

doi: 10.19920/j.cnki.jmsc.2024.06.003

公司和高管特征与上市公司违规行为^①

——基于机器学习的经验证据

何 瑛¹, 任立祺², 于文蕾³, 杜亚光¹

(1. 北京邮电大学经济管理学院, 北京 100876; 2. 中移数智科技有限公司, 北京 100080;
3. 中央民族大学管理学院, 北京 100081)

摘要: 上市公司违规问题一直备受资本市场高度关注, 对于其影响因素的单一维度变量因果关系研究固然关键, 但构建有效的整体性预测模型研究同样具有重要意义. 本研究基于内部治理视角选取重要的公司特征和高管个人特征构建上市公司违规行为预测模型, 以 2008 年—2019 年中国 A 股上市公司为样本, 利用机器学习算法 LightGBM 和 SHAP 工具, 研究两类特征对违规行为的预测能力、重要性排序及预测模式. 研究结果表明: 模型可以在一定程度上预测公司违规行为, 公司特征比高管个人特征对预测产生的影响更大. 其中, 上市公司信息透明度越高、总资产净利率越大、资产负债率越低、高管团队持股比例越高、业绩波动性越小、分析师关注度越高, 模型预测违规的倾向越低; 高管年龄偏小、公司存在董事长与 CEO 两职合一情况时, 模型预测违规的倾向增高. 大部分特征均与违规行为呈现非线性关系, 与传统理论和实证研究结论相一致. 本研究从预测视角拓展我国公司高管特征研究, 为监管部门和投资者提升监管和投资效率、企业完善内部治理机制提供经验证据.

关键词: 公司违规; 公司特征; 高管个人特征; LightGBM; SHAP

中图分类号: F272 **文献标识码:** A **文章编号:** 1007-9807(2024)06-0043-26

0 引言

上市公司违规问题一直备受监管部门和投资者等多方利益相关者的高度关注, 统计显示每年约 15.45% 的中国上市公司存在违规行为, 特大违规事件也频频发生, 诸如“凯迪生态”、“抚顺特钢”等公司违规极大地打击了投资者信心, 对提升资本市场资源配置效率产生负面效应. 为深入研究上市公司违规行为成因, 学者们围绕公司内外部环境的相关因素展开诸多探讨, 外部因素主要体现在外部监管层面, 内部因素可进一步区分为公司层面和高管层面. 公司层面, 公司治理规范程度往往影响着公司违规倾向, 目前研究发现董

事会特征, 诸如董事会规模、独立董事占比等重要的公司治理因素与公司违规行为存在显著因果关系^[1-4], “裙带关系”等非正式治理因素也深刻影响着公司违规活动^[5], 公司融资效率、经营效率、盈利能力和资金流动性等财务特征亦是与公司违规行为存在一定的相关性^[6,7]. 高管层面, 早在 20 世纪 90 年代, 就有研究指出 70% 以上的上市公司欺诈与 CEO 等高管密切相关^[8]. 高阶梯队理论和烙印理论认为, 在复杂环境和有限理性作用下, 高管的外在特征、内在特质及个人经历会塑造其日后的认知水平、行为模式与价值观, 通过烙印效应的外显从而影响个体的行为决策. 高管的外在特征(如年龄、性别、教育水平等)^[10,11]、内在

① 收稿日期: 2022-02-18; 修订日期: 2023-12-22.

基金项目: 国家社会科学基金资助项目(22&ZD148); 国家社会科学基金资助项目(20AGL015).

作者简介: 何 瑛(1973—), 女, 甘肃武威人, 博士, 教授, 博士生导师. Email: heyingen2001@126.com

特征(如过度自信、风险偏好等)和个人经历(如从军经历、学术经历、海外经历、复合型经历等)会对公司违规活动造成影响^[12,13]。

综上所述,学者们对于公司违规问题的核心影响因素进行了充分地研究,为进一步探讨如何预警并规避公司违规行为提供丰富的理论参考。然而,现阶段学术研究大多从公司或高管单一视角探究公司违规行为的影响因素,基于治理视角从公司和高管多维特征进行整体性研究或综合比较研究则相对少见。同时,相关因素在预测违规方面发挥的作用大小及区别也亟待探讨。对于单一核心影响因素的因果关系研究固然重要,对于影响因素的整体性预测模型构建同样具有重要的现实意义。已有针对上市公司违规行为的识别研究多集中于公司层面的财务指标和非财务指标^[14-16],少有模型综合考虑管理层要素对公司违规行为的预测作用,这与基于高阶梯队理论探讨公司违规影响因素的诸多研究脱节。同时,已有研究大多通过构建线性回归模型、Logistic 或 Probit 二元选择模型分析各因素对公司违规行为的影响,较少从预测角度出发验证各因素对于违规行为的预测能力,而机器学习作为结合统计学、计算机科学等多学科的应用工具,在数据发现、前景预测、理论检验等领域有较好应用,在公司治理与公司金融研究领域展现出较强优势^[17]。有鉴于此,本研究拟采用机器学习中的 LightGBM 和 SHAP 算法,基于内部治理视角选取重要的公司特征和高管个人特征构建上市公司违规行为预测模型,探究两类特征对上市公司违规行为的预测能力,并探讨不同类型特征对于预测违规的贡献程度及预测模式,一方面可以拓展和丰富高阶梯队理论及公司治理理论的研究成果,另一方面也为监管部门和投资者提升监管和投资效率、企业完善内部治理机制提供经验证据和决策参考。

本研究的贡献主要体现在三个方面:第一,综合应用高阶梯队理论与公司治理理论,对上市公司违规行为进行整体性预测研究,深入探究多维特征与公司违规问题之间的联系。本研究从预测视角出发,选取重要的影响因素(公司特征和高管个人特征)构建公司违规预测模型,从更加全面和系统的角度考察重要的多维度特征对公司违

规行为的预测性能,较原有研究更加关注高管能动性造成的影响;第二,在整体性预测研究的基础上,对公司特征和高管个人特征的预测性能进行综合比较研究,深化公司违规领域研究成果。现有应用机器学习的文章大都是通过添加足够多的特征来提高预测模型绩效,而较少考察分类特征对模型的预测作用,本研究对机器学习输出模型进一步分析,在辨析各类因素如何影响公司违规行为的同时,着重探究公司与高管特征两类因素发挥预测作用的性能大小,并进行排序对比,以更好地剖析不同维度因素的作用效果;第三,引入机器学习中的 LightGBM 算法和 SHAP 工具构建预测模型并进行归因分析,提升公司违规行为预测模型的科学性。本研究采用的 LightGBM 算法是一种基于梯度提升决策树算法改进的方法,在算法实现上有别于常用的逻辑回归、支持向量机等方法,具备运行速度更快、效率更高、准确度更高等优点。此外,传统的特征重要性度量方法不能保证特征归因的一致性和局部精确性^[19],而 SHAP 算法是基于合作博弈论的可加性解释模型,考察特征对于模型输出的边际贡献,一定程度上能够合理解决机器学习模型的“黑箱问题”,量化公司财务与非财务指标、高管特征指标在预测方面的作用效果,有利于更直观地解析特征变量对公司违规的影响模式。

1 文献综述

1.1 公司和高管特征与公司违规行为

目前已有研究大多从上市公司外部和内部探究违规行为的影响因素。外部影响因素主要发挥外部监督治理功能,比如司法独立性的提升可以加大违规公司被稽查的概率,扩大公司违规公告的负面反应^[20],证券分析师能够起到监督上市公司的作用,分析师跟踪可以减少上市公司欺诈倾向,降低公司发生违规的概率^[1]。

公司违规的内部影响因素主要涉及公司层面和高管层面,公司层面主要包括公司治理因素和公司财务因素。公司治理层面比较有代表性的影响因素是董事会特征和股权结构,从董事会特征

来看,董事会作为公司内部治理的重要环节,在经营决策过程中发挥重要作用,其与公司违规行为间存在显著的相关性^[14,23]。董事会规模与企业财务舞弊存在U型关系^[14],当董事会规模过大时,企业内部容易出现工作效率下降、管理层凌驾内控等问题,增加公司违规倾向^[2,25]。随着学术界对于非正式制度治理因素的关注程度日益提高,研究发现CEO对董事会影响力的增强以及与董事之间的“裙带关系”均会提高公司违规风险^[5,25],而董事会非正式层级能够一定程度上抑制公司违规^[26]。另外,独立董事制度的日益完善进一步促进学者们对于“独董”与公司违规之间关系的探索。研究发现独立董事比例的提高有助于提升公司治理效率,改进公司内控系统,改善信息披露程度^[27],但独立董事占比的提升并不总能遏制公司违规行为,也可能加剧管理层合谋风险,增大公司违规可能性^[28]。从股权结构方面来看,高管团队持股比例越高或大股东控制权越强,公司发生违规的可能性越小,股权激励能够促使管理者与股东利益趋同^[29,30];但行为理论研究发现,拥有股权激励的CEO可能会为提高股价和薪酬而采用非法手段进行财务重述,导致公司违规^[31]。当公司中有机构投资者参与持股时,机构投资者对公司财务活动存在一定监督作用,有利于减少企业违规倾向^[4]。

公司财务层面的影响因素包括企业的经营业绩和盈利能力、股价水平、财务杠杆水平等^[32]。研究发现杠杆率越高的企业违反贷款协议的可能性越大^[6],欠缺盈利能力和流动性的企业往往会在财务报告中出现更多的错误^[7],资金周转能力在一定程度上反应公司的竞争力,竞争失败、资产周转率表现不佳可能会刺激公司产生违规行为^[6]。根据美国证券交易委员会执法公司的统计数据可知,公司规模小的企业相对大的企业发生违规的概率更高^[33];公司所处行业低迷,或其业绩表现较差、债务危机严重时,往往更有可能发生违规^[34]。财务比率通常可以相对直接地描绘公司各方面的能力,已有研究证明财务杠杆、资产周转率、资产负债率和公司规模是影响财务报告舞弊的重要因素,在检测和预测欺诈性财务报告过程中发挥重要作用^[35,36]。

公司违规内部影响因素的另一个层面即高管层面。根据高阶梯队理论,高管是公司中制定和执行决策的重要个体,其年龄、性别、教育经历等方面的差异会影响公司的策略选择,进而影响公司绩效^[9]。诸多学者基于该理论从管理层要素角度探究公司违规行为的影响因素及作用机制。高管特征可以分为内在特征和外在特征,内在特征主要指价值观、认知水平等难以量化的心理层面特征,如已有研究发现自恋型CEO会采取比较大胆的行动,做出带有偏见的投资决策,对公司造成潜在不利后果,增大公司发生欺诈的可能性^[37,38]。外在特征主要指年龄、性别、教育背景、职业背景等易于测量和获取的人口统计学特征。中国文化倾向于将年龄与个人经验和知识直接匹配,年轻的CEO有可能因为经验不足发生违规行为^[11],而年长的高管有可能利用职权的控制力谋取私利,增大公司违规概率。性别方面,女性相比男性更加谨慎和保守,多表现出风险厌恶和规避^[39],女性CEO会降低公司违规的可能性^[10]。个人经历方面,高管的学术经历、从军或从政经历会塑造其更高尚的道德观和更强烈的社会责任感,对于公司舞弊行为具有显著抑制作用^[40]。

总体来看,关于公司违规行为的研究大多聚焦关键变量,强调单一因素对公司违规的影响。随着利益相关者对公司违规预警关注度的提高,部分学者也尝试从事前预测角度构建相应的公司行为识别模型。如洪文洲等^[41]基于公司财务和公司治理两类指标构建财务舞弊识别模型。钱苹和罗玫^[18]构建适合中国情境的财务舞弊预测模型Cscore,以帮助利益相关者通过阈值判别中国上市公司财务舞弊的可能性,并将Cscore模型与西方普遍应用的Mscore^[42]和Fscore^[43]进行比较,表明Cscore模型具备应用于中国情境的相对优势。叶康涛和刘金洋^[16]验证了非财务信息对财务舞弊行为的识别作用等。虽然上述研究强调模型的预测性,但应用传统数量统计方法构建的识别模型囿于模型本身的局限性,研究变量维度有限,研究结果多呈现为线性关系,并较难得到真实的预测效果,对公司违规行为的识别能力存在先天不足。而新兴的机器学习方法能够弥补传统研究方法在变量维度、非线性关系、变量贡献度比较等方

面的不足^[44],因此有诸多学者开始探索机器学习在公司行为预测领域的应用。

1.2 机器学习与公司行为

随着心理学、社会学、计算机科学等领域的跨学科交叉融合发展,以机器学习为代表的新兴研究工具在公司金融、公司治理领域的研究应用日益增多,已有研究肯定机器学习对于金融研究、计量经济学演进、政策评估等具有的积极意义^[52,53]。机器学习是一种以逻辑实证主义为方法论基础、以数据导向的研究模式为研究范式的数据处理方式。根据数据是否携带标签(label),机器学习可分为“有监督”学习、“无监督学习”及有监督和无监督相结合的“半监督”学习。其中,有监督学习是从给定的、带有标记的训练数据中推导预测函数的机器学习任务,在构建公司行为预测模型方面具有较强的优势,常见的预测模型包括决策树、岭回归、Lasso回归、支持向量机等。其中,决策树(decision tree)是一种基本的分类与回归方法,构建的模型具有较强可读性,广泛用于研究公司行为问题并表现出较好的预测效果。陆瑶等^[57]利用Boosting回归树算法考察多维度高管特征对公司业绩的预测能力,得到比传统模型更精准的结果。也有部分研究涉及财务舞弊预测模型构建,如应用决策树方法识别违规公司,通过引入原始财务数据构建基于集成学习(adaboost)的财务舞弊预测模型,并对比证实了集成学习相对Logistic回归模型更优越的预测性能^[55];应用GBRT(又名GBDT)算法进行会计错报检测,帮助监管部门更准确地识别存在会计错报的公司^[56];应用随机森林算法对公司不当行为的影响因素进行重要性排序,并实现对于公司不当行为的预测,证实公司治理特征与公司发生不当行为之间的显著关系^[59];通过XGboost算法构建递延所得税异动指标与上市公司违规的决策树模型,帮助各方实现对违规行为的精准预警^[61]。上述新兴研究都展现了机器学习在预测公司行为方面的优势,同时机器学习方法本身也在不断更新迭代,LightGBM算法是当前更为先进的决策树回归算法,具备运行效率高、模型精度高、处理海量数据等一系列优势,在实务界和学术界被广泛认可和应用,但目前鲜有公司治理领域文献应用LightGBM算法

构建多维度的公司违规识别模型,同时机器学习输出模型的“黑箱”问题使得研究人员难以像应用传统模型一样直观了解指标的作用,现有针对解释机器学习模型特征重要性的度量方法仍存在一定争议。

1.3 文献述评

综上所述,已有大量研究从不同视角探究公司违规行为的影响因素,并将机器学习方法应用于公司行为预测之中,为本研究提供了良好的理论基础和方法依据,但仍然具有一定的局限性:第一,截至目前,多数文献从公司或高管层面的单一维度探究公司违规行为的影响因素,导致公司违规行为研究的涉及层面较广但系统性不强,缺少基于内部治理视角,选取重要影响因素对上市公司违规行为进行的整体性预测研究和分类比较研究,尚未建立多维度变量集成且具备较强应用性的模型;第二,现有文献囿于传统计量模型的局限性,很难突破变量维度有限、关系呈现单一、结果不够直观等问题;第三,机器学习的应用极大拓展经济管理领域研究的广度与深度,但机器学习输出模型的“黑箱”问题加大理解影响因素内在预测机制的难度。基于此,本研究尝试利用机器学习中的LightGBM和SHAP构建预测模型,分析涵盖公司特征和高管个人特征在内的影响因素对预测上市公司违规行为的贡献,探究并对比两类特征对于预测的贡献度,考察公司层面与高管层面特征在预测中的影响力及预测模式。值得说明的是,尽管基于集成学习构建预测模型可以容纳高维度的变量,但理论指导对特征选择仍有重要意义^[55]。本研究通过机器学习得到的预测性模型与传统实证研究的解释性模型并非相互独立,而是基于传统实证研究已探查清晰因果关系与相关关系的基础上构建预测模型,是对现有文献观点的验证与再延伸。不论是传统的“解释”模型还是当下的“预测”模型,均对不断发展完善现阶段我国公司治理机制有重要的参考意义。

2 研究设计

2.1 研究方法

本研究在构建模型时主要应用的机器学习方

法为基于梯度提升决策树(GBDT)思想的 LightGBM 算法,GBDT 是梯度提升(gradient boosting)方法在决策树基学习器上应用实现的一类集成学习算法,其将决策树和集成学习思想有机结合,具有较强的泛化能力,训练效果好且不易产生过拟合问题,是目前各领域学科应用较广泛的机器学习方法。LightGBM 是对于 GBDT 算法框架的改进实现,由微软研发并在 GitHub 上开源,训练速度更快,占用内存消耗更低,且支持分布式能够更高效地应对大数据样本,更适合作为探讨公司及高管特征对于上市公司违规行为预测性能的主要机器学习方法。

在对 LightGBM 输出模型进行解释时,本研究发现传统的全局特征重要性度量方法尚存争议,不能保证特征归因(feature attribution)的一致性和准确性,因而选用 SHAP 对结果进行深入探讨。SHAP(SHapley Additive exPlanation)是由合作博弈论启发诞生的可加性解释模型,主要用作解释基于树的集成方法类机器学习模型输出,有效缓解输出模型身处“黑箱”的问题,为学者通过模型结果分析不同特征的预测效果及预测模式提供思路。上市公司违规行为可以被视作多种影响因素合成的结果,通过 SHAP 计算因素的 Shapley 值可以更清晰地对模型进行机理解释。SHAP 的主要目标是通过计算每个特征对预测的贡献来解释实例 x 的预测

$$y_m = y_0 + \sum_{n=1}^N g(x_{mn}) \quad (1)$$

其中 x_{mn} 表示第 m 个样本的第 n 个特征, y_m 表示模型对于第 m 个样本的预测值, y_0 是模型对样本的预测基准值(指所有样本解释变量的均值), $g(x_{mn})$ 表示为 x_{mn} 的 Shapley 值,其服从关系式

$$g(x_{mn}) = \sum_S \frac{|S|!(n-|S|-1)!}{n!} \times (f_x(S \cup \{x_n\}) - f_x(S)) \quad (2)$$

其中 f_x 表示 LightGBM 拟合训练好的学习器, $\{x_1, x_2, \dots, x_N\}$ 指所有特征的集合, S 表示不含 x_n 的特征子集, x_{mn} 对应的 Shapley 值绝对值 ($|g(x_{mn})|$) 与其对于样本预测结果的影响呈正

比关系,正负表示特征对于预测结果的增强或减弱作用。应用 SHAP 有助于进一步探讨特征对于公司违规的预测模式,而不仅局限于模型的预测性能。

2.2 变量选取及说明

通过文献回顾可知,诸多关键的公司治理、公司财务及高管特质因素均与企业违规行为存在一定联系。根据委托代理理论,现有关于董事会规模及构成、公司股权结构、外部监督与违规行为作用过程的研究较为充分,诸多财务指标亦已被证实能够有效预测财务舞弊行为,因此本研究重点选取 6 个公司治理层面特征、10 个公司财务层面特征作为公司层面的解释变量。公司治理层面的特征主要反映公司的股权结构和董事会特征及部分上市公司基本信息,公司财务层面的特征主要反映公司的盈利能力、成长能力、营运能力、偿债能力和市场表现。本研究重点关注引入高管特征对于预测模型的提升效果,根据高阶梯队理论和烙印理论,高管的年龄、性别、个人早年过往经历等外在特质会对其行为决策造成影响,部分高管特征已被证实与违规行为存在一定因果关系,因而,参考陆瑶等^[57] 相关研究,本研究重点选取 13 个高管个人特征作为高管层面的解释变量,涵盖高管先天特质、职业经历、教育背景等信息。

本研究选取 2008 年—2019 年中国 A 股上市公司为研究样本。其中,最高管理层人员个人特征数据、公司财务与治理数据、公司违规数据来自国泰安经济金融研究数据库(CSMAR)“人物特征系列——上市公司人物特征”、“股票市场系列”和“公司研究系列”等多个子数据库,本研究研究对象中的高管指公司的 CEO 个人与董事长个人。本研究考虑到发现违规行为存在一定滞后性,因此追溯证监会发布的公告,将上市公司违规行为的实际发生日期确定为上市公司违规年份。其中,本研究所指“公司违规”,涵盖财务舞弊、公司运营规范以及主体主观疏忽等多类问题,不包括“高管违规但公司不违规”的情形。本研究最终选取 1 个被解释变量和 29 个解释变量分别作为训练的标签和特征,具体的变量名称与说明如表 1 所示。

表1 变量选取及说明

Table 1 Variable selection and introduction

变量名称		变量定义及处理
Panel A: 上市公司违规		
<i>Fraud</i>	公司违规	虚拟变量, 1表示公司对应年份实际发生违规行为, 0表示没有发生
Panel B-1: 公司财务层面特征		
<i>TobinQ</i>	托宾Q值	市值/资产总计
<i>Growth</i>	营业收入增长率	(营业收入本年本期金额 - 营业收入上年同期金额) / 营业收入上年同期金额
<i>ART</i>	应收账款周转率	营业收入/应收账款期末余额
<i>Leverage</i>	资产负债率	负债合计/资产总计
<i>Asset</i>	资产总计	资产各项目之总计, 单位为百万元
<i>Size</i>	上市公司规模	当年上市公司年初总资产合计的对数
<i>ROA</i>	总资产净利润率	净利润/总资产余额
<i>StdROE</i>	业绩波动性	近三年ROE的标准差
<i>Return</i>	年个股回报率	不考虑现金红利再投资的年个股回报率
<i>Incd</i>	行业代码	参照证监会行业分类2012年版, A01~S90分别转换为0~83
Panel B-2: 公司治理层面特征		
<i>Nshrms</i>	高管团队持股占比	董事、监事、高级管理人员持股数/股本总数
<i>Nshrstt</i>	国有股占比	国有股/股本总数
<i>Bdnum</i>	董事会人数	董事(含董事长)人数
<i>Indep</i>	独董占比	独董人数/董事会人数
<i>Analyst</i>	分析师关注度	对应年份内对该公司进行过跟踪分析的分析师(团队)数量
<i>Opacity</i>	上市公司信息透明度	以深交所披露为准, 采用标签编码将A~D转换, 具体为: 1表示优秀, 2表示良好, 3表示及格, 4表示不及格, -1表示空值
Panel C: 高管个人特征		
<i>Gender</i>	性别	1表示男性, 0表示女性
<i>Age</i>	年龄	高管的年龄
<i>GTenure</i>	任期	高管的既有任期
<i>SharEnd</i>	年末持股数占比	高管的年末持股数/股本总数
<i>IsDuality</i>	兼任董事长和CEO	1表示CEO兼任董事长, 0表示不存在兼任
<i>IsCocurP</i>	股东单位兼任	1表示在股东单位兼任, 0表示没有
<i>ProBack</i>	生产运作职业经历	1表示有生产运作的职业经历, 0表示没有
<i>RDBack</i>	研发职业经历	1表示有研发的职业经历, 0表示没有
<i>MktBack</i>	市场营销与公共关系职业经历	1表示有市场营销与公共关系的职业经历, 0表示没有
<i>LawBack</i>	法律职业经历	1表示有法律的职业经历, 0表示没有
<i>OveseaBack</i>	海外经历	1表示有海外背景, 0表示没有
<i>Academic</i>	学术经历	1表示有学术背景, 0表示没有
<i>FinBack</i>	金融背景	1表示有金融背景, 0表示没有

2.3 变量描述性统计

本研究所述的违规是指公司在信息披露、公司规范运营、主体疏忽等方面存在违法违规行为, 包括财务舞弊问题等^[1]。

从2008年—2019年整体平均来看, 每年违规公司约占上市公司的15%, 其中2012年违规

公司占比最高, 达到20%左右。

本研究对样本的缺失值进行如下处理: 对于公司特征, 采用该特征所属公司此特征的平均值加以填充; 对于高管特征, 按照样本公司内该高管此特征的平均值加以补充, 其中年龄和任期按照时间顺序补充, 如若还有空缺则删除该样本, 同时

剔除金融行业数据。数据处理后本研究共获取 CEO 特征样本 21 041 条,表 3 展示了公司财务与治理特征、高管个人特征的描述性统计,为避免极端值造成影响,所有连续型变量均已进行 1% 和 99% 的缩尾(winsorize)处理,数据集不重复记录同一公司在同一年内的多次违规行为。

表 2 公司年度违规统计

Table 2 Annual violation statistics of the company

年份	公司数量/家	违规公司数量/家	没有违规公司数量/家	违规公司占比/%
2008	1 341	182	1 159	13.57
2009	1 518	215	1 303	14.16
2010	1 862	236	1 626	12.67
2011	2 190	390	1 800	17.81
2012	2 404	473	1 931	19.68
2013	2 354	390	1 964	16.57
2014	2 393	347	2 046	14.50
2015	2 623	421	2 202	16.05
2016	2 907	449	2 458	15.45
2017	3 232	456	2 776	14.11
2018	3 274	542	2 732	16.55
2019	3 437	478	2 959	13.91

公司财务特征方面,样本公司总资产净利润率的均值和中位数分别为 3% 与 4%,资产负债率的均值和中位数分别为 41% 与 39%,50% 的公司的托宾 Q 值都能达到 1.7 以上,均值约为 2.24,观测公司的盈利能力、偿债能力、市场绩效指标数据基本呈对称分布;近一半公司的营业收入增长率达到 13%,应收账款周转率高于 4.73,均值分别为 25% 和 41.73,公司在成长能力、营运能力、股票年收益率指标方面表现的差异相对加大。治理特征方面,高管团队持股比例均值为 15% 左右,约为高管个人持股的二倍,国有股持股占比均值接近 0,至少一半以上的公司会保持 8 人~9 人的董事会规模,其中,独立董事占比均值约为 38%;47% 的上市公司信息透明度可以达到良好,11% 的公司可以达到优秀;大约 11% 的上市公司会被 1 名分析师(团队)跟踪分析,10% 的公司会被至少 20 名的分析师(团队)关注。

高管个人特征方面,观测样本中高管性别特征集中,平均约有 93% 的 CEO 由男性担任;高管

年龄大多分布在 40 岁~55 岁之间,CEO 的平均年龄约为 49 岁,其中,46 岁~50 岁年龄段分布的人数最多,约占总人数的 29%,其次是 51 岁~55 岁,约占总人数的 24%,35 岁以下人数占比最少,仅为 3% 左右。CEO 平均任期长度约为 3 年,约有 67% 的 CEO 任期小于等于 3 年;CEO 年末持股占比的均值约为 7%,约有 35% 的 CEO 和董事长由一人担当,约 25% 的 CEO 有在股东单位兼任的经历。职业经历方面,CEO 普遍具备丰富的管理职能经验,约有 99% 的 CEO 曾有过市场营销与公共

表 3 样本的描述性统计分析

Table 3 Descriptive statistics of samples

指标	样本量	平均值	标准差	中位数
Panel A - 1: 公司财务层面特征				
<i>TobinQ</i>	21 041	2.24	1.59	1.72
<i>Growth</i>	21 041	0.25	0.72	0.13
<i>ART</i>	21 041	41.83	197.72	4.73
<i>Leverage</i>	21 041	0.41	0.22	0.39
<i>Asset</i>	21 041	5 936.25	14 704.96	2 426.67
<i>Size</i>	21 041	21.53	1.17	21.44
<i>ROA</i>	21 041	0.03	0.09	0.04
<i>StdROE</i>	21 041	0.12	0.33	0.03
<i>Return</i>	21 041	0.13	0.60	0.00
<i>Incd</i>	21 041	35.98	16.65	33.00
Panel A - 2: 公司治理层面特征				
<i>Nshrsms</i>	21 041	0.15	0.19	0.02
<i>Nshrstt</i>	21 041	0.01	0.07	0.00
<i>Bdnum</i>	21 041	8.30	1.54	9.00
<i>Indep</i>	21 041	0.38	0.05	0.33
<i>Analyst</i>	21 041	6.82	9.02	3.00
<i>Opacity</i>	21 041	1.18	1.50	2.00
Panel B: 高管(CEO)个人特征				
<i>Gender</i>	21 041	0.93	0.26	1.00
<i>Age</i>	21 041	48.66	6.84	49.00
<i>GTenure</i>	21 041	3.09	2.35	2.00
<i>SharEnd</i>	21 041	0.07	0.13	0.00
<i>IsDuality</i>	21 041	0.35	0.48	0.00
<i>IsCocurP</i>	21 041	0.25	0.43	0.00
<i>ProBack</i>	21 041	0.10	0.31	0.00
<i>RDBack</i>	21 041	0.22	0.42	0.00
<i>MktBack</i>	21 041	0.99	0.08	1.00
<i>LawBack</i>	21 041	0.01	0.10	0.00
<i>OveseaBack</i>	21 041	0.08	0.27	0.00
<i>Academic</i>	21 041	0.20	0.40	0.00
<i>FinBack</i>	21 041	0.07	0.25	0.00

关系的职业背景;其次,约22%的CEO有过研发背景,20%左右的CEO有过高校任教、科研机构任职、协会从事研究至少某一方面的经历;再次,约10%的CEO有生产运作方面的经历,约8%的CEO有过海外任职或求学的经历,7%的CEO曾从事过金融行业;最后,仅有不到1%的CEO有法律背景。

2.4 机器学习模型构建

已有文献检验了财务指标、非财务指标构建预测模型的性能,参照现有财务舞弊模型及陆瑶^[57]等人的研究,本研究构建两组模型,分别是公司特征模型和CEO特征模型。基于公司财务指标和非财务指标构建的模型称为公司特征模型,在此模型基础上加入CEO特征,构成CEO特征模型。本研究通过分析特征加入前后模型预测效能的提升程度,考察高管个人特征对公司违规预测的贡献。具体如表4所示。

本研究的实证研究分为三大部分。第一部分,构建基于LightGBM算法的预测模型。在划分样本数据集时,由于没有发生违规的公司占全体上市公司观测值中的多数,负类样本数量是正类样本

数量的5倍~6倍,数据存在严重的不平衡现象,因此本研究未按照一年滚动窗口期拟合模型,参照Bertomeu等^[56]研究,本研究进行如下样本划分。

表4 模型构建

Table 4 Model building

模型	特征 (解释变量)	标签 (被解释变量)
公司特征模型	<i>Nshrms</i> 、 <i>Nshrstt</i> 、 <i>Bdnum</i> 、 <i>Indep</i> 、 <i>TobinQ</i> 、 <i>Growth</i> 、 <i>ART</i> 、 <i>Leverage</i> 、 <i>Asset</i> 、 <i>ROA</i> 、 <i>StdROE</i> 、 <i>Return</i> 、 <i>Indcd</i> 、 <i>Analyst</i> 、 <i>Size</i> 、 <i>Opacity</i>	Fraud
CEO特征模型	<i>Nshrms</i> 、 <i>Nshrstt</i> 、 <i>Bdnum</i> 、 <i>Indep</i> 、 <i>TobinQ</i> 、 <i>Growth</i> 、 <i>ART</i> 、 <i>Leverage</i> 、 <i>Asset</i> 、 <i>ROA</i> 、 <i>StdROE</i> 、 <i>Return</i> 、 <i>Indcd</i> 、 <i>Analyst</i> 、 <i>Size</i> 、 <i>Opacity</i> 、 <i>Gende</i> 、 <i>Age</i> 、 <i>GTenure</i> 、 <i>SharEnd</i> 、 <i>IsDuality</i> 、 <i>IsCocurP</i> 、 <i>ProBack</i> 、 <i>RDBack</i> 、 <i>MktBack</i> 、 <i>LawBack</i> 、 <i>OveseaBack</i> 、 <i>Academic</i> 、 <i>FinBack</i>	

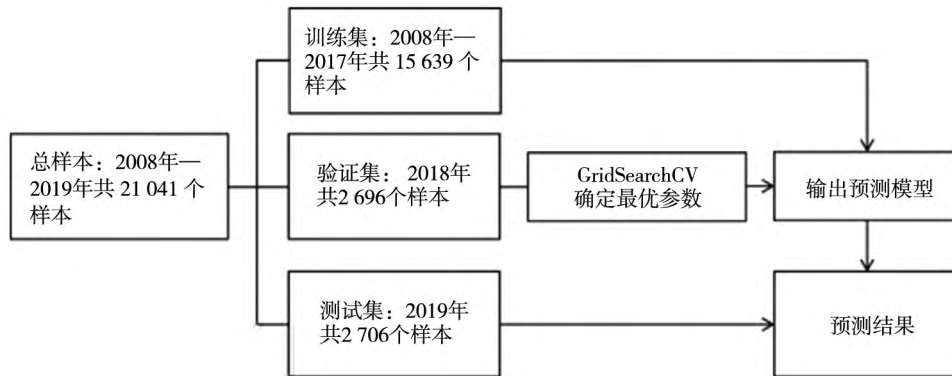


图1 样本划分

Fig.1 Samples division

其中,训练集用于构建模型,验证集用于检验模型性能,并在训练过程中监测模型拟合状态,同时,验证集和训练集可结合作用于调整超参数,测试集用于分析该模型的最终表现。考虑到公司违规行为可能存在的时滞问题,本研究将2008年—2017年的数据划分至训练集,将2018年和2019年的数据分别划分至验证集和测试集,这两个数据集中的样本与测试集中样本的时间跨度不同,有助于更

好地观察模型对未来的预测性。值得注意的是,该研究中违规公司属于少数类别,数据集的不平衡会增加预测难度,降低分类器性能,因此本研究参考Bao等^[55]研究,采用欠采样的方式解决类别不平衡问题,平衡违规公司与没有发生违规公司的样本比例使之达到1:1,之后根据训练集和验证集对数据进行调参、训练,确认预测模型,并通过测试集对模型进行测试,评估其泛化能力。

本研究通过网格搜索和交叉验证 (GridSearchCV) 确定 LightGBM 核心参数 (core parameters) 及学习控制参数 (learning control parameters) 的最优取值。核心参数主要与模型训练相关。

表 5 核心参数

Table 5 Core parameters

参数名称	参数释义
<i>objective</i>	确定模型用途, 本研究选用 binary
<i>num_boost_round</i>	指迭代次数, 一般大于 100
<i>learning_rate</i>	指梯度下降的步长, 用来确定新加入的基学习器权重
<i>num_leaves</i>	指每个基学习器的最大叶子节点, 由于 LightGBM 使用的是 leaf-wise 算法, 需要通过 <i>num_leaves</i> 调节树的复杂程度, 通常叶子节点数越大结果越精确, 但也易出现过拟合现象, 设置该参数时需要保证 $num_leaves \leq 2^{max_depth}$

学习控制参数主要与决策树的构建相关。

表 6 学习控制参数

Table 6 Learning control parameters

参数名称	参数释义
<i>min_data_in_leaf</i>	一个叶子节点中最小的数据量, 该参数的大小取决于数据的样本量及 <i>num_leaves</i>
<i>max_depth</i>	指每个基学习器的最大深度, 主要用来协调数据量与模型过拟合问题
<i>feature_fraction</i>	用于确定每棵树训练之前模型对特征子采样的比例, 用来控制过拟合
<i>bagging_fraction</i>	类似于 <i>feature_fraction</i> , 用于控制每次迭代时样本的采样比例
<i>bagging_freq</i>	与 <i>bagging_fraction</i> 同时设置使用, 指 <i>bagging</i> 使用频率
<i>lambda_l1</i> 和 <i>lambda_l2</i>	分别表示 L1 和 L2 正则化项的权重系数

本研究主要通过衡量二分类模型的评价指标 AUC (area under curve) 考察模型在三个数据集上的表现, 将其分别记为 *auc_train*、*auc_valid* 和 *auc_test*, 其计算公式为

$$AUC = \frac{\sum_{ins_i \in \text{positive class}} rank_{ins_i} - \frac{M \times (M + 1)}{2}}{M \times N} \quad (3)$$

其中 $rank_{ins_i}$ 表示第 i 个样本的序号, M 和 N 分别表示正样本和负样本的个数。AUC 值指根据模型

计算得出的 *score* 值将正样本排在负样本前面的概率, *score* 指每个测试样本属于正样本的概率。同时 AUC 值等于 ROC 曲线下面积, 通过 ROC 曲线图可以更容易地考察阈值对学习器泛化性能的影响。对于分类问题, 阈值的设定会直接影响模型输出的类别, 最佳阈值的取值通常由建模的目的决定。本研究主要考察高管特征对于模型性能的提升作用, 因而取阈值 0.5, 在公司特征模型、CEO 特征模型间横向对比准确率 (*accuracy*)、精准率 (*precision*) 及召回率 (*recall*)、Kappa 系数评价指标, 上述评价指标均由混淆矩阵计算得出。其中 *True Positives* 表示真正例, 意为样本是正时预测结果为正的个数, 即违规公司被预测为违规的个数, 记为 *TP*; *True Negatives* 表示真反例, 意为样本是负时预测结果为负的个数, 即没有违规公司被预测为没有违规的个数, 记为 *TN*; *False Positives* 表示假正例, 意为样本是负时预测结果为正的个数, 即没有违规公司被预测为违规的个数, 记为 *FP*; *False Negatives* 表示假反例, 意为样本是正时预测结果为负的个数, 即违规公司被预测为没有违规的个数, 记为 *FN*。

表 7 混淆矩阵

Table 7 Confusion matrix

true class	hypothesized class	
	0	1
0	<i>True Negatives</i>	<i>False Positives</i>
1	<i>False Negatives</i>	<i>True Positives</i>

根据混淆矩阵, 准确率指正确分类的比例, 其计算公式为

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (4)$$

精准率指预测为正的样本中正确样本所占比例, 其计算公式为

$$Precision = \frac{TP}{TP + FP} \quad (5)$$

召回率指记为正的样本中被正确预测为正的比例, 其计算公式为

$$Recall = \frac{TP}{TP + FN} \quad (6)$$

Kappa 系数是统计学中检验一致性的指标, 可以用于衡量分类精度, 取值范围为 $[-1, 1]$, 取

值越高说明模型分类的准确度越好,一般大于0时就说明模型具有一定意义,其计算公式为

$$k = \frac{P_0 - P_e}{1 - P_e} \quad (7)$$

$$P_0 = Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (8)$$

$$P_e = \frac{(FN + TN) \times (TN + FP) + (TP + FN) \times (TP + FP)}{(TP + TN + FP + FN)^2} \quad (9)$$

上述指标可以有效评估分类器对上市公司违规行为的预判效果。

第二部分,通过 SHAP 模型探究特征相对重要性的排序,分析单个特征对于上市公司违规的预测模式。本部分内容旨在通过合作博弈论思想解释第一部分 LightGBM 训练模型的输出结果,以考察不同特征对于预测结果的贡献。通过计算每个样本不同特征的 SHAP 值,可以得出特征整体对于预测的贡献度,进而从中总结得出影响力较强的特征,以及影响的正负性。如果大多数样本在某一特征上表现出一致趋势,可通过特征 SHAP 值与实际值的关系图即 SHAP 依赖图,观察到该特征的作用方向。

第三部分,通过更换 CEO 特征为董事长特征、更换机器学习实现方法、调整 LightGBM 模型参数、更换数据的时间窗口期对第一部分的分析结论进行稳健性检验。

3 实证结果分析

3.1 模型输出结果分析

通过 LightGBM 得到公司特征模型与 CEO 特征模型对于样本的训练结果,表 8 展示了模型对于违规的预测效果。

1) 通过训练结果可知,公司特征模型、CEO 特征模型均对判断上市公司是否违规发挥较好的预测作用。其中,在验证集和测试集上,CEO 特征模型的 AUC 值均达到 80% 以上,公司特征模型的 AUC 值约为 80% 左右,两组模型的预测效果均优于随机分类时 50% 的概率,说明模型具备一定分类能力。

表 8 模型(LightGBM) 输出结果

Table 8 Main results of model (LightGBM)

数据集	指标	公司特征模型(1)	CEO 特征模型(2)	提升效果(2)-(1) / %
训练集	<i>auc_train</i>	0.962 2	0.961 0	-0.12
验证集	<i>auc_valid</i>	0.799 8	0.805 4	0.55
测试集	<i>auc_test</i>	0.802 7	0.814 2	1.15
	<i>Accuracy</i>	0.664 4	0.688 5	2.40
	<i>Precision</i>	0.359 1	0.378 1	1.90
	<i>Recall</i>	0.799 3	0.792 1	-0.72
	<i>Kappa</i>	0.294 9	0.322 8	2.79

图 2 展示了两组模型的 ROC 曲线(receiver operating characteristic curve)图,图的横坐标代表 FPR(假阳性率),指模型错认为违规但没有违规公司占所有没有违规公司的比例,纵坐标代表 TPR(真阳性率),指模型识别为违规公司占所有违规公司的比例,ROC 曲线下面积即 AUC 值,ROC 曲线图反映了不受阈值设定影响情况下模型的预测性能。

2) 随着高管特征的加入,模型的预测性能得到轻微改善,具体表现为 AUC 值在验证集和测试集的得分提升,但提升十分有限。精准率(*precision*)的提升伴随着召回率(*recall*)的轻微下降,模型整体的泛化能力并未发生明显改善,这在一定程度上说明本研究选取的高管特征对于预测上市公司是否违规发挥的作用有限,且有增加模型过拟合的风险。导致预测性能提升微弱的可能原因是选取指标有限,本研究选取的高管特征主要为客观测度指标,包括年龄、性别、过往经历等,诸如高管内在特质、管理风格等较为隐性并且通常难以观测度量的指标未纳入考虑。除此之外,高管在某些特征方面的低差异性也导致个人特征在预测模型中发挥的作用有限,如 90% 以上的 CEO 是男性且拥有市场营销与公共关系的职业经历,99% 以上的 CEO 没有法律背景。在模型应用过程中,公司的财务指标和治理指标相比高管特征指标更能反映一家公司可能蕴藏的违规风险,应合理关注公司及高管特征反映出的信息。

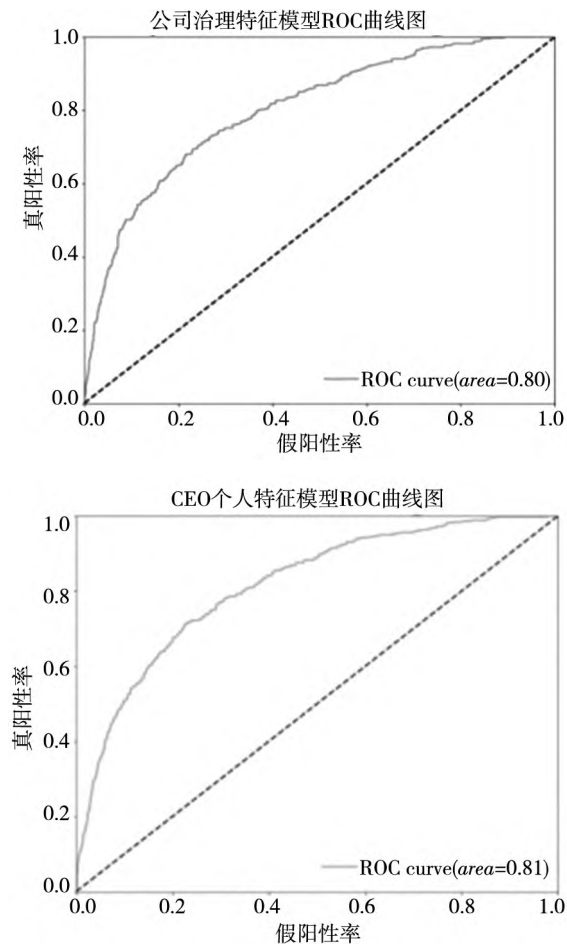


图2 模型 ROC 曲线图

Fig. 2 ROC curves of models

3) CEO 特征模型对于违规预测的准确率约为 69% , 精准率约为 38% , 召回率约为 79% , 表现为具备相对较低的查准率(精准率)和相对较高的查全率(召回率) , 公司特征模型具有类似的表现. 当阈值设为 0.5 时, 模型具有从全部违规公司中预测出近 80% 违规公司的能力, 但模型预测为违规的公司中实际发生违规行为的公司只有不到 40% , 这主要是因为样本存在不平衡现象, 实际生活中发生违规的公司占全部公司的少数, 仅为 15% 左右, 这导致测试集中包含大量负样本, 致使模型对于没有发生违规公司误判的绝对值数量远大于对于违规公司正确判断的数量, 因而相对指标取值较低. 在进一步实践过程中, 应当根据分类问题的需求对精确率和召回率提出不同的要求, 设定符合需求场景的“最佳阈值”, 如监管部门需要对存在违规嫌疑的公司进一步深究时, 应要求模型提供更高

的精准率, 而当监管过程中需要尽可能大比例、大范围地识别违规公司, 则要求模型拥有更高的召回率. 公司特征模型、CEO 特征模型的 $Kappa$ 系数约为 0.3 左右, 模型实现的分类准确度为一般水平.

3.2 特征的重要性分析

在 LightGBM 算法构建的 CEO 特征模型基础上, 本研究为进一步解析特征对于模型实现预测发挥的作用, 引入 SHAP 算法进行后续探讨. 通过 SHAP 得到模型特征的 *Shapley* 值, 如图 3 所示, 在对样本的 *Shapley* 平均绝对值进行排序后可以观测到各个特征重要性的名次, 排序越靠前的特征对于解释违规贡献的作用越大; 摘要图则在重要性排序基础上, 增加特征与其 *Shapley* 值的对应关系, 图中的每一行代表一个特征, 每个散点代表一个样本, 点在 y 轴方向的位置由特征重要性的排序决定, 在 x 轴方向的位置由样本的 *Shapley* 值决定, 通过散点的位置可以观测该实例的影响是与较高或较低的预测相关联, 颜色越深则特征观察数值越大, 越浅则反之.

由图 3 可知, 公司特征对于判别上市公司是否违规贡献的重要性约占特征整体的 89% , 其中, 财务特征占比约 56% , 治理特征占比约 33% , 与之相比, 高管个人特征对于模型判别发挥的作用较低, 其对于预测违规贡献的重要性仅占特征整体的 11.44% . 通过进一步研究可知, 特征 *Opacity* 的重要性较高, 约为 19.93% , 其在摘要图中的样本分布较为分散, 说明该特征影响较大; 其次是财务层面的 2 个特征, 分别为 *ROA*、*Leverage* , 其中特征 *ROA* 占比较高, 约为 13.72% , *Leverage* 次之, 约为 10.81% , 其余财务层面指标 *Indcd*、*StdROE*、*TobinQ* 等特征的贡献介于 4% ~ 7% 之间, *Asset*、*Growth*、*Size*、*ART*、*Return* 的贡献介于 2% ~ 4% , 5 个指标的重要性差距较小. 治理层面指标方面, *Nshrsms* 的重要性相对较高, 大于 4% , *Analyst*、*Indep*、*Bdnum*、*Nshrstt* 指标的重要性相对较小. 整体来看, 公司特征重要性低于 10% 的指标大概分为三档, 第一档的重要性高于 4% , 第二档的重要性介于 2% ~ 4% 之间, 第三档低于 1% , 同档位范围内的特征对于预测公司违规行为贡献的作用较为相似.

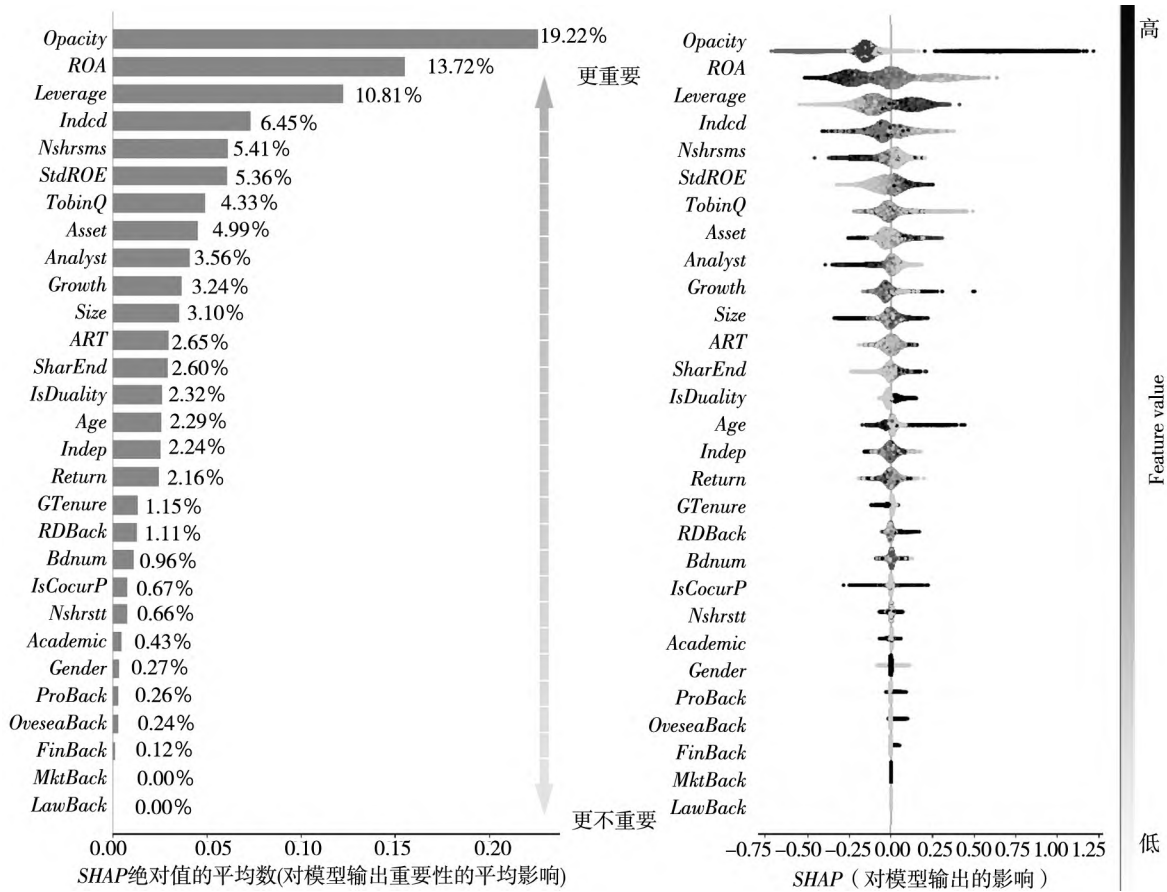


图3 模型(LightGBM)输出的重要性排序及摘要图
 Fig. 3 Importance map and summary of model(LightGBM) output

高管特征方面, *SharEnd*、*IsDuality*、*Age* 的贡献较为靠前且相近, 重要性贡献程度均介于 2% ~ 3% 区间内, 其次是 *GTenure*、*RDBack*, 最后重要性占比低于 1% 的特征有 *IsCocurP*、*MktBack*、*Academic*、*Gender*、*ProBack*、*OveseaBack*、*FinBack*、*LawBack*, 这些特征对于预测公司违规行为产生相对较低的关联。综上, 高管个人特征对于预测的贡献劣于公司特征, 其中 *SharEnd*、*IsDuality*、*Age* 是高管个人特征中贡献相对较大的指标, 其重要性占比分别约为 2.60%、2.32% 和 2.29%。

3.3 特征的预测模式分析

通过上部分的分析可知, CEO 特征模型的样本点弥漫很少出现较为均匀的颜色分布, 这说明大部分特征对于违规的解释都不是单一的线性关系, 少部分特征与模型预测公司违规行为呈现较为明显的单调相关性。借助 SHAP, 可

以进一步得到特征的 SHAP 依赖图, 与传统的部分依赖图有别, SHAP 依赖图的纵坐标是特征的 Shapley 值而非解释变量的数值。通过依赖图能够较为清晰地观测样本点的分布, 对于探讨单个特征如何影响模型的输出结果有一定参考意义。如图 4 所示, 图中的点代表某一上市公司观测值, 横轴表示该样本点的实际值, 纵轴表示该样本点对应的 Shapley 值, 记为 SHAP 值。特征的 SHAP 值表示该特征对模型预测结果的影响, SHAP 的绝对值代表影响力的大小, 符号代表影响的正负性。

3.3.1 公司特征与公司违规

在 CEO 特征模型中, 公司特征对于预测公司违规行为的影响更加显著, 本研究根据重要性排序列示公司特征 SHAP 依赖图, 其中不含行业代码。

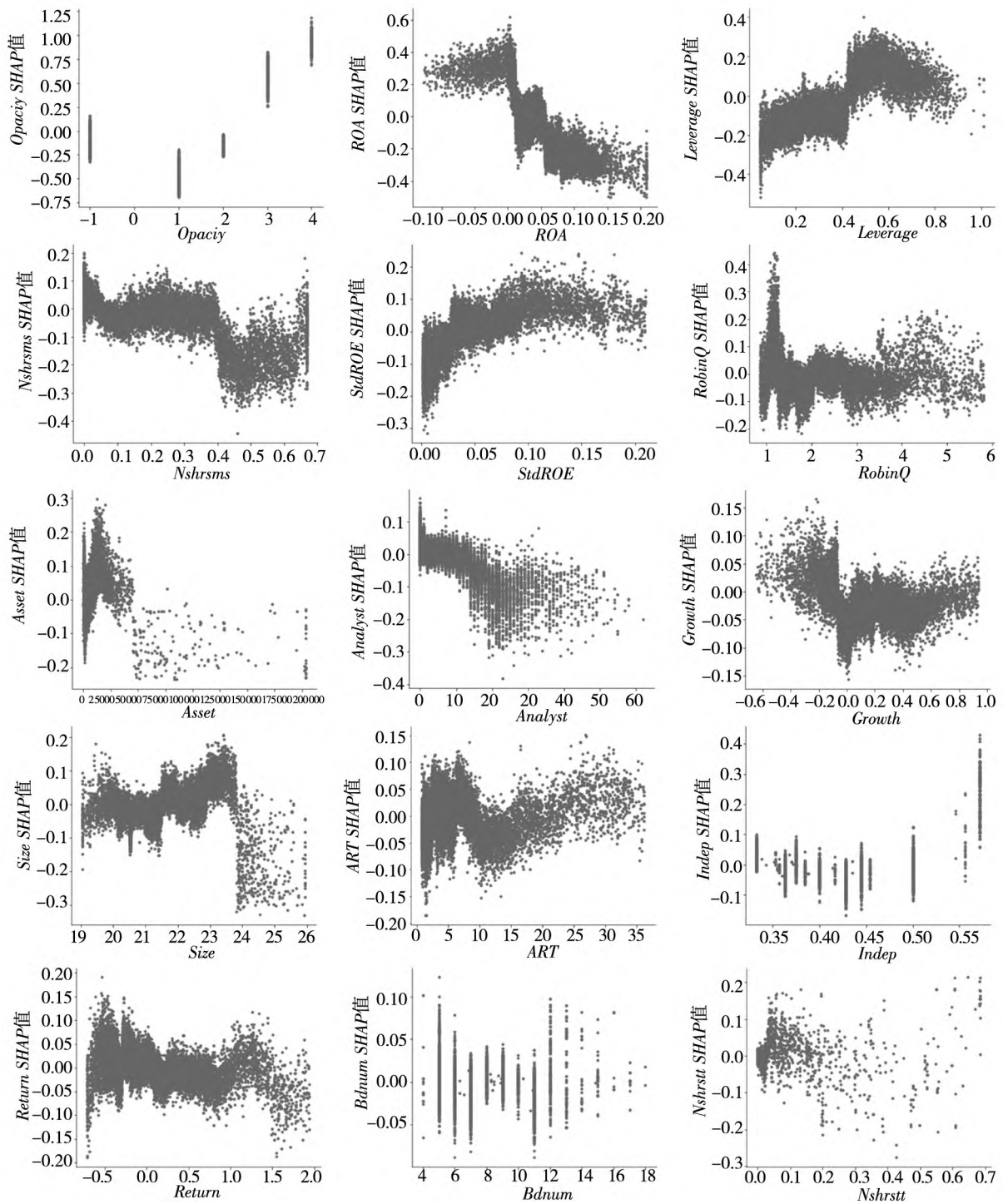


图 4 公司特征 SHAP 依赖图

Fig.4 SHAP partial dependency diagram of company characteristics

1) 特征 *Opacity* 取自深交所每年对其上市公司的信息披露透明度评价,其 SHAP 值与样本值呈正相关, x 轴方向从 1~4 表示公司信息透明度从优到劣,当年披露信息透明度为“A”的公司日

后被发现该年违规的概率较小,且随着信息透明度降低,公司该年被预测违规的倾向增大,公司信息透明度缺失值对于预测违规的影响较小. 特征 *Opacity* 评价的考核内容涵盖上市公司信息披露

的真实性、准确性、完整性、及时性等 9 个维度,该特征与公司违规行为高度相关,对于模型预测具有积极意义; 2) 特征 *ROA* 是衡量公司盈利能力的重要财务指标,其 *SHAP* 值与样本值呈负相关,依赖图中 $ROA < 0$ 的样本点大部分集中在 *SHAP* 值 > 0 的区域,说明当某一公司的 *ROA* 小于 0 时,模型更倾向于判断其发生违规行为,随着 *ROA* 逐渐增大,这种倾向性减弱。*ROA* 特征对于预测公司违规行为贡献负影响,即公司盈利能力不足时更易被判定发生违规行为; 3) 特征 *Leverage* 是衡量公司偿债能力的重要财务指标,其 *SHAP* 值与样本值呈正相关,对于预测公司违规行为贡献正影响,随着公司的 *Leverage* 增大,模型越倾向于预测其发生违规。负债融资作为公司重要的融资手段,可以在一定程度上降低公司代理成本,但随着负债增多,公司的资产负债率不断提高,其定期还本付息的偿债压力也随之增大,这可能促使公司产生违规倾向并发生违规行为,因而高负债会增大模型预测违规的倾向; 4) 特征 *Nshrms* 的 *SHAP* 值与样本值呈负相关,随着高管团队持股比例提高,对于预测公司违规行为贡献负影响,但 *SHAP* 值 > 0 区域内仍存在高比例持股的样本点,高管团队持股比例的提升有利于在一定程度上缓解委托代理问题,抑制公司违规倾向,但不能完全避免公司发生违规; 5) 特征 *StdROE* 的 *SHAP* 值与样本值呈正相关,随着公司业绩波动性的增强,模型倾向于对违规产生更高概率的预测,已有研究表明业绩波动与公司风险相关,业绩波动性越大,公司违规的可能性越高; 6) 特征 *Analyst* 的 *SHAP* 值与样本值呈负相关,其对预测公司违规产生较为明显的负向影响,根据样本分布, $x \in [1, 50]$ 内的样本点密度较大,约占总样本的 72%,超 70% 的公司会被至少一名分析师关注,证券分析师作为监督上市公司的外部力量,有效抑制公司违规倾向,随着某公司分析师跟踪程度提高,其被预测为违规的概率下降。

除上述部分特征与公司违规行为间可以观察到相对明显的单调性外,其他特征的 *SHAP* 值与样本值未呈现明显的线性关系。财务特征方面,鉴于 *SHAP* 图样本点分布可知,当企业的托宾 *Q* 值

偏小、营业收入增长率偏低、应收账款周转率偏低时,意味着企业可能遭遇运营情况欠佳、利润下降或行业竞争激烈等问题^[31],企业迫于压力易发生违规行为。特征 *Asset* 和特征 *Size* 分别从年末资产、年初资产对数的角度反映上市公司的规模状况,由分析结果可知,在一定范围内,资产规模越大,公司被预测为违规的概率也越高,当资产规模持续增加至偏大时,预测违规的概率较低,这可能是因为具备一定规模的大公司具有相对成熟规范的风险管控能力。治理特征方面,特征 *Bdnum* 和 *Indep* 的 *SHAP* 依赖图表现出一定相似性,二者的 *SHAP* 值与样本值未呈现明显的线性关系,但当董事会规模、独董占比过大或过小时,尤其当董事会规模超过 12 人,或独董占比超过 1/2 时,模型预测公司违规的倾向增强,这可能是因为过大的人数规模将导致权责不明、监督职能弱化等问题,降低董事会工作效率,而独董占比的提升有可能增加内部人合谋参与违规的风险^[2, 28]。

根据上述的预测模式分析,可以更直观地考察特征如何对预测结果产生影响,公司的财务特征、治理特征与公司违规行为间普遍存在非线性关系,仅上市公司信息透明度、*ROA*、资产负债率、高管团队持股比例、业绩波动性、分析师关注度指标对公司违规表现出相对单一的影响,公司特征的输出结果与已有研究的部分结论相印证。

3.3.2 高管个人特征与公司违规

高管特征对于模型预测违规行为产生的影响相对较小,其中,高管特征中重要性排序靠前的特征有 *SharEnd*、*IsDuality*、*Age*。

1) 特征 *SharEnd* 的 *SHAP* 值与样本值未呈现明显的线性关系,但随着 CEO 年末持股数占比的提高,部分样本点的 *SHAP* 值弥漫至大于 0 的区域,说明 CEO 年末持股数增加会提高模型预测违规的概率,造成这样分布的可能原因有两个方面,一方面,CEO 持股可以协调其与股东间的利益冲突,公司违规往往会对公司的股价造成负面影响,有损持股人利益,因而随着 CEO 持股增加,其更倾向于与股东利益保持一致,尽可能规避公司决策或将导致的违规行为; 但另一方面,CEO 持股比例反映了 CEO 所有权权利,CEO 持股比例

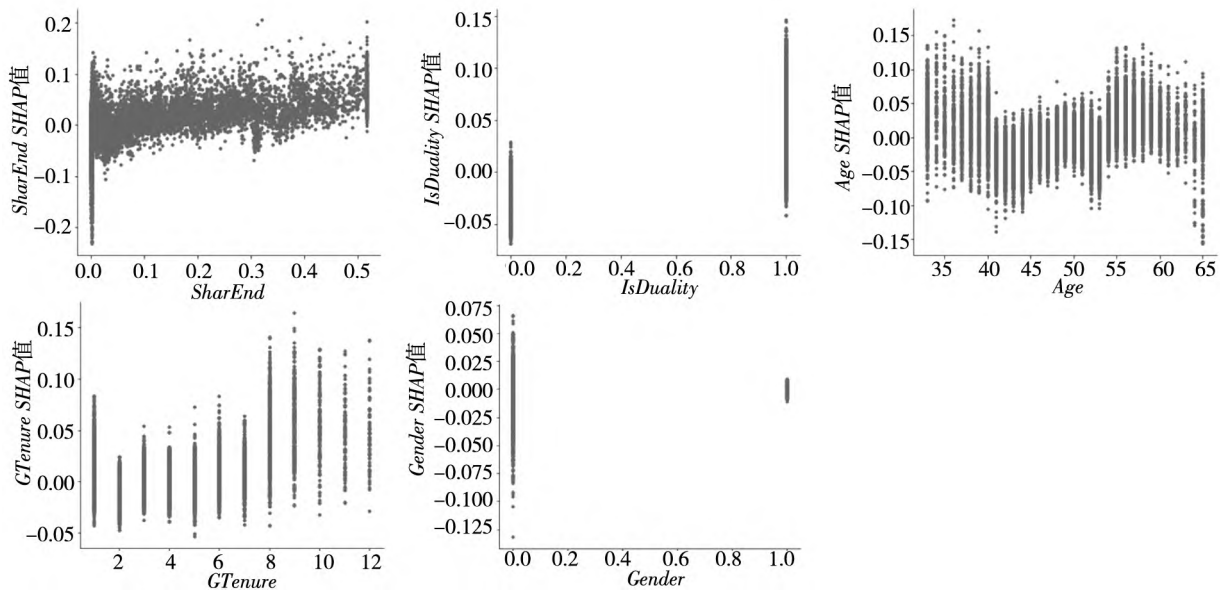


图5 高管特征 SHAP 依赖图

Fig. 5 SHAP partial dependency diagram of managerial individual characteristics

越高,其权力也就越大,对董事会的影响力也就越强,这可能导致“裙带关系”作用下,公司违规被发现的概率降低、代价变小,进而增大公司违规倾向^[25]。除此之外,研究表明股权激励在 CEO 薪酬中的比例与公司财务虚假陈呈正相关关系^[31]。在两种因素的作用下,CEO 持股比例与公司违规行为未表现出明显的线性关系; 2) 特征 Age 的 SHAP 值与样本值未呈现明显的线性关系,但当 CEO 年龄偏小或者偏大时,模型倾向于预测违规。根据前述样本分布,35 岁以下高管占比最少,约为 3% 左右,在这个年龄区间内,样本点弥散在 SHAP 值 > 0 的区域更广,模型预测违规概率更大,当年龄沿 x 轴方向不断增大时,样本点的 SHAP 值减小直至到达 0 附近后呈较均匀上下分布,当 CEO 年龄处于 45 岁 ~ 50 岁的区间时,SHAP 值上下波动的区域相对较窄,近乎 30% 的高管年龄处在这个范围内,此时年龄对于模型预测的影响力较小,当年龄继续增大时,SHAP 值分布再次变宽,模型预测违规概率再次增强。这样的分布可能更多源于年龄对于高管的内在影响与外在塑造,年轻的 CEO 容易产生过度自信心理且往往经验不足,易造成公司违规,随着 CEO 年龄增大,其将具备更丰富的管理经验和更强大的决策前信息收集能力^[51],并对风险表现出一定的厌恶

情绪,倾向于在进行财务决策时维持现有声誉并依赖过去经验,避免采取冒险决策和创新战略,这些随年纪变化形成的管理风格有利于减少失误的决策,因而处于该年纪段的高管所属公司表现出相对小的违规倾向,但年龄的持续增大也意味着 CEO 的地位与权力获得进一步体现,部分 CEO 会利用职权谋取私利,加剧代理问题,致使公司违规概率提高; 3) 我国董事长与 CEO 两职兼任情况比较普遍,根据描述性统计,超过三分之一的 CEO 同时担任公司董事长,当 IsDuality = 1 时表示公司存在两职合一现象,此时样本点的 SHAP 值相对较高,模型预测违规的倾向相对较大。根据已有研究,两职合一的管理层相对二职分离时表现出更强的盈余管理倾向,致使公司会计舞弊的可能性增大,除此之外,两职合一赋予 CEO 更大的自由量裁权和更多的寻租机会,使其可通过干预董事会作用的发挥进一步影响公司行为。

其他高管个人特征对于违规贡献的重要性占比总体约为 4.24%,其中,高管任期特征 GTenure 对于解释违规贡献的重要性相对较大,重要性占比约为 1%。特征 GTenure 的 SHAP 值与样本值未呈现明显的线性关系,但当任期偏长时,模型预测违规的概率相对较高,这可能是由于任期增加意味着管理者个人权力的增长,管理层权力越大,公

司违规的可能性也越大;特征 *Gender* 对于解释违规贡献的重要性相对小,重要性占比低于 0.5%,主要原因是我国上市公司中男性高管占据绝大多数,超过 90%,这使得性别特征对于违规表现产生的影响较不明显,难以对预测倾向产生足够贡献.特征 *RDBack*、*MktBack*、*OveseaBack*、*FinBack*、*ProBack*、*Academic*、*LawBack* 对于解释公司违规行为贡献的重要性占比共计约为 2.43%,高管的职业经历、海外背景、学术背景、金融背景对于模型预测的影响力相对较小.特征 *IsCocurP* 对于预测违规贡献的重要性占比约为 0.67%.

3.3.3 特征交互与公司违规

本研究选取对模型影响显著的特征 *Opacity*、*ROA*、*Leverage*,绘制含交互项的 *SHAP* 特征依赖图,以进一步观测重要特征存在特征交互时的预测模式.

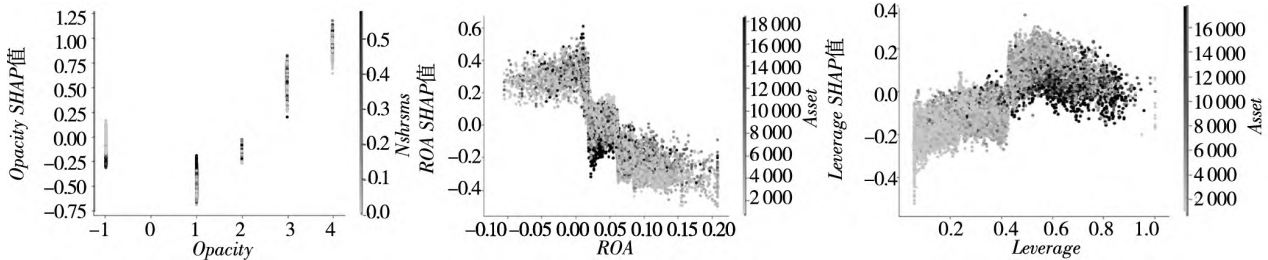


图 6 含交互项 *SHAP* 依赖图

Fig. 6 *SHAP* partial dependency diagram including interactive items

3.3.4 公司个体违规预测的举例分析

通过 *SHAP* 依赖图可以观察样本点的分布,并分析特征的预测模式与现有理论的契合情况.除此之外,根据 *SHAP* 方法还可以考察特征如何判断某公司个体的违规倾向.模型在判断某公司违规倾向的过程中会受到多种因素的共同作用,如图 7 和图

依赖图中的色散表示其他特征的真实值对应的颜色区间.通过颜色分布可以观测到其他特征与该特征的相互作用,如图 6 所示.1) 特征 *Opacity* = -1 时表示上市公司信息透明度指标缺失,此时深灰色样本点弥散在图形下方,由此可见在信息透明度不详时,高管团队持股比例越大,模型预测违规的倾向越低;2) 在特征 *ROA* ∈ [0.0, 0.05] 的区间内,部分样本点弥散至 *SHAP* 值 < 0 的区域,此时对于 *ROA* 相似的公司,公司资产越大则模型预测违规的倾向越低;3) 随着特征 *Leverage* 增大,样本点颜色逐渐由深变浅,但深灰色样本点较多分布在趋势线的下方,当 *Leverage* 超过一定范围时,对于资产负债率相似的公司,规模越大模型预测违规的倾向越低.通过 2)、3) 可知,在一定范围内,当模型识别资产负债率或总资产净利润率相似的公司时,资产规模偏大会降低模型预测违规的倾向.

8 所示.箭头的方向和长度分别表示该特征对违约倾向的作用方向与大小.其中,深灰色部分表示模型倾向预测为违规的影响因素,浅灰色部分表示模型倾向预测为没有违规的影响因素.此处分别随机选取一家预测为没有违规的公司和预测为违规的公司进行个体分析,以观察特征如何影响模型的预测倾向.

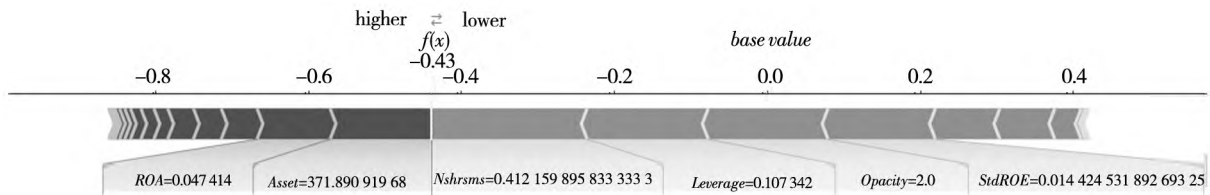


图 7 预测为未违规公司的 *SHAP* 解释示例

Fig. 7 A case study of predicting no violations based on *SHAP*

图 7 为一家 2019 年被预测为没有发生违规公司(证券代码: 300553)的 *SHAP* 特征贡献图,其被预测为没有违规的原因有高管团队持股比例大

于观测公司的平均及中间水平、公司的资产负债率仅为 10% 左右、当年披露信息透明度为“B”、业绩波动性低于观测公司的平均及中间水平等.

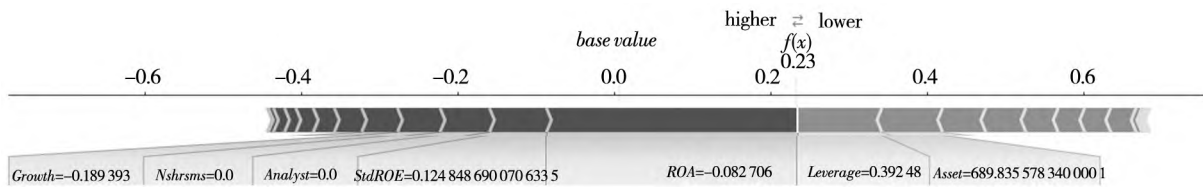


图 8 预测为违规公司的 SHAP 解释示例

Fig. 8 A case study of predicting violations based on SHAP

图 8 为一家 2019 年被预测为发生违规行为公司(证券代码: 603991) 的 SHAP 特征贡献图, 其被预测为违规的原因有业绩波动性大于观测公司的平均水平、公司被跟踪分析的分析师(团队) 数量为空、公司成长能力较差等。

综上所述, 基于 LightGBM 算法构建的公司特征模型、高管特征模型能够预测公司违规行为。根据测试集的 AUC 值, CEO 特征模型具备相对较高的分类性能, 模型在查全和查准方面存在较大差异, 其查全率相对较高, 更适合为利益相关者从大范围的公司中发现违规公司提供经验数据, 但在精准预警方面表现欠佳。在 LightGBM 输出的整体性预测模型基础上, 借助 SHAP 算法进一步分析得出各类特征对预测公司违规的贡献度排序, 公司特征比高管个人特征对预测产生的影响更大。通过公司的财务指标和治理指标更能有效预测违规倾向, 其中有关公司财务的评价维度相对全面, 反映出的连续性、差异化信息更多。具体而言, 公司特征方面, 上市公司信息透明度、总资产净利率、资产负债率三个特征对于预测违规行为的影响更大, 公司的盈利能力、偿债能力对于预测有较大的贡献; 高管个人特征方面, 高管个人的年末持股比例、董事长与 CEO 是否两职合一、高管年龄三个特征对于预测违规行为的影响更大。更进一步, 本研究通过 SHAP 依赖图对特征与预测倾向的关系进行了更深层次的探讨, 大多数公司特征均与其 SHAP 值呈现复杂的非线性关系。其中, 上市公司信息透明度、总资产净利率、资产负债率、高管团队持股比例、业绩波动性、分析师关注度对于预测公司违规倾向呈相对单一的线性关系, 预测模式的分析印证了公司治理领域关于违规行为研究多结论并存的现状。

4 稳健性检验

4.1 更换高管特征

本研究为检验上述结论的稳定性, 将 CEO 个人特征替换为董事长个人特征进行研究。董事长特征样本数量为 20 520 条, 其与 CEO 特征数据集的差异主要是因为 CEO、董事长换届时间不统一。如表 9 所示, 在上市公司观测值中, 男性董事长占比更高, 约有 94%。董事长的平均年龄与任期均略高于 CEO, 分别约为 52 岁和 3.6 年, 董事长的年末持股占比的均值约为 11%, 董事长在股东单位兼任比例也相对较高, 约为 45%。后天经历方面, 董事长与 CEO 相同均普遍具备丰富的管理职能经验, 约有 93% 的董事长曾有市场营销与公共关系的职业背景; 其次, 28% 左右的董事长有过高校任教、科研机构任职、协会从事研究至少某一方面的经历, 高于 CEO 的学术比例, 约 20% 的董事长拥有研发背景; 再次, 约 8% 的董事长有生

表 9 董事长特征的描述性统计分析

Table 9 Descriptive statistics of chairman's characteristics

指标	样本量	平均值	标准差	中位数
Gender	20 520	0.94	0.23	1.00
Age	20 520	52.17	7.73	52.00
GTenure	20 520	3.60	2.68	3.00
SharEnd	20 520	0.11	0.15	0.01
IsDuality	20 520	0.36	0.48	0.00
IsCocurP	20 520	0.45	0.50	0.00
ProBack	20 520	0.08	0.28	0.00
RDBack	20 520	0.20	0.40	0.00
MktBack	20 520	0.93	0.26	1.00
LawBack	20 520	0.01	0.09	0.00
OveseaBack	20 520	0.08	0.27	0.00
Academic	20 520	0.28	0.45	0.00
FinBack	20 520	0.10	0.30	0.00

产运作方面的经历,或海外任职、求学的经历,10%的董事长曾从事过金融行业;最后,仅有不到1%的董事长有法律背景.总体而言,董事长相较CEO具有更丰富的学术、金融背景,CEO则拥有更充分的生产运作、研发背景.

通过 LightGBM 得到董事长特征模型对于样本的训练结果,表 10 展示了模型对于违规的预测效果.

预测结果显示将 CEO 特征更替为董事长特征后本研究结论依旧稳健.在验证集和测试集上,董事长特征模型的 AUC 值均达到 80% 以上,公司特征模型的 AUC 值约为 80% 左右,模型均具备一定分类能力.加入高管特征后模型的预测性能得到轻微改善,董事长个人特征对于预测违规发挥的作用有限.两组模型的召回率均高于精准率,且

加入董事长特征后虽提升了模型查准性,但同时损失一定的查全性,模型泛化能力未得到全面的改善.

表 10 基于董事长特征模型(LightGBM)输出结果

Table 10 Main results of model(LightGBM) based on chairman's characteristics

数据集	指标	公司特征模型 (1)	董事长特征模型(2)	提升效果 (2) - (1) / %
训练集	<i>auc_train</i>	0.939 1	0.978 9	3.98
验证集	<i>auc_valid</i>	0.790 1	0.800 5	1.04
测试集	<i>auc_test</i>	0.805 3	0.810 5	0.53
	<i>Accuracy</i>	0.655 3	0.697 5	4.22
	<i>Precision</i>	0.350 2	0.382 8	3.25
	<i>Recall</i>	0.803 0	0.784 2	-1.88
	<i>Kappa</i>	0.284 0	0.330 6	4.66

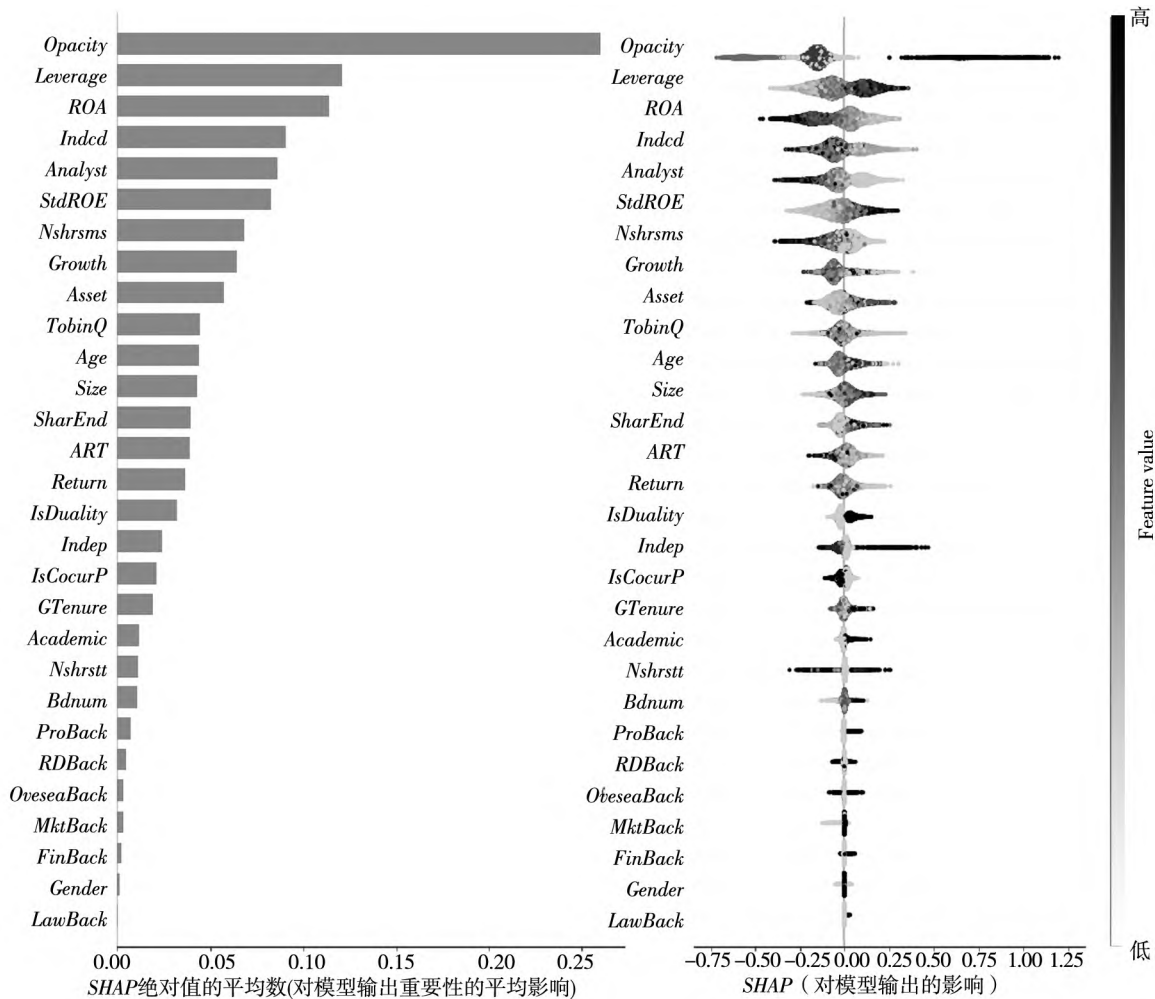


图 9 基于董事长特征模型(LightGBM)输出的重要性排序及摘要图

Fig.9 Importance map and summary of model(LightGBM) output based on chairman's characteristics

如图 9 所示,重要性方面依旧是上市公司信息透明度、总资产净利率、资产负债率三个公司层面的特征重要性排序靠前,对预测倾向的影响更大. 高管个人特征中,年龄、年末持股比例、董事长与 CEO 是否两职合一对于预测违规贡献力度较大的特征,高管任期次之,个人经历与性别对于模型预测发挥的作用较小,与 CEO 特征模型的结论相一致. 综上所述,更换 CEO 特征为董事长特征后结果依然稳健.

4.2 更换 GBDT 实现方法

XGBoost 是对于 GBDT 算法框架的又一改进实现,被广泛应用于机器学习比赛. LightGBM 与 XGBoost 的输出模型拥有相当的精度,但由于部分算法的改进及决策树生长策略的择优,LightGBM 在训练速度和内存占用方面的表现优于 XGBoost,因而本研究选用 LightGBM 为主要研究方法,选择 XGBoost 作为稳健性检验方法. 通过 XGBoost 可以得到公司特征模型、CEO 特征模型对

于数据样本的训练结果.

表 11 模型(XGBoost) 输出结果

Table 11 Main results of model (XGBoost)

数据集	指标	公司特征模型(1)	CEO 特征模型(2)	提升效果 (2) - (1) /%
训练集	<i>auc_train</i>	0.999 1	0.999 5	0.04
验证集	<i>auc_valid</i>	0.787 9	0.789 7	0.17
测试集	<i>auc_test</i>	0.788 1	0.784 7	-0.34
	<i>Accuracy</i>	0.673 7	0.685 1	1.15
	<i>Precision</i>	0.360 8	0.365 9	0.51
	<i>Recall</i>	0.754 5	0.718 6	-3.58
	<i>Kappa</i>	0.290 0	0.291 2	0.11

由表 11 结果可知,通过 XGBoost 训练得到的输出结果与 LightGBM 大体相似,公司特征模型、CEO 特征模型均具备一定的预测能力;在公司特征模型基础上加入高管个人特征后对于模型预测效果的改善作用不明显;模型的召回率高于精准率,阈值取 0.5 的情况下模型更擅长从上市公司中预测尽可能多的违规公司,但在精准预测上表现欠佳.

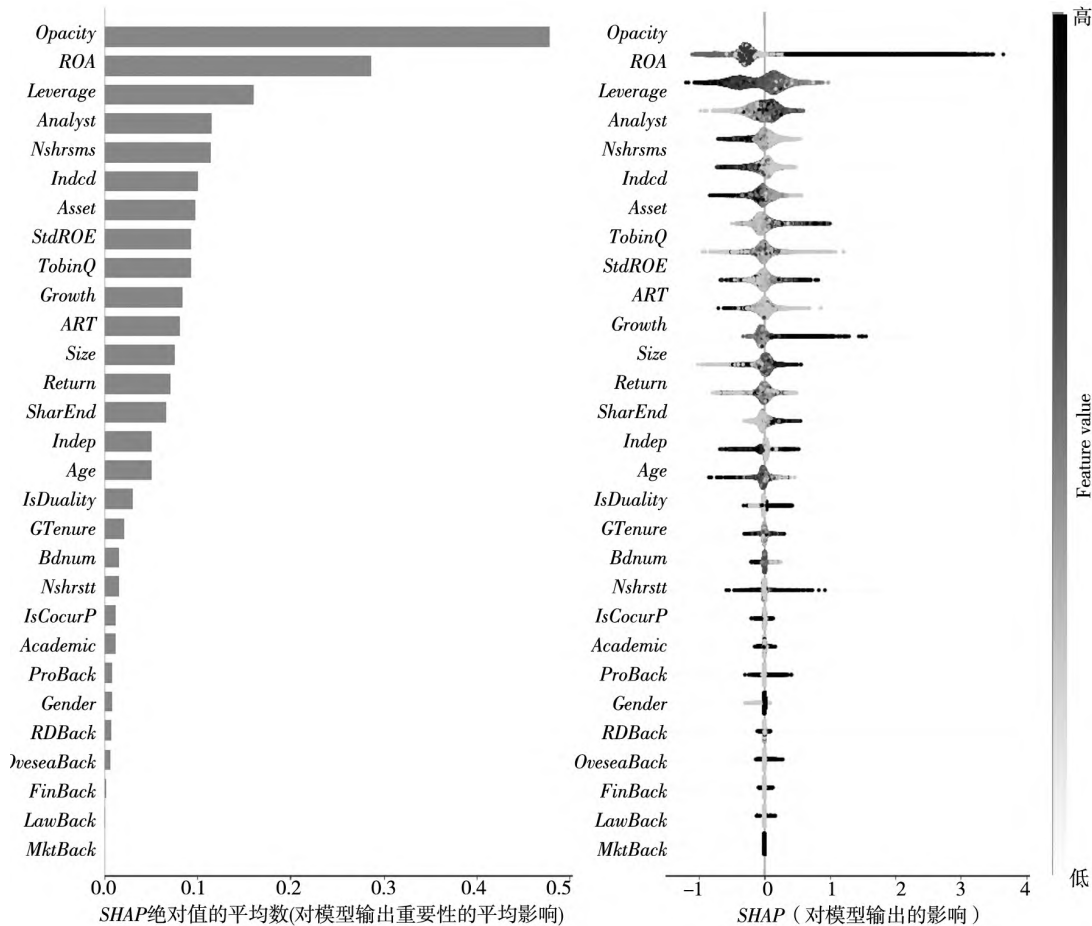


图 10 模型(XGBoost) 输出的重要性排序及摘要图

Fig. 10 Importance map and summary of model(XGBoost) output

如图 10 所示,特征重要性及样本点分布方面,公司特征对于预测的贡献更高,上市公司信息透明度、总资产净利润率、资产负债率三个特征排序靠前,影响更大,高管个人特征中,年末持股比例、董事长与 CEO 是否两职合一、年龄是对于预测违规贡献力度较大的特征,高管任期次之,个人经历与性别对于模型预测发挥的作用较小。综上所述,更换机器学习算法实现方式后结果依然稳健。

4.3 更换核心参数

模型结果对于参数的敏感性强弱在一定程度上反映出模型的稳健性大小,本研究通过更换模

型的最优核心参数和学习控制参数,观察预测结果受到的影响,进而判断模型的稳健性。本研究选取的算法中 max_depth 是决定模型性能和泛化能力的重要参数,通常其参考值设定在 3~10 之间, $learning_rate$ 是决定目标函数能否收敛到局部最小值及何时收敛到最小值的重要参数,通常设置在 0.01~0.1 之间。在稳健性检验部分,本研究将 max_depth 设为{3,10}, $learning_rate$ 设为{0.01, 0.05, 0.1} 2 个主要参数两两组合后共形成 6 种方案,将其分别代入模型以检验参数改变对于预测结果稳定性的影响,为直观说明模型训练结果,此处仅展示测试集的 AUC 、精准率、召回率。

表 12 调参后模型(LightGBM) 输出结果

Table 12 Main results of model(LightGBM) based on alternative parameters

$learning_rate$	max_depth	指标	公司特征模型 (1)	CEO 特征模型 (2)	提升效果 (2)-(1) %
Panel A: 本研究选用核心参数及输出结果					
0.05	10	auc_test	0.8027	0.8142	1.15
		$Precision$	0.3591	0.3781	1.90
		$Recall$	0.7993	0.7921	-0.72
Panel B: 更换核心参数后的输出结果					
0.05	3	auc_test	0.8010	0.8047	0.37
		$Precision$	0.3703	0.3970	2.67
		$Recall$	0.7599	0.7527	-0.72
0.1	10	auc_test	0.8037	0.8012	-0.25
		$Precision$	0.3611	0.3710	0.99
		$Recall$	0.7849	0.7832	-0.18
0.1	3	auc_test	0.7997	0.8021	0.24
		$Precision$	0.3798	0.3996	1.98
		$Recall$	0.7814	0.7455	-3.58
0.01	10	auc_test	0.8092	0.8067	-0.24
		$Precision$	0.3549	0.3780	2.30
		$Recall$	0.8047	0.7688	-3.58
0.01	3	auc_test	0.8021	0.7988	-0.33
		$Precision$	0.3879	0.3820	-0.58
		$Recall$	0.7437	0.7545	1.08

由表 12 可知,调整重要参数后,模型预测结果指标的变化幅度较小,两组模型的 AUC 基本保持在 80% 左右,且召回率均原远高于精确率,模型都实现较好预测但在查准方面表现欠佳;加入高管特征后仅有部分模型的性能轻微改善,部分模型出现 AUC 值下降情况,高管特征对于预测上市公司是否违规发挥的作用有限,前述结论具备

一定稳健性。

4.4 更换时间窗口期

前期构建模型过程中主要采用 1 年窗口期进行数据训练,为避免选择窗口期不同造成的结果差异,该部分根据 3 年时间窗口的连续型变量均值构建预测模型。研究发现前述结果具有一定稳健性。

表 13 更换时间窗口期后模型(LightGBM) 输出结果

Table 13 Main results of model(LightGBM) based on alternative rolling windows

数据集	指标	公司特征模型(1)	CEO 特征模型(2)	提升效果 (2) - (1) / %
训练集	<i>auc_train</i>	0.941 4	0.981 1	3.98
验证集	<i>auc_valid</i>	0.800 4	0.809 9	0.94
测试集	<i>auc_test</i>	0.798 5	0.793 6	-0.49
	<i>Accuracy</i>	0.727 6	0.712 9	-1.48
	<i>Precision</i>	0.406 5	0.392 3	-1.41
	<i>Recall</i>	0.697 1	0.715 1	1.79
	<i>Kappa</i>	0.342 2	0.327 6	-1.45

由表 13 可知,更换时间窗口对于模型预测公司违规的影响不大。调整窗口期后的两组模型在验证集和测试集上的 AUC 值均在 80% 左右,准确率均在 70% 以上。高管特征对模型性能的提升作用不显著,召回率有所增大的同时损失一定精准率,致使模型整体的正确率下降。模型泛化能力未得到显著改善。前述结论具备一定稳健性。

5 进一步研究

为了进一步探讨机器学习模型相对于传统预测模型的优势,本研究引入逻辑回归(Logistic)、支持向量机(SVM)、随机森林(Random Forest)、梯度提升决策树(GBDT)四种算法,基于 CEO 特征模型,考察不同机器学习方法的拟合效果与预测能力。

表 14 不同模型对比研究

Table 14 Comparative study of different models

指标		<i>auc_test</i>	<i>Precision</i>	<i>Recall</i>
模型名称	Logistic	0.747 8	0.328 1	0.754 5
	SVM	0.758 4	0.370 2	0.720 4
	Random Forest	0.79	0.381 1	0.743 7
	GBDT	0.788 3	0.377 6	0.729 4
	LightGBM	0.814 2	0.378 1	0.792 1

通过对比五个模型可知,LightGBM 在测试集的 AUC 值相对较高,比以传统算法为代表的 Logistic、SVM 高 6% 左右,比以集成学习算法为代表的 Random Forest、GBDT 高 2% 左右;精准率方面,LightGBM 与 SVM、Random Forest、GBDT 的差距不大,四个模型均有 37% 左右的查准率;召回率方面,LightGBM 高于其他算法模型。总体而言,四

种算法均可以实现对于违规行为的有效预测,并且五个模型在 0.5 的阈值下都更善于发现违规而非精准预警。其中,与 LightGBM 相似的 Random Forest、GBDT 集成学习类算法对于违规预测的表现优于传统算法。与其他集成学习类算法相比,LightGBM 除了能够在不损害精度的条件下加快训练速度,节省运算时间,还对改善模型预测性能起到一定积极作用。

6 结束语

本研究基于高阶梯队理论、烙印理论和公司治理理论,从内部治理视角选取重要的公司特征和高管个人特征构建上市公司违规行为预测模型,以 2008 年—2019 年中国 A 股上市公司为样本,引入 LightGBM 算法和 SHAP 工具,研究两类特征对违规行为的预测能力,探讨不同类型特征的重要性排序及预测模式,并举例分析模型如何基于影响因素进行预测。研究结果表明:首先,通过公司特征和高管个人特征构建的机器学习模型可以在一定程度上预测上市公司违规行为,在公司特征模型基础上加入 CEO 特征,模型在训练集上的 AUC 值有轻微改善,提升约 1%,正确率提升约 2%,但模型整体的泛化能力未得到显著优化,精准率与召回率未同时提升,高管个人特征对于预测公司违规行为发挥的作用有限;其次,在整体性预测的基础上,公司特征比高管个人特征对预测违规行为产生的影响更大。公司层面,上市公司信息透明度越高、总资产净利率越大、资产负债率越低、高管团队持股比例越高、业绩波动性越小、分析师关注度越高,模型预测公司违规的倾向越低;高管个人特征层面,年龄偏小、存在董事长与 CEO 两职合一情况时,模型预测公司违规的倾向越高。由于 CEO 经历的同质性及男性 CEO 占比较高,CEO 的个人经历与性别特征仅对部分样本造成影响,对于模型解释违规行为的贡献度较低;最后,大部分公司特征和高管个人特征均与公司违规呈现非线性关系,与传统理论和实证研究结论相一致。SHAP 依赖图显示仅有部分特征与违规行为间存在相对明显的单调性,大多数样本点分布形成的趋势较为复杂,存在区间波动,这说明

违规行为的形成由诸多因素共同作用导致,大部分影响因素不会对违规产生固定的影响力及绝对的影响方向。除此之外,通过研究公司个体预测结果,可以更加直观地探讨机器学习模型如何根据特征表现预测违规,有利于进一步理解和检验现有影响因素对微观个体的作用过程。本研究应用机器学习,将高管个人特征作为一类变量参与违规预测过程,聚焦高管特征与公司违规行为之间的关联,在分析特征预测模式的同时验证现有可解释模型形成的理论,为基于高阶梯队理论延伸和拓展我国高管特征研究提供经验借鉴,对于进一步规范和理解我国公司高管的角色定位具有一定积极意义。

本研究在拓展和丰富高阶梯队理论及公司治理理论相关研究成果的同时,为监管部门通过精细化管理提升监管效率、投资者通过前瞻预测信息提升投资效率以及企业通过完善内部治理机制减少违规行为提供了经验证据,实践参考价值十分明显。

首先,对外部监管部门而言:1) 有助于外部监管部门运用新兴机器学习方法得到的预测性结论提升监督稽查效率。对于企业违规行为的事前监督和防范是监管部门的重要职能,基于机器学习算法的企业违规预测模型具有良好的预测效果、较强的处理速度和多元的分析结果,便于监管部门及时分析和关注企业行为,采取预防措施提升管理效率;2) 为外部监管部门尽早识别企业异常行为提供新思路。传统的企业违规识别模型重点关注财务指标,本研究将研究要素拓展至公司治理特征及高管特征,对于企业内部要素的考察更加全面和细致,从而提升监管部门监督的精细度;3) 为外部监管部门完善法律法规改善信息环境提供证据。数智时代信息的载体愈加多元,信息利用效率不断提升,数据赋能的作用愈发重要。监管部门应不断健全信息披露相关法律法规,凝聚

多方力量改善信息环境,提升企业信息透明度,一方面抑制企业违规行为,另一方面为利益相关者了解企业信息、制定理性决策提供良好的环境和决策信息参考。

其次,对企业投资者而言:1) 投资者可以加强对上市公司治理水平、高管特征等非财务信息的关注,动态调整投资决策。本研究的结论对于拓宽投资考察因素有参考意义,投资者不仅可以通过关注企业财务绩效变化预测异常行为,还能通过了解公司治理机制、高管聘任等信息制定投资决策,在一定程度上缓解投资风险;2) 投资者在具备金融学、管理学等知识的基础上,可以进一步关注新兴技术在金融领域的应用,运用机器学习模型辅助投资决策。与传统计量模型相比,本研究构建的模型在预测准确度、速度与效率等方面都具有相对优势,具备一定的实用前景,为投资者运用机器学习算法跟踪企业动态、预测企业行为、辅助投资决策拓展思路。

最后,对企业管理者而言:1) 企业应不断完善公司治理机制,对自身行为进行有效地事前控制与规范。本研究发现预测违规行为不局限于企业的财务绩效;从非财务信息视角来看,公司治理层面要素对企业违规行为具有较强的预测能力。因此企业应强化内部控制机制、提升公司治理水平,通过完善内部监督制度规范企业违规行为;2) 企业应重视高质量信息披露,提升信息透明度与可靠性,加强投资者关系管理,从而保护利益相关者的权益。本研究发现信息透明度、分析师关注度等要素与企业违规行为呈现较为明显的负向关系,企业提升信息透明度一方面能够加强信息使用者的外部监督,进而约束企业行为;另一方面也能够提升信息使用者对企业的信任,促进企业与利益相关者的良性互动,有助于企业的可持续健康发展,从而遏制企业违规动机。

参考文献:

- [1] 马奔,杨耀武. 视而不见? 证券分析师与上市公司欺诈关系研究——基于考虑部分可观测的 Bivariate Probit 估计[J]. 南开经济研究, 2020, (2): 92-113.
Ma Ben, Yang Yaowu. Turn a blind eye? A research on the relationship between securities analyst and corporate fraud: Based on bivariate probit estimation method with consideration on partial observability [J]. Nankai Economic Studies, 2020, (2): 92-113. (in Chinese)
- [2] 蔡志岳, 吴世农. 董事会特征影响上市公司违规行为的实证研究[J]. 南开管理评论, 2007, (6): 62-68+92.

- Cai Zhiyue, Wu Shinong. The empirical study on behavior of committing fraud in listed companies: Evidence from the characteristics of board of directors [J]. *Nankai Business Review*, 2007, (6): 62-68+92. (in Chinese)
- [3] 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, (6): 47-51.
Liang Jie, Wang Xuan, Li Jinzhong. An empirical research on relationship between modern corporate governance and accounting information fraud [J]. *Nankai Business Review*, 2004, (6): 47-51. (in Chinese)
- [4] 陆瑶, 朱玉杰, 胡晓元. 机构投资者持股与上市公司违规行为的实证研究 [J]. *南开管理评论*, 2012, 15(1): 13-23.
Lu Yao, Zhu Yujie, Hu Xiaoyuan. Institutional shareholding and corporate fraud: Evidence from China [J]. *Nankai Business Review*, 2012, 15(01): 13-23. (in Chinese)
- [5] 陆瑶, 胡江燕. CEO与董事间“老乡”关系对公司违规行为的影响研究 [J]. *南开管理评论*, 2016, 19(2): 52-62.
Lu Yao, Hu Jiangyan. The impact of hometown connectedness between the CEO and board of directors on corporate fraud [J]. *Nankai Business Review*, 2016, 19(2): 52-62. (in Chinese)
- [6] Persons O S. Using financial statement data to identify factors associated with fraudulent financial reporting [J]. *Journal of Applied Business Research (JABR)*, 1995, 11(3): 38-46.
- [7] Kreutzfeldt R W, Wallace W A. Error characteristics in audit populations—their profile and relationship to environmental factors [J]. *Auditing—A Journal of Practice & Theory*, 1986, 6(1): 20-43.
- [8] Beasley M S, Carcello J V, Hermanson D R, et al. *Fraudulent Financial Reporting: 1987-1997: An Analysis of US Public Companies* [R]. New York: Committee of Sponsoring Organizations of the Treadway Commission, 1999.
- [9] Hambrick D C, Mason P A. Upper echelons: The organization as a reflection of its top managers [J]. *Academy of Management Review*, 1984, 9(2): 193-206.
- [10] Liao J, Smith D, Liu X. Female CFOs and accounting fraud: Evidence from China [J]. *Pacific-Basin Finance Journal*, 2019, 53: 449-463.
- [11] Troy C, Smith K G, Domino M A. CEO demographics and accounting fraud: Who is more likely to rationalize illegal acts? [J]. *Strategic Organization*, 2011, 9(4): 259-282.
- [12] Van Scotter J R, Roglio K D D. CEO bright and dark personality: Effects on ethical misconduct [J]. *Journal of Business Ethics*, 2020, 164(3): 451-475.
- [13] 何瑛, 于文蕾, 戴逸驰, 等. 高管职业经历与企业创新 [J]. *管理世界*, 2019, 35(11): 174-192.
He Ying, Yu Wenlei, Dai Yichi, et al. Career experience of executive and corporate innovation [J]. *Journal of Management World*, 2019, 35(11): 174-192. (in Chinese)
- [14] 杨清香, 俞麟, 陈娜. 董事会特征与财务舞弊——来自中国上市公司的经验证据 [J]. *会计研究*, 2009, (7): 64-70+96.
Yang Qingxiang, Yu Lin, Chen Na. Board characters and financial fraud: Empirical evidence from Chinese listed companies [J]. *Accounting Research*, 2009, (7): 64-70+96. (in Chinese)
- [15] 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.
Xiong Fangjun, Zhang Longping. Risk identification and evidence collection of financial fraud on the listed companies [J]. *Research on Economics and Management*, 2016, 37(10): 138-144. (in Chinese)
- [16] 叶康涛, 刘金洋. 非财务信息与企业财务舞弊行为识别 [J]. *会计研究*, 2021, (9): 35-47.
Ye Kangtao, Liu Jinyang. Non-financial information and detection of financial frauds [J]. *Accounting Research*, 2021, (9): 35-47. (in Chinese)
- [17] 张敏, 刘云菁, 郭金同. 财务与会计领域的大数据研究: 技术与应用 [J]. *会计与经济研究*, 2021, 35(3): 3-22.
Zhang Min, Liu Yunjing, Guo Jintong. Big data research in the field of finance and accounting: Technology and application [J]. *Accounting and Economics Research*, 2021, 35(3): 3-22. (in Chinese)
- [18] 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): 18-25.
Qian Ping, Luo Mei. Predicting accounting fraud in China [J]. *Accounting Research*, 2015, (7): 18-25. (in Chinese)
- [19] Scott M L, Gabriel G E, Su-In Lee. Consistent Individualized Feature Attribution for Tree Ensembles [DB/OL]. ArXiv, 2018(2), abs/1802.03888, <https://doi.org/10.48550/arXiv.1802.03888>.
- [20] 曹春方, 陈露兰, 张婷婷. “法律的名义”: 司法独立性提升与公司违规 [J]. *金融研究*, 2017, (5): 191-206.
Cao Chunfang, Chen Lulan, Zhang Tingting. The name of the law: Judicial independence promote and corporate violations [J]. *Journal of Financial Research*, 2017, (5): 191-206. (in Chinese)

- [21]郑登津,袁薇,邓祎璐. 党组织嵌入与民营企业财务违规[J]. 管理评论, 2020, 32(8): 228-243+253.
Zheng Dengjin, Yuan Wei, Deng Yilu. Organization embedding and financial violation[J]. Management Review, 2020, 32(8): 228-243+253. (in Chinese)
- [22]徐星美,权小锋. 中华老字号的会计基因: 基于盈余管理的视角[J]. 管理科学学报, 2022, 25(1): 81-94.
Xu Xingmei, Quan Xiaofeng. Accounting gene of Chinese time-honored brand: From the perspective of earnings management[J]. Journal of Management Sciences in China, 2022, 25(1): 81-94. (in Chinese)
- [23]Uzun H, Szewczyk S H, Varma R. Board composition and corporate fraud[J]. Financial Analysts Journal, 2004, 60(3): 33-43.
- [24]孟庆斌,李昕宇,蔡欣园. 公司战略影响公司违规行为吗[J]. 南开管理评论, 2018, 21(3): 116-129+151.
Meng Qingbin, Li Xinyu, Cai Xinyuan. Does corporate strategy influence corporate frauds[J]. Nankai Business Review, 2018, 21(3): 116-129+151. (in Chinese)
- [25]陆瑶,李茶. CEO对董事会的影响力与上市公司违规犯罪[J]. 金融研究, 2016, (1): 176-191.
Lu Yao, Li Cha. CEO Influence over the board and corporate fraud[J]. Journal of Financial Research, 2016, (1): 176-191. (in Chinese)
- [26]刘振杰,顾亮,李维安. 董事会非正式层级与公司违规[J]. 财贸研究, 2019, 30(8): 76-87.
Liu Zhenjie, Gu Liang, Li Weian. Board informal hierarchy and firm violation[J]. Finance and Trade Research, 2019, 30(8): 76-87. (in Chinese)
- [27]Marciuakaitye D, Szewczyk S H, Uzun H, et al. Governance and performance changes after accusations of corporate fraud[J]. Financial Analysts Journal, 2006, 62(3): 32-41.
- [28]邓可斌,周小丹. 独立董事与公司违规: 合谋还是抑制[J]. 山西财经大学学报, 2012, 34(11): 84-94.
Deng Kebin, Zhou Xiaodan. Independent directors and violations: Conspiracy or restrain[J]. Journal of Shanxi University of Finance and Economics, 2012, 34(11): 84-94. (in Chinese)
- [29]Armstrong C S, Jagolinzer A D, Larcker D F. Chief executive officer equity incentives and accounting irregularities[J]. Journal of Accounting Research, 2010, 48(2): 225-271.
- [30]Tutino M, Merlo M. Accounting fraud: A literature review[J]. Risk Governance and Control: Financial Markets and Institutions, 2019, 9(1): 8-25.
- [31]Harris J, Bromiley P. Incentives to cheat: The influence of executive compensation and firm performance on financial misrepresentation[J]. Organization Science, 2007, 18(3): 350-367.
- [32]何瑛,李娇,黄洁. 上市公司股份回购的内在特质、经济后果与管理反应[J]. 经济与管理研究, 2014, (11): 55-62.
He Ying, Li Jiao, Huang Jie. Inherent characteristics, economic effect and managerial response of share repurchases for listed companies[J]. Research on Economics and Management, 2014, (11): 55-62. (in Chinese)
- [33]Feroz E H, Park K, Pastena V S. The financial and market effects of the SEC's accounting and auditing enforcement releases[J]. Journal of accounting research, 1991, 29: 107-142.
- [34]Khanna V, Kim E H, Lu Y. CEO connectedness and corporate fraud[J]. The Journal of Finance, 2015, 70(3): 1203-1252.
- [35]Kaminski K A, Sterling Wetzel T, Guan L. Can financial ratios detect fraudulent financial reporting? [J]. Managerial Auditing Journal, 2004, 19(1): 15-28.
- [36]Dalnial H, Kamaluddin A, Sanusi Z M, et al. Detecting fraudulent financial reporting through financial statement analysis[J]. Journal of Advanced Management Science, 2014, 2(1): 17-22.
- [37]Li J, Tang Y I. CEO hubris and firm risk taking in China: The moderating role of managerial discretion[J]. Academy of Management Journal, 2010, 53(1): 45-68.
- [38]Rijsenbilt A, Commandeur H. Narcissus enters the courtroom: CEO narcissism and fraud[J]. Journal of Business Ethics, 2013, 117(2): 413-429.
- [39]Iqbal Z, Sewon O, Baek H Y. Are female executives more risk-averse than male executives? [J]. Atlantic Economic Journal, 2006, 34(1): 63-74.
- [40]Xiang R, Zhu W. Academic independent directors and corporate fraud: Evidence from China[J]. Asia-Pacific Journal of Accounting & Economics, 2023, 30(2): 285-303.
- [41]洪文洲,王旭霞,冯海旗. 基于 Logistic 回归模型的上市公司财务报告舞弊识别研究[J]. 中国管理科学, 2014, 22

- (S1): 351–356.
- Hong Wenzhou, Wang Xuxia, Feng Haiqi. Research on the financial report fraud detection of listed companies based on logistic regression model[J]. *Chinese Journal of Management Science*, 2014, 22(S1): 351–356. (in Chinese)
- [42] Beneish M D. The detection of earnings manipulation[J]. *Financial Analysts Journal*, 1999, 55(5): 24–36.
- [43] Dechow P M, Ge W, Larson C R, et al. Predicting material accounting misstatements[J]. *Contemporary Accounting Research*, 2011, 28(1): 17–82.
- [44] Gu S, Kelly B, Xiu D. Empirical asset pricing via machine learning[J]. *The Review of Financial Studies*, 2020, 33(5): 2223–2273.
- [45] Wille B, Wiernik B M, Vergauwe J, et al. Personality characteristics of male and female executives: Distinct pathways to success? [J]. *Journal of Vocational Behavior*, 2018, 106: 220–235.
- [46] 王磊, 杨晓倩, 张鹏程, 等. 兼听则明: 高管社会网络与企业投资决策同伴效应[J]. *管理科学学报*, 2022, 25(6): 81–99.
- Wang Lei, Yang Xiaoqian, Zhang Pengcheng, et al. Hear all parties: Executive social networks and peer effects in firm's investment decisions[J]. *Journal of Management Sciences in China*, 2022, 25(6): 81–99. (in Chinese)
- [47] Gow I D, Kaplan S N, Larcker D F, et al. CEO Personality and Firm Policies[R]. *National Bureau of Economic Research*, 2016.
- [48] Dennis W C, Ruidi Shang. Tone at the bottom: Measuring corporate misconduct risk from the text of employee reviews[J]. *Management Science*, 2022, 68(9): 6355–7064.
- [49] 江新峰, 张敦力, 李欢. “忙碌”独董与企业违规[J]. *会计研究*, 2020, (9): 85–104.
- Jiang Xinfeng, Zhang Dunli, Li Huan. “Busy” independent directors and corporate violations[J]. *Accounting Research*, 2020, (9): 85–104. (in Chinese)
- [50] 姜富伟, 马甜, 张宏伟. 高风险低收益? 基于机器学习的动态 CAPM 模型解释[J]. *管理科学学报*, 2021, 24(1): 109–126.
- Jiang Fuwei, Ma Tian, Zhang Hongwei. High risk low return? Explanation from machine learning based conditional CAPM model[J]. *Journal of Management Sciences in China*, 2021, 24(1): 109–126. (in Chinese)
- [51] Forbes D P. Are some entrepreneurs more overconfident than others? [J]. *Journal of Business Venturing*, 2005, 20(5): 623–640.
- [52] Lommers K, El Harzli O, Kim J. Confronting machine learning with financial research[J]. *The Journal of Financial Data Science*, 2021, 3(3): 67–96.
- [53] Ahrens A, Aitken C, Schaffer M E. Using Machine Learning Methods to Support Causal Inference in Econometrics[M]. *Behavioral Predictive Modeling in Economics*. Cham, Switzerland: Springer, 2021.
- [54] Ke G, Meng Q, Finley T, et al. Lightgbm: A highly efficient gradient boosting decision tree[J]. *Advances in Neural Information Processing Systems*, 2017, 30: 3146–3154.
- [55] Bao Y, Ke B, Li B, et al. Detecting accounting fraud in publicly traded US firms using a machine learning approach[J]. *Journal of Accounting Research*, 2020, 58(1): 199–235.
- [56] Bertomeu J, Cheynel E, Floyd E, et al. Using machine learning to detect misstatements[J]. *Review of Accounting Studies*, 2021, 26: 468–519.
- [57] 陆瑶, 张叶青, 黎波, 等. 高管个人特征与公司业绩——基于机器学习的经验证据[J]. *管理科学学报*, 2020, 23(2): 120–140.
- Lu Yao, Zhang Yeqing, Li Bo, et al. Managerial individual characteristics and corporate performance: Evidence from a machine learning approach[J]. *Journal of Management Sciences in China*, 2020, 23(2): 120–140. (in Chinese)
- [58] Cheng K, Huang N, Shi Z. Survey-Based Forecasting: To Average or Not to Average[M]. *Behavioral Predictive Modeling in Economics*. Cham, Switzerland: Springer, 2021.
- [59] Wang R, Asghari V, Hsu S C, et al. Detecting corporate misconduct through random forest in China's construction industry[J]. *Journal of Cleaner Production*, 2020, 268: 122266.
- [60] Chen X, Cho Y H, Dou Y, et al. Fundamental analysis of detailed financial data: A machine learning approach[J]. *Journal of Accounting Research*, 2022.
- [61] 郦金梁, 吴谣, 雷曜, 等. 有效预警上市公司违规的递延所得税异动指标和人工智能模型[J]. *金融研究*, 2020, (8): 149–168.

- Li Jinliang , Wu Yao , Lei Yao , et al. Fraud detection using deferred income tax: A machine learning perspective [J]. Journal of Financial Research , 2020 , (8) : 149 – 168. (in Chinese)
- [62] 胡楠, 薛付婧, 王昊楠. 管理者短视主义影响企业长期投资吗? ——基于文本分析和机器学习[J]. 管理世界, 2021, 37(5): 139 – 156 + 11 + 19 – 21.
- Hu Nan , Xue Fujing , Wang Haonan. Does managerial myopia affect long-term investment? Based on text analysis and machine learning [J]. Journal of Management World , 2021 , 37(5) : 139 – 156 + 11 + 19 – 21. (in Chinese)
- [63] Zhang X , Du Q , Zhang Z. A theory-driven machine learning system for financial disinformation detection [J]. Journal of Operations Management , 2022 , 31: 3160 – 3179.
- [64] Mullainathan S , Spiess J. Machine learning: An applied econometric approach [J]. Journal of Economic Perspectives , 2017 , 31(2) : 87 – 106.
- [65] Lennox C , Wu J S. A review of China-related accounting research in the past 25 years [J]. Journal of Accounting and Economics , 2022: 101539.
- [66] Liang D , Tsai C F , Lu H Y R , et al. Combining corporate governance indicators with stacking ensembles for financial distress prediction [J]. Journal of Business Research , 2020 , 120: 137 – 146.

Corporate and managerial individual characteristics and listed company violation: Evidence from a machine learning approach

*HE Ying*¹ , *REN Li-qi*² , *YU Wen-lei*³ , *DU Ya-guang*¹

1. School of Economics and Management , Beijing University of Posts and Telecommunications , Beijing 100876 , China;
2. Mobile Digital Consulting Co. , Ltd , Beijing 100080 , China;
3. School of Management , Minzu University of China , Beijing 100081 , China

Abstract: Listed companies' violations have been an important issue that attracts the attention of the capital market. While studying the causal relationship between single-dimensional variables and this issue is crucial, constructing an effective holistic prediction model is also of great significance. This paper constructs a prediction model of listed companies' violations based on important company characteristics and managerial individual characteristics from the perspective of internal governance. Using a sample of Chinese A-share listed companies from 2008 to 2019, this study introduces two machine learning algorithms, LightGBM and SHAP, to examine the predictive ability, importance ranking, and prediction mode of the two types of characteristics on violation behaviors. The results show that the model can predict corporate violations to a certain extent, and corporate characteristics have a greater impact on the prediction than managerial individual characteristics. Specifically, higher information transparency of listed companies, higher net profit margin of total assets, lower asset-liability ratio, higher managerial shareholding ratio, lower performance volatility, and higher analyst attention are associated with a lower tendency for the model to predict violations. In addition, the model predicts an increased tendency for violations when executives are young and when the chairman and CEO roles are combined. Moreover, most corporate characteristics and managerial individual characteristics exhibit a non-linear relationship in predicting corporate violations, which is consistent with the findings of traditional theoretical and empirical studies. Overall, our study enriches the research on the characteristics of corporate executives in China from a predictive perspective and provides empirical evidence for regulatory authorities and investors to improve supervision and investment efficiency and for companies to optimize internal governance mechanisms.

Key words: corporate violation; company characteristics; managerial individual characteristics; LightGBM; SHAP