Modern Management 现代管理, 2023, 13(8), 1025-1033
Published Online August 2023 in Hans. https://www.hanspub.org/journal/mm
https://doi.org/10.12677/mm.2023.138129

Hans汉斯

基于机器学习的康美药业财务舞弊甄别研究

赵　宇，赵淳宇，王梦瑶

四川师范大学商学院，四川 成都

收稿日期：2023年6月29日；录用日期：2023年7月12日；发布日期：2023年8月7日

摘　要

如何有效甄别上市公司财务舞弊行为，成为业界和学界持续关注的重要议题。本研究将最近五年受到中国证监会处罚的医药生物行业A股上市公司作为样本，以康美药业为例，基于舞弊三角理论选取24个特征，采用结合SMOTE过采样技术的随机森林分类算法模型进行测试与分析。结果表明，相较于将公司简单归类为舞弊与非舞弊两类，使用多个不同的特征集建立模型或构建多个不同算法的模型进行财务舞弊甄别研究的效果更好。

关键词

财务舞弊，机器学习，康美药业

Research on Financial Fraud Screening of Kangmei Pharmaceutical Based on Machine Learning

Yu Zhao, Chunyu Zhao, Mengyao Wang

School of Business, Sichuan Normal University, Chengdu Sichuan

Received: Jun. 29th, 2023; accepted: Jul. 12th, 2023; published: Aug. 7th, 2023

Abstract

How to effectively identify financial fraud behavior of listed companies has become an important issue of continuous concern in the industry and academia. In this study, the A-share listed companies in the pharmaceutical and biological industry that have been punished by the China Securities Regulatory Commission in the past five years are taken as samples. Taking Kangmei Pharmaceutical as an example, 24 features are selected based on the fraud triangle theory, and Random forest classification algorithm model combined with SMOTE Oversampling technology is used for testing and

文章引用: 赵宇, 赵淳宇, 王梦瑶. 基于机器学习的康美药业财务舞弊甄别研究[J]. 现代管理, 2023, 13(8): 1025-1033.
DOI: 10.12677/mm.2023.138129

analysis. The results indicate that compared to simply categorizing companies into fraudulent and non fraudulent categories, using multiple different feature sets to establish models or constructing models with multiple different algorithms for financial fraud screening research is more effective.

Keywords

Financial Fraud, Machine Learning, Kangmei Pharmaceutical

1. 引言

出于吸引投资、减少税款缴纳等目的，或面临退市的压力，部分上市公司选择向外界提供虚假的财务信息[1]。我国金融市场发展起步较晚，监管机制还不够完善，如何有效甄别上市公司是否进行了财务舞弊成为业界和学界持续关注的重要议题。近年来，从统计学模型到机器学习模型，不同的方法和技术被应用于财务舞弊甄别研究[2]。从模型的精确率、召回率、准确率和 F1 分数来看，机器学习模型效果更为显著。但在现有文献中，机器学习模型效果大多体现在平衡样本数据处理，对不平衡样本数据处理较少，或采用欠采样、随机抽样等技术，其他的采样技术有待进一步研究[3]。事实上，由于不同行业领域公司财务指标特征、上市公司数量具有较大差异，常规采样技术难以达到统计学要求，基于此，本研究在舞弊三角理论进行定性特征选取基础上[4]，试图采用结合 SMOTE 过采样技术的随机森林分类算法模型，针对近年来财务舞弊典型案例公司康美药业及其所在的医药生物行业其他公司，探索基于机器学习的财务舞弊甄别研究。

2. 相关研究综述

2.1. 财务舞弊的因素

财务舞弊的因素理论中舞弊三角理论认为财务舞弊来源于机会、压力和借口三个因素，它们相互作用，共同导致舞弊行为的发生。在存在相应的机会时，想法才会变为行动，故机会是舞弊行为的重要诱因。Gozman 和 Currie [5]指出，在股东或其他投资者给定公司管理层较高的盈利目标，或需要其弥补公司亏损时，公司管理层会承受巨大压力，舞弊的可能性也随之增加。Cressey [6]则指出，实施财务舞弊者倾向于让自己处于道德区域内，当自己或其他人为舞弊行为提供某个理由时，实施财务舞弊者会对舞弊行为进行合理化包装，错误地认为此类行为并未违规违法，这就形成了舞弊三角中的借口因素。Call 等[7]主要研究非财务类型的舞弊因素，发现公司基层员工在获取更大的权力后举报公司舞弊的概率降低。崔东颖，胡明霞[8]研究表明，在市场竞争愈发激烈时，上市公司很可能在利益驱使、内部控制不合理以及外部监管力度不足的共同作用下实施舞弊行为。

2.2. 甄别财务舞弊的方法

甄别财务舞弊的方法主要有两类。一类是以聚类分析、主成分分析和 Logistic 回归等为代表的传统统计模型。例如，Etemadi 和 Zolghi [9]研究传统的统计模型如何预测上市公司财务舞弊，经过分析后选用 Logistic 回归模型；Persons [10]则采用逐步逻辑回归(Stepwise-Logistic)的方法进一步研究此类问题。另一类是新兴的机器学习模型，比较典型的包括 SVM (Support Vector Machine，支持向量机)、MLP 神经网络(Multi-Layer Perception，多层感知器)等。Cecchini 等[11]基于支持向量机(SVM)模型，通过 SVM-FK 方

法预测上市公司财务舞弊；Bao 等[12]则使用集成学习方法处理原始数据，实验结果显示，改进后机器学习模型的预测效果获得较大提升。传统的统计模型简单易懂，计算起来较为简便，计算结果的可解释性强，然而，该类模型在非线性数据方面的表现大多不太理想。机器学习模型是近年来研究的热点，在财务舞弊甄别方面，该类模型的精确率、召回率、准确率和 F1 分数表现良好。

3. 研究设计与方法

3.1. 特征选取

特征选取是运用机器学习方法研究问题时不可或缺的一步，特征的数量和质量对研究效果影响显著[13]。

Table 1. Feature classifications, names, and definitions
表 1. 特征分类、名称及定义

编号	特征分类	特征名称	特征定义
x_1	舞弊机会	流通股占比	可流通股本/总股本
x_2		年度股东大会出席率	出席股东大会的股东持有股份/总股本
x_3		公司是否国有控股	国家控制=1，其他=0
x_4		董事会会议次数	所属年度董事会会议次数
x_5		董事会规模	所属年度董事会人数
x_6		董事长与总经理兼任情况	同一人=1，非同一人=0
x_7		独立董事占比	独立董事总人数/董事会总人数
x_8	舞弊压力	应收账款变动率	(应收账款 $_i$/资产总计 $_i$)/(应收账款 $_{i-1}$/资产总计 $_{i-1}$)
x_9		固定资产折旧变动率	[折旧发生额 $_{i-1}$/(折旧发生额 $_{i-1}$+固定资产净值 $_{i-1}$)]/[折旧发生额 $_i$/(折旧发生额 $_i$+固定资产净值 $_i$)]
x_{10}		资产质量变动率	[1 − (流动资产 $_i$+固定资产净值 $_i$)/资产总计 $_i$]/[1 − (流动资产 $_{i-1}$+固定资产净值 $_{i-1}$)/资产总计 $_{i-1}$]
x_{11}		主营业务收入变动率	主营业务收入 $_i$/主营业务收入 $_{i-1}$
x_{12}		毛利率变动率	[(主营业务收入 $_{i-1}$ − 主营业务成本 $_{i-1}$)/主营业务收入 $_{i-1}$]/[(主营业务收入 $_i$ − 主营业务成本 $_i$)/主营业务收入 $_i$]
x_{13}		应收账款周转天数	360/[主营业务收入/(应收账款+应收票据)]
x_{14}		资产质量	1 − (流动资产+固定资产净值)/资产总计
x_{15}		盈余现金流量差	(净利润 − 经营活动现金净流量)/资产总计
x_{16}		现金流动负债比	经营活动现金净流量/流动负债
x_{17}	舞弊借口	审计意见类型	标准无保留意见=0，保留意见=1，拒绝发表意见或无法表示意见=2，否定意见=3
x_{18}		应计水平	(净利润 − 经营活动产生的现金流量净额)/资产总计
x_{19}		应计方向	应计水平为正=1，其他=0
x_{20}		管理层平均年龄	所属年度已披露的高管平均年龄
x_{21}		管理者自负	高管中薪酬最高的前三名薪酬之和/高管薪酬总额
x_{22}		管理层性别比例	所属年度高管中男性所占比例
x_{23}		高管人员持股	所属年度高管人员持股数的自然对数
x_{24}		高管更迭	董事长或总经理发生变更=1，其他=0

如果将全部特征信息输入机器学习模型，则会导致程序运行时间过长，且难以反映各特征重要性水平的细致差异，因此，需要对特征信息进行约简。本研究选取定性约简的方法，在以往研究基础上，基于舞弊三角理论，对机会、压力、借口共选取了 24 个特征。其中，部分特征可直接使用，部分特征需经过简单计算再使用，如表 1 所示。

3.2. 方法选择

在众多 Bagging 集成算法中，随机森林具有较强的代表性，相较于部分机器学习模型，随机森林在处理不平衡样本时具有一定的优势[14]。根据本研究特点，在医药生物行业领域，进行财务舞弊并受到处罚的上市公司占全部公司的比重较小，财务舞弊甄别研究的样本是典型的非平衡样本。因此，本研究采用随机森林分类算法进行舞弊分类预测。基于舞弊三角理论，根据前述三类特征分别构建三个随机森林模型，分别为舞弊机会模型、舞弊压力模型和舞弊借口模型。将样本二分类为舞弊样本和非舞弊样本，通过机器学习库的对应接口得出模型对两类标签的分类概率，以分类概率的大小反映其舞弊风险水平的高低。

3.3. 数据来源

首先，通过国泰安 CSMAR 中国上市公司违规处理研究数据库，查询“处罚公告披露年度”与“涉及违规年度”字段，初步获取最近五年涉嫌财务舞弊的上市公司名单。然后，结合同花顺数据中心和中国证券监督管理委员会公开信息，在查询其处罚年度具体公告后，剔除了非财务舞弊的违规公司。最终，确认进行财务舞弊并于 2018 年至 2022 年上半年受到中国证监会处罚的医药生物行业 A 股上市公司共计 16 家，其中多家公司连续数年实施财务舞弊。

非财务舞弊公司样本选择方面，因财务舞弊并受到中国证监会处罚的上市公司在整体上市公司中占比较小，且本研究针对医药生物行业上市公司，如果按照 1:1 的比例选取对照控制样本会导致总样本数量过少，缺乏代表性，并可能会丢失相关重要信息。但如果将除财务舞弊样本之外的其他医药生物行业 A 股上市公司全部作为控制样本，很可能出现过拟合的现象。此外，部分上市公司可能由于舞弊金额不大、舞弊性质不严重等原因未被监管机构发现或处罚。因此，本研究按照 1:4 的比例配比，为每一家舞弊公司对应挑选四家非舞弊公司作为控制样本。挑选条件为：所处行业为医药生物行业，所属年度与舞弊样本相同，总资产数额与舞弊样本接近，最近五年未被 ST、未涉嫌财务舞弊，数据整体缺失值小于 5%。

经过上述筛选，确定财务舞弊公司 15 家(除康美药业)，非财务舞弊公司 60 家，共计 75 家。通过国泰安 CSMAR 数据库逐一搜索公司简称或股票代码，下载其对应年份以及前一年的系列特征。将财务舞弊公司或非财务舞弊公司每一年的系列特征及其标签(舞弊样本标签：1，非舞弊样本标签：0)作为一份样本，最终获取舞弊样本 38 份，非舞弊样本 152 份，共计 190 份。

4. 实验与分析

4.1. 实验过程

在具体的模型搭建和使用过程中，先导入 scikit-learn 中随机森林分类器的相关机器学习库，再通过 imblearn.over_sampling 导入 SMOTE，完成前期准备工作。按时间先后顺序对样本进行排序，将前 80%划分为训练集，后 20%划分为测试集，并使用 SMOTE 过采样技术对训练集进行平衡化处理。随后构建随机森林分类器模型，将模型内决策树的数量设置为 1000，向模型投入训练集进行训练再投入测试集测试模型效果。舞弊机会模型、舞弊压力模型和舞弊借口模型采用了相同的算法模型和采样技术，但相互独立、互不影响。

4.2. 评价指标选取

在 scikit-learn 的随机森林分类器模型中，对应的 RFC 接口较多，可以向使用者反馈森林中的决策树结构、模型评估对象的参数和通过测试集检验出的平均准确度等信息。想要更加全面、客观地评价模型，仅仅采用精确度作为评价指标是明显不足的。因此，本研究选取常用的四个指标来分别反映三个模型对财务舞弊行为的预测和判别能力，即精确率、召回率、准确率和 F1 分数。

为了清晰明了地阐释以上四个指标，列示混淆矩阵如表 2 所示。其中，TP (True Positive)指预测为正类且实际为正类的样本；FP (False Positive)指预测为正类但实际为反类的样本；FN (False Negative)指预测为反类但实际为正类的样本；TN (True Negative)指预测为反类且实际为反类的样本。本研究将财务舞弊设置为正类，非财务舞弊设置为反类。

Table 2. Confusion matrix
表 2. 混淆矩阵

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

精确率是指在模型预测为正类的样本中，确实为正类的样本所占比重，说明该模型预测结果为正类的可信程度。在本研究中，精确率反映模型预测为财务舞弊的样本内实际发生财务舞弊的样本所占比重。

召回率是指在实际情况为正类的样本中，被模型识别为正类的样本所占比重，说明该模型对实际为正类样本的查找能力。在本研究中，召回率反映在实际发生财务舞弊的全部样本内模型识别为财务舞弊的样本所占比重。

准确率是指在全部的预测样本中，预测结果与实际情况一致的样本所占比重，说明该模型的整体预测能力。在本研究中，准确率反映模型识别为舞弊且实际为舞弊与识别为非舞弊且实际为非舞弊的样本之和，占据全部预测样本的比重。

F1 分数是在兼顾考虑模型的精确率和召回率后，对模型的一种综合衡量，由精确率和召回率通过简单计算获得，数值介于 0 和 1 之间。

4.3. 模型评价

上节模型评价指标中，召回率可以衡量财务舞弊样本被识别出来的可能性大小。在财务舞弊甄别方面，目的是探索尽可能识别出存在舞弊行为的上市公司的模型。因此，本研究基于舞弊三角理论构建的三个模型，选取召回率最大的模型参数作为随机森林分类器模型的最佳参数。各评价指标如表 3 所示，在精确率、召回率、准确率和 F1 分数方面都表现良好，可用于进一步舞弊甄别研究。

Table 3. Confusion matrix (Unit: %)
表 3. 混淆矩阵(单位：%)

	精确率	召回率	准确率	F1 分数
舞弊机会模型	82.86	90.62	84.97	86.57
舞弊压力模型	85.71	93.75	88.25	89.55
舞弊借口模型	87.88	90.62	88.42	89.23

4.4. 实验结果

特征重要性在随机森林模型的构建过程中具有重要作用，可以凸显特征与标签的相关程度，说明样本数据的各项特征与标签关系如何。此外，基于特征重要性进行特征筛选或数据降维在现有文献研究中也常有应用[15]。本研究通过 scikit-learn 机器学习库设定的 RFC 接口，输出舞弊机会模型、舞弊压力模型和舞弊借口模型各项特征重要性，并进行绘图处理，以柱状图的形式直观反映出甄别医药生物行业上市公司财务舞弊风险的重要特征，如图 1~3 所示。

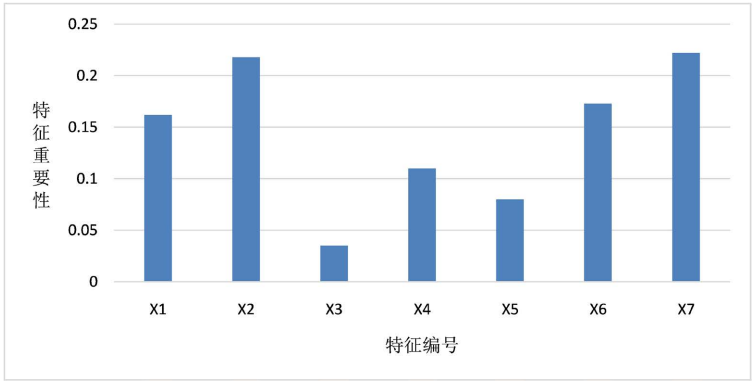

Figure 1. Importance of each feature in the fraud opportunity model
图 1. 舞弊机会模型各特征重要性

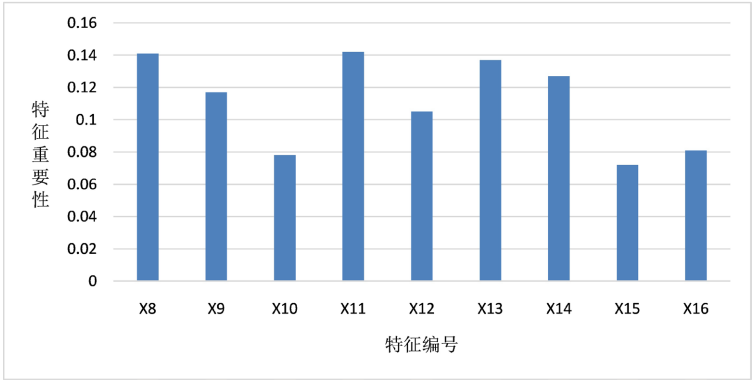

Figure 2. Importance of each feature in the fraud pressure model
图 2. 舞弊压力模型各特征重要性

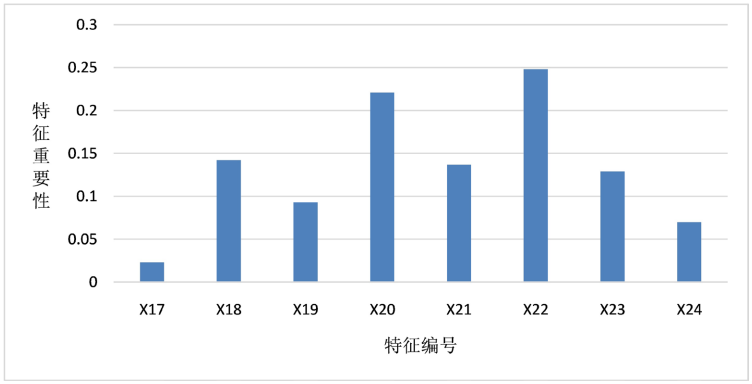

Figure 3. Importance of each feature in the fraud excuse model
图 3. 舞弊借口模型各特征重要性

4.5. 初步分析

在舞弊机会模型中，年度股东大会出席率和独立董事占比较为重要；判断上市公司的舞弊压力大小，需要重点关注其应收账款、主营业务收入的变动状况和资产质量；管理层平均年龄和性别比例则对上市公司舞弊借口这一因素存在较为显著的影响。此外，在构建模型的众多特征中，公司是否国有控股与审计意见类型这两个特征的表现与众不同，其特征重要性的数值都低于 0.05。结合样本数据与实际情况，本研究分析如下：近年来，医药生物行业上市公司中存在不少国有控股公司，但公司是否国有控股与财务舞弊风险大小的关联并不明显；在审计意见类型方面，大多数审计师事务所的行为趋于保守，在审计工作中给出的标准无保留意见较多，即便是对于存在较高舞弊风险的上市公司，也可能仅仅给出保留意见或无法表示意见。

5. 案例分析与讨论

康美药业成立于 1997 年，2001 年成功上市。随后十多年间，康美药业随中医药行业的迅速发展，顺利获取多方融资，公司快速扩张，成为国内中医药行业的代表公司。2018 年 10 月，康美药业突然被曝出涉嫌财务舞弊，随后相关监管机构紧急成立调查组介入进行调查。2019 年 4 月，康美药业公布 2018 年年报，其中涉及 300 亿资金的“会计差错更正”。2019 年 5 月，中国证监会公布对康美药业的调查结果，确认康美药业在 2016~2018 年期间多次实施财务舞弊。2021 年 4 月，康美药业向法院申请破产重整。

根据前述测试结果和分析，将康美药业的对应特征分别投入舞弊机会模型、舞弊压力模型和舞弊借口模型，使用给定接口 model.predict_proba 输出以上模型将康美药业各年样本分类为标签 1 或 0 的基评估器比例，通过该比例的数值来反映康美药业 2016~2018 年财务舞弊风险的变化情况，如图 4 所示。

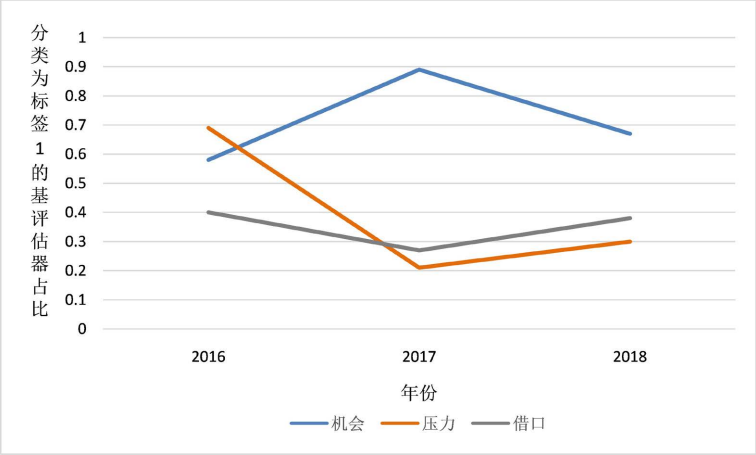

Figure 4. Opportunities, pressures and excuse risks of financial fraud of Kangmei Pharmaceutical
图 4. 康美药业财务舞弊机会、压力及借口风险

结合康美药业的实际情况，对其 2016~2018 年的舞弊机会风险、舞弊压力风险和舞弊借口风险进行如下分析。

5.1. 舞弊机会风险

一方面，康美药业的股权架构不合理，本研究通过流通股占比、年度股东大会出席率等特征反映。根据康美药业历年年度报告可知，马兴田及其妻子许冬瑾二人合计持有康美药业约 35%的股份，除此以外，马兴田与许冬瑾还通过控股金信典当行等多家公司，增强自身对康美药业的影响力和控制力。因此，

康美药业是典型的“一股独大”，股权架构非常不合理。另一方面，董事长兼任多职，独立董事的作用被严重限制，本研究通过董事长与总经理兼任情况、独立董事占比等特征反映。1997 年，马兴田创立了康美药业，伴随着康美药业的发展壮大，马兴田却始终将权力牢牢握在手中。在康美药业的董事会成员中，马兴田、许冬瑾分别担任董事长和副董事长，马兴田还同时兼任公司总经理。按照我国相关法律规定，康美药业设有一定人数的独立董事，但公司董事会的意见对独立董事的推举和任免存在极大影响。在这样的情况下，康美药业的独立董事很难发挥作用，不能起到控制公司财务舞弊风险的作用。综上所述，康美药业的舞弊机会风险一直处于较高的水平。

5.2. 舞弊压力风险

康美药业的公司运营出现问题后，管理层采用虚增利润、虚增货币资金等方式进行伪装，让众多投资者和监管机构错误地认为康美药业是“白马股”，本研究通过应收账款、主营业务收入等特征反映。2016~2018 年期间，康美药业的 88.79 亿元资金被关联方挪用，以恶意炒作自身股票的方式抬高股价，向股票市场发出虚假的利好信息，吸引投资者购买公司股票。同时，康美药业还不断虚增利润、虚增货币资金，保证财务报表上的优异数值，从而获取银行等金融机构的融资。然而，这一系列行为无异于饮鸩止渴，在舞弊压力风险数值下降的表象下，存在着更大的危机。

5.3. 舞弊借口风险

面对利益的诱惑，康美药业管理层没能坚持自身的道德底线，本研究通过管理层平均年龄、管理层性别比例等特征反映。目前，我国对财务舞弊的上市公司人员处罚力度较轻，公司管理层可能存在侥幸心理，认为实施财务舞弊行为后不会被发现或发现后的处罚较轻，为自己准备实施的违法违规行为寻找借口。同时，对康美药业进行审计工作的广东正中珠江会计师事务所也未能坚持审计的独立性，对其出具了标准无保留意见。从前述的特征重要性柱状图中可以看出，审计意见类型这一特征重要性的数值较低，说明审计师事务所的行为趋于保守，审计意见类型对于区分舞弊与非舞弊公司的作用仍然存在提升空间。

综上所述，2016~2018 年，康美药业的舞弊机会风险一直处于较高水平。2016 年，康美药业的舞弊压力风险较大，舞弊借口风险中等，随后都呈现先下降再小幅回升的状态。不难看出，如果使用单一的舞弊压力模型或舞弊借口模型进行二分类，可能无法将康美药业正确地分类为舞弊公司，扩展至其他机器学习模型也存在类似情况。一方面，在甄别公司财务舞弊时，可以使用多个不同的特征集建立模型或构建多个不同算法的模型；另一方面，随机森林分类算法模型是基评估器的“少数服从多数”，对于决策树分类比例较为接近的样本，应当持有谨慎态度。因此，在使用机器学习方法进行财务舞弊甄别研究时，相较于将公司简单归类为舞弊与非舞弊两类，综合考虑其多方面的财务舞弊风险可能会更好。

6. 结语

本研究通过国泰安 CSMAR 数据库、同花顺数据中心和中国证券监督管理委员会官网等多方渠道收集、整理数据，基于文献研究法定性选取特征，构建使用 SMOTE 过采样技术的随机森林分类算法模型，取得较为理想的测试结果。最后，将案例公司康美药业的一系列对应特征分别投入三个模型，输出其舞弊机会风险、舞弊压力风险和舞弊借口风险的变化情况，并结合公司的实际状况对其进行分析。结果表明，康美药业的 2016~2018 年的舞弊机会风险一直处于较高的水平。在现有文献研究中，基于机器学习将上市公司二分类为舞弊类型与非舞弊类型的方法存在一定的局限性，因为财务舞弊风险是普遍存在的，被简单归类为非舞弊类型的公司可能存在较高的舞弊风险。因此，在研究此类问题时，可以使用多个不

同的特征集建立模型或构建多个不同算法的模型，并综合考虑其多方面的财务舞弊风险。

基金项目

四川师范大学实验技术研究项目“财务报表舞弊甄别分析实验”(SYJS2021008)。

参考文献

[1] 刘秀兰, 吕灿, 付强. 近年来我国企业财务信息失真又趋严重的原因及对策探讨[J]. 西南民族大学学报(人文社会科学版), 2012, 33(12): 150-155.

[2] Khan, A.M.A. and Peng, J. (2022) Using Machine Learning Meta-Classifiers to Detect Financial Frauds. *Finance Research Letters*, **48**, Article 102915. https://doi.org/10.1016/j.frl.2022.102915

[3] 伍彬, 刘云菁, 张敏. 基于机器学习的分析师识别公司财务舞弊风险的研究[J]. 管理学报, 2022, 19(7): 1082-1091.

[4] Marco, S.A., Luis, U.A. and José, E.J. (2022) Predictive Fraud Analysis Applying the Fraud Triangle Theory through Data Mining Techniques. *Applied Sciences*, **12**, 3382. https://doi.org/10.3390/app12073382

[5] Gozman, D. and Currie, W. (2014) The Role of Investment Management Systems in Regulatory Compliance: A Post-Financial Crisis Study of Displacement Mechanisms. *Journal of Information Technology*, **29**, 44-58. https://doi.org/10.1057/jit.2013.16

[6] Cressey, D.R. (1953) Other People’s Money; a Study of the Social Psychology of Embezzlement. Patterson Smith Publishing Corporation, Montclair.

[7] Call, A.C., Kedia, S. and Rajgopal, S. (2016) Rank and File Employees and the Discovery of Misreporting: The Role of Stock Options. *Journal of Accounting and Economics*, **62**, 277-300. https://doi.org/10.1016/j.jacceco.2016.06.003

[8] 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(4): 6-9.

[9] Etemadi, H. and Zolghi, H. (2013) Application of Logistic Regression to Identify Fraudulent Financial Reporting. *Journal of Audit Science*, **13**, 5-23.

[10] Persons, O.S. (2011) Using Financial Statement Data to Identify Factors Associated with Fraudulent Financial Reporting. *Journal of Applied Business Research*, **11**, 38-46. https://doi.org/10.19030/jabr.v11i3.5858

[11] Cecchini, M., Aytug, H., Koehler, G.J. and Pathak, P. (2010) Making Words Work: Using Financial Text as a Predictor of Financial Events. *Decision Support Systems*, **50**, 164-175. https://doi.org/10.1016/j.dss.2010.07.012

[12] Bao, Y., Ke, B., Li, B., Yu, Y.J. and Zhang, J. (2020) Detecting Accounting Fraud in Publicly Traded US Firms Using a Machine Learning Approach. *Journal of Accounting Research*, **58**, 199-235. https://doi.org/10.1111/1475-679X.12292

[13] 赵浩, 李盼盼. 基于邻近梯度的机器学习特征选择优化方法[J]. 计算机仿真, 2020, 37(11): 289-293.

[14] An, B. and Suh, Y. (2020) Identifying Financial Statement Fraud with Decision Rules Obtained from Modified Random Forest. *Data Technologies and Applications*, **54**, 235-255. https://doi.org/10.1108/DTA-11-2019-0208

[15] 陶世银, 贺敬安. 基于 XGBoost 与特征重要性筛选的闪电预报模型构建研究[J]. 国外电子测量技术, 2022, 41(1): 99-105.