doi: 10.19920/j.cnki.jmsc.2026.02.004

基于机器学习的公司避税预测研究①

彭 章[1]，陆 瑶[2]，牛美龄[2]，周欣怡[2]
(1. 中央财经大学财政税务学院，北京 102206; 2. 清华大学经济管理学院，北京 100084)

摘要: 本文运用梯度提升回归树和随机森林算法同时考察以往文献提及的、能够影响公司避税的 6 个方面的 64 个公司内外部特征，对避税程度的预测能力. 发现避税工具特征，尤其是捐赠支出、应计盈余、员工人数、研发支出，对避税程度有较强预测能力，而公司外部特征的预测能力普遍较弱. 考察关键预测特征与避税程度的关系发现，通常避税程度会随捐赠支出、员工人数的增加而减少，随应计盈余、研发支出的增加而增加; 少数情况下避税程度与员工人数正相关，与应计盈余、研发支出相关性不强; 捐赠支出与总资产收益率、应计盈余与监事会规模、员工人数与 CEO 任期、研发支出与机构投资者持股比例存在交互作用. 且基于决策树的机器学习算法比线性算法预测能力更强. 该结果说明在税收征管和智慧税务建设过程中可以更多关注避税工具特征，制定政策时需注重降低避税工具可得性.

关键词: 公司避税预测; 机器学习; 避税工具
中图分类号: F812.42; F275 **文献标识码**: A **文章编号**: 1007 - 9807(2026)02 - 0064 - 17

0 引 言

税收是财政的基石，在国家治理中发挥着基础性、支柱性、保障性作用. 企业所得税是我国主要税种之一，是财政收入的重要来源. 虽然合理避税能节约现金流、提升公司价值，但过度追求低税负会损害国家财政并增加公司违规风险. 因此，如何抑制公司避税是政府亟待解决的现实问题. 近年来，大数据、人工智能为税收征管带来新思路，许多学者[1]认为机器学习等技术能更精确地识别税务风险，助力税收征管. 我国政府高度重视税务部门智能化发展，十四五规划中提出要建设智慧税务，推动税收征管现代化. 2021 年 3 月，中共中央办公厅、国务院办公厅印发的《关于进一步深化税收征管改革的意见》明确指出，要充分运用大数据、云计算、人工智能等现代信息技术驱动税务执法、服务、监管制度创新和业务变革，实现“以数治税”分类精准监管.

要实现对公司避税的数字化精准监管、提升税收征管效率，需全面认识公司避税行为的预测变量、事先预判其避税程度. 以往研究发掘了大量公司避税决定因素，但这类解释性研究存在三大问题: 一是部分指标，如公司治理水平，与公司避税程度的关系存在争议[2,3]; 二是一些变量与公司避税程度存在非线性关系[4]，但目前大多研究使用的线性模型难以挖掘这种关系; 三是现有研究大多关注单一变量对公司避税程度的解释，缺乏综合多维特征的预测性研究，而对公司避税的事前预测比事后分析更有意义.

机器学习是人工智能的代表性技术，在经济

① 收稿日期: 2023 - 01 - 29; 修订日期: 2024 - 05 - 20.
基金项目: 国家自然科学基金资助重点项目(72532005); 国家自然科学基金资助青年科学基金项目(72202250); 中国博士后科学基金资助项目(2022M713660); 北京工商大学数字商科与首都发展创新中心项目(SZSK202323); 2024 年度中央财经大学教育教学改革基金项目(2024ZCJG21); 2024 年文科建设“双高”计划创新方向建设专项《中国特色科技金融发展之路》; 教育部人文社会科学重点研究基地重大项目(22JJD790047); 清华大学经济管理学院“影响力”提升计划项目(2022051006).
作者简介: 彭 章(1994—)，女，江西九江人，博士，副教授. Email: pengzhang321@126.com

学研究领域也备受关注[5]. 机器学习的优势在于其强大的预测能力,能够同时处理高维度变量,且无需事先假设变量之间的关系,通过算法自动识别变量间的非线性关系. 而影响公司避税的变量众多,且这些变量与避税程度之间的关系复杂,运用机器学习算法可以筛选出其中的关键特征,考虑变量间的非线性、交互关系,弥补传统线性模型的不足. 因此,本文试图运用梯度提升回归树和随机森林算法,从公司内外部出发,同时考察避税工具特征、公司治理特征、其他公司内部特征、宏观层面特征、行业层面特征、外部微观特征六类共 64 个变量,对公司避税程度的预测能力,识别公司避税的关键预测特征,并进一步探究关键预测特征与公司避税行为的关系.

本文发现,公司避税工具特征对避税程度有较强预测能力,但公司外部特征的预测能力较弱,且相较 OLS,运用梯度提升回归树、随机森林等机器学习算法能显著提升模型预测能力. 考察特征重要性发现,避税工具特征中,捐赠支出、应计盈余、员工人数、研发支出是关键预测特征. 考察特征与避税程度的关系发现: 捐赠支出更多的公司,避税程度更低; 应计盈余与避税程度基本呈正相关,特别是当应计盈余处于 -0.05 ~0.05 时,避税程度随着应计盈余的提高显著上升,但其他区间两者相关性较弱; 员工人数与避税程度主要呈负相关,但在员工人数少于约 400 人时,两者呈正相关; 研发支出与避税程度基本呈正相关,但研发支出占总资产比例超过 0.04 后,两者相关性不明显. 考察预测模型中避税工具特征与其他特征的交互作用发现,捐赠支出与总资产收益率、应计盈余与监事会规模、员工人数与 CEO 任期、研发支出与机构投资者持股比例存在交互作用. 该结果说明,避税工具特征,特别是捐赠支出、应计盈余、员工人数、研发支出,对公司避税影响较大,而外部特征影响较小; 部分特征与避税程度之间存在非线性关系,且避税工具特征与其他类别特征在预测避税时有交互作用,机器学习算法能有效识别这些非线性、交互关系,从而更好地预测避税程度. 本研究还将避税程度预测问题转换为分类问题和排序问题,多维度评估避税预测模型的效果,发现模型在召回率、精确率、NDCG@ K、准确率等方面均表现出色. 本文对公司未来避税程度进行了预测,发现避税工具特征同样是预测未来期避税程度的关键特征.

本文有两方面学术贡献: 首先,运用机器学习算法全方位考察公司避税预测特征,利用数据驱动的方式发掘重要避税预测特征,并揭示其与避税程度之间的复杂关系. 与以往的公司避税研究仅聚焦于单一特征并依赖线性模型不同,本文考虑了非线性、交互的关系. 本文结果有助于解决当前公司避税决定因素的争议,帮助认识重要预测特征与避税程度的复杂关系. 第二,本研究拓展了机器学习在经济管理研究中的应用. 当前,大数据、机器学习等技术正在改变经济学、管理学研究范式[5-7],相关研究集中于利用机器学习分析和预测宏观变量[8]、股票市场表现[9,10]、公司业绩[11]、财务舞弊和欺诈[12,13]等方面,尚未探索机器学习在预测公司避税行为的应用. 本文填补了这一空白,为经济管理研究中的机器学习应用提供了新视角.

1　公司避税相关研究综述

1.1　公司避税衡量指标研究

公司避税行为难以直接观测,因此准确衡量避税程度是相关研究的重点. 目前,避税指标主要分为两类: 实际所得税税率(Effective Tax Rate,简称实际税率) 及其变体,账面 - 税收差异(Book Tax Difference) 及其变体.

实际税率,即企业所得税费用或企业支付所得税现金与税前利润的比值,该指标越高代表企业纳税越多,避税程度更低,基于实际税率的指标已被广泛采用[14]. 现实中,避税行为会降低实际税率,例如操纵研发费用获得高新技术企业认定能够直接享受优惠税率[15]. 但实际税率指标易受制度和公司因素影响,许多学者基于实际情况对其进行修正. 例如,Dyreng 等[16]指出,当期实际税率的计算中可能出现分母为负值等问题,采用长期实际税率可规避这些问题; Balakrishnan 等[17]认为,同行业中规模相似公司合法税收筹划的机会相近,建议以同行业中规模相似公司的实际税率均值为基准,从公司自身实际税率中剔除该均值,以更好反映公司避税激进程度; 叶康涛和刘

行[18]指出税收优惠政策使得我国公司名义所得税率各不相同,因此建议用名义税率与实际税率的差异来衡量避税程度.

账面 - 税收差异指标源于会计准则与所得税法的分离,使得企业能够利用两者差异、通过减少应纳税所得额进行避税[18],比如操纵研发费用享受研发费用税前加计扣除政策,进行避税. 该指标越高说明企业应纳税所得额相对利润更少,避税程度更高. Wilson[19]发现,存在避税行为的公司通常表现出更大的账面 - 税收差异. 虽然该指标被广泛采用[20],但也存在缺陷. Phillips 等[21]指出暂时性的账面 - 税收差异主要反映公司盈余管理情况. 因此,后续研究[22,23]通过线性回归排除盈余管理的干扰进行改进.

由此可见,会计与税务制度、公司自身情况都可能影响避税指标的准确性,故考察我国公司避税程度时,应结合制度环境和公司情况,尽可能采用多指标确保结果可靠性.

1.2 公司避税的决定因素研究

理论上,学者们从不同角度探讨了公司避税决定因素: 早期研究[24]将避税行为看作投资行为,认为公司避税程度取决于避税带来的收益、成本和风险; 随着对公司内部代理问题认识加深,Desai 和 Dharmapala[22]等学者将代理问题纳入分析框架,认为公司治理因素对企业避税行为有重要影响; Slemrod[25]则将避税行为纳入供给 - 需求框架,认为企业避税程度取决于避税工具的供给和企业对避税的需求.

实证上,大量学者也对公司避税决定因素进行研究. 公司内部,企业社会责任感[26,27]、公司治理水平[4,22]、高管特征和激励[14,20]、产权性质[28]等因素都会影响公司避税程度. 公司外部,税务部门监督和税收征管力度[29,30]、财政压力[31]、产品市场竞争[2]、外部监督和压力[32]等因素也会对公司避税水平产生显著影响.

已有文献深入探究了公司避税的决定因素,但对一些因素与避税水平的关系仍有争议. 例如,理论上治理水平高的公司绩效更好,管理者有动机提高业绩表现,导致公司治理水平与避税程度正相关[33],但大量文献[22,23,34]也指出避税行为常被管理者用于攫取私人利益,这导致公司治理水平与避税程度负相关. 部分文献还指出一些变量与公司避税程度的关系是非线性的,如 Armstrong 等[4]运用分位数回归发现 CEO 薪酬激励等公司治理变量对不同程度的避税行为影响不同.

总体来看,已有文献发掘了众多公司避税决定因素,为公司避税预测特征的选择提供了重要参考. 但现有研究尚未考察这些特征对公司避税程度的预测能力,结果存在较大争议. 且部分研究表明决定因素与避税程度的关系呈非线性,而多数文献仍使用线性模型,缺乏对非线性关系的考虑.

2 研究方法和研究设计

2.1 基于机器学习的财务预测模型

机器学习中的统计学习算法是人工智能的代表性技术,是当前实现计算机智能化的有效手段[35]. 相比传统计量经济学模型,机器学习算法有三点突出区别: 一,机器学习算法由数据驱动,无需预设变量关系,能够自动获取最优模型. 而传统计量模型由模型驱动,需事先假设因变量与自变量之间的关系. 且机器学习擅长捕捉复杂、非线性关系; 二,机器学习算法旨在构建具有强泛化能力的模型,注重模型在样本外的预测能力. 而传统计量模型更强调变量间的因果关系,通过样本内模型拟合来评估变量的解释能力,往往忽略了对预测能力的考察. 然而,预测能力和解释能力是相辅相成的,一个能够描述真实因果关系的模型应具有良好的预测能力,因此预测能力可以反映理论解释现象的能力[36],若一个基于理论的模型无法很好地对其他样本进行预测,表明该理论仍需进一步补充和修正. 因此,评估模型的预测能力至关重要; 三,当预测变量众多且高度相关时,多重共线性问题严重,传统线性计量模型难以识别重要预测变量,需借用主成分分析、动态因子分析等降维方法对数据预处理后再建模. 而机器学习算法能够采用灵活的函数形式,自动在数量众多且高度相关的变量中选择预测因子并进行降维处理[6],提升样本外预测能力并识别重要预测变量[37]. 因此在高维度、高相关性的情景下,机器学习算法更加便捷高效.

这三点突出区别使得机器学习算法适用于本

文的研究问题. 首先,公司避税特征与避税程度可能呈现复杂、非线性关系,机器学习算法能更有效地考察这些关系; 其次,目前公司避税研究集中在对单一变量的解释性研究,缺乏从预测能力角度的研究. 而预测公司避税程度对于实现精准税收征管至关重要,因此,需要考察特征的预测能力,评估预测特征的重要程度; 第三,公司避税决定因素众多且相互关联,比如产品市场竞争程度[2]和公司治理水平[4,22,34]都能够影响避税程度,但产品市场竞争也是一种外部治理机制,能够提升公司治理水平. 由于公司避税预测特征具有高维度和高相关性,因此适用于机器学习算法.

基于此,本文提出以下基于机器学习的财务预测模型

$$TaxAvo_{i,t} = g^{*}(\boldsymbol{Z}_{i,t}) \tag{1}$$

其中 $TaxAvo_{i,t}$ 为公司 i 在第 t 年度的避税程度,$Z_{i,t}$ 为 64 维预测特征向量. 基于已有文献,本文系统梳理了公司避税的决定因素,并据此选择预测特征,详细选择依据见在线附录②. 具体而言,本文将公司避税行为的预测特征分为七类: 基本预测特征、避税工具特征、公司治理特征、其他内部特征、宏观层面特征、行业层面特征、微观层面特征. 通过机器学习方法,考虑预测特征之间的相互作用,将避税程度用预测特征构成的函数 $g^{*}(\cdot)$ 进行表示. 预测特征的分类、名称和符号见表 1,详细定义见在线附录.

表 1　变量类型和名称

Table 1 Variable types and names

变量类型	变量名称(变量符号)
被预测变量(模型输出)	
避税程度指标	名义税率与实际税率之差(*TME*)、1 减去实际税率(*NETR*)、账面 - 税收差异(*BTD*)
预测特征(模型输入)	
基本预测特征(7 个)	公司规模(*Size*)、公司年龄(*Fage*)、名义税率(*Rate*)、总资产收益率(*Roa*)、经营净现金流(*OCF*)、固定资产比例(*PPE*)、杠杆率(*Lev*)
年度	年度虚拟变量(*Year*2008 - *Year*2019)
公司内部特征(43 个)	
避税工具特征	捐赠支出(*Donation*)、是否有企业捐赠金额数据(*Donation_Dummy*)、员工人数(*Emp_Size*)、员工薪酬(*Emp_Pay*)、应计盈余(*Tacc*)、可操纵应计盈余(*Absda*)、关联交易(*Trade*)、是否有关联交易数据(*Trade_Dummy*)、研发支出(*RD*)、是否有研发支出数据(*RD_Dummy*)、是否有外币业务(*Foreign*)、是否有注册地转移情况(*Registran*)
公司治理特征	是否国企(*State_Own*)、是否家族企业(*Fam_Own*)、是否发行 B 股 H 股(*BH_Share*)、股权制衡度(*Top1toTop25*)、第一大股东持股比例(*Top*1)、前十大股东持股比例(*Top*10)、股权集中度(*Top3_SQ*)、机构投资者持股比例(*Inst_Own*)、董事会规模(*Board_Size*)、董事长是否兼任总经理(*Duality*)、独立董事比例(*Board_Ind*)、董事薪酬(*Board_Pay*)、董事持股比例(*Board_Own*)、女性董事比例(*Board_Gender*)、审计委员会独立性(*Audit_Ind*)、高管薪酬(*Exec_Pay*)、高管持股比例(*Exec_Own*)、CEO 薪酬(*CEO_Pay*)、CEO 持股比例(*CEO_Own*)、CEO 任期(*CEO_Tenure*)、CEO 性别(*CEO_Gender*)、CEO 年龄(*CEO_Age*)、CEO 海外经历(*CEO_Aboard*)、CEO 金融财务背景(*CEO_Career*)、CEO 学术背景(*CEO_Resea*)、监事会规模(*Sup_Size*)、监事薪酬(*Sup_Pay*)、监事持股比例(*Sup_Own*)、代理成本(*AC*)
其他内部特征	客户集中度(*CCHHI*)、是否有客户集中度数据(*CCHHI_Dummy*)
公司外部特征(21 个)	
宏观层面特征	税收征管强度(*TE*)、是否实施“金税三期”(*Goldentax*)、最低工资(*Minwage*)、财政盈余(*Gov_Fissurp*)、财政支出增长压力(*Gov_Fispres*)、经济增长(*Gov_GDP*)、政府规模(*Gov_Size*)、经济政策不确定性(*EPU*)、金融发展程度(*Findep*)、市场化水平(*MKT*)、政府和市场关系市场化水平(*GM*)、非国有经济发展情况(*NS*)、产品市场发展情况(*CM*)、要素市场发展情况(*FM*)、中介市场和法治环境情况(*MIOLaw*)
行业层面特征	行业集中度(*Ind_HHI*)、行业内企业数量(*Ind_Size*)、行业成长性(*Ind_Growth*)
微观层面特征	审计师质量(*Big*4)、分析师关注(*Analyst*)、媒体关注(*Media*)

② 本文的在线附录见: https://github.com/mokona321/TaxAvoidancePredict 或 https://pan.baidu.com/s/1w7pLAJfkv2QNJB1W6UDDSA?pwd=1234. 后文提及在线附录均指此链接.

本文采用两种主流决策树算法: 梯度提升回归树(gradient boosting regression trees,GBRT) 和随机森林(random forest,RF) . 受篇幅所限,选择算法的理由、算法介绍和参数选择依据见在线附录. 本文使用 R 语言 gbm 工具包拟合梯度提升回归树,设置参数如下: 交互深度为 5,学习率为 0.001,最大树数量 10 000,通过 5 折交叉验证确定最优的树的数量. 在交叉验证过程中,仅使用训练集样本. 随机森林算法使用 R 语言 randomForest 工具包拟合,参数设定为: 树的数量 2 000 颗,特征数量为总特征个数的 1/3. 随机种子数设为 1.

2.2 研究设计

2.2.1 预测模型构建

为比较各类变量的预测能力,本文以 7 个基本预测特征和 12 个年度虚拟变量的预测模型为参照,即“基准模型”,分别加入公司内外部 6 类特征,通过观察加入各类特征之后模型预测能力的提升幅度来判断每类特征的预测能力. 若加入某类特征使模型预测能力显著提升,则表明该类特征对公司避税程度有较强预测能力. 最后,将所有特征纳入模型,进行预测. 这里,加入所有内部特征的模型简称为“内部模型”、加入所有外部特征的模型简称为“外部模型”、加入所有 64 个内外部特征的模型简称为“全模型”.

随机抽取三分之二家公司作为训练集,剩余三分之一家公司为测试集. 首先考察各类变量的预测能力,然后分析各个变量的重要性,再利用部分依赖图考察关键预测特征与避税程度之间的关系,以及关键预测特征的交互作用.

2.2.2 预测能力评价

遵循 Gu 等[9],使用样本外 R^2(R^2_{oos}) 作为预测能力的主要衡量指标,同时选用均方误差(mean squared error,MSE) 、平均绝对误差(mean absolute error,MAE) 作为评价指标. 评价指标计算方式见式(2) ~式(4) . 其中 $TaxAvo_{i,t}$ 为公司 i 在第 t 年的避税程度,$TEST$ 为测试集,n_T 为测试集样本数量,$\widehat{TaxAvo}_{i,t}$ 为模型对公司 i 第 t 年的避税程度预测值,$\overline{TaxAvo}_{i,t}$ 为测试集避税程度均值

$$R^2_{oos} = 1 - \frac{\sum_{i\in TEST}\left(TaxAvo_{i,t} - \widehat{TaxAvo}_{i,t}\right)^2}{\sum_{i\in TEST}\left(TaxAvo_{i,t} - \overline{TaxAvo}_{i,t}\right)^2} \quad (2)$$

$$MSE = \frac{1}{n_T}\sum_{i\in TEST}\left(TaxAvo_{i,t} - \widehat{TaxAvo}_{i,t}\right)^2 \quad (3)$$

$$MAE = \frac{1}{n_T}\sum_{i\in TEST}\left|TaxAvo_{i,t} - \widehat{TaxAvo}_{i,t}\right| \quad (4)$$

2.2.3 特征重要性与作用模式

首先,通过考察各个特征的预测能力(即特征重要性) ,筛选关键预测特征. 其次,通过部分依赖图分析关键特征与避税程度的作用模式. 部分依赖图能够展示单个或一组特征变量对模型预测结果的边际效应. 因此,本文利用单变量部分依赖图,观察预测特征与避税程度的关系; 利用双变量联合部分依赖图,展示机器学习模型中两个预测特征之间的交互作用. 特征重要性与部分依赖图的基本思想见在线附录.

3 变量和数据

3.1 变量构造

采用两个基于实际税率的指标衡量公司避税程度: 一是名义税率与实际税率之差(TME) ,通过减去行业均值对其进行行业调整. 二是使用 1 减去实际税率($NETR$) ,实际税率也进行了去行业均值调整. 这两个指标越高,表示公司实际税率越低,避税程度更高. 使用名义税率与实际税率之差是因为我国企业适用不同名义税率[18]; 考虑到各行业所得税政策存在差异,故进行去行业均值调整. 64 个预测变量的具体衡量方式见在线附录.

3.2 数据来源

公司层面数据来源于国泰安数据库,宏观经济数据来源于政府统计年鉴,经济政策不确定性数据来源于 Baker、Bloom 和 Davis 编制的世界各大经济体经济政策不确定性月度指数报告,市场化指数来源于王小鲁等[38],名义税率、最低工资数据来自 WIND 数据库,金税三期工程数据来自国家税务总局及各地政府网站.

因 2008 年后高管特征等数据更完整,且 2020 年后大规模减税降费政策使避税指标较多反映税收减免情况,本文选用 2008 年—2019 年我国 A 股上市公司作为样本. 剔除了金融业公司和数据异常样本,对连续变量在 1% 水平进行缩尾处理,避免极端值影响. 最终得到 2 596 个公司

共13 113个观测值. 数据处理、样本分布及变量描述性统计见在线附录.

4　主要结果与分析

4.1　预测能力分析

各模型在 R^2_{oos}、MSE、MAE 评价指标下的预测结果高度一致,受篇幅限制,本文使用 R^2_{oos} 作为主要评价指标进行分析,见表 2. MSE、MAE 评价结果以及选择 R^2_{oos} 原因见在线附录.

4.1.1　各类特征的预测能力对比

1) 公司内部特征. 运用 OLS 预测 *TME*、*NETR* 时,表 2 列(4)显示,在基准模型上加入避税工具特征后,R^2_{oos} 分别提升 4.19、3.59 个百分点; 加入公司治理特征后,R^2_{oos} 有所降低; 加入其它内部特征后,R^2_{oos} 变化不大; 将 3 类公司内部特征同时加入基准模型得到内部模型,与基准模型相比,R^2_{oos} 分别提升 3.92、3.53 个百分点.

表 2　基于 R^2_{oos} 评价指标的模型预测能力比较(单位: %)

Table 2 Comparison of model predictive power based on R^2_{oos} (Unit: %)

表 2.1　用 *TME* 作为公司避税程度衡量指标(单位: %)								
预测模型	OLS	GBRT	RF	OLS 算法下相比基准模型 R^2_{oos} 的提升	GBRT 算法下相比基准模型 R^2_{oos} 的提升	GBRT 算法相比 OLS 算法 R^2_{oos} 的提升	RF 算法下相比基准模型 R^2_{oos} 的提升	RF 算法相比 OLS 算法 R^2_{oos} 的提升
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
基准模型	8.19	**13.04**	11.73	—	—	4.85	—	3.54
加入公司内部特征								
基准模型 + 避税工具特征	12.38	**17.01**	16.64	4.19	3.97	4.63	4.91	4.26
基准模型 + 公司治理特征	7.57	**12.62**	11.96	−0.62	−0.42	5.05	0.23	4.39
基准模型 + 其他内部特征	8.24	**13.33**	11.50	0.05	0.29	5.09	−0.23	3.26
基准模型 + 所有公司内部特征	12.11	**16.75**	16.31	3.92	3.71	4.64	4.58	4.20
加入公司外部特征								
基准模型 + 宏观层面特征	8.22	**12.58**	11.02	0.03	−0.46	4.36	−0.71	2.80
基准模型 + 行业层面特征	8.39	**13.57**	12.53	0.20	0.53	5.18	0.80	4.14
基准模型 + 外部微观特征	8.30	**13.14**	12.18	0.11	0.10	4.84	0.45	3.88
基准模型 + 所有公司外部特征	8.48	**13.21**	12.91	0.29	0.17	4.73	1.18	4.43
加入全部公司内外部特征								
全因素模型	11.96	**16.88**	16.58	3.77	3.84	4.92	4.85	4.62
表 2.2　用 *NETR* 作为公司避税程度衡量指标(单位: %)								
预测模型	OLS	GBRT	RF	OLS 算法下相比基准模型 R^2_{oos} 的提升	GBRT 算法下相比基准模型 R^2_{oos} 的提升	GBRT 算法相比 OLS 算法 R^2_{oos} 的提升	RF 算法下相比基准模型 R^2_{oos} 的提升	RF 算法相比 OLS 算法 R^2_{oos} 的提升
预测因素与模型	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
基准模型	8.81	**13.70**	12.49	—	—	4.89	—	3.68
加入公司内部特征								
基准模型 + 避税工具特征	12.40	**16.53**	16.21	3.59	2.83	4.13	3.72	3.81
基准模型 + 公司治理特征	8.59	**12.87**	12.23	−0.22	−0.83	4.28	−0.26	3.64
基准模型 + 其他内部特征	8.78	**13.84**	12.33	−0.02	0.14	5.06	−0.16	3.55
基准模型 + 所有公司内部特征	12.34	**16.09**	16.00	3.53	2.39	3.75	3.51	3.66

续表 2

Table 2 Continues

表 2.2 用 *NETR* 作为公司避税程度衡量指标(单位:%)								
预测模型	OLS	GBRT	RF	OLS 算法下相比基准模型 R^2_{oos} 的提升	GBRT 算法下相比基准模型 R^2_{oos} 的提升	GBRT 算法相比 OLS 算法 R^2_{oos} 的提升	RF 算法下相比基准模型 R^2_{oos} 的提升	RF 算法相比 OLS 算法 R^2_{oos} 的提升
加入公司外部特征								
基准模型 + 宏观层面特征	9.03	**13.28**	11.80	0.22	−0.42	4.25	−0.69	2.77
基准模型 + 行业层面特征	8.61	**14.30**	13.37	−0.20	0.60	5.69	0.88	4.76
基准模型 + 外部微观特征	8.91	**13.61**	12.81	0.10	−0.08	4.70	0.32	3.90
基准模型 + 所有公司外部特征	8.87	**14.08**	13.64	0.06	0.38	5.21	1.15	4.77
加入全部公司内外部特征								
全因素模型	12.31	**16.78**	16.52	3.50	3.08	4.47	4.03	4.21

注:本文采取两种突出标注方式,从两个维度进行比较分析:1. 在相同评价指标体系和算法下,关注不同类别模型的预测效果,最佳结果用下划线标注;2. 在相同评价指标体系和模型下,比较不同算法的预测效果,最佳结果以粗体显示.

运用 GBRT 预测 *TME*、*NETR* 时,表 2 列(5)显示,在基准模型上加入避税工具特征后,R^2_{oos} 分别提升 3.97、2.83 个百分点;加入公司治理特征后,R^2_{oos} 有所降低;加入其它内部特征后,R^2_{oos} 稍有提升;内部模型的 R^2_{oos} 相比基准模型分别提升 3.71、2.39 个百分点. 运用 RF 预测结果与 GBRT 接近.

结果表明,无论采用何种算法或避税衡量方式,加入避税工具特征均能够显著提升预测能力,但加入公司治理特征不行,可能是因其数量多且预测能力弱,导致出现过拟合问题,预测表现不佳. 该结果说明避税工具对避税程度有重大影响,能够帮助预测公司避税行为,而公司治理对避税行为的影响较小. 总体来看,公司内部特征对避税程度的预测能力较强.

2) 公司外部特征. 表 2 列(4)显示,运用 OLS 预测 *TME*、*NETR* 时,在基准模型上加入宏观层面特征后,R^2_{oos} 稍有提升;加入行业层面特征后,R^2_{oos} 变化不大;加入外部微观特征后,R^2_{oos} 稍有提升;将 3 类公司外部特征同时加入得到外部模型,与基准模型相比,R^2_{oos} 仅提升 0.29、0.06 个百分点.

运用 GBRT 预测 *TME*、*NETR* 时,表 2 列(5)显示,加入宏观层面特征后,R^2_{oos} 有所下降;加入行业层面特征或外部微观特征后,R^2_{oos} 稍有提升;外部模型相比基准模型,R^2_{oos} 仅提升 0.17、0.38 个百分点. 运用 RF 预测结果与 GBRT 接近. 加入宏观层面特征、行业层面特征或外部微观特征后,R^2_{oos} 相比基准模型变化不大,外部模型的 R^2_{oos} 相比基准模型分别提升 1.18、1.15 个百分点.

该结果说明,公司外部特征对公司避税水平的预测能力较弱.

3) 公司内外部特征的总体预测能力. 将64 个特征加入基准模型得到全模型后预测发现,无论运用何种方法,全模型的 R^2_{oos} 相比基准模型提升 3.08 ~ 4.85 个百分点,说明全模型预测能力相较基准模型有明显提高,但与只加入避税工具特征、内部模型的预测能力相近. 该结果再次说明公司内部的避税工具特征对避税的预测能力较强.

4.1.2 机器学习算法与 OLS 的预测能力对比

表 2 显示,对比 OLS,运用 GBRT 预测 *TME* 时各模型的提升效果在 4.36 ~ 5.18 个百分点,运用 GBRT 预测 *NETR* 时各模型的提升效果在 3.75 ~ 5.69 个百分点;运用 RF 预测 *TME* 时各模型的提升效果在 2.80 ~ 4.62 个百分点,运用 RF 预测 *NETR* 时各模型的提升效果在 2.77 ~ 4.77 个百分点. 可见 GBRT、RF 算法的预测能力远好于 OLS,说明预测特征与公司避税程度的关系很可能是非线性的.

4.2 特征重要性分析

进一步分析各特征的重要性. 对每个模型,分

别运用 GBRT 算法、RF 算法,使用两个避税指标,计算四次特征重要性,对结果取均值进行排名. 表 3 展示了重要性均值位于前十的特征. 表 3 列(1) ~列(3) 列示了全模型、内部模型、外部模型中特征重要性均值的前十名,表 3 列(4) 、列(5) 分别展示了 GBRT 算法、RF 算法下全模型特征重要性的前十名. 各类模型排名前十特征重要性均值、全模型特征重要性数值、排名见在线附录.

全模型和内部模型中,除了基本避税特征外,捐赠支出、应计盈余、员工人数、研发支出这四个避税工具特征都位于前十,说明这些特征对公司避税程度有强预测能力. 该结果再次印证,避税工具变量是预测公司避税程度的重要特征. 而公司治理特征总体预测能力较差,仅董事薪酬相对重要.

表 3　特征重要性排序(前十名)

Table 3 Feature importance rank(top 10)

排序	全模型	内部模型	外部模型	GBRT	RF
	(1)	(2)	(3)	(4)	(5)
1	*Roa*	*Roa*	*Roa*	*Roa*	*Roa*
2	***Donation***	***Donation***	*Lev*	***Donation***	*Rate*
3	*Rate*	*Lev*	*PPE*	*OCF*	***Donation***
4	*Lev*	*OCF*	*Rate*	*Lev*	*Lev*
5	*OCF*	*Rate*	*Size*	*Tacc*	***RD***
6	***Tacc***	***Tacc***	***Ind_Size***	*Rate*	*PPE*
7	***Emp_Size***	*PPE*	***Ind_HHI***	***Emp_Size***	***Emp_Size***
8	*PPE*	***Emp_Size***	***MIOLaw***	*PPE*	***Tacc***
9	***RD***	***RD***	*OCF*	***RD***	*Size*
10	***Board_Pay***	***Board_Pay***	***Media***	***Board_Pay***	*OCF*

注: 表中未加粗的为公司基本特征,加粗的特征为公司内部或外部特征.

外部模型中,行业内企业数量、行业集中度、中介市场和法治环境情况、媒体关注度重要性位列前十,说明在外部特征中,行业的规模和竞争度、当地的法治环境、中介组织发展情况、媒体的关注对公司避税程度有较好预测能力.

4.3　关键特征与公司避税程度的作用模式

运用部分依赖图对关键特征,即捐赠支出、应计盈余、员工人数、研发支出,与避税程度之间的作用模式进行分析. 为增强作用模式的解释性,本文遵循 Shrestha 等[39] 提出的基于机器学习的理论构建范式,通过样本拆分、模式发现、理论解释和理论检验等步骤,将机器学习算法识别出的模式与实证研究方法相结合进行检验,目前该范式也被经济学研究接受应用[40] . 检验结果与部分依赖图表现基本一致,测试集样本回归结果见在线附录.

图 1 为捐赠支出(*Donation*,即企业捐赠金额/税前利润) 与 *TME*、*NETR* 的部分依赖图. 可见,随着捐赠支出的增加,避税程度有所降低. 运用测试集进行检验也发现两者呈显著负相关. 这可能是因为慈善捐赠、缴纳所得税都是企业社会责任的体现,慈善捐赠越多的公司更有社会责任感,避税程度越低[27] .

图 2 为应计盈余(*Tacc*,即总应计盈余/总资产) 与 *TME*、*NETR* 的部分依赖图,灰色直方图展示了 *Tacc* 的分布情况. 在应计盈余的主要取值范围内,随着应计盈余提高,避税程度有所增加. 特别地,当 *Tacc* 小于 -0.05 时,公司避税程度较低且相对平稳,当 *Tacc* 在 [-0.05,0.05] 区间,公司避税程度随着 *Tacc* 增加迅速上升,当 *Tacc* 超过 0.05 后,公司避税程度基本保持不变. 运用测试集进行分样本检验结果也显示 *Tacc* 与 *TME*、*NETR* 仅在 [-0.05,0.05] 区间呈显著正相关,其他区间不显著. 该趋势说明 *Tacc* 在 [-0.05,0.05] 内的盈余管理行为可能是出于避税动机,而当应计盈余超过 0.05 时,盈余管理行为与避税程度的关系不大.

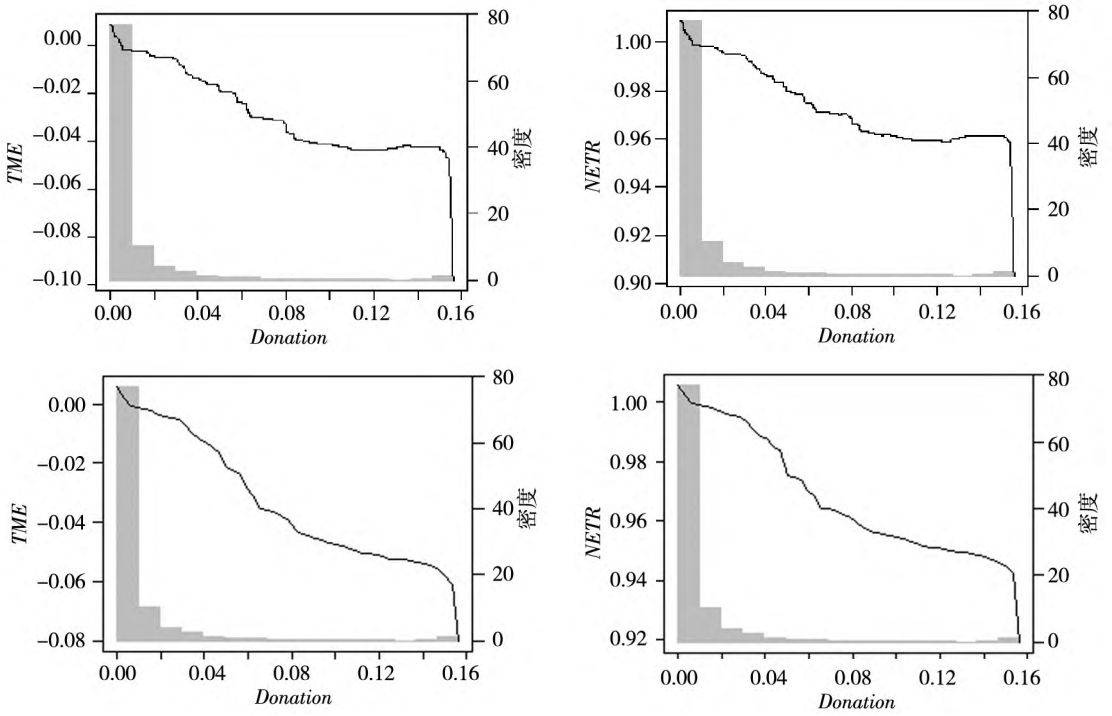

图 1　捐赠支出的部分依赖图

Fig. 1 Partial dependence plots of donation expenditure

注: 图中黑色实线为捐赠支出的部分依赖图,灰色直方图为捐赠支出指标的分布直方图,横轴代表捐赠支出,纵轴左侧代表公司避税程度,纵轴右侧代表捐赠支出指标的密度。左上图以 *TME* 为公司避税程度指标,运用 GBRT 算法; 右上图以 *NETR* 为公司避税程度指标,运用 GBRT 算法; 左下图以 *TME* 为公司避税程度指标,运用 RF 算法; 右下图为以 *NETR* 为公司避税程度指标,运用 RF 算法,下同.

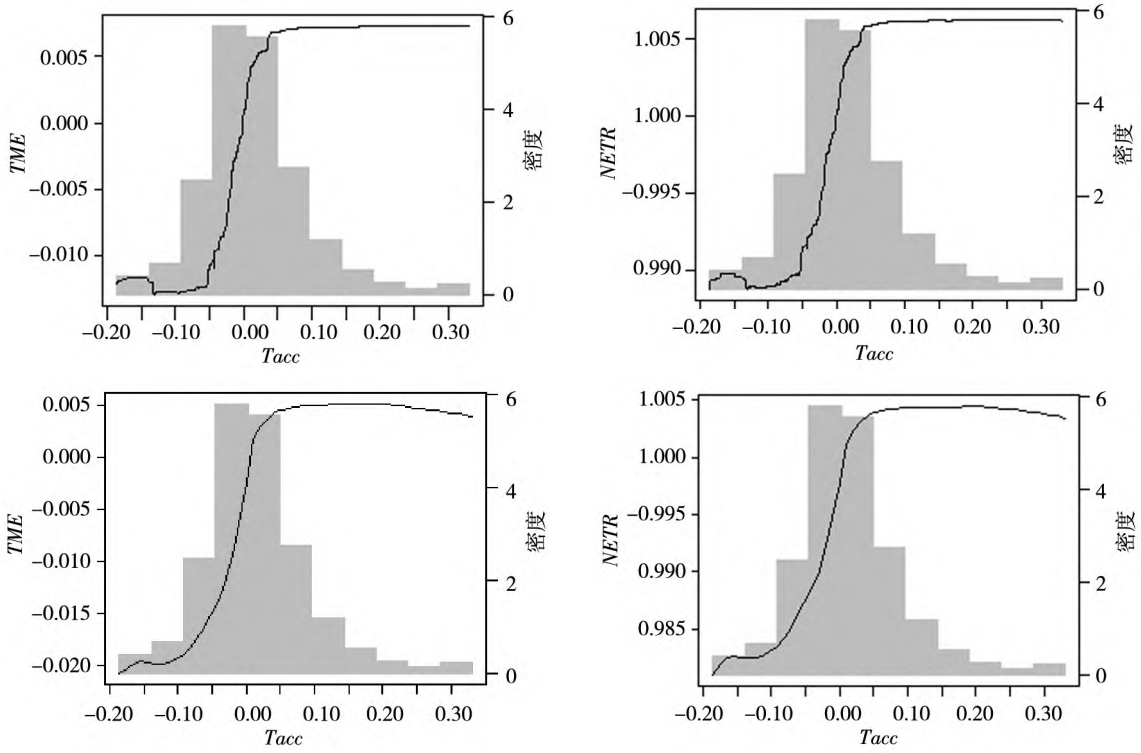

图 2　应计盈余的部分依赖图

Fig. 2 Partial dependence plots of total accruals

注: 图中黑色实线为应计盈余的部分依赖图,灰色直方图为应计盈余指标的分布直方图,横轴代表应计盈余,纵轴左侧代表公司避税程度,纵轴右侧代表应计盈余指标的密度.

图 3 为员工人数(*Emp_Size*,即公司员工人数,单位: 万人) 的部分依赖图,灰色直方图展示了 *Emp_Size* 的分布情况. 在员工人数的主要取值范围内,随着员工人数的增加,避税程度呈下降趋势. 这可能是因为当员工人数越多时,劳动力成本形成的"就业税盾"更大. 而当 *Emp_Size* 在 0 ~ 0.04 左右,即员工人数小于 400 人时,避税程度随员工人数增加有小幅度提升. 运用测试集数据进行分样本检验也呈现相同趋势. 这可能是因为员工数量小于 400 左右的基本是中小企业,员工越多、劳动力成本越高时,竞争力差、风险大的中小企业更有动机通过避税来节约现金流.

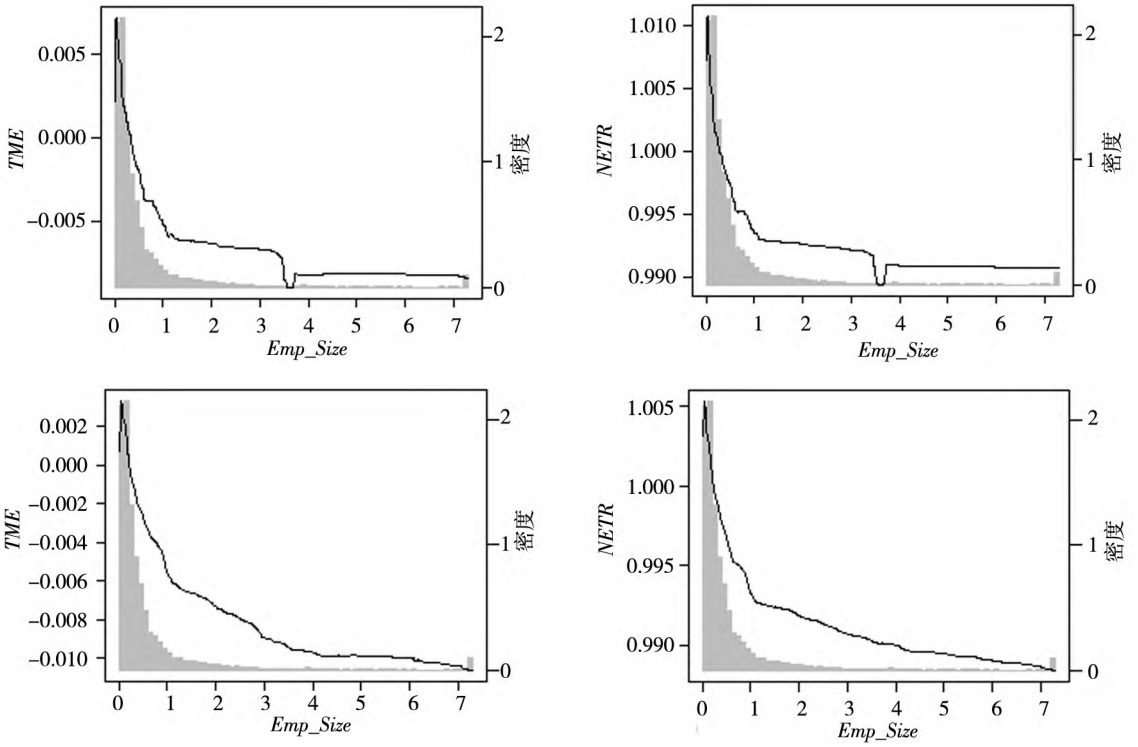

图 3　员工人数的部分依赖图

Fig. 3 Partial dependence plots of the number of employees

注: 图中黑色实线为员工人数的部分依赖图,灰色直方图为员工人数指标的分布直方图,横轴代表员工人数,纵轴左侧代表公司避税程度指标,纵轴右侧代表员工人数指标的密度.

图 4 为研发支出(*RD*,即研发支出 / 总资产) 的部分依赖图,灰色直方图展示了 *RD* 的分布情况. 在研发支出的主要取值范围内,避税程度随研发支出的提升而提高. 特别地,当 *RD* 小于 0.02 时,随着研发支出的提高,避税程度迅速提升; 当 *RD* 在 [0.02,0.04] 内,随着研发支出的提高,避税程度也有所提升. 该正相关可能源于我国在企业研发创新方面的税收优惠,如研发费用加计扣除等,很多企业会利用这些优惠,操纵研发支出以达到避税目的[15]. 而当 *RD* 超过 0.04 后,GBRT 算法下两者正相关不明显,甚至在 RF 算法下避税程度反而随研发支出增加稍有减少,这可能是因为大量研发的目的是创新而非避税,故此时避税程度不会随着研发支出的增加而大幅提升. 运用测试集数据进行检验也呈现类似趋势.

4.4　关键特征的交互作用分析③

图 5 是捐赠支出(*Donation*) 和总资产收益率(*Roa*,即息税前利润 / 总资产) 的联合部分依赖图. *Donation* 的主要取值范围是(0.00,0.13] (约占全体样本的 77.42%) ,*Roa* 的主要取值范围是

③　由于避税工具特征重要性高,本文重点关注避税工具特征。鉴于虚拟变量交互作用的信息量有限,未将其纳入考量,最终绘制 1 344 个(7 × 48 × 4) 双变量联合部分依赖图. 为简洁起见仅展示其中重要且有意义的交互作用分析,并仅展示 GBRT 的部分依赖图,RF 的部分依赖图与 GBRT 一致,详见在线附录.

[0.03,0.15](约占全体样本的83.83%),可见在*Donation*和*Roa*集中分布区间,随着捐赠支出增加、总资产收益率提高,避税程度增加.且当*Roa*较高时,捐赠支出与避税之间的负相关性更弱.这可能是因为随着盈利能力增强,避税行为给公司节约的现金流越多、边际收益越大,导致即使是愿意承担社会责任的公司也会有更强的动机避税.

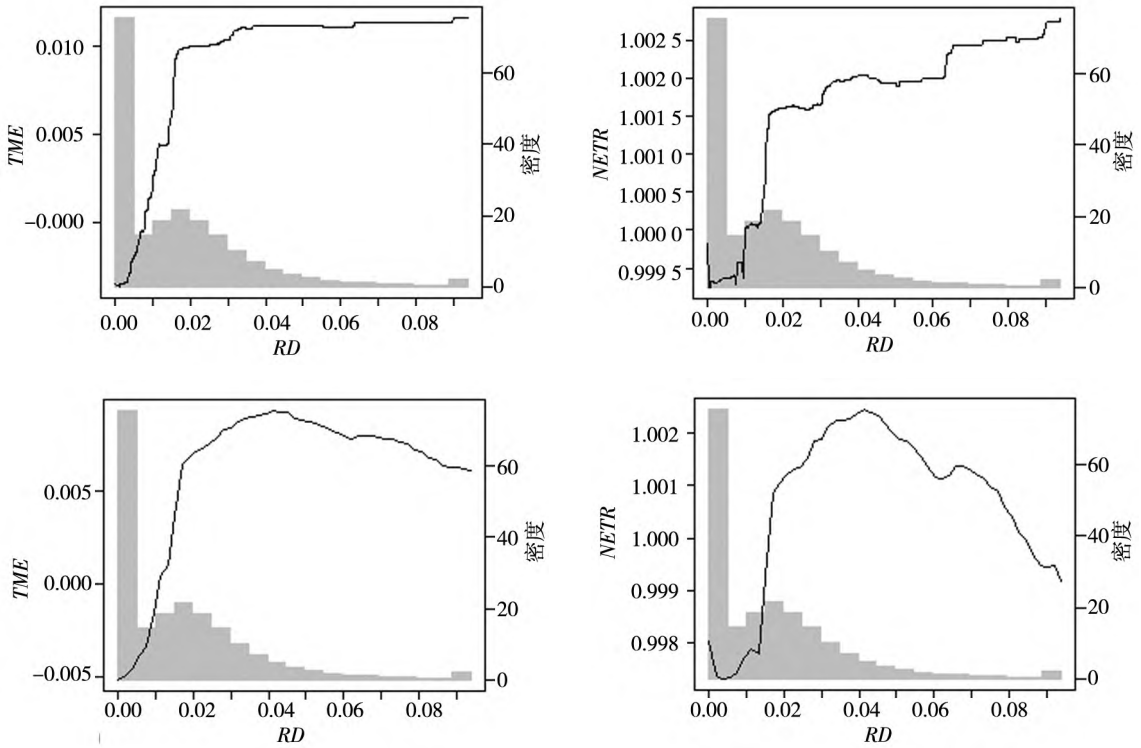

图 4　研发支出的部分依赖图

Fig. 4 Partial dependence plots of R&D expenditure

注:图中黑色实线为研发支出的部分依赖图,灰色直方图为研发支出指标的分布直方图,横轴代表研发支出,纵轴左侧代表公司避税程度指标,纵轴右侧代表研发支出指标的密度.

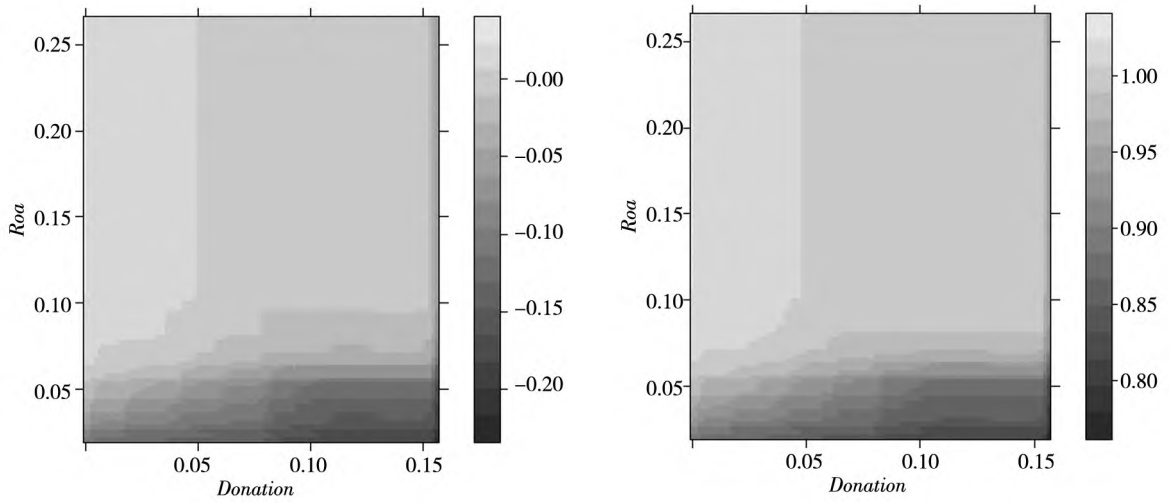

图 5　企业捐赠规模和总资产收益率的部分依赖图

Fig. 5 Partial dependence plots of *Donation* and *Roa*

注:横轴代表企业捐赠规模,纵轴代表总资产收益率,右侧色卡代表公司避税程度,颜色越浅,公司避税程度越高.

图6是应计盈余(*Tacc*)和监事会规模(*Sup_Size*,即监事会总人数的自然对数)的联合部分依赖图. *Tacc* 的主要取值范围是 [-0.10,0.20](约占全体样本的91.58%),*Sup_Size* 的主要取值范围是 [1.00,1.90](约占全体样本的93.13%). 可见在 *Tacc* 和 *Sup_Size* 集中分布区间,随着应计盈余增加、监事人数减少,公司避税程度升高. 且在监事会规模较小时,避税程度随应计盈余增加而上升的趋势更明显,这可能是因为监事会规模越大、监督作用越强,公司利用盈余管理来进行避税的难度越大.

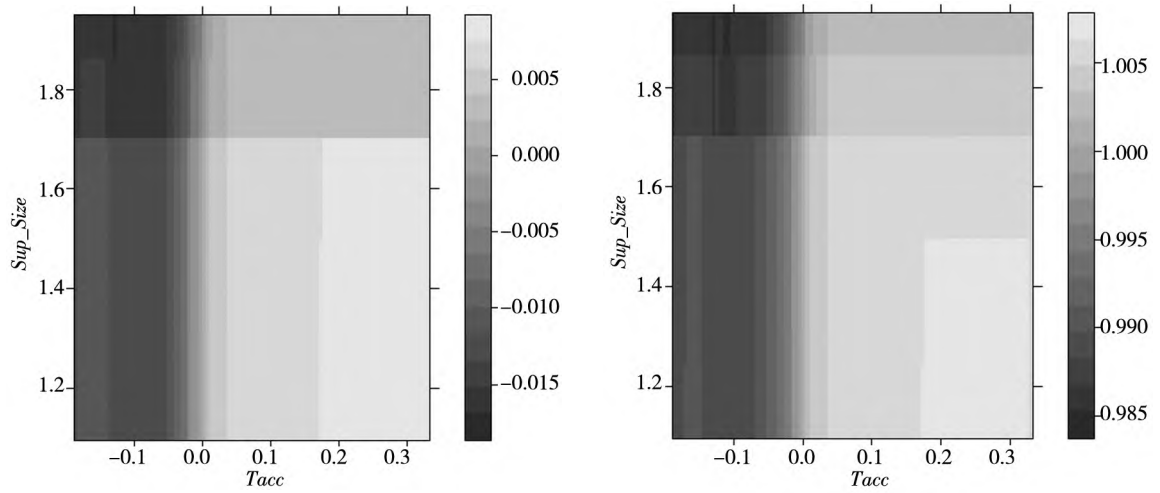

图6　应计盈余和监事人数的部分依赖图

Fig. 6 Partial dependence plots of *Tacc* and *Sup_Size*

注: 横轴代表应计盈余,纵轴代表监事人数,右侧色卡代表公司避税程度,颜色越浅,公司避税程度越高.

图7是员工人数(*Emp_Size*)和CEO任期(*CEO_Tenure*,即按月计算的CEO任职时间加一的自然对数)的联合部分依赖图. *Emp_Size* 主要取值范围是 [0.04,2.45](约占全体样本的89.86%),*CEO_Tenure* 主要取值范围是 [1.60,4.80](约占全体样本的89.32%),可见在 *Emp_Size* 和 *CEO_Tenure* 集中分布区间,基本上随着员工人数减少、CEO任期增加,公司避税程度增加. 且公司避税程度随着员工人数的增加而减少的现象在CEO任期较短的公司中更明显,这可能是因为任期较长的CEO对避税技巧了解更全面,能够使用更隐蔽的方式避税,使得员工人数增加引起的外部关注和监督抑制避税的作用更弱.

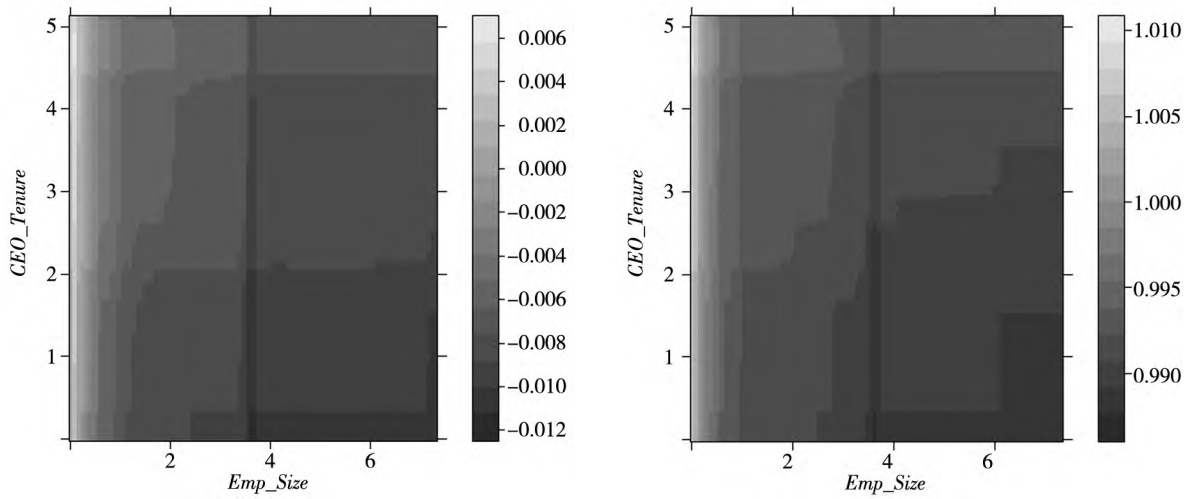

图7　员工人数和CEO任期的部分依赖图

Fig. 7 Partial dependence plots of *Emp_Size* and *CEO_Tenure*

注: 横轴代表员工人数,纵轴代表CEO任期,右侧色卡代表公司避税程度,颜色越浅,公司避税程度越高.

图 8 是研发支出(*RD*)和机构投资者持股比例(*Inst_Own*,即机构投资者持股数/总股数)的联合部分依赖图. *RD* 的主要取值范围是[0.00,0.04](约占全体样本的 90.31%),*Inst_Own* 的主要取值范围是[0.10,0.80](约占全体样本的 93.13%),可见在 *RD* 和 *Inst_Own* 集中分布区间,随着研发投入和机构投资者持股比例的增加,避税程度会升高. 且在机构投资者持股比例较高时,公司避税程度随研发支出增加而上升的趋势更明显. 这可能是因为利用研发创新税收优惠政策进行避税的风险较低,而机构投资者会鼓励高管利用低风险的避税工具以节约公司现金流.

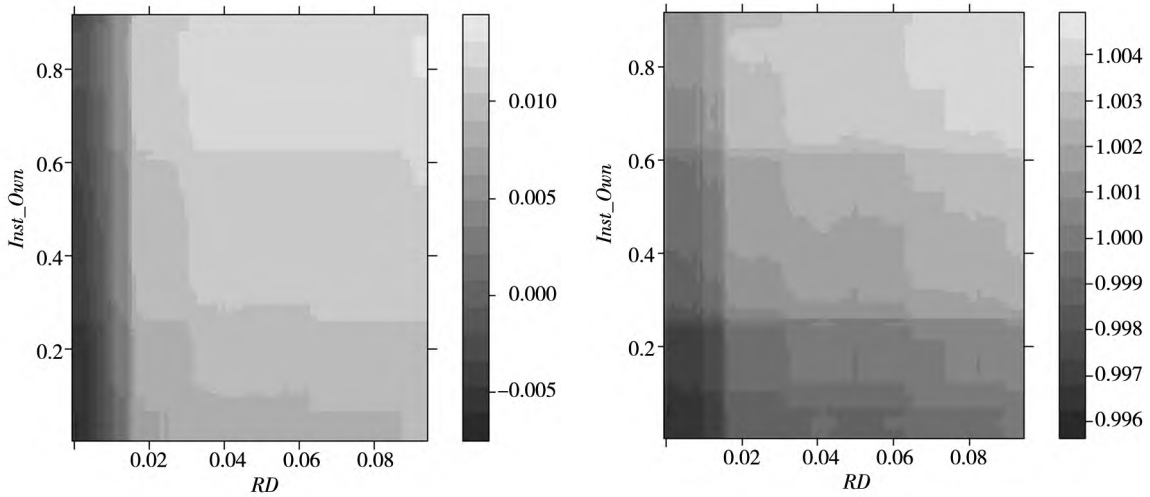

图 8　研发投入和机构投资者持股比例的部分依赖图

Fig. 8 Partial dependence plots of *RD* and *Inst_Own*

注: 横轴代表研发投入,纵轴代表机构投资者持股比例,右侧色卡代表公司避税程度,颜色越浅,避税程度越高.

5　拓展性分析

本文进一步考察公司避税预测模型的效果和应用范围. 首先,关注模型在税务稽查情境中的性能. 通过将回归问题转化为分类问题和排序问题,本文根据公司避税程度高低定义了高避税公司和低避税公司,计算了召回率(*Recall*)、精确率(*Precision*)和归一化折损累积收益指标(NDCG@K)指标,进一步揭示了模型在正确识别高避税公司和减少误判低避税公司方面的实际表现. 结果表明,与类似研究比较,本文的公司避税预测模型在各方面均表现更出色,验证了其在税务稽查情境中的有效性. 其次,本文关注公司未来期避税行为,通过利用第 $T-1$ 期 ~ 第 $T-3$ 期的预测特征水平值和相应的预测特征变化量,对第 T 期公司避税程度进行预测. 本文发现,随着时间跨度的增加,过去期特征水平值和变化量在预测未来避税行为时的信息价值逐渐减弱,并且避税工具特征同样是识别公司未来期避税行为的关键预测特征. 受篇幅限制,结果见在线附录.

6　稳健性检验

本文进行了 9 个稳健性检验确保结果可靠: 1) 更换 GBRT、RF 算法中的重要参数重新预测; 2) 重新随机抽样划分训练集和测试集; 3) 换用 LASSO 和 XGBoost 算法进行预测; 4) 使用行业调整的账面 - 税收差异作为避税衡量指标重新预测; 5) 考虑到税收征管实践中更关注公司避税的等级分类,我们划分高避税、低避税公司,利用梯度提升决策树和随机森林算法进行分类预测; 6) 采用"减法"策略,在全模型上分别剔除公司内外部 6 类特征重新预测; 7) 利用过去三年数据作为训练集建立模型进行滚动预测; 8) 利用包含缺失值的样本重新进行预测; 9) 设置随机种子数为 2 ~ 101,进行 100 次重复预测. 以上结果见在线附录.

7　结束语

本文运用梯度提升回归树算法、随机森林算法,考察 64 个公司内外特征对避税程度的预测能力. 研究发现: 1) 避税工具特征对避税程度预测能力强,尤其是捐赠支出、应计盈余、员工人数、研发支出的预测能力尤为突出,而公司外部特征的预测能力普遍较弱; 2) 考察单变量部分依赖图可知,对于大部分样本来说,避税程度会随着捐赠支出、员工人数的增加而减少,随着应计盈余、研发支出的增加而增加. 而少数情况下,应计盈余、员工人数、研发支出与公司避税程度之间的关系则有所不同. 具体地,当应计盈余占总资产比值极端化(小于 - 0. 05 或大于 0. 05) 时,应计盈余与公司避税之间相关性不强; 当员工人数极少(少于 400 人) 时,公司避税程度反而随着员工人数的增加而上升; 当研发支出占总资产比例过高(超过 4%) 时,避税程度随着研发支出的增加而增加的现象不明显; 3) 考察双变量部分依赖图可知,捐赠支出与总资产收益率、应计盈余与监事会规模、员工人数与 CEO 任期、研发支出与机构投资者持股比例存在交互作用. 具体地,捐赠金额为 0 或很低且 *ROA* 很高的企业避税程度较高; 应计盈余占总资产比值达到 0. 05 以上且监事人数较少的企业避税程度较高; 员工人数在 400 人左右且 CEO 任期长的公司避税程度较高; 研发支出与总资产比值在 4% 左右且机构投资者持股比例高的公司避税程度较高; 4) 即使是预测未来期避税程度或数据缺失等复杂情境,机器学习模型的预测效果依然远好于线性模型.

本文结果对政策制定者、税收征管部门、智慧税务系统建设均有政策启示. 对政策制定者而言,本文发现六类特征中避税工具特征对于公司避税程度预测能力最强,故要抑制公司避税,最重要的是完善相关法规政策,降低避税工具的可得性. 对税收征管部门来说,研究结果说明,可重点关注捐赠金额为 0 或很低且 *Roa* 很高、应计盈余占总资产比值大于 0. 05 且监事人数少、员工人数约 400 人且 CEO 任期长、研发支出与总资产比值约 4% 且机构投资者持股比例高的企业; 在智慧税务系统的建设中,本文发现加入公司避税工具特征的机器学习模型能够更准确地预测避税程度,因此建议采用纳入公司避税工具指标的机器学习模型评估公司税收风险.

本文对大数据与人工智能方法在税务管理中的实践也有一定参考价值. 第一,税务实践中常有数据缺失问题,部分机器学习算法如 GBRT 能自动处理缺失数据且预测效果较好,因此数据缺失时可考虑使用此类机器学习算法; 第二,考虑到避税指标的多样性和机器学习模型存在一定随机性,实践中可采取同时建立多个模型、对多个模型取均值或取最优的方式来增强结果的稳健性和可信度,并及时结合实际稽查结果评估模型效果,持续优化模型.

参 考 文 献:

[1] 姬颜丽, 王文清. 大数据背景下税收风险识别精准度存量研究——基于机器学习的视角 [J]. 财政研究, 2020, (9): 119 - 129.
Ji Yanli, Wang Wenqing. The stock of research on accurate identification of tax risk under the background of big data technology: Based on machine learning [J]. Public Finance Research, 2020, (9): 119 - 129. (in Chinese)

[2] Cai H, Liu Q. Competition and corporate tax avoidance: Evidence from Chinese industrial firms [J]. Economics Journal, 2009, 119(537): 764 - 795.

[3] Armstrong C S, Blouin J L, Larcker D F. The incentives for tax planning [J]. Journal of Accounting and Economics, 2012, 53(1 - 2): 391 - 411.

[4] Armstrong C S, Blouin J L, Jagolinzer A D, et al. Corporate governance, incentives, and tax avoidance [J]. Journal of Ac-

counting and Economics, 2015, 60(1): 1 – 17.

[5] 洪永淼, 汪寿阳. 大数据如何改变经济学研究范式? [J]. 管理世界, 2021, 37(10): 40 – 55, 72, 56.
Hong Yongmiao, Wang Shouyang. How is big data changing economic research paradigms? [J]. Management World, 2021, 37(10): 40 – 55, 72, 56. (in Chinese)

[6] 刘景江, 郑畅然, 洪永淼. 机器学习如何赋能管理学研究? [J]. 管理世界, 2023, (9): 191 – 215.
Liu Jingjiang, Zheng Changran, Hong Yongmiao. How can machine learning empower management research? [J]. Management World, 2023, (9): 191 – 215. (in Chinese)

[7] 任之光, 李 金, 赵海川, 等. 大数据驱动的管理决策: 研究范式与发展领域 [J]. 管理科学学报, 2023, 26(10): 152 – 158.
Ren Zhiguang, Li Jin, Zhao Haichuan, et al. Big data-driven management and decision sciences: The research paradigm and directions [J]. Journal of Management Sciences in China, 2023, 26(10): 152 – 158. (in Chinese)

[8] Goulet Coulombe P, Leroux M, Stevanovic D, et al. How is machine learning useful for macroeconomic forecasting? [J]. Journal of Applied Econometrics, 2022, 37(5): 920 – 964.

[9] Gu S, Kelly B, Xiu D. Empirical asset pricing via machine learning [J]. Review of Financial Studies, 2020, 33(5): 2223 – 2273.

[10] 李 斌, 龙 真. 中国股票市场可预测性研究: 基于机器学习的视角 [J]. 管理科学学报, 2023, 26(10): 138 – 158.
Li Bin, Long Zhen. Return predictability in the Chinese stock markets: A machine learning perspective [J]. Journal of Management Sciences in China, 2023, 26(10): 138 – 158. (in Chinese)

[11] 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): 119 – 139.
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): 119 – 139. (in Chinese)

[12] 张学勇, 施 懿. 基于元学习的财务舞弊识别研究 [J]. 管理科学学报, 2023, 26(10): 95 – 113.
Zhang Xueyong, Shi Yi. Financial fraud recognition model based on meta-learning [J]. Journal of Management Sciences in China, 2023, 26(10): 95 – 113. (in Chinese)

[13] 李建平, 孙 灏, 常闫芃, 等. 考虑审计要素多重语义关联的财务欺诈识别 [J]. 管理科学学报, 2024, 27(3): 58 – 70.
Li Jianping, Sun Hao, Chang Yanpeng, et al. Financial statement fraud identification considering the multiple-dimensional semantic associations of auditing elements [J]. Journal of Management Sciences in China, 2024, 27(3): 58 – 70. (in Chinese)

[14] Dyreng S D, Hanlon M, Maydew E L. The effects of executives on corporate tax avoidance [J]. The Accounting Review, 2010, 85(4): 1163 – 1189.

[15] Chen Z, Liu Z, Serrato J C S, et al. Notching R&D investment with corporate income tax cuts in China [J]. American Economic Review, 2021, 111(7): 2065 – 2100.

[16] Dyreng S D, Hanlon M, Maydew E L. Long-run corporate tax avoidance [J]. The Accounting Review, 2008, 83(1): 61 – 82.

[17] Balakrishnan K, Blouin J L, Guay W R. Tax aggressiveness and corporate transparency [J]. The Accounting Review, 2019, 94(1): 45 – 69.

[18] 叶康涛, 刘 行. 公司避税活动与内部代理成本 [J]. 金融研究, 2014, (9): 158 – 176.
Ye Kangtao, Liu Hang. Corporate tax avoidance and internal agency costs [J]. Journal of Financial Research, 2014, (9): 158 – 176. (in Chinese)

[19] Wilson R. An examination of corporate tax shelter participants [J]. The Accounting Review, 2009, 84 (3): 969 – 999.

[20] Armstrong C S, Blouin J L, Larcker D F. The incentives for tax planning [J]. Journal of Accounting and Economics, 2012, 53(1-2): 391-411.

[21] Phillips J, Pincus M, Rego S. Earnings management: New evidence based on deferred tax expense [J]. The Accounting Review, 2003, 78(2): 491-521.

[22] Desai M A, Dharmapala D. Corporate tax avoidance and high-powered incentives [J]. Journal of Financial Economics, 2006, 79(1): 145-179.

[23] Desai M A, Dharmapala D. Corporate tax avoidance and firm value [J]. Review of Economics and Statistics, 2009, 91(3): 537-546.

[24] Chen K P, Chu C C. Internal control versus external manipulation: A model of corporate income tax evasion [J]. The Rand Journal of Economics, 2005, 36(1): 151-164.

[25] Slemrod J. The economics of corporate tax selfishness [J]. National Tax Journal, 2004, (57): 877-899.

[26] Hoi C K, Wu Q, Zhang H. Is corporate social responsibility (CSR) associated with tax avoidance? Evidence from irresponsible CSR activities [J]. The Accounting Review, 2013, 88(6): 2025-2059.

[27] Lanis R, Richardson G. Corporate social responsibility and tax aggressiveness: An empirical analysis [J]. Journal of Accounting and Public Policy, 2012, 31(1): 86-108.

[28] Bradshaw M, Liao G, Ma M S. Agency costs and tax planning when the government is a major shareholder [J]. Journal of Accounting and Economics, 2019, 67(2-3): 255-277.

[29] Hoopes J L, Mescall D, Pittman J A. Do IRS audits deter corporate tax avoidance? [J]. The Accounting Review, 2012, 87(5): 1603-1639.

[30] **范子英, 田彬彬. 税收竞争、税收执法与企业避税 [J]. 经济研究, 2013, 48(9): 99-111.**
Fan Ziying, Tian Binbin. Tax competition, tax enforcement and tax avoidance [J]. Economic Research Journal, 2013, 48(9): 99-111. (in Chinese)

[31] **李广众, 贾凡胜. 财政层级改革与税收征管激励重构——以财政"省直管县"改革为自然实验的研究 [J]. 管理世界, 2020, 36(8): 32-50.**
Li Guangzhong, Jia Fansheng. Reform of fiscal hierarchy and reconstruction of tax enforcement incentive: Evidence from the "Province-managing-county" fiscal reform in China [J]. Management World, 2020, 36(8): 32-50. (in Chinese)

[32] Chen S, Schuchard K, Stomberg B. Media coverage of corporate taxes [J]. The Accounting Review, 2019, 94(5): 83-116.

[33] Chen S, Huang Y, Li N, et al. How does quasi-indexer ownership affect corporate tax planning? [J]. Journal of Accounting and Economics, 2019, 67(2-3): 278-296.

[34] Desai M A, Dyck A, Zingales L. Theft and taxes [J]. Journal of Financial Economics, 2007, 84(3): 591-623.

[35] **李　航. 统计学习方法(第 2 版) [M]. 北京: 清华大学出版社, 2019.**
Li Hang. Statistical Learning Methods [M]. Beijing: Tsinghua University Press, 2019. (in Chinese)

[36] Shmueli G. To explain or to predict? [J]. Statistical Science, 2010, 25(3): 289-310.

[37] **洪永淼, 汪寿阳. 人工智能新近发展及其对经济学研究范式的影响 [J]. 中国科学院院刊, 2023, 38(3): 353-357.**
Hong Yongmiao, Wang Shouyang. Impacts of cutting-edge artificial intelligence on economic research paradigm [J]. Bulletin of Chinese Academy of Sciences, 2023, 38(3): 353-357. (in Chinese)

[38] **王小鲁, 胡李鹏, 樊　纲. 中国分省份市场化指数报告(2021) [M]. 北京: 社会科学文献出版社, 2021 年.**
Wang Xiaolu, Hu Lipeng, Fan Gang. Marketization Index of China's Provinces: NERI Report 2021 [M]. Beijing: Social Sciences Academic Press (China), 2021. (in Chinese)

[39] Shrestha Y R, He V F, Puranam P, et al. Algorithm supported induction for building theory: How can we use prediction models to theorize? [J]. Organization Science, 2021, 32(3): 856-880.

[40] Ludwig J, Mullainathan S. Machine learning as a tool for hypothesis generation [J]. Quarterly Journal of Economics, 2024, 139(2): 751 – 827.

Predicting corporate tax avoidance based on machine learning

PENG Zhang[1], *LU Yao*[2], *NIU Mei-ling*[2], *ZHOU Xin-yi*[2]

1. School of Public Finance and Taxation, Central University of Finance and Economics, Beijing 102206, China;
2. School of Economics and Management, Tsinghua University, Beijing 100084, China

Abstract: This study utilizes Gradient Boosting Regression Trees and Random Forest Agorithms to explore the predictive power of 64 features of six dimensions, which previous studies have identified as determinants of corporate tax avoidance. The results show that features of tax shelters have strong predictive power. Specifically, donation expenditure, total accruals, employment, and R&D expenditure emerge as the most influential predictors. However, external features have limited predictive power. Examining the relationships between key features and tax avoidance, this paper finds that, in general, donation expenditure and employment are negatively related to tax avoidance levels, while total accruals and R&D expenditure are positively associated with tax avoidance levels. However, in a few cases, employment is positively correlated with tax avoidance, and the correlations between tax avoidance and total accruals or R&D expenditure are weak. Interactive effectsexist between donation expenditure and return on total assets, total accruals and supervisory board size, employment and CEO tenure, and R&D expenditure and institutional ownership. Tree-based machine learning algorithms have greater predictive power than linear regressions. Our results suggest that more attention should be paid to features of tax shelters in tax collection procedures and the design of intelligent tax systems, and that restricting tax shelters is crucial in policy making.

Key words: corporate tax avoidance prediction; machine learning; tax shelters