同时，基于决策树等进行指标筛选，并利用Bagging、GBDT、支持向量机三种方法分别构建模型。模型识别效果较好且具有一定的可行性，丰富了利用机器学习识别财务舞弊这一领域。

（2）本文基于我国上市公司的真实数据建立财务舞弊识别模型。因财务舞弊披露年度相较于财务舞弊年度可能会有所推迟，本文可以帮助投资者更好地判断财务报表的真实性，降低因未识别出上市公司财务舞弊而做出错误决策的风险。

1.3 研究内容与研究方法

1.3.1 研究内容

本文基于现有国内外财务舞弊相关文献，研究利用机器学习算法识别财务舞弊。考虑到我国于2007年1月1日发布了新会计准则，本文选取2007年至2019年出现财务舞弊问题的上市公司作为舞弊样本，并依照一定标准构建配对样本。本文从财务、非财务两个方面构建指标体系。在对数据进行过采样处理后，本文利用多种机器学习算法构建模型，并依据不同评价指标评价各模型财务舞弊识别效果。

本文分为以下五个部分:

第一部分为问题提出，包括本文选题背景、研究目的及意义、研究内容及方法、本文的创新点等内容。

第二部分为现状描述，主要介绍了现有国内外关于财务舞弊问题的一些研究成果，为本文研究提供思路及方向。

第三部分为理论基础，对包括财务舞弊动因理论、有限理性理论在内的财务舞弊相关理论等进行了阐述。

第四部分为实证分析，主要介绍了数据来源、指标构建、指标筛选、模型选取、模型评价等内容。本文从国泰安数据库中选取2007年至2019年财务舞弊样本及非舞弊样本，从财务、非财务两个方面构建指标体系，并利用支持向量机、Bagging、GBDT三种方法分别构建财务舞弊识别模型，依据模型的评价指标比较各模型分类效果。

第五部分为总结与展望，对论文进行简单总结，同时，提出本文的不足之处以及未来可能的研究方向。

1.3.2 研究方法

在通过文献研究法对财务舞弊识别相关文献进行梳理总结后，本文利用机器学习算法进行财务舞弊识别模型的构建。具体如下：

（1）文献研究法

本文通过文献研究法对现有研究成果进行梳理总结，将财务舞弊识别相关文献分为模型构建与指标选取两个方面，为本文财务舞弊识别模型的构建提供思路。

（2）实证研究法

本文选取2007年1月1日至2019年12月31日违规类型中存在虚列资产与虚构利润的62家舞弊公司，并依据配对样本与舞弊样本细分行业相同等标准构建配对样本。本文选取偿债能力、发展能力等六项财务指标以及治理结构、股权性质、内部控制三项非财务指标构建指标体系。此外，本文利用SMOTE(Synthetic Minority Oversampling Technique)过采样方法进行样本重建，利用Bagging、GBDT、支持向量机三种方法分别构建模型，并

依据召回率等指标评价模型效果。

1.4 本文的创新点

本文的创新点如下:

（1）在模型构建方面，通过对已有文献的分析发现，Logistic 回归、支持向量机等方法在财务舞弊识别中运用较为广泛。本文分别基于支持向量机、Bagging、GBDT 构建财务舞弊识别模型，并依据所得结果对各模型识别效果进行比较研究。本文所选三种算法既包括在财务舞弊识别中运用较为广泛的模型，也有较为新颖的探索。

（2）在样本及指标选取方面，本文以我国 2007 年至 2019 年上市公司的年报数据作为样本的选取范围，研究跨度较长。此外，以往学者们大多在基于财务指标的基础上，考虑引入特定非财务指标并判断其对于模型识别效果的影响。本文选取偿债能力、发展能力等六项财务指标以及治理结构、股权性质、内部控制三项非财务指标构建指标体系，更为全面地考虑了相关影响因素。

第 2 章　国内外研究现状

2.1 国外关于财务舞弊问题的研究

2.1.1 模型构建

在模型构建方面，Beneish（1997）基于财务数据，利用 Probit 回归建立财务舞弊识别模型，模型预测精度为 75%[1]。Fanning 和 Cogger（1998）基于公开数据，在财务指标的基础上引入董事会信息等非财务指标，并利用人工神经网络构建财务舞弊识别模型。研究发现，相较于 Logistic 回归等模型，此方法对于财务舞弊的鉴别力更优[2]。Lin，Hwang 及 Becker（2003）对财务收益及其趋势变化进行分析，并构建模糊神经网络模型。此模型可以对人脑思维模式进行模拟，减少了审计人员的操作失误[3]。Kotsiantis，Koumanakos 及 Tzelepis 等（2006）以希腊公司为样本，基于 Logistic 回归、决策树等多种方法构建单一舞弊识别模型，并通过类似于投票的方式构建综合舞弊识别模型。结果表明，对于财务舞弊问题，与单一模型相比，综合模型的鉴别力相对更优[4]。Ophir Gottlieb 等（2006）选取美国的上市公司作为样本，分别利用支持向量机、Logistic 回归、贝叶斯建立模型。结果表明，支持向量机、Logistic 回归具有较好的识别效果[5]。Kirkos，Spathis 及 Manolopoulos（2007）以希腊制造业的上市公司财务报告为研究对象，选取若干财务指标并筛选至 10 个后，利用贝叶斯、决策树及神经网络三种方法构建模型。结果发现，贝叶斯方法具有最佳的识别能力[6]。Pediredla，Ravi 及 Rao 等（2011）分别利用 BP 神经网络、概率神经网络等六种方法构建识别模型。结果发现，六种方法中识别效果最优的为概率神经网络[7]。Gill 及 Gupta（2012）分别利用决策树、神经网络等方法构建识别模型。研究表明，神经网络模型具有最佳的识别能力[8]。Alden，Bryan 及 Lessley 等（2012）分别基于遗传算法、Logistic 回归构建模型，结果表明，基于遗传算法的模型效果更优[9]。

2.1.2 指标选取

在指标选取方面，Loebbecke 和 Willingham（1988）构建了 L/W 模型。此模型将舞弊风险划分为企业内部的组织管理、管理层的舞弊动机、管理层的价值观三大类，共计 46 个舞弊风险因素[10]。Beaver（1996）基于 79 家财务舞弊公司及相同数量的非舞弊公司，利用现金流量指标对企业是否舞弊进行判断，研究发现现金流量与债务总额指标可作为舞弊识别的有效因素[11]。Beasley（1996）基于相同数量的舞弊样本与配对样本，利用 Logistic 回归研究发现独立董事所占比例越低，公司越易存在财务舞弊行为[12]。Green 和 Choi（1997）

选取规模指标及比率指标，基于神经网络构建模型。研究发现，比率指标的舞弊识别模型训练效果要优于规模指标[13]。Dechow 等（2011）以 1982 年至 2005 年受到美国证监会处罚的 676 家公司为样本进行实证分析。指标数量为 22，包含表外活动、应计质量等方面。其后建立 F-score 舞弊识别模型，模型的识别准确率较高，在实践中得到了一定的应用[14]。

2.2 国内关于财务舞弊问题的研究

2.2.1 模型构建

在模型构建方面，岳殿民（2008）以 2002 年至 2006 年 90 家财务舞弊公司为研究对象，运用 Apriori PT 等关联规则算法对舞弊的手段进行研究，并基于结果制定识别财务舞弊的方法[15]。邓庆山（2009）以 1999 年至 2006 年上市公司数据为样本选取范围，将 1999 年至 2002 年数据划分为训练集，其余年份数据划分为测试集。从分类和聚类两个角度构建虚假财务报告识别模型，其中，在分类方面，选取神经网络等四种分类方法，发现朴素贝叶斯及神经网络的识别效果相对更好。在聚类方面，选取的模型为 V-KSOM 模型，结果发现此聚类模型对于识别虚假财务报告可行[16]。李秀枝（2010）基于 2004 年至 2009 年我国上市公司数据研究财务舞弊的特征以及如何识别财务舞弊，利用 LibSVM 算法分别构建调整利润和非调整利润的财务舞弊识别模型。结果表明，针对调整利润的模型，RBF 核函数效果更优，而针对非调整利润的模型，线性核函数效果更优[17]。金花妍（2013）分别基于支持向量机与 Logistic 回归构建财务舞弊识别模型并检验模型泛化能力，其中支持向量机模型测试精度都在 94%以上[18]。任朝阳（2016）结合案例推理的方式构建舞弊识别模型，发现相较于其他 Logistic 回归模型，非线性主成分 Logistic 回归模型的效果更优[19]。冯炳纯（2019）分别选取 2009 年至 2016 年 226 家舞弊上市公司和非舞弊上市公司，选择 Relief 和 Boruta 两种算法进行特征选择，同时选择支持向量机、随机森林等四种分类算法。结果表明，当 Relief 算法与随机森林算法结合使用时，模型的识别效果最好[20]。王威（2020）基于 2010 年至 2017 年 60 家舞弊上市公司和 120 家非舞弊上市公司的年报数据，对比稀疏组 Lasso-logistic、Lasso-logistic 等四种模型的识别效果，发现稀疏组 Lasso-logistic 模型总准确率最高[21]。

2.2.2 指标选取

在指标选取方面，曹利（2004）选取 30 家舞弊上市公司作为舞弊样本，55 家行业相同规模相近的非舞弊上市公司作为配对样本。基于舞弊动机、治理结构及财务指标三个角度构建指标体系，并构建 Logistic 回归模型。结果发现，筹资动机强烈、治理结构薄弱、

财务异常的企业更易出现财务舞弊问题[22]。陈庆杰（2012）在财务指标、审计意见指标等的基础上引入经理人性别、受教育年限等 9 个经理人特征指标，所选取的模型为神经网络。研究结果表明，相较于基于传统指标的模型，引入经理人特征后模型识别效果更优[23]。张曾莲，高雅（2017）选取 61 家舞弊、非舞弊上市公司，引入自愿性信息披露这一综合指标，运用 Logistic 回归构建模型。其中，自愿性信息按性质可分为预测、战略、拓展财务及关键非财务四个方面。模型总准确率为 73.8%[24]。刘志洋，韩丽荣（2018）利用 2007 年至 2015 年制造业上市公司数据研究依据历史信息识别财务舞弊，将 2007 年至 2013 年数据划分为训练集，2014 年及 2015 年数据划分为测试集，舞弊样本及配对样本各 152 个。在指标选取方面，财务指标均采用 2005 年至 2015 年各指标的标准差。采用线性、非线性主成分方法对数据进行降维，采取 Logistic 回归、支持向量机方法对数据进行分类。结果表明，相较于静态数据，利用异常历史信息识别财务舞弊的效果更佳[25]。

2.3 文献总结

通过梳理国内外相关文献可知，学者们对于财务舞弊问题的研究角度也有所不同。有学者侧重于财务舞弊识别模型构建的方法选取，也有学者侧重于财务舞弊识别模型的指标选取。从方法选取角度来看，Logistic 回归、支持向量机的运用较为广泛。从指标选取角度来看，国内外学者大多在基于财务指标的基础上，考虑引入特定非财务指标并判断其对于模型识别效果的影响。基于现有文献，在方法选取方面，本文考虑分别基于支持向量机、Bagging、GBDT 构建财务舞弊识别模型。在指标选取方面，本文从财务指标与非财务指标两方面构建指标体系，综合考虑财务与非财务两方面因素对于财务舞弊识别的影响。

第 3 章　理论基础

3.1 财务舞弊相关理论

3.1.1 舞弊动因理论

舞弊是一种带有欺骗性、预谋性的故意行为，其依靠一些不正当的隐蔽手段操控财务信息，使财务信息无法确切地反映企业真实的价值。财务舞弊是指故意虚假陈述、遗漏重要事实或数据，以误导财务报表使用者并使其决策错误的行为，其即是一种带有欺骗性、预谋性的故意行为。舞弊动因理论是研究财务舞弊问题的前提及识别防范舞弊行为的理论依据。在学者的不断研究中，舞弊动因理论现已较为成熟，其中的冰山理论、三角形理论等应用较为广泛。

（1）冰山理论

1989 年 Bologna 提出冰山理论。该理论认为，财务舞弊如一座冰山，海平面以上为舞弊结构，以下为舞弊行为。舞弊结构容易识别，其主要为财务资源等内部管理问题。舞弊行为则主要为价值观等个体因素，不易识别且易被忽视，因而危险性更高。冰山理论认为，人们看到的只是显性的舞弊结构部分，而隐性的舞弊行为是真正危险的部分。相较于客观存在的舞弊结构部分，舞弊行为部分更应在识别财务舞弊时引起人们的关注。

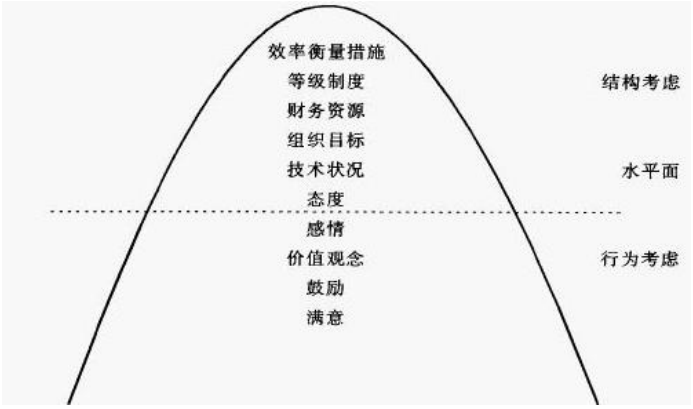

图 3.1 舞弊冰山理论

（2）三角形理论

1995 年 Albrecht 提出三角形理论。相较于冰山理论，该理论将舞弊结构归类为机会，舞弊行为归类为个人压力和借口。该理论认为，在上述三因素的综合作用下，财务舞弊产生。压力、机会及借口缺一不可，只有当三因素均存在时，舞弊才会发生，三者共同构成了舞弊三角形。其中，压力既包括公司压力，如公司需达到投资者的预期业绩、公司的收入减少影响股价、公司的现金流紧张等，也包括个人压力，如财产意外损失、存在高额负

债、即将失业等。机会因素是指舞弊行为可以不被发现或免于惩罚的条件，如公司缺乏内部控制、信息不对称、工作质量难以辨认等，这些都为财务舞弊创造了客观条件。借口因素是指舞弊者为使舞弊行为符合道德准则，而对舞弊行为进行的自我合理化。企业财务舞弊常用的借口包括保护投资者利益、帮助公司缓解财务困境等。

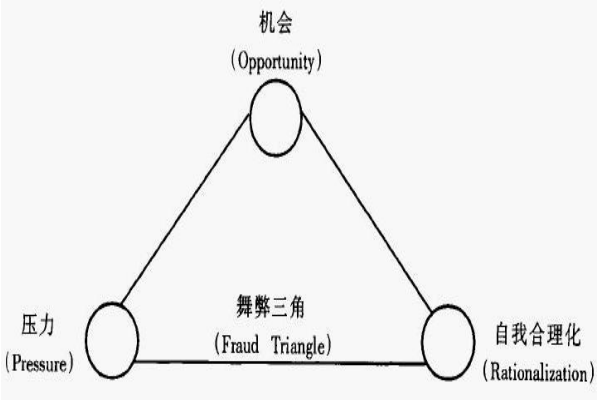

图 3.2 三角形理论

（3）GONE 理论

GONE 理论 1993 年由 Bologna 提出。GONE 分别为四个舞弊因素英文的首字母。相较于舞弊三因素理论，该理论将舞弊压力归类为需要，舞弊借口归类为贪婪，并增加了暴露这一因素。该理论认为，舞弊的风险程度由四因素共同决定，四者的重要程度相同。其中，贪婪因素是指舞弊者满足个人利益的欲望，是驱动舞弊者仅考虑个人利益进行舞弊的内在力量。机会因素与舞弊三因素理论中的机会因素类似。需要因素是指舞弊者所承受的外部压力，如财产意外损失、存在高额负债、即将失业等。暴露因素一方面是指舞弊暴露的可能性，另一方面是指暴露后对于相关人员的惩罚程度。贪婪和需要为个人主观因素，是企业财务舞弊存在的内因。机会和暴露为外在客观因素，是企业财务舞弊存在的外因。

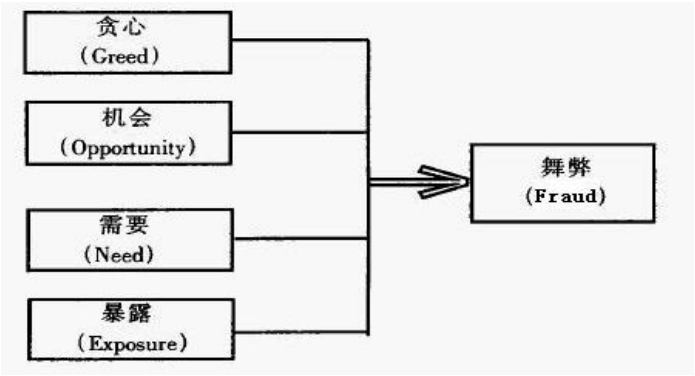

图 3.3 GONE 理论

（4）舞弊风险因子理论

Bologna 等人提出了舞弊风险因子理论。相较于舞弊四因素理论，该理论将机会、暴露归类为一般风险因子，贪婪、需要归类为个别风险因子。该理论认为，当上述两个因素

同时存在，并且舞弊者认为可以获得利益时，财务舞弊就会产生。其中，舞弊的机会、被揭发的可能性等属于一般风险因子，能够被控制。舞弊的动机、个人品质等属于个别风险因子，难以被控制。

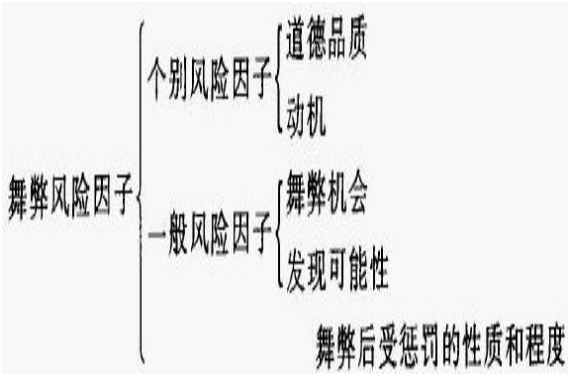

图 3.4 舞弊风险因子理论

上述四理论对于舞弊动因的分析逐渐有了内和外、整体和个体的区分，对舞弊因素的阐述越发具体全面。此外，上述四理论同时强调舞弊是各个因素共同导致的结果，缺失任何一个因素都不足以促使舞弊发生。可见，导致财务舞弊发生的因素是复杂的，而在进行舞弊防范时，只需破坏其中一个因素就可以阻止财务舞弊的发生，这为识别财务舞弊提供了思路。

3.1.2 有限理性理论

经济学家阿罗最初提出了“有限理性”这一概念。阿罗认为，多个复杂因素共同决定了个人的有限理性，如知识的不完整性、环境的复杂性以及对于困难的预测程度等。由于经历、性格等的差异，不同人对同一情况的判断和预测都会不一致。此外，任何人的能力都是有限的，其无法在各个领域都达到专业水平，这便会在一定程度上降低其决策的准确性。在非个人交换的形式中，任何人都会遇到陌生未知的领域。在不熟悉的复杂环境中，实现完全理性是非常困难的。1978 年，西蒙将有限理性进行了理论提升，提出了“社会人”这一概念。西蒙认为，“经济人”概念为完全理性假设，存在一定问题。此概念假定个人具有有序稳定的偏好体系，该体系能够为其准确定制行动的方案。然而，在实际中，决策者所掌握的信息都是有限的。交易次数越多则不确定性越高，决策者所掌握的信息也就越不完全，因而其无法寻找到全部方案并完全地预测各方案结果。此外，决策者的决策目标通常模糊复杂。因而，在决策过程中，决策者大多使用满意准则，而非最优准则。考虑到成本、环境等多方面因素，决策者会在各个备选方案中选择满意方案。

西蒙进一步对程序理性和结果理性进行了区分。程序理性不过度强调结果的准确性，而是注重过程合理与否，其指行为程序符合标准。基于程序理性，只要保证了过程符合标准，那么结果相应也会在合理范围内，不会出现过大的偏差。结果理性则考虑结果是否达

到既定目标，而不强调目标实现的过程。西蒙认为，复杂未知的环境中，个人不应过度强调结果的实现，而应当依据理性的行为程序完成既定目标。在合理完成行为程序的情况下，程序理性可以促使结果理性。程序理性与结果理性之间的区别对财务舞弊行为的分析具有重要的指导作用。财务信息反映了企业的财务状况、经营状况等，其是公开的，并且具备结果理性的特性。信息使用者依据此信息制定经济决策，其着眼点在于结果，而不是程序理性，忽略了结果的形成过程中对于程序理性的控制。企业业绩评价体系中对财务指标的过分强调在某种意义上成为了财务舞弊的诱因。各方对于企业的评估大多依据特定财务指标，而忽略了非财务指标，且不考虑产生此结果的程序是否合乎规定。此种重视结果理性忽视程序理性的行为容易导致财务舞弊行为的发生。该理论可为后文非财务指标的选取提供理论基础。

3.1.3 委托代理理论

委托代理理论是企业实施内部控制的理论基础，为契约理论的分支。契约理论认为，企业是由多个契约组成的统一体。企业由契约联结而成，是各利益主体达成的共识，以使其参与企业的经营活动。契约的主体签署一系列的契约后构成一个整体，将其资本、信息等投入于企业，并期望获得相应回报。就企业而言，构成企业契约的主体包括股东、管理者、债权人等。各利益主体依据财务信息达到正确判断企业经营情况的目的。财务信息在契约的签订过程中意义重大，各主体在签订契约前需依据企业财务信息判断是否与其签订契约。管理者和委托人签订了契约并承担受托责任，当经营出现问题时，为避免承担责任，管理者进行财务舞弊。此外，企业内部存在多个契约关系，不同契约的主体也存在差别，各主体间博弈与合作共存。不同契约间甚至同一契约之中都难以避免利益的冲突，因而无法使全部契约主体的利益都得到满足。出于降低履行契约所需成本的目的，财务舞弊行为发生。该理论可为后文财务指标的选取提供理论基础。

市场经济的高速发展使得企业的规模、交易范围等迅速扩大，企业所有者的能力、时间等产生了局限，无法完全亲自地对企业活动进行控制。因而企业所有者委托具备专业知识与能力的人行使自身权利，代为经营企业。委托代理关系就此产生。委托代理双方对于信息的持有程度不平衡。管理者处于信息优势地位，掌握大量的经营状况相关信息。当相关约束机制缺乏时，管理者为获取更多利益，会将信息优势转化为自身利益，使得只有其自身掌握企业财务信息的真实状况。其中，相关约束机制包括内部控制等，内部控制的一大出发点在于防止财务舞弊，内部控制完善可在一定程度上保障财务信息的真实准确。而当内部控制等相关约束机制缺乏时，相较于利益相关者的资源，财务信息成为了管理者滥

用权力的一种手段。基于此，所有者会考虑制定相关激励机制，以减弱信息不对称带来的不良影响。然而，激励机制的制定、运作等所需成本均较高，且其无法脱离财务信息独立运行。当企业所有者不愿承担激励机制的成本时，激励机制则难以维持运作。在利益驱动下，管理者控制财务信息，做出有损所有者等利益的行为。该理论可为后文内部控制等指标的选取提供理论基础。

3.2 机器学习模型介绍

3.2.1 可用于财务舞弊识别的模型选取

对于上市公司财务舞弊识别，在构建指标体系的基础上，通过构建一些规则识别其是否存在财务舞弊行为，在本质上是一种分类问题。在解决分类问题方面，利用机器学习算法是较为可行高效的方法。机器学习通过设计模型与方法，从已知数据集中学习及寻找数据间的相互联系，并由此对未知数据的规律进行预测与分类。其中，支持向量机在处理样本量较少且维度较高的分类问题方面具有独特优势，其核心思想是构造分割面将不同类型的样本分隔开，同时使距离分割面最近的样本与该分割面的距离最大，可用于二分类问题。其次，Logistic 回归原理简单且分类高效，目前已被大量运用于财务舞弊识别领域。而集成学习依据特定的集成策略组合多个学习器,克服了基学习器的缺点,优化了算法的性能，本文考虑将集成学习与当下在财务舞弊识别领域应用较为广泛的 Logistic 回归相结合构建财务舞弊识别模型。集成学习中的 Bagging 算法分类性能较好，具有结构简单、运算效率快等优点，是集成学习的代表算法。其能够有效地解决小样本分类问题，适用于训练集样本量较小的情况。此外，GBDT 算法可灵活处理包括离散型、连续型变量在内的各类型数据。在参数调整较少的情况下，GBDT 算法预测精度相对较高。综上，本文分别利用支持向量机、Bagging、GBDT 构建财务舞弊识别模型。

3.2.2 支持向量机

支持向量机是一种针对小样本、少样本的机器学习方法，自 20 世纪 90 年代产生以来发展迅速，是一种广义线性分类器，其核心思想是构造分割面将不同类型的样本分隔开，同时使距离分割面最近的样本与该分割面的距离最大，其中，距离分割面最近的样本即为支持向量。支持向量机在处理样本量较少且维度较高的分类问题方面具有独特优势，可用于二分类问题。对于二分类问题，给定样本集 D 以及类标签 y_i，支持向量机的任务即是寻找一个能够将样本分割为两种类型的超平面，可能存在多个满足此条件的超平面，应当寻

找那个能够最大化超平面两侧空白区域的分类超平面。在样本空间内，超平面可由下式表示:

$$\omega^{\mathrm{T}}x+b=0 \tag{3.1}$$

其中，ω即法向量，b即位移量。法向量表示超平面的方向，位移量表示超平面与原点间的距离。超平面记为(ω,b)。当$y_{\mathrm{i}}=+1$时，$\omega^{T}x_{i}+b>0$，当$y_{\mathrm{i}}=-1$时，$\omega^{T}x_{i}+b<0$。使式中等号成立的样本点即为支持向量，两不同类支持向量至超平面的距离和称为间隔，表示为$\gamma=\frac{2}{\|\omega\|}$。寻找间隔最大的超平面即最大化$\frac{2}{\|\omega\|}$，也即

$$\min_{\omega,\mathrm{b}}\frac{1}{2}\|\omega\|^{2} \tag{3.2}$$

$$s.t.y_{i}(\omega^{T}x_{i}+b)\geq 1,\mathrm{i}=1,2,...,\mathrm{m}.$$

由上述内容可知，寻找最大间隔超平面的问题即为约束最优化求解的问题。相应的拉格朗日函数为$L(\omega,b,\alpha)=\frac{1}{2}\|\omega\|^{2}+\sum_{i=1}^{m}\alpha_{i}(1-y_{i}(\omega^{T}x_{i}+b))$。

求解最优参数ω^{*}与b^{*}，得

$$\omega^{*}=\sum_{j=1}^{m}a_{j}^{*}y_{j}x_{j}$$

$$b^{*}=\frac{1}{y_{\mathrm{i}}}-\sum_{j=1}^{m}y_{j}a_{j}^{*}(x_{j}^{T}x_{i}) \tag{3.3}$$

其中，下标$j\in\{\mathrm{j}\,|\,\mathrm{a}_{\mathrm{j}}^{*}>0\}$。

以上建立于样本线性可分的假定下。对于线性不可分的情况，为使样本在新的坐标空间中线性可分，则需要将样本从原始的坐标空间变换到更高维的坐标空间，$\Phi(\mathrm{x})$表示x由原始空间至新坐标空间的变换，此时，超平面可以表示为:

$$f(x)=\omega^{T}\Phi(\mathrm{x})+\mathrm{b} \tag{3.4}$$

类似于以上，即有

$$\min_{\omega,\mathrm{b}}\frac{1}{2}\|\omega\|^{2} \tag{3.5}$$

$$s.t.y_{i}(\omega^{T}\Phi(x_{i})+b)\geq 1,\mathrm{i}=1,2,...,\mathrm{m}.$$

与线性可分情况不同的是，在线性不可分情况下求解需用到核函数，即

$$k(x_i, x_j) = \Phi(\mathrm{x_i})^T \Phi(\mathrm{x_j}) \tag{3.6}$$

最终，得

$$\begin{aligned} f(x) &= \omega^T \Phi(\mathrm{x}) + \mathrm{b} \\ &= \sum_{i=1}^{m} a_i y_i k(x_i, x) + \mathrm{b} \end{aligned} \tag{3.7}$$

3.2.3 Bagging

集成学习依据特定的集成策略组合多个学习器,克服了基学习器的缺点,优化了算法的性能，其又名多分类器系统。集成学习可划分为两个类别,一类是通过 Bagging 集成策略组合多个相互间不存在依赖关系的基学习器，一类是通过 Boosting 集成策略组合多个必须串行生成、相互间存在强依赖关系的基学习器。

Bagging 的分类效果普遍较好,结构简单,并行框架,因此运算效率快。Bagging 的基学习器并行式生成，之间不存在强依赖性，训练此算法的复杂度与单个基学习器同阶。Bagging 基于自主采样法,在确定采样集后分别训练基学习器并将其结合，此算法又称套袋法。对于回归，结合方式通常为简单平均法，对于分类，则通常通过简单投票法进行结合。

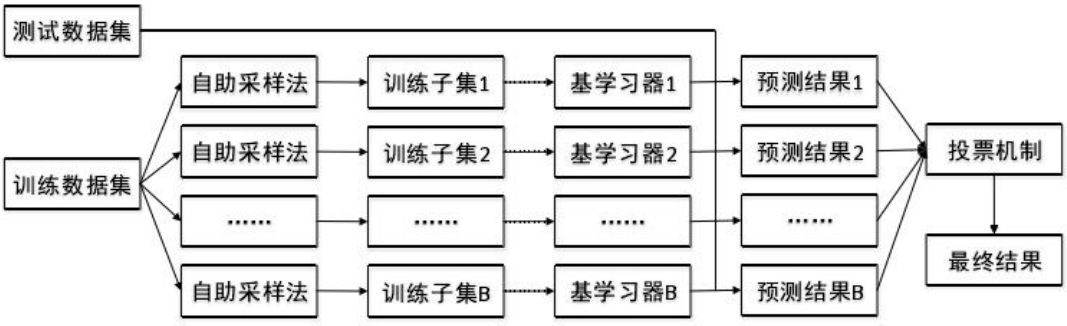

图 3.5 Bagging 算法原理

Bagging 算法的优点在于其能够有效地解决小样本分类问题，适用于训练集样本量较小的情况。此外，当训练集产生微小变化时，Bagging 模型的结果不会因此产生过大变化，其可以较好地处理不稳定算法对于训练集敏感的情况。Bagging 算法的局限性在于其在样本域处理数据，当存在大量的训练数据时，个体分类器基于挖掘的信息能够完成对于全部样本的高精度分类，此时 Bagging 算法的作用则不明显。此外，通过有放回的随机抽样方法产生新训练集会使一些样本无法被抽取，而一些样本则被重复采样。在此情况下，被忽略的样本空间内分类器难以分类正确，而特定子空间内分类器精度很高。

3.2.4 GBDT

Boosting 集成是一个不断迭代提升的过程，复杂度通常高于 Bagging。Boosting 的基学习器串行式生成，之间存在强依赖性。GBDT 算法是其中的典型算法。当基学习器为决策

树时，提升方法被称为提升树，其完成学习优化的方式为加法模型与前向分布算法。然而，对一般损失函数，优化存在一定困难，针对此问题，学者提出了 GBDT 算法。GBDT 算法以回归树为基学习器，沿负梯度方向拟合回归树，使残差不断减小。相比传统算法，GBDT 所具有的优势在于可灵活处理离散型与连续型变量，此外，与传统算法相比，GBDT 算法具有更高的预测精度和更短的调整时间。

GBDT 算法如下：

输入：训练数据集 $T=\{(x_1,y_1),(x_2,y_2),...,(x_n,y_n)\}$，其中 $x_i \in x=R^n, i=1,2,...,n, y_i \in R$，损失函数 $L(y,f(x))$。

输出：回归树 $f_M(x)$。

（1）初始化，估计使损失函数极小化的常数值：

$$f_0(x)=\arg\min\sum_{i=1}^{N}L(y_i,c) \tag{3.8}$$

（2）对 $m=1,2,...,N$ 有：

①对 $i=1,2,...,N$，计算损失函数负梯度，作为残差估计，平方损失情况下，负梯度即是残差，其他损失函数情况下，其约等于残差值：

$$r_{m,i}=\frac{\partial L(y_i,f_{m-1}(x_i))}{\partial f_{m-1}(x_i)} \tag{3.9}$$

②对 $r_{m,i}$ 拟合单个决策树，获取第 m 棵树叶节点域：

$$R_{m,j}, j=1,2,...,J$$

③对 $j=1,2,...,J$，估计回归树节点区域，得到残差近似值：

$$c_{m,j}=\arg\min\sum_{i=1}^{N}L(y_i,f_{m-1}(x_i)+c) \tag{3.10}$$

④更新回归树

$$f_m(x)=f_{m-1}+\sum_{j=1}^{J}c_{m,j}I(x\in R_{m,j}) \tag{3.11}$$

（3）获得回归树

$$f_M(x)=\sum_{i=1}^{M}\sum_{j=1}^{J}c_{m,j}I(x\in R_{m,j}) \tag{3.12}$$

第 4 章　实证分析

4.1 实证研究设计思路

企业进行财务舞弊时通常会同向地调整收入与费用,一些财务指标不会表现出异常。舞弊手段专业性、隐蔽性的逐渐增强使得信息之间可能存在隐蔽信息，财务指标正常且无异常变动也并不一定意味着企业不存在财务舞弊行为，此时利用传统方法识别企业的财务舞弊行为则会存在困难。机器学习通过设计模型与方法，从已知数据集中学习及寻找数据间的联系，并基于此预测未知数据的规律。其可以从数据中将宝贵的隐藏信息提取出来，利用机器学习挖掘企业财务信息有助于报表使用者做出更全面、系统、科学的判断。

本文基于机器学习对财务舞弊识别模型进行研究比较。本文以 2007 年 1 月 1 日至 2019 年 12 月 31 日我国上市公司数据作为样本选取范围，财务舞弊样本为存在虚列资产与虚构利润的上市公司，配对样本依据配对样本与舞弊样本细分行业相同等标准构建。财务舞弊样本数为 136 个，配对样本数为 803 个。若对原始数据集直接建模，会使模型难以学习少数类特征信息，无法准确地对少数类标签进行预测。对于样本非平衡问题，考虑到欠采样会造成样本的减少，本文利用 SMOTE 过采样方法进行样本重建。

在指标选取方面，本文以尽量全面为原则初步选取指标，以使指标体系尽可能地反映出上市公司的整体情况。考虑到财务指标容易被操纵，本文对于上市公司的财务指标进行总体分析，所选财务指标包含盈利能力、偿债能力、经营能力等六个方面。考虑到非财务指标可以客观地反映出公司的真实经营情况，本文所选非财务指标包含治理结构、股权性质、内部控制三个方面。其后对数据进行指标筛选，避免冗余特征影响分类器性能。

在模型选取方面，支持向量机是一种针对小样本、少样本的机器学习方法，其在处理样本量较少且维度较高的分类问题方面具有独特优势，可用于二分类问题。Bagging 算法的分类性能较好，具有结构简单、运算效率快等优点，是集成学习的代表算法。GBDT 算法可灵活处理包括离散型、连续型变量在内的各类型数据。在参数调整较少的情况下，GBDT 算法预测精度相对较高。本文分别利用支持向量机、Bagging、GBDT 构建财务舞弊识别模型，并对各模型依据召回率等指标进行评估比较。

图 4.1 实证研究设计思路图

4.2 数据来源

本文主要研究财务报告舞弊。财务报告舞弊是指公司故意虚假陈述、遗漏财务信息以操纵财务报告的编制，进而误导财务报表使用者并使其决策错误。财务报告作为上市公司向报表使用者定期传递公司财务状况与经营成果等信息的载体，是报表使用者评估公司质量、了解公司运营状况的有效依据。其中的数据化信息主要为财务信息，如公司特定时点的资产信息、一定时期内的利润信息等。非数据化信息主要为重大事项等信息。绝大部分财务报告舞弊为财务数据舞弊。

考虑到我国于 2007 年 1 月 1 日颁布了新会计准则，为避免 2007 年前后上市公司财务报告数据统计不一致，本文以 2007 年 1 月 1 日至 2019 年 12 月 31 日我国上市公司数据作为样本选取范围，财务舞弊信息来源于国泰安数据库中的违规处理数据库。该数据库包含在上交所、深交所上市的存在违规行为的企业发布的公告、监管机构发布的公告等。该数据库将公司违规类型划分为 16 类，其中，非财务类型的违规包括操纵股价、违规炒作等，财务类型的违规包括虚列资产、虚构利润等。其中，虚列资产是将未来无法产生利润的甚至不存在的内容计为资产。例如，企业对于损失或费用暂时无法承担时，将其列为递延资产或待摊费用。虚构利润一般通过调整收入、费用等方式实现。企业通常通过虚增或虚减收入、虚增或虚减费用等方式操纵利润。企业进行财务舞弊时通常会同向地调整收入与费用，因而一些财务指标不会表现出异常，存在一定隐蔽性。企业操纵利润的目的其一在于平滑各年度利润，其二在于增加当年亏损，提前确认未来亏损，营造出企业未来年度盈利

能力较高的假象。

参考现有研究针对财务舞弊违规类型的选取，本文选取违规类型中存在虚列资产与虚构利润的上市公司作为初步筛选样本。由于金融行业上市公司的财务报告在一些方面不同于其他行业，本文所研究的上市公司不包括金融行业上市公司。同时,本文剔除 ST、退市样本，剔除违规公告缺失、数据不完整样本。部分企业连续多年存在财务舞弊行为，对于在 2007 年至 2019 年间多次发生财务舞弊的公司，本文将每一舞弊年度均作为研究对象。考虑到年度财务报告可以更全面地反映公司当年的财务情况，本文选取的财务舞弊样本均为年度财务报告舞弊。在阅读各初选企业违规公告后，本文最终选取 62 家出现财务舞弊问题的上市公司，舞弊样本数为 136 个。

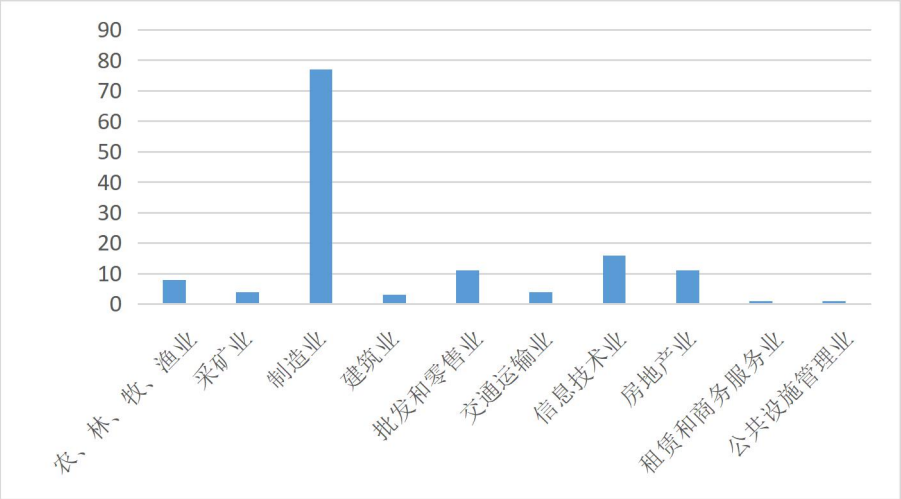

图 4.2 舞弊样本行业分布

在财务舞弊公司行业分布方面，依据 2012 版上市公司行业分类，在本文所选的 136 个财务舞弊样本中，制造业的舞弊企业为 37 家，舞弊样本数为 77 个，占全部舞弊样本的近 57%，远超其他行业。信息技术业的舞弊企业为 9 家，舞弊样本数为 16 个，占全部舞弊样本的近 12%。房地产业的舞弊企业为 4 家，舞弊样本数为 11 个，占全部舞弊样本的近 8%。批发和零售业的舞弊企业为 4 家，舞弊样本数为 11 个，占全部舞弊样本的近 8%。其中，制造业舞弊样本信息如表 4.1 所示。

表 4.1 制造业舞弊样本信息

细分行业	企业名称	舞弊年份
农副食品加工业	佳沃股份	2011
造纸和纸制品业	安妮股份	2008
金属制品业	恒星科技	2012，2013
石油加工、炼焦和核燃料加工业	长春燃气	2014
	太化股份	2014
化学纤维制造业	神马股份	2014

橡胶和塑料制品业	风神股份	2011，2012
化学原料和化学制品制造业	山西路桥	2015，2016
	圣济堂	2016
	新纶科技	2016，2017，2018
纺织业	众和股份	2016
非金属矿物制品业	四通股份	2016
	金刚玻璃	2015，2016，2017
	三峡新材	2011，2012，2016，2017
	中兵红箭	2014，2015，2016
专用设备制造业	千山药机	2015，2016
	中创环保	2016
	鞍重股份	2013，2014，2015
	金自天正	2016
有色金属冶炼和压延加工业	银邦股份	2015，2016，2017
	赣锋锂业	2016
电子设备制造业	长园集团	2016，2017
	浪潮信息	2009
	分众传媒	2007，2008，2009，2010，2011
	协鑫集成	2011，2012
	金亚科技	2014
	超华科技	2014
	联建光电	2014，2015
医药制造业	益佰制药	2013，2014，2015，2016，2017，2018
	延安必康	2015，2016，2018
	康芝药业	2011，2012
	尔康制药	2015，2016
电气机械和器材制造业	惠而浦	2015，2016
	理工环科	2015
运输设备制造业	康尼机电	2015，2016，2017
	江苏国信	2013，2014
	洪都航空	2015

除制造业外，本文舞弊样本同时涉及采矿业、建筑业、房地产业等其他行业。其他行业舞弊样本信息如表 4.2 所示。

表 4.2 其他行业舞弊样本信息

所属行业	细分行业	企业名称	舞弊年份
农、林、牧、渔业	农业	北大荒	2011，2012，2013
		神农科技	2014，2015，2016
	林业	福建金森	2015
	渔业	獐子岛	2017
采矿业	煤炭开采和洗选业	昊华能源	2015，2016，2017，2108
建筑业	建筑装饰业	嘉寓股份	2010，2011，2012
批发和零售业	零售业	南纺股份	2007，2008，2009，0210

		上海物贸	2008，2009，2010，2011
	批发业	三木集团	2012，2013
		龙宇燃油	2014
交通运输、仓储和邮政业	水上运输业	淮河能源	2011，2012，2013，2014
信息传输、软件和信息技术服务业	软件和信息技术服务业	迪威迅	2010，2011，2012
		海峡创新	2018
		亚联发展	2009，2010，2011，2012
		大智慧	2013
		中国海防	2015
	电信、广播电视卫星传输服务	欢瑞世纪	2013，2016
	互联网和相关服务	实益达	2017，2018
		利欧股份	2016
		海联讯	2011
房地产业	房地产业	高新发展	2007，2009
		亚太实业	2012，2013，2014
		中国高科	2014，2015
		济南高新	2014，2015，2016，2017
租赁和商务服务业	商务服务业	联建光电	2016
水利、环境和公共设施管理业	生态保护和环境治理业	科融环境	2017

部分企业曾连续多年出现财务舞弊，其中，制造业中的益佰制药甚至持续六年存在财务舞弊行为。究其原因，我国 2013 年至 2018 年间对于财务舞弊企业的罚款最高为 60 万，部分企业的罚款金额低至 30 万，过低的舞弊成本在一定程度上导致了企业多次进行财务舞弊。针对此现象，2020 年 3 月起，我国新证券法启用。针对财务舞弊行为设置了 100 万元至 1000 万元的处罚额度，情节严重者将被勒令退市。

在配对样本选取方面，首先，配对样本与舞弊样本细分行业相同且年份相同，本文依据 2012 版上市公司行业分类确定上市公司所属细分行业。第二，配对样本在 2007 年 1 月 1 日至 2019 年 12 月 31 日未因财务舞弊受到处罚。第三，配对样本不属于 ST 及 PT 公司。第四，配对样本与舞弊样本的资产规模相差不超过 30%。本文最终确定配对样本数为 803。本文舞弊样本类标签为 0，配对样本类标签为 1。部分配对样本信息如表 4.3 所示。

表 4.3 部分配对样本信息

企业名称	所属行业	年份
太龙药业	医药制造业	2011，2012
上海凯宝	医药制造业	2011，2012
科新机电	专用设备制造业	2013，2014，2015
达刚控股	专用设备制造业	2013，2014，2015
吉林敖东	医药制造业	2016，2018

大东方	零售业	2010
英特集团	批发业	2012，2013
中远海特	水上运输业	2013
正虹科技	农副食品加工业	2011
三六五网	互联网和相关服务	2017，2018
冀东水泥	非金属矿物制品业	2014
天坛生物	医药制造业	2014，2015，2016，2018
广汇汽车	医药制造业	2011
万东医疗	专用设备制造业	2013，2016
迈克生物	医药制造业	2018
胜宏科技	电子设备制造业	2015，2016
健帆生物	专用设备制造业	2016
世纪鼎利	软件和信息技术服务业	2012，2018
光大嘉宝	房地产业	2017
上海能源	煤炭开采和洗选业	2018
苏大维格	电子设备制造业	2016
雪浪环境	专用设备制造业	2014，2016
赛微电子	电子设备制造业	2016，2017
雪榕生物	农业	2016
东方电热	电气机械和器材制造业	2015
东土科技	电子设备制造业	2016，2017
迪瑞医疗	专用设备制造业	2014，2016
聚飞光电	电子设备制造业	2015，2016，2017
全志科技	电子设备制造业	2015，2016，2017
佳创视讯	软件和信息技术服务业	2012

4.3 指标初选

4.3.1 指标选取依据

本文依据财务舞弊的原因与手段选取指标。财务舞弊行为存在的根本原因在于利益驱动，舞弊收益大于舞弊成本使企业选择进行财务舞弊。基于上文有限理性理论，财务信息反映了企业的财务状况、经营状况等，其是公开的，并且具备结果理性的特性。信息使用者着眼点在于结果，而不是程序理性，忽略了结果的形成过程中对于程序理性的控制，各方对于企业的考核大多依据特定财务指标，忽视了非财务指标。此外，基于上文契约理论，各利益主体依据财务信息达到正确判断企业经营情况的目的。财务信息在企业契约的签订过程中意义重大，财务信息主要为财务报告中的数据化信息，如公司特定时点的资产信息、一定时期内的利润信息等。各主体在签订契约前需依据企业财务信息判断是否与其签订契约。股东依据企业财务信息决定派发股利的数额，投资者依据企业财务信息评估经营业绩，

减少决策的不确定性。基于此，上市公司可能出于使特定财务指标达标、满足各利益主体的期望等目的实施财务舞弊。由于财务指标容易被操纵，本文对于上市公司的财务指标进行总体分析，所选财务指标包含盈利能力、偿债能力、经营能力等六个方面。其中，利润是公司管理者经营业绩的集中体现，盈利能力反映了企业在一段时期内获利的能力。偿债能力是公司长久生存的关键，可直接体现公司的经营能力。经营能力反映了企业资产的管理运用效率，是企业充分利用资产获取利润的能力。

在财务舞弊手段方面，企业通常通过将待摊费用长期挂账等方式虚增资产、少记欠款等方式虚减负债，通过虚增或虚减收入、虚增或虚减费用等方式操纵利润。且上市公司舞弊手段的专业性、隐蔽性逐渐增强，企业进行财务舞弊时通常会同向地调整收入与费用,因而一些财务指标不会表现出异常,存在一定隐蔽性。而对于非财务指标，管理层通常难以进行操纵，其不易被篡改，可以客观地反映出公司的真实经营情况。因而本文同时选取治理结构、股权性质以及内部控制三方面非财务指标。其中，公司治理决定了公司未来的发展,其在监督、减少财务舞弊方面具有重要的作用。股权结构合理能够弥补企业治理的缺陷,提升企业的治理效率。内部控制质量低会增加财务报告的噪音,降低财务报告的质量。

4.3.2 指标选取

参考国内外学者的研究，同时依据可比性、可得性等指标选取原则，本文从九个方面选取指标。企业财务舞弊情况受多种因素影响，各因素间存在一定关联。为使财务舞弊识别结果客观准确，本文以尽量全面为原则初步选取指标，以使指标体系尽可能地反映出上市公司的整体情况，为后续的指标筛选工作做充分准备。同时，为了最大程度地避免公司规模效应所带来的影响，本文所选财务指标均为比率指标。初步选取的指标个数为 68 个，其中财务指标个数为 54 个，数据来源于国泰安数据库中的财务指标分析数据库。非财务指标个数为 14 个，治理结构、股权性质以及内部控制指标分别来源于国泰安数据库中的治理结构、股权性质以及内部控制数据库。指标说明如表 4.4 所示。

表 4.4 指标说明表

	指标名称	指标代码	特别说明
盈利能力	资产报酬率	X1	息税前利润与资产总额比
	总资产净利润率	X2	净利润与总资产余额比
	流动资产净利润率	X3	净利润与流动资产余额比
	净资产收益率	X4	净利润与股东权益余额比
	息税前利润与资产总额比	X5	
	长期资本收益率	X6	收益总额与长期资本额比
	营业毛利率	X7	营业毛利额与营业收入比

	营业成本率	X8	营业成本与营业收入比
	营业利润率	X9	营业利润与营业收入比
	营业净利率	X10	净利润与营业收入比
	总营业成本率	X11	营业总成本与营业总收入比
	销售费用率	X12	销售费用与营业收入比
	管理费用率	X13	管理费用与营业收入比
	财务费用率	X14	财务费用与营业收入比
	销售期间费用率	X15	销售期间费用与营业收入比
	成本费用利润率	X16	利润总额与成本费用比
	资产减值损失与营业收入比	X17	
	息税折旧摊销前营业利润率	X18	息税折旧摊销前净利润与营业总收入比
	息税前营业利润率	X19	息税前净利润与营业收入比
	归属于母公司净资产收益率	X20	归属于母公司所有者的净利润与归属于母公司所有者权益合计期末值比
每股指标	每股收益	X21	净利润本期值与实收资本本期期末值比
	归属于母公司每股收益	X22	归属于母公司所有者的净利润本期值与实收资本本期期末值比
	息税前每股收益	X23	息税前净利润本期值与实收资本本期期末值比
	息税折旧摊销前每股收益	X24	息税折旧摊销前净利润本期值与实收资本本期期末值比
	每股营业利润	X25	营业利润本期值与实收资本本期期末值比
	每股未分配利润	X26	未分配利润期末值与实收资本本期期末值比
	每股经营活动现金净流量	X27	经营活动现金净流量本期值与实收资本本期期末值比
	每股投资活动现金净流量	X28	投资活动现金净流量本期值与实收资本本期期末值比
	每股筹资活动现金净流量	X29	筹资活动现金净流量本期值与实收资本本期期末值比
	每股企业自由现金流量	X30	企业自由现金流量本期值与实收资本本期期末值比
	每股股东自由现金流量	X31	股东自由现金流量本期值与实收资本本期期末值比
	每股折旧和摊销	X32	折旧和摊销本期值与实收资本本期期末值比
	每股现金净流量	X33	现金及现金等价物本期净增加额与实收资本本期期末值比
偿债能力	经营活动净现金流量与流动负债比	X34	
	资产负债率	X35	负债合计与资产总计比

	长期借款与总资产比	X36	
	有形资产负债率	X37	负债总额与扣除无形资产的总资产比
	有形资产带息债务比	X38	负债总额与有形资产比
	长期资本负债率	X39	非流动负债与长期资本比
	长期负债权益比率	X40	长期负债与所有者权益比
经营能力	应收账款与收入比	X41	
	存货与收入比	X42	
	固定资产与收入比	X43	
	总资产周转率	X44	营业收入与资产总额期末余额比
现金流能力	营业收入现金含量	X45	销售商品、提供劳务所收现金与营业收入比
	营业收入现金净含量	X46	经营活动净现金流量与营业总收入比
	全部现金回收率	X47	经营活动净现金流量与资产总计期末余额比
	现金再投资比率	X48	经营活动净现金流量与再投资资产比
发展能力	资本保值增值率	X49	所有者权益合计本期期末值与期初值比
	资本积累率	X50	所有者权益合计本期增加值与期初值比
	总资产增长率	X51	总资产本期增加值与期初值比
	可持续增长率	X52	权益报酬率*留存收益率
	所有者权益增长率	X53	所有者权益本期增加值与期初值比
	每股净资产增长率	X54	每股净资产本期增加值与期初值比
治理结构	独立董事人数	X55	
	监事总规模	X56	监事（含监事主席）
	高管人数	X57	年报披露的高管人员总人数
	年薪披露方式	X58	上市公司的年薪披露方案
	未领取薪酬董事人数	X59	包括仅领取津贴的董事（不包括独立董事）
	未领取薪酬监事人数	X60	包括仅领取津贴的监事（不包括独立监事）
	四委设立个数	X61	审计、战略、提名、薪酬与考核委员会设立个数
	独立董事与上市公司工作地点一致性	X62	1 代表相同，2 代表不同，3 代表不确定
股权性质	第一大股东持股比例	X63	
	前十大股东持股比例	X64	
内部控制	是否披露内控评价报告	X65	1 代表是，2 代表否，3 代表不确定
	是否出具内控评价报告结论	X66	1 代表是，2 代表否，3 代表不确定
	内部控制是否有效	X67	1 代表是，2 代表否，3 代表不确定

	内部控制是否存在缺陷	X68	1 代表是，2 代表否，3 代表不确定

（1）财务指标

财务指标可以对企业的财务状况进行综合评估，是反映企业经营成果的重要载体。财务指标易被企业操纵，本文在财务指标方面进行总体的分析，其中财务指标为基于财务报表的衍生计算。利润是公司管理者经营业绩的集中体现，盈利能力反映了企业在一段时期内获利的能力。偿债能力是公司长久生存的关键，可直接体现公司的经营能力。经营能力反映了企业资产的管理运用效率，是企业充分利用资产获取利润的能力。本文从盈利能力、偿债能力、经营能力等六个方面选取财务指标。

①盈利能力

盈利能力是描述企业管理水平的重要指标，反映了企业在一段时期内获利的能力。利润是投资者获取收益的主要资金来源、公司管理者经营业绩的集中体现，是公司内外高度关注的核心问题。盈利能力受企业管理效率的直接影响，较好的资产运营效益反映了企业较强的规避风险能力及盈利能力。证监会对于企业上市、新股增发等行为都有明确严格的限制，为获得投资者的信任，企业可能会存在粉饰盈利指标等财务舞弊行为。基于此，本文在盈利能力方面选取营业净利率、财务费用率等二十个指标。

②每股指标

每股指标主要包含每股收益指标、每股未分配利润指标等。其中，每股收益指标可综合反映企业的盈利能力，是一段时间内企业税后利润占总股数的比例。每股未分配利润可综合反映企业的留存与股利分配情况，是企业当期未分配利润占总股本的比例，基于此，本文在每股指标方面选取每股收益、每股未分配利润等十三个指标。

③偿债能力

偿债能力可直接体现公司的经营能力，是否具备偿还其债务的能力是公司长久生存的关键。分析偿债能力可以评价企业对其长短期债务的偿还能力。三角形理论指出，需要达到投资者的预期业绩、收入减少影响股价、现金流紧张等公司压力为舞弊的其中一个动机。根据三角形理论，当企业利用资产与经营所得难以偿还其债务时，为防止陷入财务危机，企业有动机通过财务舞弊的方式筹集资金。基于此，本文在偿债能力方面选取资本负债率等七个指标。

④经营能力

经营能力指公司充分利用资产获取利润的能力。分析经营能力可以对企业资产的管理运用效率等作出评价，其决定了公司获利、偿还债务的能力。企业常通过虚假增加资产的

方式进行财务舞弊，通过虚增资产，企业的固定资产、应收账款等项目数额增加，应收账款与收入比等经营指标增大。基于此，本文在经营能力方面选取总资产周转率等四个指标。

⑤现金流能力

现金流反映了企业获取现金的能力、偿还债务的能力、收益的质量等内容，是企业在一段时期内现金及现金等价物流入流出数。基于此，本文在现金流能力方面选取现金再投资比率等四个指标。

⑥发展能力

发展能力反映了企业在未来一段时期内发展的变动趋势，是企业扩大规模、提高市场占有率的潜在能力。在竞争激烈的市场经济下，一个公司的价值很大程度上取决于其未来的市场占有率、销售收入。对于公司盈利能力、偿债能力等的提升，最终都是为了使公司的发展能力得到提高，进而使公司能够持久发展。舞弊企业通过虚提资产增长率等指标使投资者认为其发展潜力巨大。然而，企业增长速度过快会在一定程度上使得企业的内部结构产生不确定性。基于此，本文在发展能力方面选取总资产增长率等六个指标。

表 4.5 为部分指标描述性统计分析结果。本文所选财务指标均为比率指标，指标均值与标准差大多处于 0 至 1 之间,其中均值与标准差处于 0 至 1 的指标约占全部指标的 71%。

表 4.5 部分指标描述性统计分析结果

	X1	X2	X3	X4	X5	X6	X7
mean	0.0618	0.0472	0.0797	0.0642	0.0617	0.1034	0.3339
std	0.0631	0.0594	0.1298	0.1519	0.0631	0.1109	0.1968
min	-0.2652	-0.3092	-1.2552	-2.0915	-0.2652	-0.5097	-0.3002
25%	0.0302	0.0154	0.0274	0.0333	0.0300	0.0506	0.1876
50%	0.0559	0.0440	0.0711	0.0728	0.0559	0.0963	0.3022
75%	0.0881	0.0733	0.1261	0.1164	0.0881	0.1519	0.4434
max	0.4756	0.3820	1.0585	0.6205	0.4756	1.1549	0.9479

（2）非财务指标

管理层通常难以对非财务指标进行操纵，其不易被篡改，可以客观地反映出公司的真实经营情况。其中，公司治理主要为股东等利益相关者在权责分配中的利益关系，决定了公司未来的发展，其在监督、减少财务舞弊方面具有重要的作用。股权结构可以直观地反映企业控制权的分布情况与集中度，股权结构合理能够弥补企业的治理缺陷，提升企业的治理效率。内部控制的一大出发点在于防止财务舞弊，其是企业管理的一个有效工具。内部控制质量低会增加财务报告的噪音，降低财务报告的质量。本文将治理结构、股权性质以及内部控制三方面的非财务指标引入舞弊识别指标体系。

①治理结构

在治理结构方面，企业的治理结构存在缺陷会增加财务舞弊行为产生的可能性。董事会、监事会、管理层等作为企业治理结构的重要组成部分，是影响治理效率的关键因素。董事会作为股东与管理层之间的桥梁，是企业治理体系的核心。其有效决策将会使企业的价值提高，董事会的规模、独立性等是影响其治理效果的重要因素。独立董事可对公司的业务进行客观独立的判断，能够提高董事会对于管理层的监管效率，在我国上市公司治理中发挥有效作用。若董事会中缺乏独立董事，则容易发生舞弊行为。监事会则执行监督职能，对于公司的财务经营状况进行监督，是公司的监督机构。企业监事规模的大小一定程度上反映了财务舞弊行为存在的可能性。此外，高管人员的数量较少时权力相对集中，重大决策由少数人员制定，基于舞弊动因理论，当舞弊动机及机会存在时，财务舞弊存在的可能性较高。而高管人员的数量较高时权力相对分散，在此情况下，企业进行财务舞弊的可能性较低。在此方面，本文将上述独立董事人数、监事总规模、高管人数等八个指标引入指标体系，数据基于公司的年报、中报、季报、临时公告等。

②股权性质

在股权性质方面，股权结构可以直观地反映企业控制权的分布情况与集中度，是企业治理结构的基础。合理的股权结构能够弥补企业治理中所存在的缺陷，使企业治理效率得到提升。股权集中度过高时，大股东与中小股东对于信息的持有程度不平衡，大股东具有信息优势。其能够基于信息优势，实施舞弊行为并侵占中小股东权益，常用方式为虚构关联交易等。大股东与中小股东间的利益冲突明显，由于股权分散、监督成本过高等原因，中小股东难以对大股东进行监督。在此情况下，企业更有可能存在财务舞弊行为。在此方面，本文将第一大及前十大股东持股比例引入指标体系，数据基于公司股东情况披露。

③内部控制

在内部控制方面，基于委托代理理论，委托人和代理人的行为准则都是为了自身利益最大化,两者的目标函数不一致。委托人将相关决策权赋予代理人,以期其提供对自身有利的服务。然而,代理人所追求的则是最大限度地满足自身需要,包括空闲的时间、可观的报酬等。当内部控制等相关约束机制缺乏时，管理者为获取更多利益，会将信息优势转化为自身利益，做出有损所有者利益的行为。内部控制不是局部范围内的控制，而是对企业各项业务的全面控制，其对于公司的发展起着关键作用。内部控制的一大出发点在于防止财务舞弊，其是企业管理的一个有效工具。内部控制完善可在一定程度上保障财务信息的真实准确，相反地，内部控制质量低会增加财务报告的噪音，降低财务报告的质量。本文在

此方面选取内控评价报告是否披露等四个指标，数据基于公司对于内部控制有效性的评价。四个指标均为分类变量，其中，1 代表是，2 代表否，3 代表不确定。

4.4 模型评估方法

模型训练完成后，即需对模型进行性能评价。在机器学习分类算法评价中，混淆矩阵可以准确地对模型进行性能评价。在混淆矩阵中，预测类别与实际类别间的比较可以反映出正确分类的样本数量，本文主要运用准确率、查准率、召回率、F1 度量值等指标评价模型效果。对于二分类问题，依据预测类别与实际类别的组合可将样本划分为真阳（TP）、假阳（FP）、真阴（TN）、假阴（FN）四个部分，这四部分即构成混淆矩阵，其中真阳（TP）、真阴（TN）为模型正确分类的样本。混淆矩阵的行表示实际类别，列表示预测类别，如表 4.6 所示。

表 4.6 二分类混淆矩阵

实际类别 \ 预测类别	0	1
0	TP	FN
1	FP	TN

依据混淆矩阵计算以下指标:

准确率指模型正确识别的样本占全部样本的比例，表达式如下:

$$Accuracy = \frac{TP + TN}{TP + TN + FN + FP} \tag{4.1}$$

查准率又名精确率，是指在预测类别为真的样本中，真阳性样本所占比例。即真实舞弊样本占预测为舞弊样本的比例，表达式如下:

$$\mathrm{Precision} = \frac{TP}{TP + FP} \tag{4.2}$$

召回率又名查全率，是指在实际类别为真的样本中，真阳性样本所占比例。即在全部舞弊样本中，模型成功识别的样本所占比例。因财务报告舞弊的特殊性，模型应旨在尽可能地识别出舞弊样本，故召回率为本文的一个主要评价指标。表达式如下:

$$\mathrm{Recall} = \frac{TP}{TP + FN} \tag{4.3}$$

F1 度量值对查准率与召回率综合考虑，是两指标的调和指标，表达式如下:

$$F_1 = \frac{2 * Precision * Recall}{Precision + Recall} \tag{4.4}$$

以假阳率为横轴，真阳率为纵轴构建坐标系，不同阈值对应多组真阳率与假阳率，基于此绘制的曲线为 ROC 曲线。ROC 曲线涉及真阳率、真阴率、假阳率三个指标，各指标定义如下:

真阳率即为查全率，又称敏感度，表达式如下：

$$TPR = \frac{TP}{TP + FN} \tag{4.5}$$

真阴率又称特异度，是指在实际类别为假的样本中，真阴性样本所占比例，表达式如下:

$$TNR = \frac{TN}{TN + FP} \tag{4.6}$$

假阳率是指在实际类别为假的样本中，假阳性样本所占比例，表达式如下:

$$FPR = \frac{FP}{TN + FP} \tag{4.7}$$

ROC 曲线越靠近左上角，模型的效果越好。曲线下的面积即为 AUC，相应地，AUC 的值越大，模型的效果越好，该值可定量地描述模型分类效果。

4.5 样本重建与指标筛选

4.5.1 样本重建

本文的财务舞弊样本数为 136，配对样本数为 803，财务舞弊样本与配对样本比约为 1:6，存在样本非平衡问题。若对原始数据集直接建模，会使模型预测偏于多数类样本，难以学习少数类特征信息，无法准确地对少数类标签进行预测，可能会导致模型的效果不佳，而少数类样本也即财务舞弊样本正是研究中所关注的对象，因而需要对样本非平衡问题进行处理。重采样是处理样本非平衡问题的一个主要方法，具体可分为欠采样与过采样。

欠采样主要针对多数类样本，其通过在多类样本中抽选部分样本并将其删除的方式，实现样本非平衡率的降低。此方法的优点在于操作相对简单，缺点在于易遗漏重要的数据信息。与欠采样不同，过采样通过增加少类样本数的方式，实现数据集由不平衡至平衡的转变，其主要针对少数类样本。其中，随机过采样对现有的少类样本进行随机复制或简单

旋转，以缓解样本不平衡程度。此方法易于实现，但缺点在于易导致过拟合。与随机过采样不同，SMOTE 方法对样本不再进行简单地复制，其利用插值法为少类样本增添新的样本，以平衡少类样本与多类样本的比例。SMOTE 方法的基本思想在于依据策略为少类样本增添人工合成样本，其是过采样的一种经典方法。SMOTE 算法的步骤为：

（1）依据 K 近邻算法，对各少类样本确定其在全部少类样本中的 K 个近邻；

（2）依据样本不平衡程度设置过采样率 R，在各少类样本与其的 K 个近邻中按照下式进行随机线性插值，以获得 R 个新少类样本；

$$x_{newlj} = x + rand(0,1)(x_i - x), l = 1,2,...,K, j = 1,2,...,R \quad (4.8)$$

其中，x 为一个少数类样本，$rand(0,1)$ 为 $(0,1)$ 间的一个随机数，x_i 为 x 的第 l 个近邻。

（3）将由上述方法得到的合成数据加入至原始数据集中得到新的数据集。

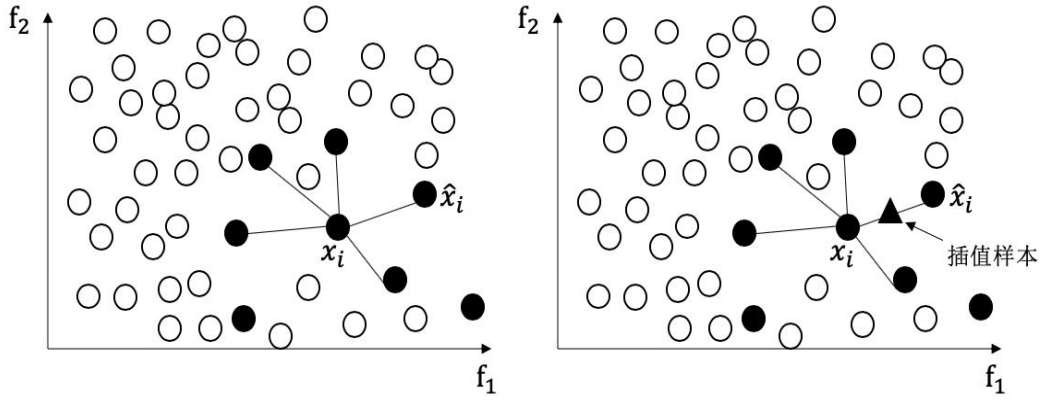

图 4.3 SMOTE 线性插值

考虑到欠采样会造成样本的减少，本文利用 SMOTE 方法进行样本重建，过采样后财务舞弊样本与配对样本比为 1:1。

4.5.2 指标筛选

指标筛选具体为从初始特征集合中选择信息价值高的特征，并依据选择后的特征训练模型，使得选择后的特征可以在特定评价指标上达到最优。其能够保留强特征作为训练分类器的样本特征，从而尽可能保留了最多数据集的原始信息。一方面，指标筛选在降低数据维度、提高模型性能的同时，能够缩短模型的运行时间。另一方面，指标筛选能够降低学习难度，有助于理解数据的特征与结构。本文分别基于决策树和相关性进行指标筛选。

决策树在解决分类等问题时较为常用，是一种基于树形结构的经典算法，其在解决分类与回归问题时分别为分类树和回归树。决策树的核心思想是依据分类条件对样本分类并生成决策节点，直至样本无法继续进行分类。决策树采用由根向叶的构造方式生成一棵树。根节点及各内部节点对应一个属性测试，依据测试结果，样本被依次划分至各子节点中，

从根节点至叶节点的每一条路径分别对应一条分枝规则。将此过程不断递归，直至节点下的样本全属于同一类别或无法划分，此时这些节点被称为叶节点，分别对应一个决策结果。整棵决策树即对应一组分类的表达规则。

ID3、CART 以及 QUEST 等都为决策树的典型算法，其中，ID3 算法为后来相关算法的提出打下了基础。决策树可用于评价特征重要性。接近根节点的特征重要性相对更高。本文基于决策树进行指标筛选，并记为指标筛选 1，结果如表 4.7 所示。

表 4.7 基于决策树的指标筛选结果

指标代码	指标名称
X3	流动资产净利润率
X11	总营业成本率
X12	销售费用率
X13	管理费用率
X14	财务费用率
X26	每股未分配利润
X28	每股投资活动现金净流量
X32	每股折旧和摊销
X36	长期借款与总资产比
X38	有形资产带息债务比
X45	营业收入现金含量
X51	总资产增长率
X60	未领取薪酬监事人数
X64	前十大股东持股比例

X3-X51 为财务指标。其中，X3、X11、X12、X13、X14 为盈利能力指标，X26、X28、X32 为每股指标，X36、X38 为偿债能力指标，X45 为现金流能力指标，X51 为发展能力指标。X60、X64 为非财务指标。其中，X60 为治理结构指标。监事会作为公司的监督机构，对于公司的财务经营状况进行监督，执行监督职能。监事会成员薪酬来源于上市公司会降低其监督效果，为避免受到薪酬方面的干扰，其成员薪酬一般不在上市公司领取。当监事会规模一定时，未领取薪酬监事人数越低，企业出现财务舞弊的可能性越高。X64 为股权性质指标。股权集中度越高，企业出现财务舞弊的可能性越高。

此外，本文基于相关性剔除冗余指标，并记为指标筛选 2。本文初始指标数为 68 个，存在一些相关性较强的指标，如 X9（营业利润率）与 X16（成本费用利润率）两指标的相关系数为 0.80，X11（总营业成本率）与 X17（资产减值损失/营业收入）两指标的相关系数为 0.87。冗余指标会对模型性能产生一定的负面影响，因此需要将冗余指标剔除。本文将相关性阈值设定为 0.7，相关性大于 0.7 则表明两指标的相关度较高。对于相关性较强的指标，分析指标含义，在降维的同时确保留下的指标尽可能全面地代表数据特征。如每

股指标中的 X29（每股筹资活动现金净流量）与 X33（每股现金净流量）之间有较高相关性，考虑到每股筹资活动现金净流量所反映的内容更为详尽，本文保留 X29（每股筹资活动现金净流量），剔除 X33（每股现金净流量）。剔除冗余指标后剩余指标个数为 36，结果如表 4.8 所示。其中，X1-X49 为财务指标，X55-X68 为非财务指标。

表 4.8 基于相关性的指标筛选结果

指标代码	指标名称
X1	资产报酬率
X7	营业毛利率
X9	营业利润率
X12	销售费用率
X13	管理费用率
X14	财务费用率
X28	每股投资活动现金净流量
X29	每股筹资活动现金净流量
X30	每股企业自由现金流量
X31	每股股东自由现金流量
X32	每股折旧和摊销
X34	经营活动净现金流量与流动负债比
X35	资产负债率
X36	长期借款与总资产比
X38	有形资产带息债务比
X41	应收账款与收入比
X42	存货与收入比
X43	固定资产与收入比
X44	总资产周转率
X45	营业收入现金含量
X46	营业收入现金净含量
X47	全部现金回收率
X48	现金再投资比率
X49	资本保值增值率
X55	独立董事人数
X56	监事总规模
X57	高管人数
X58	年薪披露方式
X59	未领取薪酬董事人数
X60	未领取薪酬监事人数
X61	四委设立个数
X62	独立董事与上市公司工作地点一致性
X63	第一大股东持股比例
X64	前十大股东持股比例
X67	内部控制是否有效
X68	内部控制是否存在缺陷

4.6 上市公司财务舞弊识别模型构建

本文将样本按照 7:3 的比例随机划分为训练集与测试集。基于第三章相关理论同时结合上文所选样本及指标，本文利用支持向量机、Bagging、GBDT 三种方法分别构建财务舞弊识别模型，并对各模型进行评估比较。支持向量机、Bagging、GBDT 三种模型工作原理相同。首先对训练集进行学习，挖掘其特征规律并建立模型。训练集主要为模型的学习样本，其识别结果不足以评价模型效果的好坏，测试集能够较为客观地评价模型的识别能力。因此，本文的混淆矩阵及模型评估指标均基于测试集数据。

4.6.1 基于支持向量机的财务舞弊识别模型构建

支持向量机在处理样本量较少且维度较高的分类问题方面具有独特优势，可用于二分类问题。支持向量机是一种针对小样本、少样本的机器学习方法，其核心思想是构造分割面将不同类型的样本分隔开，同时使距离分割面最近的样本与该分割面的距离最大。通过支持向量机找到的超平面将舞弊样本与非舞弊样本作以区分，是判断样本是否进行舞弊的分类器。

在核函数方面，核函数是某个高维空间的内积，在支持向量机中起着重要作用。常用核函数包括线性核、多项式核、高斯核等，详细地：

线性核：$k(x_i,x_j)=x_i^T x_j$，

多项式核：$k(x_i,x_j)=(x_i^T x_j)^d$，$d\geq 1$为多项式次数，

高斯核：$k(x_i,x_j)=\exp(-\frac{\|x_i-x_j\|^2}{2\sigma^2})$，$\sigma>0$为高斯核带宽。

选用不同的核函数，会产生不同的支持向量机算法，得到不同的分类结果。公司是否存在财务舞弊行为与指标间存在着非线性关系，本文所研究的财务舞弊识别为非线性问题。本文选取的核函数类型为径向基函数与多项式函数，并将相应的支持向量机分别记为支持向量机 1 和支持向量机 2。

在参数方面，支持向量机包含 C 与 gamma 两个重要参数。其中，参数 C 表示允许分类误差存在的程度。C 越大则模型泛化能力越弱，较高的惩罚因子通过增加模型的自由度选择更多的支持向量，以确保所有样本都可以被正确分类，支持向量个数会对模型训练速度产生影响。C 越小则模型泛化能力越强，越能够容忍训练时的误差，较低的惩罚因子使分界面平滑。参数 gamma 指对低维样本进行高维度映射。参数值越大表示影射维度越高，但

也越容易引起过拟合。交叉验证法为模型参数选择方法之一，其大致思路为模型训练前先保留部分训练样本，并将其作为测试样本以评价模型。本文基于 5 折交叉验证进行网格搜索寻找最优参数，其基本原理为，首先网格搜索遍历全部参数组合，其后运用上述交叉验证法得到识别准确率最高的参数组合。表 4.9 和表 4.10 分别为径向基核函数的支持向量机混淆矩阵与性能评估。表 4.11 和表 4.12 分别为多项式核函数的支持向量机混淆矩阵与性能评估。

表 4.9 支持向量机 1 混淆矩阵

指标筛选 1		预测类别	
		0	1
实际类别	0	211	17
	1	48	206
指标筛选 2		预测类别	
		0	1
实际类别	0	223	5
	1	19	235

在基于决策树的指标筛选下，当参数 C 为 12，gamma 为 1.0 时，支持向量机模型效果最优。依据表 4.6 混淆矩阵，在利用上述模型评估指标公式计算后得到，模型准确率为 86.51%，查准率为 81.47%，召回率为 92.54%，F1 度量值为 86.57%。在基于相关性的指标筛选下，当参数 C 为 12，gamma 为 0.2154 时，支持向量机模型效果最优。模型准确率为 95.02%，查准率为 92.15%，召回率为 97.81%，F1 度量值为 94.90%。

表 4.10 支持向量机 1 性能评估

	准确率（%）	查准率（%）	召回率（%）	F1 度量值（%）
指标筛选 1	86.51	81.47	92.54	86.57
指标筛选 2	95.02	92.15	97.81	94.90

相较于基于决策树的指标筛选，在基于相关性的指标筛选下，径向基核函数的支持向量机模型在准确率、查准率、召回率及 F1 度量值方面均更优。

表 4.11 支持向量机 2 混淆矩阵

指标筛选 1		预测类别	
		0	1
实际类别	0	209	19
	1	58	196
指标筛选 2		预测类别	
		0	1
实际类别	0	215	13
	1	71	183

在基于决策树的指标筛选下，当参数 C 为 12，gamma 为 1.0 时，支持向量机模型效果最优。模型准确率为 84.02%，查准率为 78.28%，召回率为 91.67%，F1 度量值为 84.45%。

在基于相关性的指标筛选下，当参数 C 为 12，gamma 为 0.0464 时，支持向量机模型效果最优。模型准确率为 82.57%，查准率为 75.17%，召回率为 94.30%，F1 度量值为 83.66%。

表 4.12 支持向量机 2 性能评估

	准确率（%）	查准率（%）	召回率（%）	F1 度量值（%）
指标筛选 1	84.02	78.28	91.67	84.45
指标筛选 2	82.57	75.17	94.30	83.66

基于决策树指标筛选下的多项式核函数支持向量机模型在准确率、查准率、F1 度量值方面优于基于相关性指标筛选下的模型，但在召回率方面低于基于相关性指标筛选下的模型。召回率反映了模型成功识别的样本占全部舞弊样本的比例，可见基于决策树指标筛选下的多项式核函数支持向量机模型存在舞弊样本遗漏的情况。

此外，在同一指标筛选方法下，径向基核函数的支持向量机模型各指标均高于多项式核函数，识别效果整体优于多项式核函数。其中，基于相关性指标筛选下的径向基核函数支持向量机模型识别效果最优，舞弊样本识别准确率为 97.81%，非舞弊样本识别准确率为 92.52%。

4.6.2 基于 Bagging 的财务舞弊识别模型构建

Bagging 集成策略组合多个相互间不存在依赖关系的基学习器，其分类效果普遍较好。Bagging 结构简单,为并行框架,因此运算效率快，训练此算法的复杂度与单个基学习器同阶。而对于基学习器，Logistic 回归原理简单且分类高效，其思想来源于统计学中的线性回归。Logistic 回归应用非常频繁，目前已被大量运用于财务舞弊识别领域。本文构建以 Logistic 回归为基学习器的 Bagging 模型。以下为 Logistic 回归模型介绍。

Logistic 回归可应用于二分类问题，其思想来源于统计学中的线性回归。对于二分类问题，可以利用单位阶跃函数将线性回归模型的预测值转化为类别，然而此函数为非连续函数，因而通过 Sigmoid 函数确定分类类别，Sigmoid 函数的形式为:

$$y=\frac{1}{1+e^{-z}} \tag{4.9}$$

在该函数中，当 z 趋向正无穷时，y 趋向于 1，z 趋向负无穷时，y 趋向于 0。将线性回归模型与该函数相结合便可得到如下 Logistic 回归模型：

$$y=\frac{1}{1+e^{-(wx+b)}} \tag{4.10}$$

依据 y 的取值判定分类类别，当 y 的取值大于阈值时，类别为 1，当 y 的取值小于阈值时，类别为 0。y 的取值越小，类别为 0 的概率越高，相反地，y 的取值越大，类别为 1 的概率相应越高。

表 4.13 和表 4.14 分别为 Bagging 模型的混淆矩阵与性能评估。

表 4.13 Bagging 模型混淆矩阵

指标筛选 1		预测类别	
		0	1
实际类别	0	179	49
	1	84	170
指标筛选 2		预测类别	
		0	1
实际类别	0	170	58
	1	79	175

在基于决策树的指标筛选下，Bagging 模型的准确率为 72.41%，查准率为 68.06%，召回率为 78.51%，F1 度量值为 72.91%。舞弊样本识别准确率为 78.51%，非舞弊样本识别准确率为 66.93%。在基于相关性的指标筛选下，Bagging 模型的准确率为 71.57%，查准率为 68.27%，召回率为 74.56%，F1 度量值为 71.28%。舞弊样本识别准确率为 74.56%，非舞弊样本识别准确率为 68.90%。

表 4.14 Bagging 模型性能评估

	准确率（%）	查准率（%）	召回率（%）	F1 度量值（%）
指标筛选 1	72.41	68.06	78.51	72.91
指标筛选 2	71.57	68.27	74.56	71.28

基于决策树指标筛选下的 Bagging 模型在准确率、召回率、F1 度量值方面优于基于相关性指标筛选下的模型，但在查准率方面低于基于相关性指标筛选下的模型。查准率反映了真实舞弊样本占预测为舞弊样本的比例，可见基于决策树指标筛选下的 Bagging 模型存在舞弊高估的情况。此外，支持向量机模型的识别效果要优于 Bagging 模型。

4.6.3 基于 GBDT 的财务舞弊识别模型构建

集成学习依据特定的集成策略组合多个学习器,克服了基学习器的缺点,优化了算法的性能。GBDT 算法是其中 Boosting 集成的典型算法。GBDT 算法以回归树为基学习器,沿负梯度方向拟合回归树,使残差不断减小。相比传统算法，GBDT 所具有的优势在于可灵活处理离散型与连续型变量，此外，与传统算法相比，GBDT 算法具有更高的预测精度和更短的调整时间。本文构建基于 GBDT 的财务舞弊识别模型。GBDT 模型涉及的参数主要包括 n_estimators、learning_rate、max_depth 等，依次为弱学习器最大迭代数、学习率、最大深度。弱学习器最大迭代数也即最大的弱学习器个数。学习率也称步长，为各弱学习器的权重缩减系数，取值范围为 0 至 1。最大深度即指决策树最大深度。综合比较各参数不同取值下的模型效果后，在主要超参数的取值范围中搜索最优参数组合。在基于决策树的指标筛选下，本文的弱学习器最大迭代数为 50，学习率为 0.3，决策树最大深度为 4。在

基于相关性的指标筛选下，本文的弱学习器最大迭代数为50，学习率为0.5，决策树最大深度为4。表4.15和表4.16分别为GBDT模型的混淆矩阵与性能评估。

表4.15 GBDT模型混淆矩阵

指标筛选1		预测类别	
		0	1
实际类别	0	217	11
	1	47	207
指标筛选2		预测类别	
		0	1
实际类别	0	218	10
	1	44	210

在基于决策树的指标筛选下，GBDT模型的准确率为87.97%，查准率为82.20%，召回率为95.18%，F1度量值为88.22%。在指标重要性方面，排在前五名的指标分别为X3(流动资产净利润率)、X32(每股折旧和摊销)、X14(财务费用率)、X13(管理费用率)、X36(长期借款与总资产比)。在基于相关性的指标筛选下，GBDT模型的准确率为88.80%，查准率为83.21%，召回率为95.61%，F1度量值为88.98%。在指标重要性方面，排在前五名的指标分别为X1(资产报酬率)、X38(有形资产带息债务比)、X32(每股折旧和摊销)、X36(长期借款与总资产比)、X14(财务费用率)。

表4.16 GBDT模型性能评估

	准确率（%）	查准率（%）	召回率（%）	F1度量值（%）
指标筛选1	87.97	82.20	95.18	88.22
指标筛选2	88.80	83.21	95.61	88.98

相较于基于决策树的指标筛选，在基于相关性的指标筛选下，GBDT模型在准确率等各方面均更优。舞弊样本识别准确率为95.61%，非舞弊样本识别准确率为82.68%。此外，GBDT模型的识别效果要优于Bagging模型。

依据基于GBDT的财务舞弊识别模型，X1(资产报酬率)、X3(流动资产净利润率)、X13(管理费用率)、X14(财务费用率)、X32(每股折旧和摊销)、X36(长期借款与总资产比)、X38(有形资产带息债务比)对于财务舞弊的识别较为重要。其中，X14(财务费用率)、X32(每股折旧和摊销)、X36(长期借款与总资产比)多次进入模型指标重要性的前五名。X14(财务费用率)在不同指标筛选方法下分别位于指标重要性的第三名和第五名，X32(每股折旧和摊销)分别位于指标重要性的第二名和第三名，X36(长期借款与总资产比)分别位于指标重要性的第五名和第四名。

表4.17 重要指标

指标代码	指标名称
X1	资产报酬率

X3	流动资产净利润率
X13	管理费用率
X14	财务费用率
X32	每股折旧和摊销
X36	长期借款与总资产比
X38	有形资产带息债务比

4.6.4 对比分析

在利用支持向量机、Bagging、GBDT 三种方法分别构建财务舞弊识别模型的基础上，本部分整合上述结果，对比分析各模型性能。各模型性能评估指标如下。

表 4.18 基于决策树指标筛选后各模型性能评估指标比较

模型	准确率（%）	查准率（%）	召回率（%）	F1 度量值（%）	AUC
支持向量机 1	86.51	81.47	92.54	86.57	0.9332
支持向量机 2	84.02	78.28	91.67	84.45	0.8870
Bagging	72.41	68.06	78.51	72.91	0.7973
GBDT	87.97	82.20	95.18	88.22	0.9481

表 4.18 为基于决策树进行指标筛选后各模型效果。结果显示，GBDT 模型各性能评估指标均最高，整体优于其他模型，其次为径向基核函数的支持向量机模型，两模型各指标差异较小。

表 4.19 基于相关性指标筛选后各模型性能评估指标比较

模型	准确率（%）	查准率（%）	召回率（%）	F1 度量值（%）	AUC
支持向量机 1	95.02	92.15	97.81	94.90	0.9785
支持向量机 2	82.57	75.17	94.30	83.66	0.8808
Bagging	71.57	68.27	74.56	71.28	0.7939
GBDT	88.80	83.21	95.61	88.98	0.9520

表 4.19 为基于相关性进行指标筛选后各模型效果。结果显示，径向基核函数的支持向量机模型的准确率、查准率、召回率、F1 度量值以及 AUC 值均最高，整体优于其他模型，其次为 GBDT 模型。两模型召回率指标差异最小，可见 GBDT 模型对于舞弊样本的识别效果较好。

在模型识别效果方面，本文利用支持向量机、Bagging、GBDT 三种方法分别构建财务舞弊识别模型，各模型识别准确率均在 70%以上。其中，径向基核函数的支持向量机模型与 GBDT 模型识别效果较好。在基于相关性进行指标筛选后，径向基核函数的支持向量机模型效果最好，模型准确率为 95.02%，召回率为 97.81%。在基于决策树进行指标筛选后，GBDT 模型效果最好，模型准确率为 87.97%，召回率为 95.18%。因财务报告舞弊的特殊性，模型应旨在尽可能地识别出舞弊样本，故召回率为本文的一个主要评价指标。在基于较少指标的情况下，GBDT 模型的召回率较好。因而在综合考虑数据获取成本等因素后，本文建议利用 GBDT 模型对财务舞弊进行识别。而当数据易得时，则可考虑使用精度更高的支持

向量机模型。此外，除径向基核函数的支持向量机模型外，基于不同指标筛选结果下的 GBDT 模型、Bagging 模型、多项式核函数的支持向量机模型在准确率、查准率、召回率、F1 度量值以及 AUC 值各方面差异均较小。在基于决策树的指标筛选下，多项式核函数的支持向量机模型与 Bagging 模型的准确率更高。故基于决策树进行指标筛选后得到的每股折旧和摊销、财务费用率、长期借款与总资产比等 14 个指标代表性较好。

第 5 章　总结与展望

5.1 研究总结

利益相关者了解公司运营状况的一个有效依据为财务报告，其反映了企业在一定时期内的财务状况和经营成果。然而，一些上市公司或是出于躲避监管的目的，或是出于公司管理层个人利益需要，通过各种手段进行财务舞弊。且因企业财务舞弊手段不断更新且隐蔽性增强，财务舞弊隐蔽时长最长达七年，最短为一年以内，大部分财务舞弊公司的隐蔽时长为两年，财务舞弊披露年度相较于财务舞弊年度会有所推迟。此外，企业进行财务舞弊时通常会同向地调整收入与费用,一些财务指标不会表现出异常。舞弊手段专业性、隐蔽性的逐渐增强使得信息之间可能存在隐蔽信息，财务指标正常且无异常变动也并不一定意味着企业不存在财务舞弊行为，此时利用传统方法识别企业的财务舞弊行为则会存在困难。财务舞弊识别模型可降低由于未正确判断企业财务状况而做出错误决策的风险，本文基于机器学习对财务舞弊识别模型进行比较研究。

本文的主要工作包括前期文献的阅读整理、数据的搜集处理、模型算法的选择、对数据进行实证分析等。本文选取 2007 年 1 月 1 日至 2019 年 12 月 31 日违规类型中存在虚列资产与虚构利润的 62 家上市公司。对于在 2007 年至 2019 年间多次发生财务舞弊的公司，本文将每一舞弊年度均作为研究对象，最终确定舞弊样本数为 136。本文依据配对样本与舞弊样本细分行业相同等标准构建配对样本，最终确定配对样本数为 803。在模型构建方面，支持向量机在处理样本量较少且维度较高的分类问题方面具有独特优势，可用于二分类问题。Bagging 算法的分类性能较好，具有结构简单、运算效率快等优点。GBDT 算法可灵活处理包括离散型、连续型变量在内的各类型数据。在参数调整较少的情况下，GBDT 算法预测精度相对较高。本文利用支持向量机、Bagging、GBDT 三种方法分别构建财务舞弊识别模型，并对模型通过交叉验证确定参数，从而对模型进行了优化。本文主要运用准确率、查准率、召回率、F1 度量值等评价模型效果，模型评估指标均基于测试集数据。通过研究本文得到以下结论:

（1）在模型识别效果方面，本文利用支持向量机、Bagging、GBDT 三种方法分别构建财务舞弊识别模型，各模型识别准确率均在 70%以上。其中，径向基核函数的支持向量机模型与 GBDT 模型识别效果较好。在基于相关性进行指标筛选后，径向基核函数的支持向量机模型效果最好，模型准确率为 95.02%，召回率为 97.81%。在基于决策树进行指标筛

选后，GBDT 模型效果最好，模型准确率为 87.97%，召回率为 95.18%。因财务报告舞弊的特殊性，模型应旨在尽可能地识别出舞弊样本，故召回率为本文的一个主要评价指标。在基于较少指标的情况下，GBDT 模型的召回率较好。因而在综合考虑数据获取成本等因素后，本文建议利用 GBDT 模型识别财务舞弊。而当数据易得时，则可考虑使用精度更高的支持向量机模型。

（2）在指标重要性方面，本文依据可比性、可得性等原则，从财务与非财务两方面选取 68 个指标。其中，在非财务指标方面，可重点关注前十大股东持股比例及未领取薪酬监事人数。依据基于 GBDT 的财务舞弊识别模型，在财务指标方面，盈利能力中的资产报酬率、流动资产净利润率、管理费用率、财务费用率，偿债能力中的长期借款与总资产比、有形资产带息债务比以及每股折旧和摊销对于财务舞弊的识别较为重要。此外，每股折旧和摊销、财务费用率以及长期借款与总资产比对于财务舞弊的识别更为重要。每股折旧和摊销反映了每股中固定资产、无形资产等各项折旧及摊销，财务费用率是财务费用与营业收入的比值，反映了上市公司的盈利能力，长期借款与总资产比是长期借款与资产总额的比值，反映了上市公司的偿债能力。

5.2 相关建议

5.2.1 对于上市公司的建议

董事会、监事会、管理层等作为企业治理结构的重要组成部分，是影响治理效率的关键因素。上市公司应保证董事会、监事会的独立性，保持内部有效沟通，提高信息透明度。此外，合理的股权结构能够弥补企业治理中所存在的缺陷，使企业治理效率得到提升。上市公司应采取多形式的股权激励机制，合理分散股权结构。

5.2.2 对于投资者的建议

投资者可依据财务舞弊识别模型进行决策。当数据易得时，可使用支持向量机模型。当数据获取成本较高时，可使用 GBDT 模型。此外，投资者可重点关注每股折旧和摊销、财务费用率、长期借款与总资产比等指标。

5.3 研究不足与展望

5.3.1 研究不足

本文在基于机器学习比较上市公司财务舞弊识别模型过程中，存在以下不足之处：

（1）由于企业财务舞弊手段不断更新且隐蔽性增强，因此难以排除非舞弊样本中存在实际进行了财务舞弊但未被发现的情况，这可能会掩盖两种类别间的差别，对本文模型识别结果存在一定影响。

（2）舞弊企业往往呈现多年或者连续舞弊的风险，而财务报表各年数据又存在一定的关联关系，本文在构建模型时暂未考虑时间因素。

5.3.2 研究展望

财务舞弊在对企业自身的有序发展产生阻碍的同时，也会损害相关者利益，降低人们对于监管制度等的信心，破坏经济的健康和稳定。识别上市公司是否实施财务舞弊对于投资者等非常重要，但也存在较多困难。首先识别过程需要搜集大量的信息，其后需要对所得信息进行分析，这都会耗费非常多的时间成本。投资者等在进行决策时可参考基于机器学习的财务舞弊识别模型，降低由于未正确判断企业财务状况而做出错误决策的风险。机器学习可以将数据中宝贵的隐藏信息提取出来，是财务报表使用者了解和评估企业的一个有效辅助工具，在财务舞弊识别领域对其的应用将更为频繁，建模方法的实用价值将更高，本文对于未来的研究有以下展望：

（1）对财务舞弊样本进行细分。目前，包括本文在内的大部分研究都只是将样本简单地划分为舞弊样本与非舞弊样本两类，未来可以考虑依据财务舞弊的严重程度、企业受罚力度等信息对舞弊样本进行进一步的划分，深入对于舞弊样本的研究。

（2）对舞弊企业异常数据与时间的联系进行分析。一些上市公司多年或连续几年持续存在财务舞弊行为，未来可以考虑将时间因素引入识别模型，尝试挖掘异常指标与时间的关联。

参考文献

引文文献

[1]Beneish M.D..Detecting GAAP Violation:Implications for Assessing Earnings Management among Firms with Extreme Financial Performance[J].Journal of Accounting and Public Policy,1997,16(3):271-309.

[2]Fanning Kurt M.,Cogger Kenneth O..Neural network detection of management fraud using published financial data[J].Intelligent Systems In Accounting, Finance and Management,1998, 7:21-41.

[3]Lin J.W.,Hwang M.I,Becker J.D..A Fuzzy Neural Network for Assessing the Risk of Fraudulent Financial Reporting[J].Managerial Auditing Journal,2003(18):657-665.

[4]Kotsiantis S., Koumanakos E., Tzelepis D.,et al. Forecasting Fraudulent Financial Statements using Data Mining[J]. International Journal of Computational Intelligence, 2006,3(2):104-110.

[5]Kotsiantis S., Koumanakos E., Tzelepis D.,et al. Forecasting Fraudulent Financial Statements using Data Mining[J]. International Journal of Computational Intelligence, 2006,3(2):104-110.

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

[7]Pediredla R.S.,Ravi V.,Rao G.R.,et al.Detection of Financial Statement Fraud and Feature Selection Using Data Mining Techniques[J].Decision Support Systems,2011(50):491-500.

[8]Gill N.S., Gupta R.. Analysis of Data Mining Techniques for Detection of Financial Statement Fraud[J].The IUP Journal of Systems Management,2012,10(1):7-15.

[9]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.

[10]Loebbecke J.K.,Willinghan J..Review of SEC Accounting and Auditing Enforcement Releases[Z].Working Paper,University of Utah,1988.

[11]Beaver W.H..Financial Ratios As Predictors of Failure[J].Journal of Accounting Research,1966, (Supplement) :71-111.

[12]Beasley M..An Empirical Analysis of the Relation between the Board of Director Composition and Financial Statement Fraud[J].The Accounting Review,1996,71:443-465.

[13]Green B.P.,Choi J.H..Assessing the risk of management fraud through neural network technology[J].Auditing:A Journal of Practice & Theory,1997,16(1):14-28.

[14]Dechow P.,Weili G.,Larson C..Predicting Material Accounting Misstatements[J]. Contemporary Accounting Research,2011,28(1):17-82.

[15]岳殿民. 中国上市公司会计舞弊模式特征及识别研究[D]. 天津财经大学, 2008.

[16]邓庆山. 基于数据挖掘技术的上市公司会计信息失真识别研究[D]. 江西财经大学, 2009.

[17]李秀枝. 我国上市公司财务报告舞弊特征及识别研究[D]. 中国矿业大学, 2010.

[18]金花妍. 基于内部控制视角的财务舞弊治理研究[D]. 东北财经大学, 2013.

[19]任朝阳. 中国上市公司会计舞弊识别与治理研究[D]. 吉林大学, 2016.

[20]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.

[21]王威. 稀疏组 Lasso-logistic 回归模型在财务报告舞弊识别中的应用研究[J]. 数学的实践与认识, 2020(09):49-58.

[22]曹利. 中国上市公司财务报告舞弊特征的实证研究[D]. 复旦大学, 2004.

[23]陈庆杰. 基于经理人特征的财务报告舞弊识别模型的改进研究——来自中国上市公司

的实证检验[J].经济问题,2012(08):118-122.
[24]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(09):172-175.
[25]刘志洋,韩丽荣.财务报告舞弊识别效率改善研究——基于分类技术改进和数据信息优化兼容视角[J].财经问题研究,2018(01):99-107.

阅读型文献

[26]秦江萍.会计舞弊的市场反应与识别:理论分析与经验证据[M].北京:经济科学出版社,2006.
[27]陈玉雪.论股权结构与财务报告信任关系的重建——基于25家财务舞弊上市公司的实证分析[J].中国注册会计师,2019(12):30-34.
[28]陈邑早,张莹,孔晨.组织认同与亲组织财务报告舞弊决策——多重中介效应分析[J].经济管理,2020(09):176-192.
[29]丁德臣.集成随机森林和支持向量机的商业银行财务困境预测研究[J].数学的实践与认识,2020(02):290-300.
[30]黄世忠.上市公司财务造假的八因八策[J].财务与会计,2019(16):4-11.
[31]孔晨.风险偏好、决策情绪与CEO财务报告舞弊行为[J].经济与管理,2020(01):86-92.
[32]马广奇,张保平.企业研发创新影响财务舞弊风险吗[J].财会月刊,2019(24):7-18.
[33]马晓君,董碧滢,王常欣.一种基于PSO优化加权随机森林算法的上市公司信用评级模型设计[J].数量经济技术经济研究,2019(12):165-182.
[34]王芙蓉.会计信息可比性对财务舞弊风险识别的影响研究——基于财务信息使用者的视角[J].财会通讯,2020(11):31-34.
[35]吴芃,卢珊,杨楠.财务舞弊视角下媒体关注的公司治理角色研究[J].中央财经大学学报,2019(03):51-69.
[36]向鸿鑫,杨云.不平衡数据挖掘方法综述[J].计算机工程与应用,2019(04):1-16.
[37]许金叶,施旖旎.资本市场财务舞弊与产权性质的关系——基于本福德定律的财务数据测试[J].财会通讯,2019(02):37-41.
[38]张力派,程晨,陈玲玲.大数据时代对上市公司财务舞弊的影响——研究综述及展望[J].管理现代化,2020(05):122-129.
[39]张志恒,邓启兰.基于支持向量机的会计信息失真识别研究[J].财会通讯,2019(13):10-14.
[40]郑丽萍,赵杨.上市公司财务舞弊的成因与治理研究——以瑞幸咖啡公司为例[J].管理现代化,2020(04):4-6.
[41]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.
[42]蔡霞.基于混合模型对财务舞弊预警研究[D].西南财经大学,2019.
[43]韩小芳.财务舞弊公司董事会后续治理及其对外部审计的影响 ——基于中国上市公司的实证检验[D].东北财经大学,2010.
[44]王嘉欣.机器学习方法在上市公司财务舞弊预测问题中的应用[D].山东大学,2019.
[45]向琳.基于权重Borderline-SMOTE-RF财务报表舞弊识别研究[D].中南财经政法大学,2019.
[46]Albrecht W.S.,Romney M.B.,Cherrington D.J.,et al.Red-flagging management fraud:A validation[J].Advances in Accounting,1986(3):323-333.
[47]Bologna J.The one minute fraud auditor[J].1989,8(1):29-31.
[48]Chrysovalantis Gaganis.Classification Techniques for the Identification of Falsified

Financial Statements:A Comparative Analysis[J].Intelligent Systems in Accounting,Finance and Management,2009(16):207-229.
[49]Treadway Committee.Fraud Commission Issues Final Report[J].Journal of Accountancy,1987 (11):34.

后记

本论文从选题、撰写初稿至论文定稿，得到了导师及答辩老师等的帮助与指导。在此要由衷地感谢各位老师。

首先要感谢我的导师，从论文选题至定稿的各个过程都离不开老师的悉心指导，老师的意见为我提供了非常大的帮助，老师的循循善诱使我受益匪浅。

其次要感谢答辩老师们，老师们对我提出的宝贵意见使我的论文得到了进一步的完善，老师们严谨的治学态度是我学习的榜样。

最后要感谢身边的家人朋友对我的关怀与支持。