基于 XGBoost 的上市公司财务舞弊预测模型研究

周卫华　翟晓风　谭皓威

（中国财政科学研究院）

研究目标：探讨如何利用大数据和机器学习方法对上市公司财务数据和非财务数据进行分析和挖掘，并应用于上市公司财务舞弊识别和预测。**研究方法：**提出一种基于机器学习方法的上市公司财务舞弊预测模型 Xscore，对上市公司财务舞弊进行预测。**研究发现：**Xscore 模型能够提高模型预测的准确率，在准确率、召回率、AUC 指标、KS 值、PSI 稳定性等方面均优于 Fscore 模型和 Cscore 模型，更适合我国上市公司财务舞弊预测。**研究创新：**基于 2000～2020 年中国上市公司数据集为观测样本，通过 Benford 定律、LOF 局部异常法、IF 无监督学习法，解决了机器学习应用于财务舞弊识别研究时普遍面临的灰色样本问题，甄选兼具领域特性和统计特征的特征变量；首次将 XGBoost 集成学习方法应用到上市公司财务舞弊预测分析中，有效提高了上市公司财务舞弊准确率。**研究价值：**本文将 XGBoost 集成学习方法引入上市公司财务舞弊识别领域，有助于促进人工智能、机器学习在会计学中的研究与应用，为促进上市公司披露高质量的财务信息和维护资本市场秩序提供参考。

关键词　XGBoost　机器学习　财务舞弊　预测模型

中图分类号　F239.1；F23　　**文献标识码**　A

DOI:10.13653/j.cnki.jqte.2022.07.009

一、问题的提出

建立高标准的资本市场是推动我国经济高质量发展、加快构建新发展格局、推进国家治理体系和治理能力现代化的坚实基础。提升上市公司财务信息披露质量和防范上市公司财务舞弊对于维护资本市场秩序和社会公众利益具有重要作用。近几年，党中央、国务院在严肃财经纪律、加大财会监督、规范财务审计秩序、提升会计信息质量、遏制财务造假等方面给予极大的关注。从法律层面，2019 年修订的《中华人民共和国证券法》大幅提高了对财务舞弊行为的惩戒力度。2021 年 8 月国办发布《关于进一步规范财务审计秩序促进注册会计师行业健康发展的意见》，要求会计师事务所坚决落实“看门人”职责履行，有效发挥注册会计师审计鉴证作用。2022 年 3 月财政部、证监会发布《关于进一步提升上市公司财务报告内部控制有效性的通知》，要求加强财务舞弊风险的评估与控制。2021 年 11 月 12 日，广州中级人民法院判决康美药业股份有限公司因年报等虚假陈述侵权赔偿证券投资者损失 24.59 亿元，正中珠江会计师事务所及直接责任人员承担全部连带赔偿责任，13 名公司高管及 5 名独立董事分别承担 20％、10％和 5％的连接赔偿责任。康美药业案例的判决结果表明

上市公司管理层、会计师事务所、独立董事的法律责任得到进一步压实。然而，传统财务舞弊识别方法仅对财务报表数据进行简单分析和建模，很难发现上市公司背后财务舞弊的事实。这对于公司管理层、会计师事务所、独立董事的任职要求和风险规避提出巨大挑战，目前在理论界和实务界已引起极大争议。因此，如果能找到简单且有效的财务舞弊识别和预测方法，对于监管者、管理层、审计师和投资者都具有重要的参考价值。

本文的研究目标就是基于大数据和机器学习方法，根据中国上市公司披露的财务数据和非财务数据，构建一个符合中国资本市场特征的上市公司财务舞弊预测模型。BAO 等（2020）首次利用集成学习的方法进行美国上市公司财务舞弊行为识别，研究发现使用 RUSBoost 集成学习的上市公司舞弊识别模型显著优于传统 Logistic 回归模型和支持向量机模型。与 BAO 等（2020）研究不同，首先，本文首次使用 XGBoost 集成学习方法进行模型构建，该方法比 RUSBoost 集成学习方法更为领先，更适合财务舞弊识别研究①。其次，采用中国上市公司样本进行研究更契合我国资本市场特征和模型适用性。再次，本文充分融合会计领域知识和机器学习方法优势，利用财务数据和非财务数据样本特征进行模型训练。最后，本文基于 XGBoost 方法构建的财务舞弊预测模型 Xscore 具有良好的样本外预测能力。

从国内外已有的研究文献来看，主流研究主要集中于影响财务舞弊的各类特征因素（曾月明和许素，2019；任朝阳和李清，2017；卢馨等，2015；吴革和叶陈刚，2008），统计模型多以 Logistic 模型和线性回归模型为主（张曾莲和高雅，2017；熊方军和张龙平，2016；钱苹和罗玫，2015；洪荭等，2012）。上述研究利用因财务舞弊被中国证监会处罚的上市公司为样本，采用的特征变量各自不同，检验样本所处的时间段也不相同，研究目标主要集中在财务舞弊特征因素的可解释性，而非样本外的预测性。而且以 Logistic 回归为代表的线性模型对特征变量约束较多，往往难以拟合非线性的复杂问题导致准确率受限。由于无法判断现有财务舞弊预测模型的优劣，在实务中投资者仍然缺少简单、实用且可靠的财务造假预测模型（钱苹和罗玫，2015）。钱苹和罗玫（2105）尽管通过对比分析评估西方资本市场的 Mscore 模型和 Fscore 模型，提出适合中国情境的 Cscore 模型，但是由于受到线性回归模型的约束和限制，Cscore 模型在样本外的预测能力并未得到进一步证明。

基于大数据和机器学习的研究方法应用于财务舞弊预测模型的构建已引起学术界的极大兴趣（张力派等，2020；BAO 等，2020；BROWN 等，2020）。机器学习模型一般分为有监督学习和无监督学习，财务舞弊预测模型属于有监督学习。监督学习模型在应用中有几个关键点：首先，监督学习样本集包括舞弊样本集和非舞弊样本集，舞弊样本按照证监会处罚为依据一般比较容易构造，而非舞弊样本受到灰色样本②干扰容易出现选择偏误问题。如果监督学习样本集有偏甚至错误，那么所学习构造的模型自然也不准确。其次，在特征指标选取方面，必须兼具领域知识和统计特性，无论是局限于财务指标而忽略非财务指标或者不经筛选将大量特征值全部扔到模型中，既可能会出现模型过拟合问题也可能导致“维灾难”，这将会影响模型的泛化能力。最后，所构建的监督学习模型应在学习效果和泛化能力上取得平衡，以确保模型在样本外具有较强的适应性，真正实现较好的预测效果。

① 相比 RUSBoost 方法，XGBoost 方法通过在目标函数中加入正则项，用以权衡目标函数的下降和模型的复杂程度，有效避免过拟合，更适用于正负样本极不均衡的机器学习场景。

② 灰色样本指的是由于证监会认定财务舞弊存在滞后性或虽然实际舞弊但尚未发现和披露的样本。

本文选取 2000 年至 2018 年 A 股上市公司为样本学习集，2019 年至 2020 年为样本外预测集，构建了基于 XGBoost 集成学习算法的中国上市公司财务舞弊预测模型。本文主要研究贡献体现在：第一，所构建的财务舞弊模型具有更好的预测能力，利用样本外预测集进行预测验证，结果显示预测准确度达 85％，召回率为 79％，精准率为 42％，模型稳定性为 0.09，均明显优于 Cscore 模型和 Fscore 模型。第二，基于上述模型提出一种比率缩放的评分映射方法——Xscore 财务舞弊评分卡，实证检验具有较高的预测精准率和负样本捕获率，评分前 5％的高分段舞弊命中率高达 78％，前 20％的高分段舞弊捕获率达 64％。第三，在特征变量选取上从公司治理、财会监督、财务指标、企业运营等四个角度确定特征变量，同时从缺失率、信息量、相关度、稳定性以及迭代回归显著性五个统计特性筛选指标，最终精选 27 个特征变量。第四，为避免非舞弊样本选择偏误和正负样本不均衡问题，通过灰色样本剔除和 SMOTE 过采样方法提高模型的可靠性以及预测召回率。第五，针对机器学习模型的“黑盒”特性和可解释性不强问题，通过样本特征重要性和 SHAP 方法对模型预测结果进行全局归因分析和局部归因分析，以可视化方式提升模型的可解释能力。

二、文献回顾

1. 基于经验推理的财务舞弊特征识别方法

对财务报告舞弊的征兆、动机与手段识别的早期探索，主要基于专家判断、舞弊案例研究、舞弊样本描述性统计等基础方法经归纳推理后的经验判断。Loebbecke 等（1989）将 46 项财务舞弊诱因归纳为公司产生舞弊的内部条件、管理层舞弊动机和管理层道德观念三大类舞弊因子，建立 L/W 模型并用 77 个舞弊案例的详细资料进行验证。章美珍（2002）在银广夏舞弊案例中发现舞弊的征兆与财务指标和对外经营政策存在矛盾、未充分披露的关联交易、盈余减去经营活动产生的现金流为负、薄弱的内控环境与制度密切相关。陈慧璇和朱君（2013）对 163 份舞弊处罚公告进行多维度描述性统计，从行业集中度、舞弊手段、舞弊的并发性与持续性总结出统计规律。黄世忠等（2020）通过描述 2007～2019 年 113 家上市公司财务舞弊样本的特征分布与舞弊手段，从财务报表和非财务报表角度提出了一系列有效识别财务报表舞弊的异常特征。受限于研究方法，传统的舞弊研究停留在舞弊特征与手段的说明，缺乏结合多种舞弊影响因素的量化分析，研究成果较为局限。

2. 基于经典统计模型的财务舞弊动因分析方法

单变量分析法与多变量分析法是最早使用的量化方法。Beaver（1966）应对企业财务危机问题时采用了单变量分析法，通过非参数统计的二分类检验方法来确定分割点，以达到最大化降低错误分类的效果。许多学者借鉴 Beaver 的方法，将单变量分析应用到财务舞弊的预测中（Hansen 等，1996；Deshmukh 和 Millet，2011）。Altman（1968）最早将多变量分析法用于财务舞弊预测模型中，利用组内公司数量及错误分类成本和先验概率，计算得到 Z-Score 值阈值，用于划分舞弊与非舞弊样本。Ofori（2016）研究发现安然公司的财务舞弊在 1997 年就能被修正的 Z-Score 值预测。但是两种方法均存在一定缺陷，单变量分析法结论片面，单个变量的选取完全左右了研究结果，而多变量分析法受限于严格的假设条件，自变量分布性质需满足正态分布等假设，因此这两种方法均未得到广泛应用。

近 20 年来，以舞弊成因理论起点，寻找适合的舞弊特征因素，构建 Logistic 回归模型探索舞弊内在机理的研究模式逐渐成为主流。Beneish（1999）选取 8 个财务指标作为特征

变量，并使用 Logistic 回归建立了 Mscore 模型，成功预测了安然公司财务舞弊。为进一步发掘特征变量的预测能力。Dechow 等（2011）研究发现模型仅涉及财务指标作为特征变量预测准确性最高，并最终筛选出 5 个财务指标建立了 Fscore 模型。钱苹和罗玫（2015）通过梳理文献整理适用于中国市场的变量列表，并采用向后逐步剔除法筛选变量，并最终构建 Cscore 模型。洪荭等（2012）认为财务指标与非财务指标需要结合考虑，以进一步揭露不同类型特征变量如何联动诱导舞弊，基于 GONE 理论筛选相关变量进行 Logistic 回归，发现管理层、治理层以及增发配股相关的不良因素会增加舞弊的可能性。熊方军和张龙平（2016）在 Logistic 回归模型中加入更多与治理层相关的变量，结果进一步证实了非财务指标的重要性。张曾莲和高雅（2017）基于舞弊三角理论创新性加入自愿信息披露这一综合性质变量，并发现其与舞弊显著相关。Logistic 回归方法在因果推断的解释性与模型结果的稳定性方面存在优势，但其样本外的预测能力并不高。因为舞弊机理涉及多特征相互作用比较复杂，一方面难以完全用线性模型简单拟合；另一方面特征变量选择、假设设定存在一定的主观性。

3. 基于大数据和机器学习方法的财务舞弊预测方法

随着大数据与人工智能技术的发展，以大数据与机器学习方法为代表的人工智能方法在财务舞弊研究中逐渐受到学界关注。Kotsiantis 等（2006）利用决策树来预测财务舞弊，发现 6 个关联度较高的财务指标。刘君和王理平（2006）构建了财务报告舞弊识别的径向基概率神经网络模型，该模型以财务指标和股权结构指标为特征变量，判正率达到 80％，显著优于线性模型。Cecchini 等（2010）提出了一种基于原始财务数据映射特征的支持向量机财务舞弊预测模型，该模型并未采用通过原始数据计算的财务指标，而是直接将原始财务数据映射为同一年内更广泛的一组比率和不同年间比率的变化，其预测结果准确性优于基于财务指标的预测模型。Ravisankar 等（2011）对比了多层前馈神经网络、支持向量机、遗传算法、Logistic 回归及概率神经网络，概率神经网络和遗传算法的舞弊识别准确率更优。李双杰和陈星星（2013）使用 BP 神经网络方法构建了上市公司利润操纵识别模型，通过数据包络分析（DEA）方法降低模型的第二类错误，总体识别率从 70％以上提升至 85％以上。BAO 等（2020）将 Ada Boost 集成学习与 RUS 欠采样数据处理方法相结合，同时引入了一种更适用于舞弊预测任务的模型评估指标。实验结果证明，该模型在财务舞弊预测方面优于 Dechow 等（2011）基于财务比率的 Logistic 回归模型，以及 Cecchini 等（2010）将原始财务数据映射为更广泛比率集的单核支持向量机模型。从已有研究来看，基于机器学习的模型能够利用更广泛更原始的数据，更好地拟合非线性关系，从而能够提升模型的识别准确率。

综上所述，随着机器算力、数据可得和统计模型的不断提升，基于大数据和机器学习的方法是财务舞弊研究的主要方向。但是与基于解释的计量经济模型不同，机器学习模型往往更为技术化和工程化，表现出一定的“黑盒”特征。本文的研究试图突破传统机器学习研究的局限性，既注重特征变量的理论解释性，也充分利用统计学习特性，构造一个适合我国上市公司舞弊识别的模型并予以应用。

三、一种新上市公司财务舞弊预测模型的设计

1. 样本筛选与数据集构建

（1）样本选择。本文选择财务舞弊样本数据来源于 CSMAR 中国上市公司财务年报数

据库和违规事件库的违规信息数据库，从中选取监管机构发布的处理文件中涉及“虚构利润”“虚列资产”“虚假记载（误导性陈述）”等上市公司违规数据作为财务舞弊样本集。本文选择 2000 年至 2020 年的样本为时间观测窗口。数据集划分训练样本、测试样本和时间外样本。训练样本和测试样本的比例是 7∶3，时间外验证样本是整个建模样本中时间切片最后的一段样本，时间外样本用来检验模型对未来样本的预测能力和稳定性。考虑到公司违规行为被认定存在一定时滞，训练样本集和测试样本集选择 2000～2018 年为观测窗口，同时剔除金融行业数据。在非舞弊样本集方面，为避免潜在舞弊尚未发现的公司数据偏误，本文剔除曾有任意违规记录的数据样本。最终确定 3006 个舞弊样本数据集和 15693 个非舞弊样本数据集。

表 1　　实验样本数据集分布

数据集	划分依据	样本数（个）	舞弊（个）	非舞弊（个）	舞弊占比（%）
训练集	2000～2018 年样本随机划分：训练集 70%；测试集 30%	10117	1756	8361	17.36
测试集		4336	752	3584	17.34
时间外样本	2019～2020 年样本	4246	498	3748	11.73
总计		18699	3006	15693	16.08

（2）灰色样本剔除。“灰色样本”指非舞弊样本中实际存在舞弊行为但未被发现的样本。如果将灰色样本作为舞弊公司的对照组，则数据可靠性必然存疑，由此训练的模型实际效果就要受到质疑（岳殿民，2008）。本文借助 Benford 定律（Kinnunen 和 Koskela，2003）、LOF 局部异常因子法（Breunig 等，2000）和无监督学习 IF 孤立森林算法（Ding 和 Fei，2013），将非舞弊样本集中的灰色样本剔除，从而提高模型的可靠性并解决舞弊样本集与非舞弊样本集不均衡问题。

首先，Benford 定律是指大量自然数据集中首位数字存在一定的统计分布规律。如果某财务指标的首位数字分布与 Benford 定律不一致，则认为该财务指标受到人为操纵的可能性较大。本文利用相关性系数和卡方拟合优度检验构建 Benford 风险因子，并剔除非舞弊样本中首位数字和第二位数字触发 Benford 风险因子的样本。

任务 1：计算概率分布与相关系数。将 Benford 定律首位、第二位数字的分布概率分别定义为 X，以上市公司为单位统计样本内该公司资产负债表、利润表、现金流量表的财务数据的首位和第二位数字的分布情况，定义为 Y，计算 X 与 Y 之间的相关系数 r，r 越接近 0，说明上市公司财务数据首位、第二位数字的分布越偏离 Benford 定律理论分布值，则造假概率较高。

任务 2：卡方拟合优度检验。本文将卡方检验作为 Benford 定律的假设检验理论，来检测观测样本数值分布与 Benford 定律理论分布的符合程度。以上市公司的资产负债表、利润表、现金流量表财务数据为观测样本，计算 χ^2 统计量并与 χ^2 标准值比较：如果样本组 χ^2 统计量超过了 χ^2 标准值，则“拒绝”原假设，表明实际的财务数据不符合 Benford 定律的理论分布，财务舞弊的可能性较高。

任务 3：Benford 风险因子判断。本文借鉴岳娇（2020）基于 Benford 定律的“财务舞弊敏感指标”，提出 Benford 风险因子，Benford 风险因子主要有以下三个：首位数字与 Benford 定律的相关系数 $r_1<0.9$；第二位数字“0”的卡方分布超过 0.18；卡方检验（置信

度 95%）结果为“拒绝”。三者之间为“或”的关系，财务数据只要符合其中一条，即触发了 Benford 风险因子判断，该样本将被剔除。剔除后，非舞弊样本从原来的 8361 条减少到 7902 条，训练集中舞弊样本占比由 17.36%提升到 18.18%。

其次，*LOF* 异常检测法。*LOF* 算法引入了一种刻画数据密度的方法——局部可达密度，以衡量样本的异常程度。根据局部可达密度的定义，如果一个样本点在样本空间中，且距离其他样本点比较远，则它的局部可达密度就比较小。因为数据分布不均匀、密度不同，可能导致一些虽然稠密但相对稀疏的点被视为异常。因此需要计算样本点 p 与点 p 的 k－近邻的相对密度，从而引出了局部异常因子（*LOF*）的概念：

$$LOF_k(p)=\frac{\sum_{o\in N_{k(p)}}\frac{lrd(o)}{lrd(p)}}{|N_k(p)|}=\frac{\frac{\sum_{o\in N_{k(p)}}lrd(o)}{|N_k(p)|}}{lrd(p)} \tag{1}$$

局部异常因子公式含义为：样本点 p 的局部相对密度（局部异常因子）LOF_k（p），为样本点 p 的 k－近邻平均局部可达密度与样本点 p 的局部可达密度 lrd（p）的比值。可以看出，样本点 p 的局部可达密度相比于其 k－近邻的平均局部可达密度越小，其异常程度越大。本文使用 LOF 模型在训练集上进行学习，将异常分超过 0.7 百分位的样本从训练集的非舞弊样本中剔除。剔除后的非舞弊样本从原来的 7902 条减少到 5531 条，训练集中舞弊样本占比由 18.18%提升到 24.10%。

最后，IF 异常检测法。孤立森林（Isolation Forest，IF）是一种基于空间随机划分思想的集成算法，是由多棵二权树并行得到，再将输出结果进行加权平均，通过多棵孤立树组合，增强模型稳定性，得到最终 IF 模型。IF 算法最终输出 IF 异常分，定义为：

$$Score(x_i)=2^{\frac{E(h(x_i))}{C(n)}} \tag{2}$$

其中，E（h（x_i））表示 x_i 在所有孤立树上的路径长度的均值，n 表示一个颗粒数上训练样本的个数，C（n）表示用 n 个样本训练的二权树的平均路径长度，作为归一化项。从 IF 异常分计算方式来看，数据 x 在多棵孤立树中的平均路径长度越短，得分越接近 1，表明数据 x 越异常；数据 x 在多棵孤立树中的平均路径长度越长，得分越接近 0，表明数据 x 越正常。本文使用 IForest 模型在训练集上进行学习，将异常分超过 0.7 百分位的样本从训练集的非舞弊样本中剔除。剔除后的非舞弊样本从原来的 5531 条减少到 3871 条，训练集中舞弊样本占比由 24.10%提升到 31.21%。

（3）SMOTE 过采样。SMOTE（Synthetic Minority Oversampling Technique，SMOTE）的思想是对少数类样本进行分析，然后在现有少数类样本之间进行插值，人工合成新样本，并将新样本添加到数据集中进行训练，经常应用于非平衡样本的处理中，解决了随机过采样通过重复复制原来少数类样本导致的过拟合现象。本文选用轻度提升向量机（Light Gradient Boosting Machine，LightGBM）对样本进行清洗，然后使用 SMOTE 算法对分类可信度较高的样本进行过采样，最终将采样后的样本与旧样本合并返回。通过 SMOTE 生成新的舞弊样本 2115 条，最终训练集的非舞弊样本与舞弊样本的比例 1∶1。

2. 特征选择与筛选

财务舞弊特征变量的选择是构建高效和准确的机器学习模型的前提。从理论上说，在大数据环境下机器学习模型可以投入足够多的特征变量，但是过多的特征变量容易造

成“维灾难”，易导致过拟合现象。通过梳理已有文献，选择特征变量是一种可行的思路。特征变量选择应兼具领域知识和统计特性，同时可采用特征组合和特征聚合的方式构造衍生特征变量。本文在借鉴已有文献基础上，选择特征变量主要从公司治理、财会监督、财务指标、企业运营四个方面分析确定。公司治理主要关注公司的股权特征、董事会、监事会和管理层等治理环境；财会监督分别从内部控制和外部审计的角度分析选择指标；财务指标是从公司的偿债能力、盈利能力、经营能力、现金流能力和发展能力五个角度分析选择指标；企业运营主要从销售收款、采购付款、资金流动、关联交易出发，关注财务数据之间的勾稽关系是否符合预期，通过趋势对比和同业对比，构建反映企业运营异常的指标。

（1）特征变量选择[①]。第一，公司治理反映公司的管理水平和内部制衡程度。在股权特征方面，已有文献主要从股权性质、股权集中度和制衡度、股权流通度、持股人性质展开；在治理环境方面，董事会治理指标中的独立董事比例、二职合一、董事会持股比例、董事会规模及董事会成员特征等都是学者们关注的重要因素；监事会治理指标中，监事会规模、监事会会议次数与财务报告舞弊正相关（刘立国和杜莹，2003；陈关亭，2007）。

第二，企业的财会监督包括内部监督和外部监督，分别以内部控制和外部审计为典型监督环境。在内部控制层面，研究表明，内部控制与财务舞弊具有高度相关性，高阶经理人舞弊较易发生在缺乏有效内部控制制度的机构（陈玟伶，2010）。迪博内控指数是依据迪博中国上市公司内部控制指数体系计算出的各上市公司的内部控制指数评分，是内部控制评价的综合指标（周守华等，2013）。在外部审计层面，潜在舞弊的上市公司更可能被审计师出具非标准无保留审计意见（陈国欣等，2007）。已有文献主要从事务所规模、审计任期、审计意见类型、是否变更事务所、审计轮岗、审计师工作经验和行业专长等方面考察了外部审计揭露财务报告舞弊的影响因素。

第三，财务指标主要从公司的偿债能力、盈利能力、经营能力、发展能力和现金流能力几个角度展开。偿债能力指标反映企业的债务偿还能力，是衡量企业能否长期经营，稳定发展的重要指标；盈利能力指标用来评价公司业绩、公司价值及创造价值的能力；经营能力反映公司资产的周转利用情况与经营管理水平，是公司健康良性发展的重要基础；发展能力指标可以衡量一个公司的成长性，成长阶段以及生命周期，反映公司的发展前景与投资价值；现金流能力指企业在生产管理活动中，有关经营、筹资、投资三大领域的现金流入与流出，可以用于反映公司的收益质量。

第四，财务活动贯穿于企业运营的全过程，财务舞弊一般都会引起财务循环信息异常。根据会计恒等式，这些舞弊方法一定会导致现金流量表或资产负债表科目的异常。每次资金在体内外循环都会因为融资成本和税务成本的发生导致循环资金的减损，一旦开始舞弊，异常科目会逐渐增加，占比会越来越高。首先，伪造净利润需要有大额资金的流动循环，资金的来源和归还环节暴露出的异常也是识别舞弊的重要突破口。其次，通过关注伪造现金流涉及的会计科目的勾稽关系，本文分析了销售收款环节和采购付款环节的财务数据勾稽关系，分别提出了识别财务舞弊线索的异常指标。最后，资金的循环离不开关联方或行为关联方的配合，还需要关注关联交易的异常情况。

① 本文共选择 211 个原始特征变量，由于受到篇幅限制，未在正文中进行报告。

（2）衍生特征变量。首先，特征聚合。通过对每个变量多个时间节点的取值进行聚合运算，作为样本的衍生特征。财务舞弊作为事后结论，先有舞弊动机，再发生舞弊行为，最终反映到财务报告中，即在实际进行舞弊行为数年前就有征兆。特征聚合采取“动静结合”策略。静态追溯是指计算样本公司的基础特征指标在第 t 年的前 3 年、前 2 年和前 1 年的年度财务数据，将以前年度静态财务数据作为新的特征来预测第 t 年的财务舞弊行为。动态追溯是指对于连续型变量，计算样本公司的基础特征指标在第 t 年的变动幅度与第 t－1 年变动幅度的差值，以及第 t－1 年的变动幅度与第 t－2 年变动幅度的差值；对于离散型变量，把第 t 年特征数据相较于第 t－1 年、第 t－1 年相较于第 t－2 年是否发生变化以及变化的方向，作为第 t 年新的特征，用财务指标的动态变动信息预测第 t 年的财务舞弊行为。

其次，特征组合。通过算法自动特征交叉。财务比率指标本质上是依靠专家经验总结出的财务数据的特征组合，而通过算法的自动特征交叉，可以考虑到尚未被专家归纳且有效的特征组合，对于提高模型拟合度具有重要作用。

最后，特征离散。特征离散化处理是一种将连续型变量离散分组，或将字符型变量赋值为数值型变量的方法。它能使模型结果更加稳定、提高模型灵活度，避免模型数据中极端值的影响。常用的离散化处理方法有 One-Hot 编码和 WOE（Weight of Evidence）编码：One-Hot 编码是将每个离散的分类变量作为一个特征，用 0 或 1 对样本进行标注，如果该特征有 n 个分类则会映射成 n 个特征；WOE 编码是对离散变量分箱处理，表示的是当前分组中舞弊样本占样本中所有舞弊样本的比例 p_{y_i} 与当前分组中非舞弊样本占样本中所有非舞弊样本的比例 p_{n_i} 之间的差异，这个差异用二者比值取对数表示，WOE 越大，这种差异越大。公式如下：

$$woe_i = \ln\left(\frac{p_{y_i}}{p_{n_i}}\right) \tag{3}$$

对于短文本类型的变量进行转换时，WOE 映射的效果相比于 One-Hot 编码效果更好。此外，对连续型变量进行 WOE 编码能够弱化极值对模型稳定性的影响，增加模型鲁棒性。本文选用卡方分箱进行 WOE 编码。

表 2　　特征变量筛选过程　　单位：个

处理阶段	阶段描述	处理标准	剔除特征数	保留特征数
特征准备	原始特征	四大变量类型	—	211
	特征衍生	按特征聚合、特征组合和特征离散衍生特征变量	—	1266
初步筛选	缺失率筛选	剔除缺失率＞70％的变量	46	1220
	IV 值筛选	剔除 IV 值＜0.02 的变量	106	1114
	相关性筛选	剔除相关系数＞0.7 的变量	569	545
	分箱离散	卡方分箱并为每个变量 WOE 编码	—	545
	PSI 筛选	剔除 PSI＞0.02 的变量	358	187
	IV 值筛选	剔除 IV 值＜0.02 的变量	22	165
	相关性筛选	剔除相关系数＞0.7 的变量，保留 IV 值较大的	97	68

（续）

处理阶段	阶段描述	处理标准	剔除特征数	保留特征数
迭代筛选	逐步回归 OLS 双向消除（同步比较）	AIC（p 值 0.2）	44	24
		BIC（p 值 0.2）	58	10
		AUC（p 值 0.2）	65	3
		KS（p 值 0.2）	65	3
综合筛选	模型调优	一致性、显著性、相关性、稳定性、变量意义综合筛选	0（剔除 3 个，补充 3 个）	27

表 3　　最终确定的特征变量

变量类型	风险类型	变量名称	变量解释
公司治理	治理环境	监管层总人数	年报披露的董事会、监事会、高级管理人员的总人数
	股权特征	境内发起人法人股股数占未流通股份比例	境内发起人法人股股数占未流通股份比例
财会监督	外部审计	会计师事务所是否变更	是否变更年报审计会计师事务所
	内部控制	迪博内控指数$_{\Delta t-(t-1)}$	迪博内控指数较上年变化值
企业运营	资金来源	存贷双高$_{t-1}$	上年是否存贷双高
		股权质押比例$_{t-1}$	上年股权质押比例
	资金归还	长期资产与现金支付异常度$t-1$	上年长期资产与现金支付异常度
	关联交易	其他应收款与营业总收入增幅异常	其他应收款增幅－营业收入增幅
		关联交易依赖度	关联交易金额占营业收入比重
	采购付款	应付账款增幅与存货和收入增幅异常$_{t-3}$	三年前应付账款增幅与存货和收入增幅异常
	销售收款	存货增幅与营业收入增幅差$_{\Delta(t-1)-(t-2)}$	上年存货增幅与营业收入增幅差较前年变化值
财务指标	盈利能力	长期资本收益率$_{\Delta t-(t-1)}$	长期资本收益率较上年增加值
		现金与利润总额比$_{t-1}$	上年现金与利润总额比
		净资产收益率	净资产收益率
		净利润与利润总额比$_{t-3}$	三年前净利润与利润总额比
	现金流能力	营运指数$_{\Delta t-(t-1)}$	营运指数较上年增加值
		筹资活动股东现金净流量$_{t-2}$	两年前筹资活动股东现金净流量
	经营能力	股东权益周转率	股东权益周转率
		非流动资产周转率$_{t-1}$	上年非流动资产周转率
	发展能力	营业利润增长率$_{\Delta t-(t-1)}$	营业利润增长率较上年增长值
		营业利润增长率	营业利润增长率
		每股经营活动产生的净流量增长率$_{t-2}$	两年前每股经营活动产生的净流量增长率
		可持续增长率$_{\Delta(t-1)-(t-2)}$	上年可持续增长率较前年的变化值
		净资产收益率增长率$_{t-1}$	上年净资产收益率增长率
		经营活动产生的净流量增长率$_{t-1}$	上年经营活动产生的净流量增长率
	偿债能力	现金比率	当年现金比率＝现金及现金等价物期末余额/流动负债
		权益乘数$_{\Delta(t-1)-(t-2)}$	上年权益乘数较前年增长值

（3）特征变量筛选。特征数量在特征衍生和离散化后共 1266 个。为减少运行效率低和过拟合情况，采用特征工程方法进行筛选，剔除对建模帮助不显著的特征变量，具体从 5 个角度进行：特征缺失率、特征信息量、特征相关度、特征稳定性 PSI、递归特征删除。特征变量筛选过程和最终确定特征变量如表 2 和表 3 所示。

3. Xscore 模型构建

（1）Xscore 模型结构。本文基于 XGBoost 算法来构建 Xscore 模型。XGBoost 算法是以 CART 为基分类器的集成学习方法之一，是对 GBDT 算法的改进和扩展。XGBoost 算法的内核是提升树方法，通过引入正则项和列抽样的方法提高了模型稳健性，同时对损失函数进行了二阶泰勒展开，使求解最优解时效率更高。

在上市公司财务舞弊模型研究中，每个样本数据点由 x_i（主体 i 的若干个特征指标）和 y_i（模型输出结果）组成，设 $S=\{(x_i, y_i)\}$（$i=1, 2, 3, \cdots, n$，$x_i \in \mathbb{R}^m$，$y_i \in \mathbb{R}$），其中 n 为样本个数，m 为样本的维度数量。设决策树的数量为 h（$h=1, 2, 3, \cdots, t$），模型损失函数为：

$$Obj^{(t)} = \sum_{i=1}^{n} l(y_i, \widehat{y_i^{(t)}}) + \sum_{h=1}^{t} \Omega(f_h) \tag{4}$$

其中 $\widehat{y_i^{(t)}} = \sum_{h=1}^{t} f_h(x_i) = \widehat{y_i^{(t-1)}} + f_t(x_i)$，$f_h \in E$，其中 E 为回归树的集合空间，f_h 为回归树 x_i 为第 i 个数据的特征向量，y_i 为真实值，$\widehat{y_i^{(t)}}$ 为预测值。在引入正则项后，对损失函数进行泰勒展开，并去掉常数项，可得到简化的目标函数为：

$$Obj^{(t)} \approx \sum_{i=1}^{n} \left[g_i f_t(x_i) + \frac{1}{2} w_i f_t^2(x_i) \right] + \Omega(f_t) \tag{5}$$

其中 g_i 为 $\alpha_{\hat{y}(t-1)} l(y_i, \widehat{y^{(t-1)}})$，$w_i$ 为 $\alpha^2_{\hat{y}(t-1)} l(y_i, \widehat{y^{(t-1)}})$，进而对样本进行训练自动化调参完成模型的构建。

（2）Xscore 模型参数学习。模型训练时需要指定模型参数，常用的调参策略是基于随机搜索、遗传算法、贝叶斯优化等形式实现。本文根据自定义的规则实现了模型自动化参数搜索。期望模型的训练集 KS 值和时间外样本 KS 值足够接近，且时间外样本集的 KS 值足够大。前者用于保证模型的跨时间稳定性不会太差，而后者用于保证模型的精度足够高。KS 值的分配权重 w 可以根据实际情况进行调节。当稳定性较差时，应更多关注两者 KS 值的差值，需要调高 KS 值权重 w。

本文的参数搜索方案使用一种针对目标 KS 值的贪心搜索方法。每次只考虑单个参数进行向前和向后搜索，当对目标值 KS 值有提高时，继续搜索，否则停止该方向的搜索。最终确定的模型参数如下表：

表 4　　Xscore 模型训练参数

参数	参数名称	参数值
learning _ rate	学习率	0.05
max _ depth	子树最大深度	3
n _ estimators	最大迭代次数	300
min _ chid _ weight	子节点权重阈值	1

（续）

参数	参数名称	参数值
sub _ sample	训练样本的采样比例	0.8
scale _ pos _ weight	调整正负样本权重	1
reg _ lambda	L2 正则化系数	300

（3）Xscore 处理效果。数据、特征和模型是机器学习的三个关键因素。为了分析每一步数据处理的效果，本文详细报告每一步数据处理后模型在测试集样本上的表现，以验证每一步数据处理的实际意义。由于 Logistic 模型对数据预处理的反应更敏感，能更好地分析每一步数据处理的效果，本文同时分析对比数据处理过程中在 Logistic 和 Xscore 上的模型表现。

第一，原始样本在特征衍生后，在保证样本量不变的情况下，召回率和精准率都有所降低。这是因为特征衍生为模型带来更多信息的同时，也增加了无关特征的噪音对模型训练的干扰，导致模型过拟合。因此，对衍生变量进行筛选，剔除对模型训练无效甚至起干扰作用的特征变量，保留信息含量更高的特征变量是非常必要的。

第二，特征离散化处理后，召回率和精准率都有提高，尤其是 Logistic 模型效果提升显著。这是因为对于 Logistic 线性回归模型而言，WOE 离散化编码为模型提供了更多的非线性信息，提高了模型对负样本的响应程度。一般而言，召回率和精准率是此消彼长的，但离散化后 Logistic 和 Xscore 模型的这两个指标都有提升，说明离散化对提高模型预测的查全度和查准度都有显著效果。

第三，特征筛选后，训练集特征维度降低，而模型效果损失却相对较小，甚至有所提升。在衍生特征根据缺失率、特征信息 IV 值、相关度和稳定性筛选后，特征维度从 1266 维精简成 68 维，筛选后特征数量是原来特征的 5%，虽然 Logistic 模型的召回率有所下降，但是精准度却有提升，AUC 从 0.85 下降到 0.80，这说明 1266 个特征剔除其中的 95%，仅能导致 0.05 的 AUC 下降，5%的特征就能表达大部分信息。更进一步，在根据 AIC 和 BIC 最小化以及 KS 和 AUC 最大化迭代筛选后，保留的 27 个特征并没有导致召回率和精准率的大幅下降，甚至在 Logistic 模型中相较于 68 维特征 AUC 有 0.003 的提升，KS 区分度提升了 0.01，Xscore 的 AUC 提升了 0.004。

第四，灰样本剔除的主要目的是提高模型训练数据的质量，减少噪音数据干扰，增强模型的稳健性。本文对比了三种灰样本剔除法：Benford 风险因子法、LOF 局部异常度筛选法、IF 孤立森林无监督学习法。SMOTE 过采样主要为了解决样本不均衡导致的过拟合现象。研究发现，经过灰色样本剔除和 SMOTE 过采样后，普遍出现召回率提升而精准率下降的现象，与灰色样本剔除的最初目的一致，训练的模型对于非舞弊样本的要求更高，预测结果更加谨慎。与此同时，模型的整体效果并没有显著降低，Logistic 和 Xscore 模型的 AUC 降低不到 0.01，KS 降低不到 0.02。因此，灰样本剔除和均衡化处理虽然损失一定的精准率，但是提高召回率对于财务舞弊识别更有意义。在财务舞弊预测的应用场景中，由于召回率的成本是投资的直接损失，而精准率的成本可由转投其他股票来降低，为尽可能降低犯第一类错误的概率，一般期望模型不遗漏任何舞弊样本，选择高召回率的模型。根据不同的应用场景，我们在训练模型时可以有选择或适度地进行灰色样本剔除和样本均衡化处理。

表 5 Xscore 模型处理效果

处理阶段	处理步骤	训练集样本（行，列）	Logistic		Xscore	
			召回率	精准率	召回率	精准率
0 初始样本	原始样本	10117，211	0.1011	0.5241	0.7732	0.3763
1 特征加工	特征衍生	10117，1266	0.0918	0.4792	0.7686	0.3710
	原始特征离散	10117，211	0.4694	0.7278	0.7735	0.3843
	衍生特征离散	10117，1266	0.4934	0.6386	0.7726	0.3976
	衍生离散特征初筛	10117，68	0.3391	0.6522	0.2832	0.6265
	特征迭代筛选	10117，27	0.3511	0.6423	0.2965	0.6445
2 灰样本剔除	Benford 因子筛选	9658，27	0.3577	0.6374	0.3138	0.6327
	LOF 异常度筛选	7608，27	0.4348	0.5514	0.3684	0.6128
	IF 异常度筛选	7608，27	0.4348	0.5514	0.3684	0.6128
	3 种灰度剔除交集	5968，27	0.7048	0.3891	0.5904	0.4476
	3 种灰度剔除迭代	5627，27	0.7247	0.3530	0.6104	0.4371
3 样本均衡	SMOTE 过采样	7742，27	0.8019	0.3094	0.7154	0.3871

4. Xscore 模型评价

（1）Fscore 模型与 Cscore 模型。为了全面评估模型在实际应用中的效果，本文以西方资本市场常用的 Fscore（Dechow 等，2011）模型和基于中国上市公司数据的 Cscore（钱苹和罗玫，2015）模型为对照，分析 Xscore 模型在财务舞弊预测中的提升效果。其中 Fscore 模型和 Cscore 模型公式如下所示。

$$Fscore = -7.893 + 0.79RSST\_ACC^{***} + 2.518CH\_RC^{***} + 1.191CH\_INV^{***} + 1.979SOFT\_AS^{***} + 0.171CH\_CS^{***} - 0.932CH\_ROA^{***} + 1.029ISSUE^{***} \quad (6)$$

$$Cscore = -0.983 - 2.261TATA^{***} - 2.495CH\_CS^{***} + 5.075OTHREC^{***} + 0.797LOSS^{***} - 0.059SD\_VOL^{***} - 3.198H5INDEX^{***} - 4.298INSTIU^{***} + 0.888ISSUE^{***} + 1.184STKCYC^{***} \quad (7)$$

为了测试模型的泛化能力，检验模型的稳定性，本文在时间外样本上进行测试。时间外样本未参与模型的训练和优化过程，一方面有效避免数据窥探问题，测试结果更具说服力，另一方面能反映用历史数据训练的模型在预测未来时的稳定性表现。将相同公司相同年份的时间外样本数据，分别在 Fscore、Cscore、Xscore 中进行检验。

（2）Xscore 模型评价指标分析。为了全面有效对比模型性能，本文选择分类任务常用指标搭建评价体系。包括：准确率（Accuracy）是所有预测正确的样本占所有的样本的比例；精准率（Precision）表示预测出的舞弊样本中确实是舞弊样本的比例，反映模型的可信度；召回率（Rrecall）表示真正的舞弊公司中被模型正确预测的比例，反映犯第一类错误的概率；F1 分数反映召回率和精准率的综合影响；TPR 即 Recall 反映真正类率，FPR 为假正类率反映被错误分类为正类的非正类比例；KS 值则定义为 KS 曲线中 TPR 与 FPR 差值的最大值；PSI 可衡量测试样本及模型训练样本评分的分布差异，其计算逻辑是用实际占比与预期占比的比值取对数作为加权，对实际占比与预期占比的差值累加得到。上述指标公式

定义如下：

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{8}$$

$$Precision = \frac{TP}{TP + FP} \tag{9}$$

$$Recall = \frac{TP}{TP + FN} \tag{10}$$

$$F_1 = \frac{2 \cdot precision \cdot recall}{precision + recall} \tag{11}$$

$$TPR = \frac{TP}{TP + FN} \tag{12}$$

$$FPR = \frac{TN}{FP + TN} \tag{13}$$

$$KS = MAX(|TPR - FPR|) \tag{14}$$

$$PSI = \sum_i (p^i_{target} - p^i_{base}) * ln\left(\frac{p^i_{arget}}{p^i_{base}}\right) \tag{15}$$

表 6　　Xscore、Cscore、Fscore 模型时间外样本测试结果汇总

Model	Accuracy	AUC	Recall	Precision	F1	AP	KS	PSI
Xscore	0.8486	0.8956	0.7851	0.4218	0.2036	0.6401	0.6486	0.0883
Cscore	0.5450	0.6969	0.5635	0.1430	0.1381	0.1590	0.3030	0.1146
Fscore	0.3573	0.6193	0.5635	0.1024	0.1379	0.1141	0.1841	0.1152

从表 6 的测试结果研究发现，Xscore 模型在多项评价指标中表现均优于 Fscore 和 Cscore。基于中国数据训练的 Cscore 模型在 KS、AUC、AP 等指标上要优于西方数据训练的 Fscore 模型。从综合 KS、PSI、Recall、AUC、AP 等评价指标来看，Xscore 模型表现明显优于 Cscore 模型和 Fscore 模型。Cscore 模型和 Fscore 模型均是基于 Logistic 模型回归的结果，因为 Logistic 模型的参数估计是基于线性回归理论的，当因变量和自变量之间存在明显的线性关系时，Logistic 模型的预测效果很好。但财务舞弊问题错综复杂，因变量与自变量之间关系并不仅仅是线性的，这种非线性关系就无法用 Logistic 模型解释。Xscore 模型是基于树模型的集成算法，可以实现特征的自动交叉组合，从而能更好地拟合非线性特征，模型训练前进行了大量数据预处理工作，显著提高了模型预测能力。

四、Xscore 上市公司财务舞弊预测模型的应用

1. Xscore 模型特征贡献度分析

本文引入贡献度评价方法提升 Xscore 模型的可解释性。贡献度评价方法需要满足一致性和精确性的条件。一致性是指当改变某个特征边际贡献度时，其余特征的边际贡献度不会发生改变。精确性是指所有特征重要性综合应当等于该模型的总体重要性，即各个特征的影响与常数项之和等价于模型的输出结果。本文从模型全局归因方法和局部归因方法两个角度

分别对特征重要性进行分析。

（1）全局归因分析。XGBoost 模型的特征归因方法 Gain、Cover 和 Weight 都属于全局归因方法，反映当删除一组特征时，模型的预期精度发生的变化，即特征对全局精度的影响。其中，Gain 反映特征变量对准确率的影响，用特征带入决策树枝干后模型准确率的提高程度来衡量；Cover 反映特征变量对观测对象的覆盖程度，用与特征相关的观察对象的数量来衡量；Weight 反映特征变量在模型中被使用的频率，用特征变量在所有决策树中被使用的次数衡量。

表 7　　Xscore 模型特征的 Gain、Weight 和 Cover 值

特征变量	Gain	Weight	Cover
净资产收益率	0.201	0.077	0.036
营运指数$_{\Delta t-(t-1)}$	0.085	0.009	0.066
现金比率	0.078	0.027	0.051
可持续增长率$_{\Delta(t-1)-(t-2)}$	0.072	0.017	0.061
长期资本收益率$_{\Delta t-(t-1)}$	0.053	0.038	0.056
营业利润增长率	0.052	0.059	0.028
净利润与利润总额比$_{t-3}$	0.047	0.061	0.043
股权质押比例$_{t-1}$	0.041	0.074	0.039
权益乘数$_{\Delta(t-1)-(t-2)}$	0.041	0.028	0.051
关联交易依赖度	0.039	0.085	0.035
净资产收益率增长率$_{t-1}$	0.038	0.038	0.039
筹资活动股东现金净流量$_{t-2}$	0.030	0.088	0.030
存货增幅与营业收入增幅差$_{\Delta(t-1)-(t-2)}$	0.027	0.035	0.042
营业利润增长率$_{\Delta t-(t-1)}$	0.024	0.036	0.030
内部控制指数$_{\Delta t-(t-1)}$	0.023	0.048	0.035
现金与利润总额比$_{t-1}$	0.022	0.034	0.034
其他应收款与营业总收入增幅异常	0.021	0.022	0.050
境内发起人法人股股数占未流通股份比例	0.019	0.030	0.043
非流动资产周转率$_{t-1}$	0.018	0.038	0.033
会计事务所是否变更	0.016	0.040	0.036
股东权益周转率	0.016	0.038	0.029
存贷双高$_{t-1}$	0.010	0.002	0.025
经营活动产生的净流量增长率$_{t-1}$	0.010	0.024	0.028
应付账款增幅与存货和收入增幅异常$_{t-3}$	0.008	0.024	0.032
高级管理人员数	0.006	0.025	0.028
每股经营活动产生的净流量增长率$_{t-2}$	0.001	0.003	0.009
长期资产与现金支付异常度$_{t-1}$	0.001	0.002	0.008

特征重要性 Gain、Cover 和 Weight 是在整个训练集上计算得到的重要度期望，更改某个特征的影响会直接导致其他特征重要性发生变化，难以满足一致性要求。为此本文引入 SHAP 方法，它具有局部精确、缺失无影响、一致性等多个优点，对集成模型的特征贡献具有较好的解释性。Shap 值定义为所有可能的特征组合子集中，特征值的平均边际贡献，即边际贡献度的加权平均值。Shap 值的计算需要遍历所有的特征组合子集，通常使用一种近似的采样算法，以提升求解的效率。对于每个集合 S 抽取 M 次样本，每次迭代的过程中，除集合 S 外的特征均用抽取的样本替代，从而计算基于当前特征子集 S 得到的模型预测结果的期望值。对每个特征值 i 的 Shap 值计算公式表示为：

$$shap_i = \frac{1}{M}\sum_{m=1}^{M}(\hat{f}(x_{+i}^m) - \hat{f}(x_{-i}^m)) \tag{16}$$

其中，$\hat{f}(x_{+i}^m)$ 和 $\hat{f}(x_{-i}^m)$ 都是样本特征值的预测结果，不在子集中的特征值被来自随机样本 z 中的特征值替换。区别是 $\hat{f}(x_{+i}^m)$ 中特征 i 的值不会被随机替换，而 $\hat{f}(x_{-i}^m)$ 中的特征 i 的值会被 z 随机替换。SHAP 是一种利用 Shap 值可加性的特征贡献度量方法。Shap 将模型的预测结果表示为所有输入特征的影响总和，其公式表示为：

$$g(x') = shap_0 + \sum_{i=1}^{M} shap_i \tag{17}$$

图 1 是训练集中所有样本所有特征值的 Shap 影响，图中每行代表一个特征，横坐标轴为 Shap 值，其中每一个点代表一个样本，颜色越深说明特征本身数值越大，颜色越浅表明特征本身数值越小。图 2 是取每个特征的 Shap 值的绝对值的均值，作为该特征的重要性。我们发现，Xscore 模型预测结果贡献度 Shap 值最大的前 5 个特征分别是净资产收益率、上年股权质押率、关联交易依赖度、前年筹资活动股东现金净流量、长期资本收益率本年较上年变化值，对应公司财务舞弊特征体系中财务指标反映的盈利能力、现金流能力和发展能力，企业运营异常指标中的资金来源和关联交易。

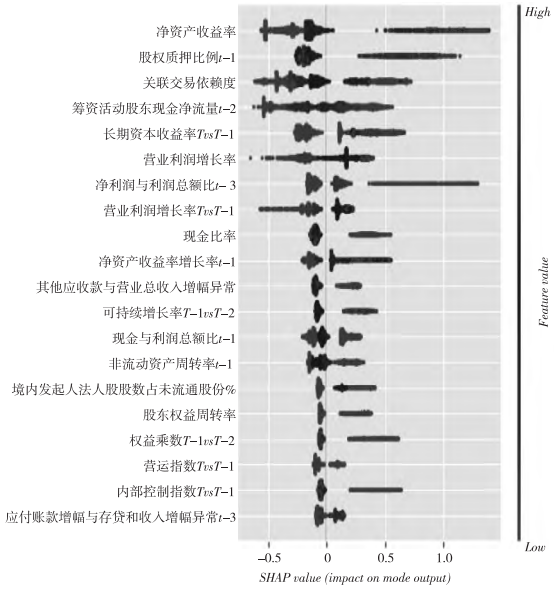

图 1 Xscore 模型特征的 Shap 影响

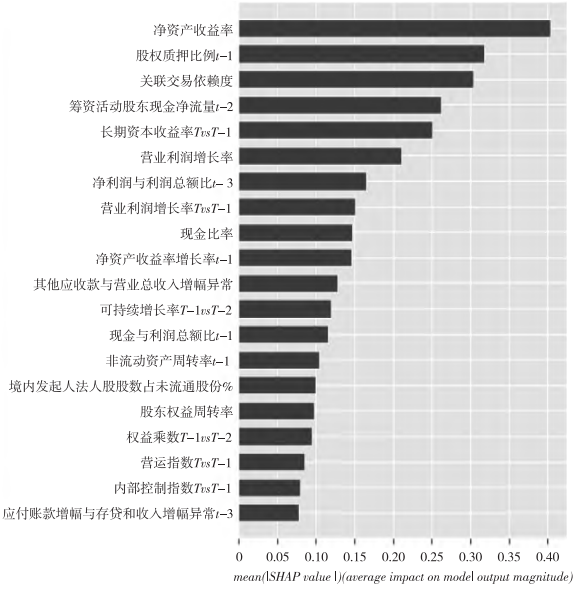

图 2 Xscore 模型特征 Shap 全局均值

（2）局部归因分析。SHAP 方法由于具有局部精确的性质，对于每个样本的每个特征的影响与常数项之和等价于模型的输出结果，因此可以在单个样本上求得每个特征对当前模型预测结果的影响。本文在此选取经典舞弊案例——康得新 2018 年年报，用 Xscore 模型对其当年的财务舞弊特征值进行预测，模型预测结果为舞弊，同时使用 SHAP 图进行可视化解释。如图 3 所示康得新 2018 年最终值 3.23。浅灰表示特征贡献是负数，深灰表示特征贡献是正数。图中深灰长度越长，预测结果为舞弊的概率越高。与全局归因结论一致，净资产收益率、股权质押比例、关联交易依赖度等特征变量对模型预测结果有重要影响。

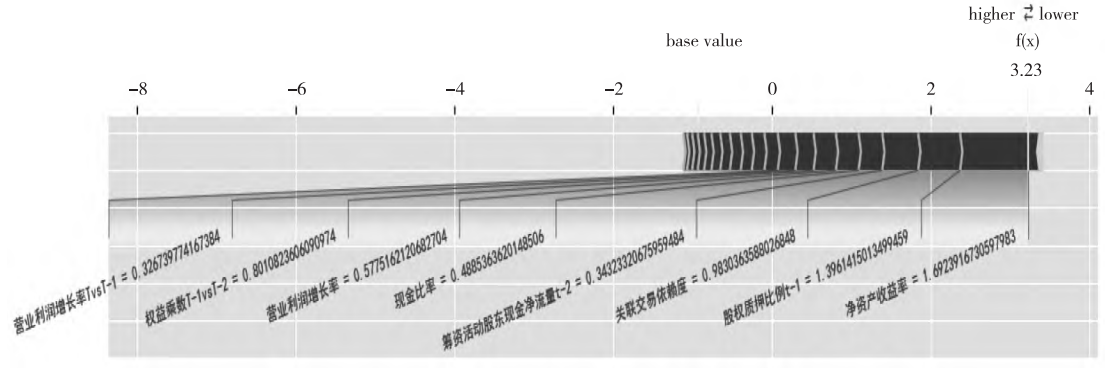

图 3　康得新 2018 年 Xscore 模型 SHAP 图

2. Xscore 评分卡应用分析

（1）Xscore 评分卡定义。为了提高 Xscore 模型的易用性，本文引入信用评分工具，通过统一的评分映射将 Xscore 模型的输出归一化。参考贷款机构对客户的信用评分工具的映射方法，本文提出评价上市公司财务舞弊风险的 Xscore 评分卡。Xscore 评分卡满足四条规则：属性点数为正；总分为正；分数某个值代表特定的舞弊风险；分数差别代表统一的舞弊风险变化。由于模型输出的概率值是负样本（舞弊样本）的概率，即分值越接近 1，等于负样本的概率越大，本文设计的 Xscore 评分卡公式如下所示：

$$Xscore = 600 + 50 \times \log_2\left(\frac{pred}{1 - pred} - lag\right) \tag{18}$$

其中，$pred$ 为集成模型的输出，lag 为基础分 600 对应似然概率阈值。本文假设当前期望模型的 20%分位点等于基础分，将模型在测试集上排在 20%分位的样本对应的概率值 0.86 作为 Xscore 评分卡的 lag。模型使用者也可以根据使用场景，定义不同的 lag。如果更关注精准率，且模型又有较强的负样本捕获能力时，则可以选择更靠前的 5%分位点的概率输出作为 Xscore 评分卡的 lag。Xscore 评分卡可以提升 Xscore 模型在上市公司财务舞弊识别预测中的易用性，即 Xscore 评分越高，财务舞弊风险就越高。

（2）Xscore 评分卡评价。关于 Xscore 评分卡的应用效果，本文统计分析 2019～2020 年上市公司样本中不同分数段的命中率和捕获率。命中率是指在该分数段上的样本中真正存在舞弊的样本的比例，反映预测的精准度。捕获率是指该分数段上的样本能覆盖全部舞弊样本的比例，反映预测的查全率。首先，Xscore 评分在前 5%分数段样本中命中率是 78%，这意味着如果只关注 Xscore 评分最高的前 5%的样本，命中真正舞弊样本的比率高达 78%。相较于总体样本中舞弊样本占比不到 12%的比例，Xscore 评分卡能高效地关注到财务舞弊

风险更高的公司，具有较高的舞弊预测精准度。如果关注前 10%预测评分的样本，命中率近 56%，关注前 20%评分段的样本，命中率 37%。其次，Xscore 评分在前 5%分数段样本中捕获率是 33%，前 10%分数段捕获率 48%，前 20%分数段捕获率 64%，这意味着如果对所有样本根据 Xscore 评分，只关注评分最高的前 20%的样本就可以捕获 64%的舞弊样本。可见，关注 Xscore 评分高分数段的样本对舞弊样本具有较强的正负样本排序能力，可以显著提高财务舞弊预测的精准度和查全度（见表 8）。

表 8　　Xscore 评分卡在时间外样本的表现　　单位：个，%

预测分数段	样本数	实际舞弊	命中率	捕获率
前 5%	213	166	77.93	33.33
前 10%	425	237	55.76	47.59
前 20%	850	318	37.41	63.86

（3）Xscore 评分卡应用。财务舞弊从发生到被监管机构认定为舞弊并公告需要一定的过程，舞弊的暴露具有滞后性。因此，本文从时间外样本筛选出 Xscore 评分处于前 2%分数段，虽未被监管机构披露为舞弊，但 Xscore 评分最高的 3 个样本，利用 SHAP 对这些样本中的 3 个样本实例的模型预测结果进行解释。SHAP 方法提供了单一模型预测的可解释性，可用于误差分析，找到对特定实例预测的解释（见表 9）。

表 9　　前 2% Xscore 分数段中未被认定为舞弊的样本　　单位：个

证券代码	证券简称	年份	证监会行业分类	Xscore
600767	运盛医疗	2019	软件和信息技术服务业	832
300353	东土科技	2019	计算机、通信和其他电子设备制造业	822
300264	佳创视讯	2020	软件和信息技术服务业	819

利用 SHAP 方法可以根据边际期望极大地增强集成模型的可解释性。图 4 到图 6 展示了前 2% Xscore 分数段中未被认定为舞弊的前 3 个样本（运盛医疗 2019 年、东土科技 2019 年、佳创视讯 2020 年）的模型预测 SHAP 图。SHAP 图展示的每个特征都各自有其贡献，将模型的预测结果从基本值分别推动到最终的取值，其中运盛医疗 2019 年的最终取值 3.25，东土科技 2019 年 3.12，佳创视讯 2020 年 3.07。浅灰表示该特征对终值的贡献是负数，深灰则表示贡献是正数，条形长度表示贡献大小。图中三个样本均为高分样本，较高的

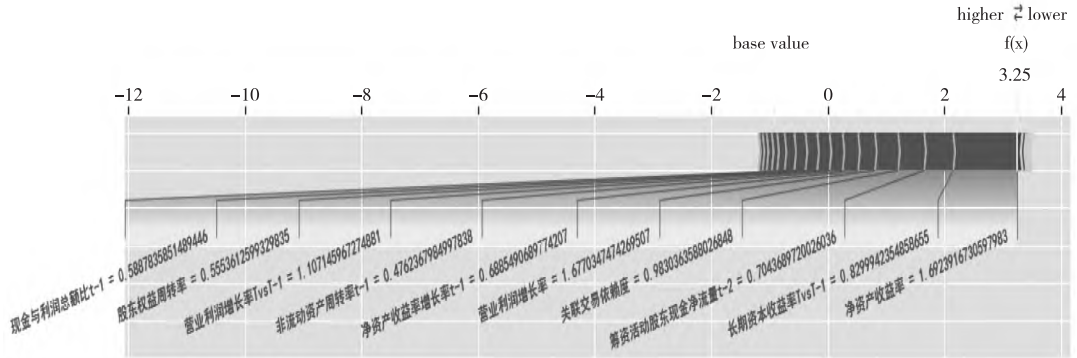

图 4　运盛医疗 2019 年 SHAP 图

最终值主要由深灰特征贡献。其中，运盛医疗 2019 年样本特征按贡献度从高到低排列分别是净资产收益率、长期资 本收益率较上年增加值、筹资活动股东现金净流量前年值；东土科技 2019 年样本特征按贡献度从高到低排列分别是净资产收益率、关联交易依赖度、股权质押比例上年数；佳创视讯 2020 年高贡献度的特征分别是净资产收益率、筹资活动股东现金净流量前年值、股权质押比例上年值。本文在特征重要性分析部分发现的净资产收益率、股权质押比例、关联交易依赖度等重要特征，在具体样本实例预测中起到了关键作用。值得注意的是，在不同的样本预测中，特征的贡献程度不一，我们可以从特征的贡献出发分析不同样本公司高风险的原因，从而指导财务舞弊人工分析核查工作中关注的侧重点。

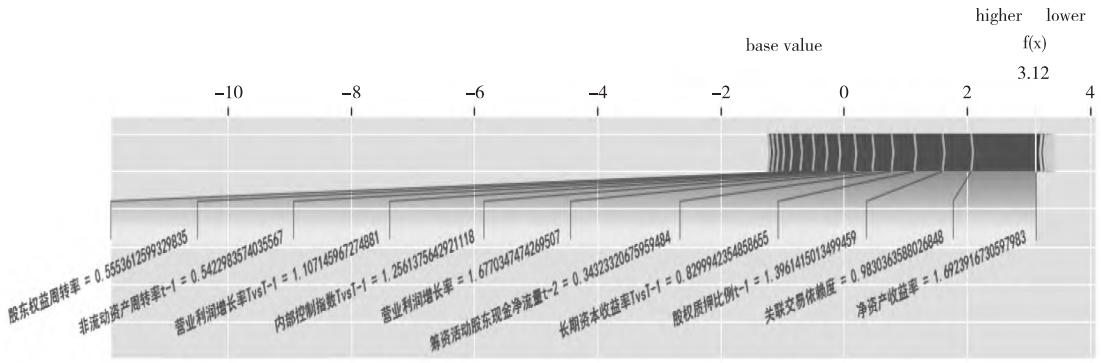

图 5　东土科技 2019 年 SHAP 图

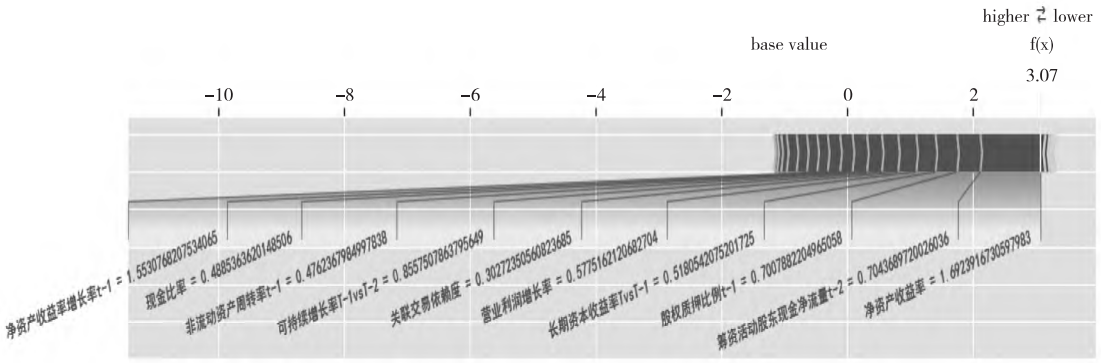

图 6　佳创视讯 2020 年 SHAP 图

五、结　　论

大数据与机器学习方法为上市公司财务舞弊预测提供了一个研究框架。本文试图通过基于 XGBoost 方法构建适用于中国上市公司特征的财务舞弊预测模型 Xscore。上市公司财务舞弊识别预测对于公司决策、投资者保护、政府监管都具有重要意义，然而由于上市公司财务舞弊手段隐蔽性强、财务数据与非财务数据复杂度高，长期以来缺乏简单实用且可靠的财务舞弊识别方法。本文按照大数据与机器学习模型研究逻辑，以 2000～2020 年 A 股上市公司为研究样本构造训练集、测试集和时间外样本集（预测集），从公司治理、财会监督、财务指标、企业运营四个方面选取刻画企业画像的特征指标体系，通过特征聚合、特征组合、

特征离散等特征工程方法扩展、筛选和降维确定最优的 27 个特征指标，利用 XGBoost 算法进行模型训练，最终构建 Xscore 模型。经过模型比较、分析与应用，主要得到以下结论：第一，Xscore 模型显著优于传统财务舞弊预测模型 Cscore 和 Fscore。Xscore 具有更好的预测精准率、召回率、正负样本区分度和稳定性，在时间外样本中的预测总体召回率 79%，精准率 42%，AUC 达 0.90，区分度 KS 达 0.65，模型稳定性 PSI 为 0.09。第二，基于 SHAP 方法对 Xscore 模型特征贡献度进行可视化分析，通过全局归因和局部归因提升 Xscore 模型的可解释性，使得 Xscore 模型既具有优秀的预测能力也具备可解释性。第三，通过定义和应用 Xscore 评分卡，进一步提升 Xscore 模型的易用性，结合 SHAP 方法使得 Xscore 模型预测结果更为直观，实证检验前 5%高分段样本的舞弊样本命中率高达 78%，前 20%的高分段能捕获总体样本中 64%的舞弊样本，这意味着拒绝 20%的 Xscore 高分段样本就可以过滤掉 64%的舞弊样本。

本文的研究结论表明机器学习方法在上市公司财务舞弊预测领域具有一定的应用价值。我们认为大数据与机器学习在上市公司财务舞弊预测中将会产生积极的影响，将推动上市公司财务信息披露质量和防范上市公司财务舞弊发挥重要作用。同时，机器学习方法可能对传统会计理论提出重大挑战，尤其是关于会计信息含量的概念，机器学习可以分析和挖掘高维且非线性的复杂数据（包括财务数据与非财务数据），这是否意味着通过机器学习可以显著提升会计信息的微观信息含量和宏观信息含量？从而推动会计信息含量的实证研究和重新评估？总的来说，基于大数据与机器学习模型的预测性将显著优于传统的计量实证模型，这对未来会计学、金融学与经济学理论的发展需要进一步的思考和探索。

参 考 文 献

[1] Altman E. I., 1968, *Discriminant Analysis and the Prediction of Corporate Bankruptcy* [J], Journal of Finance, 23 (4), 589～609.

[2] Bao Y., Ke B., Li B., Yu J., Zhang J., 2020, *Detecting Accounting Fraud in Publicly Traded U. S. Firms Using a Machine Learning Approach* [J], Journal of Accounting Research, 58 (1), 199～235.

[3] Beaver W. H., 1966, *Financial Ratios As Predictors of Failure* [J], Journal of Accounting Research, 4, 71～111.

[4] Beneish M., 1999, *The Detection of Earnings Manipulation* [J], Financial Analysts Journal, 55 (5), 24～36.

[5] Breunig M. M., Kriegel H. P., Ng R. T., Sander J., 2000, *LOF: Identifying Density-Based Local Outliers* [C], Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data, 93～104.

[6] Brown N. C., Crowley R. M., Elliott W. B., 2020, *What Are You Saying? Using topic to Detect Financial Misreporting* [J], Journal of Accounting Research, 58 (1), 237～291.

[7] Cecchini M., Aytug H., Koehler G. J., Pathak P., 2010, *Detecting Management Fraud in Public Companies* [J], Management Science, 56 (7), 1146～1160.

[8] Dechow P. M., Ge W., Larson C. R., Sloan R. G., 2011, *Predicting Material Accounting Misstatements* [J], Contemporary Accounting Research, 28 (1), 17～82.

[9] Deshmukh A., Millet I., 2011, *An Analytic Hierarchy Process Approach To Assessing The Risk Of Management Fraud* [J], Journal of Applied Business Research, 15 (1), 87.

[10] Ding Z., Fei, M., 2013, *An Anomaly Detection Approach Based on Isolation Forest Algorithm for Streaming Data using Sliding Window* [J], 3rd IFAC Conference on Intelligent Control and Automation

Science，46 (20)，12～17.

[11] Zolotareva E.，2021，*Aiding Long-Term Investment Decisions with XGBoost Machine Learning Model* [J/OL]，arXiv：2104.09341.

[12] Hansen J. V.，Mcdonald J. B.，Messier W. F. J.，Bell T. B.，1996，*A Generalized Qualitative-Response Model and the Analysis of Management Fraud* [J]，Management Science，42 (7)，1022～1032.

[13] Kinnunen J.，Koskela M.，2003，*Who Is Miss World in Cosmetic Earnings Management? —A Cross-National Comparison of Small Upward Rounding of Net Income Numbers Among Eighteen Countries* [J]，Journal of International Accounting Research，2 (1)，39～68.

[14] Kotsiantis S. B.，Zaharakis I. D.，Pintelas P. E.，2006，*Machine Learning：A Review of Classification and Combining Techniques* [J]，Artificial Intelligence Review，26 (3)，159～190.

[15] Loebbeck J. K.，Eining M. M.，Willingham J. J.，1989，*Auditors' Experience with Material Irregularities：Frequency，Nature，and Detectability* [J]，Auditing：A Journal of Practice and Theory，9，1～28.

[16] Ofori E.，2016，*Detecting Corporate Financial Fraud Using Modified Altman Z-Score and Beneish M-Score. The Case of Enron Corp* [J]，Research Journal of Finance and Accounting，7 (4)，59～65.

[17] Ravisankar P.，Ravi V.，Raghava Rao，G.，Bose I.，2011，*Detection of Financial Statement Fraud and Feature Selection Using Data Mining Techniques* [J]，Decision Support Systems，50 (2)，491～500.

[18] Zhang Y.，Chen L.，2021，*A Study on Forecasting the Default Risk of Bond Based on XGboost Algorithm and Over-Sampling Method* [J]，Theoretical Economics Letters，11 (2)，258～267.

[19] 陈关亭：《我国上市公司财务报告舞弊因素的实证分析》[J]，《审计研究》2007 年第 5 期。

[20] 陈国欣、吕占甲、何峰：《财务报告舞弊识别的实证研究——基于中国上市公司经验数据》[J]，《审计研究》2007 年第 3 期。

[21] 陈慧璇、朱君：《我国上市公司财务报告舞弊特征分析》[J]，《税务与经济》2013 年第 2 期。

[22] 陈玟伶：《企业内部控制与财务舞弊治理的研究》[D]，暨南大学硕士学位论文，2010 年。

[23] 洪荭、胡华夏、郭春飞：《基于 GONE 理论的上市公司财务报告舞弊识别研究》[J]，《会计研究》2012 年第 8 期。

[24] 黄世忠、叶钦华、徐珊、叶凡：《2010～2019 年中国上市公司财务舞弊分析》[J]，《财会月刊》2020 年第 14 期。

[25] 李双杰、陈星星：《基于 BP 神经网络模型与 DEA 模型的中国上市公司利润操纵研究》[J]，《数理统计与管理》2013 年第 3 期。

[26] 刘君、王理平：《基于概率神经网络的财务舞弊识别模型》[J]，《哈尔滨商业大学学报（社会科学版）》2006 年第 3 期。

[27] 刘立国、杜莹：《公司治理与会计信息质量关系的实证研究》[J]，《会计研究》2003 年第 2 期。

[28] 卢馨、李慧敏、陈烁辉：《高管背景特征与财务舞弊行为的研究——基于中国上市公司的经验数据》[J]，《审计与经济研究》2015 年第 6 期。

[29] 钱苹、罗玫：《中国上市公司财务造假预测模型》[J]，《会计研究》2015 年第 7 期。

[30] 孟生旺、王海淘：《基于机器学习算法的个体索赔准备金评估模型》[J]，《保险研究》2019 年第 9 期。

[31] 任朝阳、李清：《上市公司会计舞弊风险指数影响因素研究》[J]，《当代经济科学》2017 年第 5 期。

[32] 吴革、叶陈刚：《财务报告舞弊的特征指标研究：来自 A 股上市公司的经验数据》[J]，《审计研究》2008 年第 6 期。

[33] 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 年第 10 期。

[34] 岳娇：《奔福德定律与上市公司财务舞弊识别研究》[D]，浙江大学硕士学位论文，2020 年。

[35] 岳殿民：《中国上市公司会计舞弊模式特征及识别研究》[D]，天津财经大学博士学位论文，2008 年。

[36] 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 年第 5 期。

[37] 章美珍：《财务报告舞弊端倪甄别及治理对策》[J]，《当代财经》2002 年第 5 期。

[38] 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 年 9 期。

[39] 曾月明、许素：《IPO 会计舞弊影响因素研究》[J]，《管理学报》2019 年第 10 期。

[40] 周守华、胡为民、林斌、刘春丽：《2012 年中国上市公司内部控制研究》[J]，《会计研究》2013 年第 7 期。

Research on Financial Frauds Prediction Model of Chinese Public Companies with XGBoost

Zhou Weihua　Zhai Xiaofeng　Tan Haowei

(Chinese Academy of Fiscal Sciences)

Research Objectives: To explore how to use big data and machine learning methods to analyze and mine financial and non-financial data of listed companies, and apply them to the identification and prediction of financial fraud of listed companies. **Research Methods**: A machine learning method-based financial fraud prediction model Xscore is proposed to predict financial fraud of listed companies. **Research Findings**: Xscore model can improve the accuracy of model prediction, and outperforms Fscore model and Cscore model in terms of accuracy, recall, AUC index, KS value and capture rate, which is more suitable for financial fraud prediction of listed companies in China. **Research Innovations**: Based on the data set of Chinese listed companies from 2000 to 2020 as the observation sample, we solve the gray sample problem commonly faced when machine learning is applied to financial fraud identification research by Benford's law, LOF local anomaly method, IF unsupervised learning method, and select feature variables with both domain characteristics and statistical features; for the first time, we apply XGBoost integrated learning method The XGBoost integrated learning method is applied to the analysis of financial fraud prediction of listed companies for the first time, which effectively improves the accuracy of financial fraud of listed companies. **Research Value**: This paper introduces XGBoost integrated learning method into the field of financial fraud identification of listed companies, which helps to promote the research and application of artificial intelligence and machine learning in accounting, and provides reference for promoting the disclosure of high-quality financial information of listed companies and maintaining the order of capital market.

Key Words: XGBoost; Machine Learning; Financial Frauds; Prediction Model

JEL Classification: G11

(责任编辑：白延涛)