基于机器学习的公司违规预测研究

李 莹 曲晓辉

摘要：公司违规研究一般采用传统的线性回归模型处理历史数据，本文则构建了基于机器学习算法的公司违规预测模型并实施检验。研究发现：(1) 通过比对分析可知，广为使用的线性回归模型不能充分挖掘数据信息并进行有效预测，机器学习的树模型（Random Forest和GBDT）和神经网络模型（RNN和LSTM）的预测效果更优；(2) 采用SHAP方法探寻公司违规的重要预警因子及其贡献度，发现公司治理相关变量对违规预警具有重要贡献，传统线性回归模型过分强调财务数据的预警能力，弱化了公司治理因子对违规预警的贡献；(3) 进一步采用SHAP方法将Random Forest和GBDT模型的运作过程和贡献分布可视化，挖掘重要因子对公司未来违规的影响机理，并计算主要影响因子的警惕阈值。本研究实现了公司违规行为预测研究方法的创新，对投资者和监管机构的决策优化具有借鉴意义。

关键词：违规预测；机器学习；传统预测模型；因子重要性；可视化
中图分类号：F275.5　**文献标志码：**A　**文章编号：**2095-8838（2022）04-0054-13

一、引言

公司违规是资本市场研究的一大重要问题。监管机构、投资者、审计师和分析师如何更早地发现公司违规行为，并进行重点关注或跟踪调查，对资本市场的健康发展至关重要。然而，现有传统预测方法大多基于线性回归模型，从主观认知或局部经验出发确定影响要素，很难准确预警公司违规行为，事前发现和事中监管面临巨大挑战。因此，如何构建具有事前提示功能的公司违规预警模型，是学术界与实务界的重要研究课题。

近年来，人工智能技术的发展为公司违规预警研究提供了新的突破口。机器学习作为人工智能的代表性技术，是借助高级的数学方法和新型的算法从大数据中寻找出有用的数据并进行挖掘的技术。在经管领域，机器学习能够为统计预测问题提供高维模型和大量候选模型，并能够进行有效算法的筛选（Gu等,2020）。机器学习在数据结构、数据交互、模型结构和因子识别等四方面均显著优于基于计量经济学的传统预测模型（主要是线性回归模型）。具体而言，(1)在数据结构方面，经管领域数据具有非结构化、高维度和稀疏的特点，而传统线性回归模型复杂度低，无法有效捕捉高维信息，样本外预测能力差。对比之下，机器学习具备降维及变量选择能力，能够减少数

收稿日期：2022-05-05
基金项目：教育部人文社会科学重点研究基地重大项目（16JJD790035）；江苏高校哲学社会科学研究项目（2021SJA0356）；江苏高校优势学科建设工程资助项目（PAPD）；深圳市人文社会科学重点研究基地哈尔滨工业大学（深圳）大数据会计与决策研究中心基金（KP191001）
作者简介：李　莹，南京审计大学社会审计学院讲师；
曲晓辉，哈尔滨工业大学（深圳）经济管理学院/厦门大学会计发展研究中心教授，博士生导师。

据维度并压缩预测变量之间的冗余信息，使预测结果更稳定。（2）在数据交互方面，传统线性回归模型无法确定数据之间复杂的交互作用，很难进行有效的预测。相比之下，机器学习可以检测和利用数据中的高阶交互作用进行数据分析，进而提高预测模型的准确性。（3）在模型结构方面，传统线性回归模型具有明确的形式和参数，侧重因果分析，模型的表达能力弱。相比之下，机器学习模型具有复杂的模型结构和较强的非线性模拟能力，可逼近任意函数，模型表达能力强，预测精度更高。（4）在因子识别方面，传统线性回归模型蕴含多重假定，容易造成检验结论的不稳定和不准确。相比之下，机器学习放宽了模型假定，能够在数以百计的因子中，通过复杂的模型结构识别因子重要性差异。

总的来说，传统线性回归模型难以提供高维数据处理和函数形式选择方面的建议（De Prado，2018），造成基于传统预测模型的公司违规预警模型的样本外预测精度普遍不高。目前，学者们基于传统方法构建的违规预测模型，往往从主观认知或局部经验出发确定公司违规的影响因素（单华军，2010；Chen 等，2016）。为了追逐研究变量在原假设下统计意义上“显著”的低P-value值，造成一些被证明的公司违规影响因素在实际中可能并非十分重要。已有研究表明，董事会结构、内部控制水平、制度环境、盈余质量、业绩表现等均与公司违规行为密切相关（Beneish，1997；张翼和马光，2005；Johnstone等，2011；冯旭南和陈工孟，2011）。但哪些因子能够在实践中为公司违规预测提供显著有效的预警信号，却缺乏深入的研究和探讨。基于此，本文在构建基于机器学习方法的预警模型基础上，进一步检验模型的因子重要性，并对模型和数据进行可视化，从而加强人们对公司违规行为的辨识能力以及对公司违规行为的理解。

本文主要有三方面的贡献。第一，随着人工智能的发展，机器学习方法已经在许多领域取得令人满意的成果。然而在资本市场领域，学者们仍然较多基于研究样本推导公司违规的因果关系，如许多文献采用线性回归方法构建预测模型（吴世农和卢贤义，2001；Dechow等，2011；洪荭等，2012）。本文创新性地融合人工智能与公司违规研究，将机器学习方法应用于预测公司违规，扩展了公司违规研究数据分析与建模工具箱，强调了机器学习在财务研究中的重要价值，为此类问题在大数据环境下的发展提供了新方法和新思路。第二，本文通过对基于机器学习的公司违规模型的特征因子进行重要性分析，发现公司违规记录、盈利能力及外部治理水平是公司违规发生的重要预警指标。相比于内部公司治理，外部公司治理为公司违规预测模型提供了显著重要且易于取得的有用信息，该结论为以往未充分考虑公司外部治理特征的预测模型提供了补充证据。第三，机器学习复杂的模型使人们难以理解其性质，为此本文采用SHAP方法将模型运作过程可视化，打开机器学习“黑箱”，深入挖掘和分析公司违规预测模型中因子的影响机理及有效作用范围，打破了以往机器学习模型在因子效度检验方面缺乏经济依据和解释力的局面，提供了数据间经济关系解释的可能性。

本文后续部分主要包括以下几部分：首先，介绍经典的传统预测模型和机器学习算法；其次，研究设计中描述样本与数据来源以及特征选择依据，并对模型性能进行了度量；再次，分别从模型效果和因子重要性两个方面对模型展开分析；然后，为了在一定程度上打开机器学习的黑箱，本文进一步将模型运作过程可视化并计算影响阈值；最后取得结论。

二、传统预测模型与机器学习方法

（一）传统预测方法中常用模型：逻辑回归Logistic和岭回归Ridge

传统预测方法如逻辑回归的本质是线性分类器，其优点在于形式简单、易于建模。但由于线性模型在处理高维数据时表现欠佳，本文将其作为参考用来比较和强调更复杂方法的优势。具体地，本文在此主要分析常用的传统预测模型：逻辑回归Logistic和岭回归Ridge。

逻辑回归Logistic利用原始预测变量x、参数向量ω和偏置项b的广义线性函数进行估计，函数一般描述形式为：

$$g(x;\omega)=\frac{1}{1+e^{-(\omega^{T}x+b)}} \quad (1)$$

逻辑回归模型的目标函数L为交叉熵损失函数，公式为：

$$L(\omega)=-(y\ln g(x;\omega)+(1-y)\ln(1-g(x;\omega))) \quad (2)$$

$L(\omega)$目标函数可以通过回避模型优化过程中的复杂非线性关系，从而加速模型构建过程。

岭回归Ridge的目标函数是在逻辑回归的基础上加入L_2的正则项。加入正则项是防止模型过拟合的一种通用解决思路，即在保证目标函数最小化的同时，尽可能采用简单的模型以提高预测精度。Ridge的目标函数公式为：

$$L(\omega;\lambda) = -(y\ln g(x;\omega) + (1-y)\ln(1-g(x;\omega))) + \frac{1}{2}\lambda\omega^T\omega \quad (3)$$

（二）机器学习方法

1. 树模型：随机森林 Random Forest 和梯度提升树 GBDT

树模型作为机器学习的主流算法之一，更擅长处理对交互关系没有先验假设时的计算问题。决策树模型不同于传统线性模型的逻辑，属于完全非参数模型：

$$g(x;\theta,K,L) = \theta_k 1_{\{x\in C_k(L)\}} \quad (4)$$

其中，K 为决策树的叶子节点数量，L 为决策树的深度，$C_k(L)$ 为被划分在叶子节点 k 下的输入特征 x 的集合，θ_k 为节点 k 所对应的类别。

决策树的目标是寻找用来划分结果的最佳分割点，目标函数计算公式如下：

$$E(k) = -\sum_{m=1}^{M} P_m(k) \times \ln(P_m(k)) \quad (5)$$

$$L(K) = \sum_{k=1}^{k} |C_k(L)| \times E(k) \quad (6)$$

其中，M 代表类别数量，$P_m(k)$ 代表当前节点 k 中属于 m 类的比例。

决策树的优势在于可以同时处理多种特征，并构建潜在的非线性关系。但也正是由于决策树能够对特征数据及其函数关系进行充分学习，使其更易出现过拟合。因此，本文采用两个集成类树，随机森林 Random Forest 和梯度提升树 GBDT，通过将不同的树合并为一棵树来缓解过拟合问题。集成学习包括对多个弱学习器独立学习的 Bagging 学习法，以及将弱学习器提升为强学习器的 Boosting 学习法。随机森林 Random Forest 是 Bagging 方法的一个扩展变体，即在以决策树为基分类器构建 Bagging 集成的基础上，进一步在决策树的训练过程中引入随机属性选择（周志华，2015）。梯度提升树 GBDT（Gradient Boosting Decision Tree）是 Boosting 模型，采用分类与回归树（Classification and Regression Trees，CART）作为基分类器，通过每轮迭代产生一个弱分类器来拟合上一轮分类器的残差，从而得到偏差较小的预测结果。

2. 神经网络：循环神经网络 RNN 和长短期记忆网络 LSTM

在机器学习算法中，基于神经网络的深度学习得到了广泛关注与应用。财务领域中经济个体的变量在不同时点的取值并不具备完全的独立性，数据的时序性难以被线性模型、树模型和传统神经网络模型所捕获，此时，循环神经网络（Recurrent Neural Networks, RNN）的优势就凸显出来了。与传统神经网络不同，循环神经网络 RNN 主要是处理和预测序列数据，其链式属性能够将数据按时间轴展开，记录并提取长时期的历史信息，模型如下：

$$g(x^t;\ \omega_h,\omega_i,\omega_o) = Softmax(\omega_o \times h^t(x^t;\omega_h,\omega_i)) \quad (7)$$

其中，ω_h，ω_i，ω_o 为 RNN 单元输入特征、隐藏特征、输入的对应的权重参数，h^t 为 t 时刻 RNN 中传递的隐藏特征，公式如下：

$$h^t(x^t;\omega_h,\omega_i) = Relu(\omega_i \times x^t + \omega_h \times h^{t-1}) \quad (8)$$

Relu 为激活函数：

$$Relu(x) = \begin{cases} 0 & if\ x<0 \\ x & otherwise \end{cases} \quad (9)$$

RNN 使用带有 L2 正则化项的目标函数：

$$L(\omega;\lambda) = -(y\ln g(x;\omega) + (1-y)\ln(1-g(x;\omega))) + \frac{1}{2}\lambda\omega^T\omega \quad (10)$$

从理论上说，RNN 能够对任意长度的序列数据进行建模，但在实际应用中很难学习到长时间信息的关联。这是因为 RNN 为了降低模型的复杂度，一般假设当前状态仅与前几个时点的历史状态相关，随着时间序列期间的增加，距离当前时点较远时点的信息很难传递到当前时点。而长短期记忆网络 LSTM（Long Short Term Memory），一种特殊结构的 RNN，则能够记录长时期的历史信息，可以解决长期依赖问题。LSTM 的模型如下：

$$g(x^t;\ \omega_f,\omega_i,\omega_o,\omega_c) = Softmax(\omega_o \times h^t(x^t;\omega_f,\omega_i,\omega_c)) \quad (11)$$

其中，ω_f，ω_i，ω_o，ω_c 分别为 LSTM 单元中的遗忘门、输入门、输出门、细胞状态的对应权重参数。LSTM 的目标函数同模型（10）。

三、研究设计

（一）样本与数据来源

西方的理论与模型有时并不能完美契合我国情景（钱苹和罗玫，2015）。因此，基于我国数据与情景构建公司违规预测模型十分必要。本文样本区间选择 2007~2017。公司违规数据主要来自 CSMAR 数据库，同时，为了保证违规数据的完整性和全面性，本文还采用 Wind 数据库予以补充。由于违规行为从发生到被发现并公告的平均时间约为两年（Dyck 等，2010），因此本文将训练集和测试集的样本区间设置一年间隔（实际年报公告的发布时间间隔为两年），以 2007~2015 年的样本为训练集，2017 年的样本为测

表1　违规公司年度统计

年份	违规公司合计	上市公司合计	违规公司比例
2007	143	1 005	14.23%
2008	183	1 076	17.01%
2009	233	1 204	19.35%
2010	252	1 480	17.03%
2011	366	1 778	20.58%
2012	418	1 823	22.93%
2013	335	1 607	20.85%
2014	301	1 701	17.70%
2015	384	1 973	19.46%
2016	304	2 133	14.25%
2017	234	2 314	10.11%
合计	3 153	18 094	17.43%

试集。

本文尝试建立因违规行为被监管部门公开批评、谴责或处罚的公司预测模型。具体地，将虚构利润、虚列资产、虚假记载（误导性陈述）、推迟披露、重大遗漏、披露不实、欺诈上市、出资违规、擅自改变资金用途、占用公司资产、内幕交易、违规买卖股票、操纵股价、违规担保、一般会计处理不当等行为视为公司违规行为。表1列示了违规公司的年度分布情况。可以看到，在本文2007~2017年样本期间，共有3 153个违规样本，占全样本的17.43%。这里，不重复计算同一公司同年发生的违规行为。

（二）特征因子选择

虽然已有大量文献研究公司违规的影响因素，但这些研究样本区间、研究对象、模型设计等均存在较大差异，致使人们难以从全局角度科学辨别和比较公司违规的影响因子。同时，现有研究结论大多通过计量经济学的实证研究而得，由于对低 P-value 值的过度追求或统计检验手段不够严谨，导致研究成果不能为监管部门的预防和治理工作提供足够坚实可靠的科学证据。

因此，本文借鉴舞弊四因子理论，从贪婪（Greed）、机会（Opportunity）、需要（Need）和暴露（Exposure）四个方面选取特征因子。其中，贪婪和需要属于个体风险因子，即从个人和组织角度阐释舞弊行为发生的因素；机会和暴露属于一般风险因子，即从内部和外部环境角度阐释舞弊行为发生的因素。具体来说，个体风险因子与违规动机相关。上市公司财务报表是企业诸多契约的核心依据，上市公司高管薪酬、发行新股或债务融资、盈余质量、业绩表现等，均与报表数字息息相关（Beneish，1997；张翼和马光，2005；陈丽英等，2012）。一般风险因子则与组织环境有关，新兴资本市场法律法规不够健全，给“内部人”违规提供了可乘之机。因此，许多学者致力于探索公司内外部治理特征与公司违规行为之间的关系。已有研究表明，董事会结构（Beasley，1996；陈国进等，2005；蔡志岳和吴世农，2007；冯旭南和陈工孟，2011）和内部控制（单华军，2010；Johnstone等，2011）等公司内部治理对公司违规行为具有显著影响。公司外部治理环境如市场竞争（滕飞等，2016）、制度环境（Kedia和Rajgopal，2011；刘英明，2012）、法治水平（徐尧等，2017；曹春方等，2017）、审计质量（王霞和张为国，2005）、分析师关注（Dyck等，2010；Chen等，2016）、股票发行（洪荭等，2012）等与公司违规行为关系密切。综上，本文选择了47个特征因子，如表2所示。

（三）模型运用

在模型训练过程中，需要解决两个问题：第一，类别不平衡问题。在本文样本中，公司违规的平均数为0.1743，意味训练集中违规公司（正类样例）占全样本的17.43%，没有违规的公司（反类样例）占全样本的82.67%。因此本文存在类别不平衡（class-imbalance）问题。类别不平衡问题的解决方案主要有两个：欠采样（Rescaling）和过采样（Oversampling）。欠采样是通过丢弃部分反类样例，使得样本中正类、反类例数目接近，再进行学习。过采样是通过增加正类样例，使得样本中正类、反类例数目接近，再进行学习。考虑到训练样本数量有限，本文对训练集采用过采样的方法来缓解类别不均衡问题。第二，避免过拟合是设计分类器过程中的一个核心任务。过拟合是指学习时选择的模型所包含的参数过多，以致于出现这一模型对已知数据预测得很好，但对未知数据预测很差的现象（李航，2012）。对于过拟合问题，本文通过在模型中加入正则化项（Regularizer）或罚项（Penalty Term）来缓解过拟合问题。

（四）性能度量

性能度量（Performance Measure）是对学习方法的泛化能力（学习到的模型对未知数据的预测能力）进行评估。在对比不同模型能力时，使用不同的性能度量指标，会得到不同的评价结果。为了保证结论的稳健性和可靠性，本文采用多个性能度量指标。

借鉴Larcker和Zakolyukina（2012），本文采用了受试者操作特征ROC（Receiver Operating Characteristics）曲线下面积AUC（Area under the Curve）作为本文的性能度量

表 2 特征因子及其具体解释

	变量名称	变量解释
个体风险因子	cash	现金及交易性金融资产
	Receivables	应收账款和应收票据之和
	Inventories	存货
	ShortInvest	短期投资，包括交易性金融资产和衍生金融资产
	PPE	固定资产
	LongInvest	长期投资，包括可供出售金融资产、持有至到期投资、长期应收款和长期股权投资
	assets	总资产
	DebtLiability	流动负债，等于短期票据和一年内到期的长期借款之和
	IncomeTaxesPay	应付所得税
	CurrentLiability	流动负债合计
	LongDebt	长期负债合计
	Liability	总负债
	CommonEquity	实收资本
	PreferStock	优先股
	RetainEarnings	留存收益
	Sales	主营业务收入
	COGS	营业成本
	Interest	利息及相关费用
	IncomeTaxes	所得税费用
	NetIncome	净利润
	Clsprc	年报发布日股票收盘价
	Yclsprc	年末股票收盘价
	issue	是否发行新股
一般风险因子	first	第一大股东持股数量
	ratio	第二大到第五大股东持股数量之和除以第一大股东持股数量
	nationratio	公司国有股所占比例
	TopExec	公司管理层持股比例
	parttime	哑变量，公司董事长是否同时任 CEO
	BoardSize	公司董事会人数
	Independent	独立董事人数
	soe	哑变量，产权性质，国有企业为 1，否则为 0
	FGscore	市场化总指数评分
	factorscore	要素市场的发育程度评分
	productscore	产品市场的发育程度评分
	govscore	政府与市场的关系评分
	legalscore	市场中介组织的发育和法律制度环境评分
	NSOEeconomics	非国有经济的发展评分
	IsDisclosingEvaRep	披露内控评价报告为 1，否则为 0
	IsValid	哑变量，内部控制有效为 1，否则为 0
	IsDeficiency	哑变量，内部控制存在缺陷为 1，否则为 0
	AnaAttention	分析师关注，当年跟踪该公司的分析师（团队）的数量
	Big4	哑变量，审计的事务所是否为国际“四大”会计师事务所，是为 1，否则为 0
	age	上市年龄
	v_lag	哑变量，上年是否违规
	male_m	CEO 性别，若为男性取值为 1，若为女性，取值为 0
	degree_m	CEO 的学历水平，中专及中专以下为 1、大专为 2、本科为 3、硕士为 4、博士为 5
	oversea_m	CEO 的海外背景。若有海外求学或任职经历，取值为 1，否则为 0

指标。ROC源于二战中敌机检测的雷达信号分析技术，经过心理学和医学检测领域的应用，后被引入到机器学习领域（Spackman，1989）。ROC曲线是研究学习器泛化性能的有力指标。ROC曲线的纵轴是“真正例率（True Positive Rate）”，横轴为“假正例率（False Positive Rate）”。在进行学习器的比较时，采用ROC曲线下各部分的面积之和更为合理，面积越大，分类器效果越好。这个面积被称为AUC。

本文还采用了信息检索评价指标NDCG@k（Normalized Discounted Cumulative Gain at the position k），即将预测模型的效果看成一个排序质量问题，将每个返回的推荐结果相关性的分值累加后作为整个推荐列表的得分然后对排名靠后的推荐结果进行“打折处理”。该评价方法包含两个假设：（1）相比于没有违规的公司样本，违规样本的分数更高；（2）当违规公司样本在排序列表中的排序靠前时，将会被赋予更高的权重，分数更高。因此，推荐结果的相关性越高，在推荐列表中的排序越靠前，推荐效果越好，DCG的值越大。具体来说，$DCG@k = \sum_{i=1}^{k} \frac{2^{rel}-1}{log_2(i+1)}$，其中，当第i个观测值在排序列表中是真违规，$rel_i$被赋值为1，否则为0；k表示预测模型在测试集中预测的违规概率最高的k个观测值。本文将k定义为本文样本中每年违规公司数量。由于DCG在不同的排序列表之间很难进行横向对比，为了对测试集中的公司进行评估，不同公司的排序列表评估分数需要进行归一化处理，即NDCG（Normalized Discounted Cumulative Gain）。当所有的违规公司样本都在模型返回结果排序中的靠前位置，则此序列的DCG为Ideal DCG（IDCG）。NDCG@k则被定义为$NDCG@k = \frac{DCG@k}{IDCG@k}$，由于DCG的值介于（0，IDCG]，因此NDCG的值介于（0，1]。

最后，本文还采用F1-Score作为性能度量指标之一。F1-Score常被用作分类任务的最终测评方法，等于精确率（Precision）和召回率（Recall）的调和平均数，最大为1，最小为0。对于二分类问题，我们一般按学习器在测试数据集上预测的正确与否划分出四种情况：真正例TP（True Positive），即将正例预测为正例；假正例FP（False Positive），即将反例预测为正例；真反例TN（True Negative），即将反例预测为反例；假反例FN（False Negative），即将正例预测为反例。基于此，精确率 $Precision = \frac{TP}{TP+FP}$；召回率 $Recall = \frac{TP}{TP+FN}$；$F1\text{-}Score = 2\times \frac{Precision\times Recall}{Precision+Recall}$。

综上，为了合理评估公司违规预测模型的泛化能力，本文分别采用AUC、NDCG@k和F1-Score三个指标进行模型评估。

四、结果与分析

基于前述样本特征选择和模型构建，本部分首先报告了传统预测模型和机器学习模型的预测结果并展开分析；同时，为了更好地比较模型的运作过程，本部分采用加性解释模型SHAP（SHapley Additive exPlanation）计算基于机器学习模型和传统预测方法线性回归的因子重要性数值，以明确违规预测的重要影响因子，并试图揭示造成模型性能差异的成因。

（一）模型预测结果

表3报告了传统预测模型（Logistic和Ridge）、树模型（Random Forest和GBDT）和神经网络（RNN和LSTM）构建公司违规预测模型的性能度量结果。

由表3可知，采用传统预测方法的线性回归模型Logistic和Ridge的AUC为0.54，约等于随机猜测（AUC=0.5）。相比之下，树模型Random Forest和GBDT的AUC显著高于前者，分别为0.71和0.72；神经网络模型RNN和LSTM的AUC值为0.74，比随机猜测高出0.24。在用NDCG@k作为性能度量指标时，Logistic和Ridge的预测效果较差；树模型和神经网络模型的预测效果显著优于前者，是线性模型预测精度的两倍有余；神经网络模型表现最好。在F1-Score度量方式下，线性模型Logistic和Ridge的性能依然较差；树模型Random Forest和GBDT的预测效果最好，分别达到了0.45和0.46，神经网络模型的性能表现依然优秀（为0.43和0.42）。总的来说，基于机器学习方法构建的公司违规预测模型效果更好，体现了机器学习方法在财务会计领域预测问题中的优越性和稳定性。

（二）特征因子重要性分析

机器学习模型的一个重要局限在于其更偏向于结果导向。在研究公司违规预测模型时，模型性能固然重要，但理解影响模型精度的决定因素更加重要。通过对模型特征因子重要性的研究，不仅有助于模型的进一步优化，更有助于提升对公司违规现象的理解。

为了对模型因子重要性进行解读分析，本文采用起源于合作博弈论的加性解释模型SHAP。SHAP将所有特征因子视为“贡献者”，其优势在于，不仅能够反映特征的影响力，还能反馈这种影响的正负方向，从而帮助我们了解模型如何运作，以防止出现错误的结论。但SHAP不能应用于神经网络模型，因此，对于神经网络，本文借鉴Dimopoulos等（1995）的方法，通过模型对变量变化的敏感

表 3 模型预测结果评估

		AUC	NDCG@k	F1-Score
线性模型	Logistic	0.54	0.18	0.19
	Ridge	0.54	0.18	0.20
树模型	Random Forest	0.71	0.40	0.45
	GBDT	0.72	0.40	0.46
神经网络模型	RNN	0.74	0.45	0.43
	LSTM	0.74	0.44	0.42

性来衡量特征因子的重要性。模型对第 j 个变量的重要性定义为输入变量 j 的偏导数的平方和，τ_1 为训练集，z_j 表示输入变量向量的第 j 个元素，计算公式如下：

$$SSD_j = \sum_{i,t\in\tau_1}\left(\frac{\partial g(z;\theta)}{\partial z_j}\Big|_{z=z_{i,t}}\right)^2 \quad (12)$$

表 4 报告了机器学习模型和传统预测模型的因子重要性数值，每个模型特征因子重要性数值之和为 1。根据表 4 中机器学习模型的因子重要性情况，贡献较大的公司违规预测因子基本表现一致，本文将这些因子划分为四类。一是公司上年是否违规。公司上年是否违规（v_lag）在机器学习每个模型中均排名第一，说明企业的违规“前科”是模型判断未来公司是否发生违规的重要依据。已经明确存在的客观违规行为说明企业存在较高的违规风险和较差的公司治理现状，因此对公司未来违规行为的发生具有显著预测力。二是公司外部治理相关变量。在排名前 6 的因子中，除了排名第一的 v_lag，其余 5 个均为公司外部治理特征因子。分析师关注（AnaAttention）作为一种重要的外部治理机制（Du，2014），对违规行为的预测具有较大贡献。市场化水平，如要素市场的发育程度评分（factorscore）、市场中介组织的发育和法律制度环境评分（legalscore）、产品市场的发育程度评分（productscore）、政府与市场的关系评分（govscore）和非国有经济的发展评分（NSOEeconomics）与未来公司违规行为的发生高度关联。以上发现强调了公司外部治理的重要作用。三是公司内部治理相关变量。与陈国进等（2005）、杨道广和陈汉文（2015）的发现一致，第一大股东持股比例（first）、股权制衡（ratio）、内部控制（IsDisclosingEvaRep）等内部治理对公司违规具有预测能力。但董事会结构等其他公司内部治理指标，如国有股所占比例、管理层持股比例、独立董事人数、两职合一、董事会人数等并未出现在因子重要性排名前 20 中。四是利润相关指标。贡献排名前 10 的因子中，包括留存收益（RetainEarnings）、所得税费用（IncomeTaxes）和净利润（NetIncome）三个财务报表相关变量，均来自利润表。这与压力理论一致，当公司盈利能力较差时，违规动机会伴随业绩压力的增加而增强（Merton，1938；贺小刚等，2015）。本文结论体现了盈利能力对公司违规预警的重要作用，但财务报表数据对公司违规预测能力有限，而公司治理，尤其是外部公司治理，对公司未来违规预测具有重要贡献，本研究补充了以往仅关注财务信息的公司违规相关文献（Cecchini 等，2010；Dechow 等，2011）。

表 4 还报告了基于传统预测方法的线性回归模型的因子重要性排序。通过与机器学习模型对比，可以发现：第一，传统预测模型的线性回归所学习到的特征数量较少。重要性为 0 的特征，代表算法没有在这组特征中学到信息。基于传统预测模型的线性回归的重要性因子中仅有 20 个因子不为 0，说明传统预测模型的线性回归仅学习了 20 个特征因子的信息。相比之下，机器学习模型 Random Forest、RNN 和 LSTM 在这组数据中学到了全部特征的信息。造成传统预测模型的线性回归信息遗失的原因可能在于线性回归对高维数据的处理能力有限。当变量数较大时，线性回归的效率急剧下降，而机器学习模型在降维、惩罚项和泛函数等技术上的突破，在解决上述问题上具有天然的优越性。因此，采用机器学习中高级的统计工具是更为科学稳健的选择。第二，在线性回归模型学习到的 20 个特征中，除了第一大股东持股比例（first）和优先股（PreferStock），其余均为财务报表相关变量。特征学习维度单一，说明传统预测模型的线性回归过度强调财务报表数据对违规的预测能力，遗失了其他数据所蕴含的信息。总的来说，基于传统预测模型得到的结论，一方面容易达到预测瓶颈，难以持续优化，另一方面可能得出错误结论，造成认知偏差。

（三）可视化分析

机器学习算法的模型结构和框架非常复杂，人们难以理解其性质，可解释性差使其常被视为“黑箱”。但黑箱问题并非完全无解。一方面，传统预测模型的线性回归和机

表 4　因子重要性排序：机器学习 vs. 传统预测模型

	机器学习					传统预测模型		
	变量名	GBDT	Random Forest	RNN	LSTM	变量名	Logistic	Ridge
1	v_lag	0.7146	0.2825	0.7789	0.8125	Liability	0.2952	0.2941
2	AnaAttention	0.0069	0.0336	0.0354	0.0317	CurrentLiability	0.1551	0.1543
3	factorscore	0.0148	0.0297	0.0154	0.0125	assets	0.0887	0.0887
4	legalscore	0.0104	0.0180	0.0363	0.0328	sales	0.0843	0.0856
5	productscore	0.0055	0.0203	0.0220	0.0180	COGS	0.0657	0.0668
6	govscore	0.0145	0.0161	0.0116	0.0067	LongDebt	0.0589	0.0586
7	NSOEeconomics	0.0239	0.0226	0.0048	0.0030	IncomeTaxes	0.0390	0.0390
8	RetainEarnings	0.0413	0.0500	0.0019	0.0016	first	0.0366	0.0366
9	IncomeTaxes	0.0248	0.0261	0.0016	0.0014	PPE	0.0347	0.0348
10	NetIncome	0.0948	0.0701	0.0011	0.0010	DebtLiability	0.0340	0.0340
11	IsDisclosingEvaRep	0.0167	0.0076	0.0038	0.0021	CommonEquity	0.0255	0.0255
12	PPE	0.0063	0.0137	0.0022	0.0016	inventories	0.0199	0.0199
13	Yclsprc	0.0011	0.0226	0.0017	0.0014	RetainEarnings	0.0153	0.0154
14	IncomeTaxesPay	0.0021	0.0169	0.0016	0.0015	cash	0.0151	0.0151
15	age	0.0076	0.0221	0.0008	0.0015	NetIncome	0.0146	0.0143
16	ratio	0	0.0156	0.0037	0.0032	Receivables	0.0082	0.0082
17	first	0	0.0254	0.0020	0.0016	IncomeTaxesPay	0.0039	0.0039
18	ShortInvest	0.0005	0.0046	0.0025	0.0021	PreferStock	0.0023	0.0023
19	FGscore	0	0.0141	0.0028	0.0032	LongInvest	0.0022	0.0022
20	CommonEquity	0.0019	0.0137	0.0016	0.0013	ShortInvest	0.0007	0.0007
21	issue	0	0.0019	0.0173	0.0132	oversea_m	0	0
22	nationratio	0	0.0073	0.0063	0.0042	IsDisclosingEvaRep	0	0
23	IsValid	0	0.0080	0.0048	0.0039	v_lag	0	0
24	IsDeficiency	0	0.0044	0.0101	0.0075	Clsprc	0	0
25	Clsprc	0.0058	0.0160	0.0008	0.0008	IsDeficiency	0	0
26	parttime	0	0.0059	0.0031	0.0051	age	0	0
27	male_m	0	0.0121	0.0025	0.0028	Big4	0	0
28	degree_m	0	0.0123	0.0023	0.0027	issue	0	0
29	TopExec	0.0048	0.0251	0.0001	0.0004	AnaAttention	0	0
30	sales	0	0.0147	0.0019	0.0015	parttime	0	0
31	COGS	0	0.0127	0.0020	0.0017	legalscore	0	0
32	Interest	0	0.0000	0.0030	0.0025	soe	0	0
33	inventories	0.0015	0.0154	0.0002	0.0006	Interest	0	0
34	oversea_m	0	0.0082	0.0020	0.0015	IsValid	0	0
35	LongInvest	0	0.0136	0.0015	0.0014	TopExec	0	0
36	DebtLiability	0	0.0181	0.0008	0.0007	Independent	0	0
37	Receivables	0	0.0137	0.0011	0.0010	BoardSize	0	0
38	assets	0	0.0141	0.0010	0.0009	male_m	0	0
39	Independent	0	0.0049	0.0015	0.0012	factorscore	0	0
40	PreferStock	0	0.0000	0.0015	0.0015	FGscore	0	0
41	CurrentLiability	0	0.0152	0.0005	0.0006	degree_m	0	0
42	cash	0	0.0136	0.0006	0.0008	NSOEeconomics	0	0
43	LongDebt	0	0.0135	0.0007	0.0007	nationratio	0	0
44	soe	0	0.0024	0.0008	0.0009	govscore	0	0
45	Liability	0	0.0118	0.0004	0.0005	productscore	0	0
46	Big4	0	0.0049	0.0009	0.0002	ratio	0	0
47	BoardSize	0	0.0047	0.0007	0.0005	Yclsprc	0	0

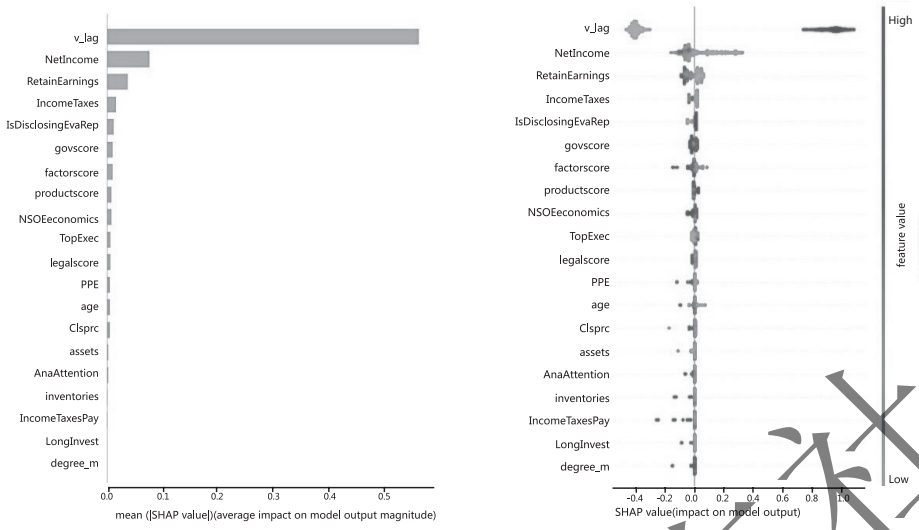

图 1　基于 GBDT 的特征因子贡献分布

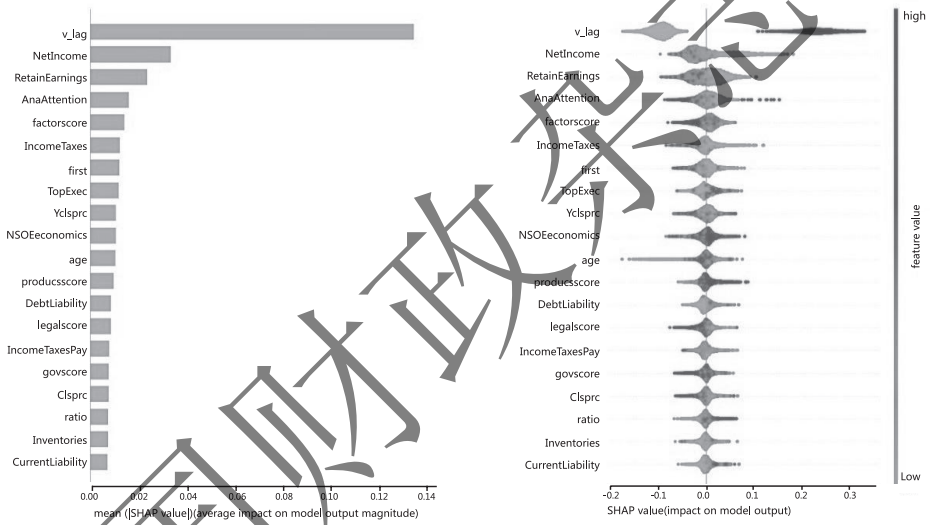

图 2　基于 Random Forest 的特征因子贡献分布

器学习的树模型具有明显的可解释性。另一方面，在建模过程中有意识进行处理，引入假设机制，也能在很大程度上避免黑箱问题。

本文采用 SHAP 方法将基于 GBDT 和 Random Forest 的公司违规预测模型的运作过程和贡献分布可视化。图 1 和图 2 分别展示了 GBDT 和 Random Forest 的特征因子贡献分布。每一行代表一个特征，一个点代表一个样本。Featurevalue 代表特征的大小，浅灰色方向代表该特征的数值小，深灰色方向代表该特征的数值大。横坐标为 SHAP 值，代表每个特征所分配的数值，当 SHAP value 大于 0，说明该特征提升了贡献值，起正向作用；当 SHAP value 小于 0，说明该特征使预测值降低，起反向作用。

根据图 1 和图 2 可知，上年是否违规 v_lag 对公司被预测为违规公司具有非常大的贡献。v_lag 的深灰色样本的 SHAP value 均大于 0，说明上年违规显著提高了公司未来被预测为违规的可能性。但上年违规并非意味着未来一定会被预测为违规。从两个图中也可以看出，较好的公司治理对公司违规有着显著制约作用。例如，外部治理因子 factorscore 和 legalscore 的大部分深灰色样本落在 SHAP value 小于 0 的一侧，说明高水平的外部治理能够抑制公司违规行为的发生，所在城市要素市场的发育程度越低、市场中介组织的发育和法律制度环境水平越差，公司被预测为违规的可能性越高；分析师关注 AnaAttention 的深灰色样本大部分落在 SHAP value 小于 0 的一侧，说明较多的分析师关注减少了第二年违规发生的可能，与桂爱勤和龙俊雄（2018）的发现一致；公司内部治理因子股东持股数量

对公司违规行为具有显著的影响，例如，第一大股东持股数量first的深灰色样本的SHAP value大多小于0，与之相应的，股权制衡ratio的浅灰色样本基本落在SHAP value小于0的一侧，说明较高的股权集中度有利于抑制公司违规行为，与陈国进等（2005）的发现一致。公司利润指标所得税费用IncomeTaxes深灰色样本的SHAP value大多小于0，浅灰色样本基本落在SHAP value大于0的一侧，说明业绩较差的公司更可能在未来发生违规。

此外，股票价格与公司违规之间具有一定的关联性，年报发布日股票收盘价Clsprc的深灰色样本大多落在SHAP value小于0的一侧，说明股价与公司违规具有负相关关系。结合表4基于机器学习算法的因子重要性排序中，年末股票收盘价Yclsprc排名第13，亦说明了股票价格对公司违规的重要预警作用。现存文献大多探讨公司违规对股票价格的影响（Feroz等,1991；杨忠莲和谢香兵，2008; 宋献中等，2017），但实际上，股票价格也可以是公司违规预测的重要参考指标。股票价格在一定程度上反映了公司的内在价值，市场可能在公司违规前已经知悉公司的违规可能，并将这种对风险认知反映在股票价格上，从而能够辅助判断公司未来违规的可能性。因此监管部门可以利用资本市场“看不见的手”辅助判断违规可能，以进行重点调查和跟踪。

（四）影响阈值

实证研究基于理论做出假设，探讨社会关系中各因素之间的因果关系及影响机制，但并不能明确变量对公司违规产生有效影响的具体数值范围。例如，有文献研究表明分析师关注能够抑制公司违规行为（Du, 2014; Chen等，2016），但分析师跟踪数量达到多少会对公司违规产生有效抑制作用，现有研究并没有给出答案。相比之下，机器学习方法在处理此类问题时具有较大优势。因此，本文利用SHAP计算Random Forest模型，分别计算上年是否违规v_lag、分析师跟踪人数AnaAttention、要素市场发育程度评分factorscore、所得税费用IncomeTaxes和第一大股东持股数量first五个重要因子的影响阈值。

考虑到上年是否存在违规记录对违规预测的重要影响，本文以此为出发点，分别从上年无违规记录和上年有违规记录两个层面展开分析，然后按照因子重要性排序，依次计算余下四个因子的影响阈值。由于因子值与其所对应的SHAP value并非完全线性相关，本文采用如下计算方式：

对于任一连续的数值型因子x，分别计算该因子在SHAP value为正和SHAP value为负时所对应的因子特征均值$\bar{x}_+$、$\bar{x}_-$，计算公式如式（13）（14）所示：

$$\bar{x}_+ = \frac{1}{N}\sum_{i=0}^{N} x_i ,\quad for\ \forall\ shap(x_i) > 0 \tag{13}$$

$$\bar{x}_- = \frac{1}{M}\sum_{i=0}^{M} x_i ,\quad for\ \forall\ shap(x_i) < 0 \tag{14}$$

其中，N代表该因子对应的SHAP value为正的样本数量，M代表该因子对应的SHAP value为负的样本数量。

当x取值区间为公式（15）（16）所示时，SHAP value为正的概率较大：

$$x \in [\bar{x}_+, x_{max}],\ if\ \bar{x}_+ > \bar{x}_- \tag{15}$$

$$x \in [x_{min}, \bar{x}_+],\ if\ \bar{x}_+ < \bar{x}_- \tag{16}$$

当x取值区间为公式（17）（18）所示时，SHAP value为负的概率较大：

$$x \in [x_{min}, \bar{x}_-],\ if\ \bar{x}_+ > \bar{x}_- \tag{17}$$

$$x \in [\bar{x}_-, x_{max}],\ if\ \bar{x}_+ < \bar{x}_- \tag{18}$$

对于类别型因子，可直接对该因子每一类所对应的SHAP value进行统计，进而得到类别因子取不同值时的违规倾向。

为了得到严格限值条件下的因子取值，需根据因子重要性顺序依次计算每个因子的取值区间，排在后面的因子的样本同时需满足前面因子的取值区间，则第j个因子x^j的$\bar{x}_+^j$、$\bar{x}_-^j$计算公式如式（19）（20）所示：

$$\bar{x}_+^j = \frac{1}{N}\sum_{i=0}^{N} x_i^j ,\quad for\ \forall\ shap(x_i^j) > 0\ and\ \ k = 1, \dots j-1,\ x_i^k \in [x_L^k, x_R^k] \tag{19}$$

$$\bar{x}_-^j = \frac{1}{M}\sum_{i=0}^{M} x_i^j ,\quad for\ \forall\ shap(x_i^j) < 0\ and\ \ k = 1, \dots j-1,\ x_i^k \in [x_L^k, x_R^k] \tag{20}$$

其中，$[x_L^k, x_R^k]$为第k个因子的取值区间，k<j。因子x^j取值区间的计算方式与公式（15）~（18）相同，在此不再赘述。

此外，违规比例p^j为在前j个因子取值均符合上述约束的样本集合φ^j中，违规样本集合φ_+^j所占的比例。计算公式如下：

$$p^j = \frac{|\varphi_+^j|}{|\varphi^j|} \quad for\ \forall\ k = 1, \dots j,\ x_i^k \in [x_L^k, x_R^k] \tag{21}$$

最终得到5个因子的影响阈值，结果如表5所示。Panel A报告了上年没有发生违规和上年发生违规两种情况下，分析师跟踪人数AnaAttention、要素市场发育程度评分factorscore、所得税费用IncomeTaxes和第一大股东持股数量first这四个因子在什么数值范围内能够有效降低公司未来违规发生的概率。Panel B报告了上年没有发生违规和上年发生违规两种情况下，分析师跟踪人数AnaAttention、要素市场发育程度评分factorscore、所得税费用IncomeTaxes

表 5　随机森林模型下基于 SHAP 计算的因子对公司违规产生有效影响的取值区间

Panel A 改善区

无违规记录				有违规记录			
变量名	符号	数值	违规概率	变量名	符号	数值	违规概率
v_lag	=	0	0.35	v_lag	=	1	0.85
AnaAttention	>	9.94	0.30	AnaAttention	>	10.38	0.82
factorscore	>	6.46	0.28	factorscore	>	7.03	0.74
IncomeTaxes	>	312.45	0.18	IncomeTaxes	>	118.93	0.71
first	>	1 305.64	0.10	first	>	450.40	0.45

Panel B 恶化区

无违规记录				有违规记录			
变量名	符号	数值	违规概率	变量名	符号	数值	违规概率
v_lag	=	0	0.35	v_lag	=	1	0.85
AnaAttention	<	3.16	0.37	AnaAttention	<	2.47	0.87
factorscore	<	5.21	0.39	factorscore	<	4.95	0.89
IncomeTaxes	<	5.96	0.44	IncomeTaxes	<	3.89	0.90
first	<	67.38	0.49	first	<	51.00	0.92

Panel C 描述性统计

变量名	样本数	平均值	最小值	25 分位	50 分位	75 分位	最大值
v_lag	18 209	0.17	0	0	0	0	1
AnaAttention	18 209	7.81	0	1	4	12	66
factorscore	18 209	6.31	−1.21	4.7	5.86	7.32	12.23
IncomeTaxes	18 209	137.95	−1 889	7.04	20.50	63.29	49 330
first	18 209	528.67	4.48	61.99	129.7	286.1	158 200

和第一大股东持股数量 first 这四个因子在什么数值范围内将显著提高公司未来违规发生的概率。同时，为了方便对比，Panel C 报告了这 5 个因子的描述性统计结果。

根据 Panel A 结果，相比于没有违规记录的企业，有违规记录的企业需要获得更多的分析师关注、所在城市有更高的要素市场发育程度，模型才会降低其次年违规发生的预测概率，说明外部公司治理质量的提高是降低未来违规发生概率的决定因素。此外，Panel A 改善区的各项数值远高于 Panel B 恶化区的各项数值，说明恶化区企业若想有效降低未来违规发生概率，需要系统提高企业的盈利能力和内外部公司治理质量。

五、研究结论

公司违规行为不仅严重挫败投资者信心，阻碍资本市场健康发展，还会弱化资本市场资源配置功能。由于现存文献所构建的公司违规预测模型往往难以应用于实践中，为了使事前监管更加有效并最大限度地减少各方损失，亟需找到提高预测效果的突破口。人工智能浪潮的兴起和迅猛发展，为精准预测带来了机遇和前景。机器学习方法作为人工智能的代表技术之一，其重要性和可行性已经得到广泛认可和应用。因此，本文基于 GONE 理论进行特征选取，在传统预测模型（Logistic 和 Ridge）、树模型（Random Forest 和 GBDT）和神经网络模型（RNN 和 LSTM）的基础上进行公司违规行为预测。

借鉴舞弊四因子理论，本文分别从个体风险因素和一般风险因素两方面提取 47 个特征。根据 AUC、NDCG@k 和 F1-Score 三个检验方法评价的模型预测效果，发现相比于传统预测模型的线性回归，基于机器学习算法所构建的公司违规预测模型具有更强的预测能力。进一步地，为了提高模型的可解释性，本文进行了如下拓展分析：（1）通过采用 SHAP 方法量化每个特征对模型所做预测的贡献，并对模型运作过程可视化，以提高模型的可解释性。研究结果表明，机器学习模型中因子对未来违规行为的预测贡献值从高到低排序，依次为公司上年是否违规、公司治理相关变量（尤其是外部治理）以及利润指标。其中，虽然违规记录是判断未来违规的重要依据，但要素市场的发育程

度和法律制度环境水平越高、分析师关注越多、股权集中度越高、盈利能力越强、股票价格越高，则模型判断公司未来发生违规的可能性越低。(2)通过对比传统预测模型的线性回归和机器学习模型因子贡献值情况，发现在同样数据样本中传统预测模型的线性回归学习到的信息较为有限，体现了传统预测模型的线性回归的局限性。在47个特征因子中，传统预测模型的线性回归仅学到20个，且学习到的特征维度较为单一，除了第一大股东持股比例和优先股以外，其余均来自财务报表数据，并没有获取其他特征因子所蕴含的信息。对比而言，机器学习对高维数据和因子交互关系方面的信息挖掘，均展现了其强大的统计性能。(3)为了定量确认重要因子对公司违规起显著作用的具体数值，本文利用SHAP计算了Random Forest模型中影响方向较为明确的5个重要因子对公司违规产生影响的取值区间，为投资者和监管方等提供了参考。由于模型精度并非达到百分之百，而且未来事项的变化不能完全预估，因此，本文获得的取值区间并非绝对准确，但是，本文所采用的方法仍然具有较好的理论和应用价值。

预测，对于人类而言充满诱惑。对未来的准确预测是国家、政府、企事业单位和个人制定战略与决策的重要依据和不懈追求。随着大数据时代的到来，人工智能和机器学习技术的优势在各领域中逐渐凸显，相信机器学习能够在未来为理论研究和应用研究带来更多新的发现。

主要参考文献：

[1] 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，2007，10(6)：62-68．

[2] 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)：195-210．

[3] 陈国进，林辉，王磊．公司治理，声誉机制和上市公司违法违规行为分析[J]．南开管理评论，2005，(6)：35-40．

[4] 陈丽英，李婉丽，吕怀立．盈余重述归因分析——资产负债表膨胀的角度[J]．南开管理评论，2012，15(6)：34-43．

[5] 冯旭南，陈工孟．什么样的上市公司更容易出现信息披露违规——来自中国的证据和启示[J]．财贸经济，2011，(8)：51-58．

[6] 桂爱勤，龙俊雄．分析师跟踪对上市公司违规行为影响的实证分析[J]．统计与决策，2018，(10)：171-173．

[7] 贺小刚，邓浩，吴诗雨，梁鹏．赶超压力与公司的败德行为——来自中国上市公司的数据分析[J]．管理世界，2015，(9)：104-124．

[8] 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，(8)：84-90．

[9] 李航．统计学习方法[M]．北京：清华大学出版社，2012．

[10] 刘英明．基于制度理论视角的财务报告舞弊行为研究[J]．财政研究，2012，(9)：74-77．

[11] 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．

[12] 单华军．内部控制、公司违规与监管绩效改进——来自2007-2008年深市上市公司的经验证据[J]．中国工业经济，2010，(11)：140-148．

[13] 滕飞，辛宇，顾小龙．产品市场竞争与上市公司违规[J]．会计研究，2016，(9)：32-40．

[14] 杨道广，陈汉文．内部控制、法治环境与守法企业公民[J]．审计研究，2015，(5)：76-83．

[15] 王霞，张为国．财务重述与独立审计质量[J]．审计研究，2005，(3)：56-61．

[16] 吴世农，卢贤义．我国上市公司财务困境的预测模型研究[J]．经济研究，2001，(6)：46-55．

[17] 徐尧，刘峰，王亚．法治环境、政治关联与违规查处——来自A股市场的经验证据[J]．当代财经，2017，(6)：80-89．

[18] 张翼，马光．法律，公司治理与公司丑闻[J]．管理世界，2005，(10)：113-122．

[19] 周志华．机器学习[M]．北京：清华出版社．2015．

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

[21] Beasley, M. S. An Empirical Analysis of the Relation between Board of Director Composition and Financial Statement Fraud[J]. Accounting Review, 1996, 71 (4) : 443-465.

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

[23] Cecchini, M., Aytug, H., Koehler, G. J., Pathak, P. Detecting Management Fraud in Public Companies[J]. Management Science, 2010, 56: 1146-1160.

[24] Chen, J., Cumming, D., Hou, W., Lee, E. Does the External Monitoring Effect of Financial Analysts Deter

Corporate Fraud in China？[J]. Journal of Business Ethics, 2016, 134 (4) : 727-742.

[25] Dechow, P. M., Weili, G. E., Larson, C. R., Sloan, R.G. Predicting Material Accounting Misstatements[J]. Contemporary Accounting Research, 2011, 28: 17-82.

[26] De Prado, M. L. Advances in Financial Machine Learning[M]. John Wiley & Sons, 2018.

[27] Dimopoulos, Y., Bourret, P., Lek, S. Use of Some Sensitivity Criteria for Choosing Networks with Good Generalization Ability[J]. Neural Processing Letters, 1995,2: 1-4.

[28] Du, X. Does Religion Mitigate Tunneling？ Evidence from Chinese Buddhism[J]. Journal of Business Ethics, 2014,125 (2) : 299-327.

[29] Dyck, A., Morse, A., Zingales, L. Who Blows the Whistle on Corporate Fraud？ [J]. Journal of Finance, 2010, 65 (6) : 2213-2253.

[30] Gu, S., Kelly, B., Xiu, D. Empirical Asset Pricing via Machine Learning[J]. The Review of Financial Studies, 2020, 33 (5) : 2223-2273.

[31] Johnstone, K., Li, C., Rupley, K. H. Changes in Corporate Governance Associated with the Revelation of Internal Control Material Weaknesses and Their Subsequent Remediation[J]. Contemporary Accounting Research, 2011,8 (1) : 331-383.

[32] Kedia, S., Rajgopal, S. Do the SEC's Enforcement Preferences Affect Corporate Misconduct？[J]. Journal of Accounting and Economics, 2011,51 (2) : 259-278.

[33] Larcker, D., Zakolyukina, A. A. Detecting Deceptive Discussions in Conference Calls[J]. Journal of Accounting Research, 2012, 50: 495-540.

[34] Merton, R. K. Social Structure and Anomie[J]. American Sociological Review, 1938, 3 (5) :672-682.

Corporate Fraud Prediction based on Machine Learning

LI Ying, QU Xiao-hui

Abstract: The traditional regression model is usually used to investigate corporate fraud. This paper develops a corporate fraud prediction model using a machine learning approach. (1) We find that our machine learning models, the tree model, and neural network, yield much better prediction performance than the commonly used method of linear regression. (2) We use SHAP method to calculate the importance values and contribution level of each feature, and find corporate governance, especially external corporate governance, has an important contribution to the corporate fraud predictions. By contrast, we find that the linear model overemphasizes the predictive ability of financial data and weakens the contribution of corporate governance factors to corporate fraud prediction, emphasizing the application value of the machine learning method in the field of financial accounting. (3) Furthermore, we use SHAP method to visualize the GBDT and Random Forest model process to analyze the impact mechanism, and calculate the threshold of the main influence factors. This paper contributes to the methodology and offers decision-making useful information for investors and market supervision.

Key words：corporate fraud prediction; machine learning; traditional prediction models; factor importance; visualization

（责任编辑 王安琪）