

《基于机器学习的公司避税预测研究》

在线附录

本附录作为正文内容的补充，方便读者更深入地了解本文的研究方法和结果。

以下是附录中各部分内容的简要介绍：

1. 附录 1 说明了公司避税预测特征的选择依据。
2. 附录 2 介绍了本文算法选择依据、算法基本思路、特征重要性和部分依赖图的计算方式和基本思想。
3. 附录 3 展示了本文数据处理过程、所使用的样本分布情况以及预测变量的详细定义和描述性统计情况。
4. 附录 4 展示了使用 MSE、MAE 作为评价指标时的预测结果。
5. 附录 5 展示了各类模型的特征重要性数值和排序。
6. 附录 6 展示了关键特征与公司避税程度的作用模式验证内容。
7. 附录 7 展示了基于随机森林算法的双变量联合部分依赖图。
8. 附录 8 展示了在分类排序评价体系下，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表现。
9. 附录 9 展示了使用公司预测特征对未来期避税行为预测的结果。
10. 附录 10 展示了本文稳健性检验的结果。

目录

附录 1：公司避税预测特征选择依据	1
1. 公司内部特征	1
2. 公司外部特征	2
附录 2：算法选择、算法介绍、特征重要性与作用模式分析的基本思想	3
1. 算法选择依据	3
2. 算法介绍与参数说明	3
3. 特征重要性与作用模式分析的基本思想	4
附录 3：样本分布、变量定义与描述性统计	5
附表 A-1 样本分布	5
附表 A-2 变量定义及计算方式	5
附表 A-3 描述性统计	7
附录 4：MSE、MAE 评价指标的预测结果	9
附表 A-4 MSE、MAE 评价指标下模型预测能力结果	9
附录 5：各类模型的特征重要性数值和排序	10
附表 A-5 加入各子类特征的模型中重要性排名前十的特征	10
附表 A-6 加入各类特征模型的特征重要性均值（前十名）	10
附表 A-7 全模型所有特征重要性排序及特征重要性数值	11
附录 6：关键特征对公司避税程度的作用模式验证	15
附表 A-8 捐赠支出对公司避税程度的回归结果	15
附表 A-9 应计盈余对公司避税程度的回归结果	15
附表 A-10 员工人数对公司避税程度的回归结果	15
附表 A-11 研发支出对公司避税程度的回归结果	16
附录 7：基于随机森林算法的双变量联合部分依赖图	17
附图 A-1 捐赠支出和总资产收益率的部分依赖图	17
附图 A-2 应计盈余和监事会规模的部分依赖图	18
附图 A-3 员工人数和 CEO 任期的部分依赖图	19
附图 A-4 研发支出和机构投资者持股比例的部分依赖图	20
附录 8：分类排序评价体系下的预测能力	21
附表 A-12 分类评价体系下的预测能力结果	22
附录 9：使用公司预测特征水平值对未来期避税行为预测	24
附表 A-13 加入特征变化量的未来期避税行为预测结果（单位：%）	25
附表 A-14 未来期避税行为预测结果（单位：%）	26
附录 10：稳健性检验结果	27
1. 更换模型参数	27
附表 A-15 更换 GBRT 算法参数的样本外 R 方（ R^2_{oos} ）	27

附表 A-16 更换 RF 算法参数的样本外 R 方 (R^2_{oos})	27
2. 更换样本	28
附表 A-17 更换训练集和测试划分方式后的样本外 R 方 (R^2_{oos})	28
3. 更换算法	28
附表 A-18 更换算法后的样本外 R 方 (R^2_{oos})	29
4. 更换公司避税程度衡量指标	29
附表 A-19 运用 <i>BTD</i> 作为避税衡量指标的样本外 R 方 (R^2_{oos})	29
5. 预测高避税公司与低避税公司	30
附表 A-20 预测是否为高避税公司的 AUC 值	30
6. 采用“减法”方式构造预测模型	31
附表 A-21 采用“减法”方式构造预测模型的样本外 R 方 (R^2_{oos})	31
7. 以三年期进行滚动预测	32
附表 A-22 三年期滚动预测的样本外 R 方 (R^2_{oos}) 均值	32
8. 考虑预测变量存在缺失的情况	32
附表 A-23 包含预测变量缺失样本进行预测的样本外 R 方 (R^2_{oos})	33
9. 考虑机器学习算法的随机性	33
附表 A-24 100 次模型预测能力结果	33
参考文献	35

附录 1：公司避税预测特征选择依据

我们在已有文献的基础上，系统梳理了公司避税的决定因素，并基于此选择预测特征。

首先，我们选取了 7 个基本特征，包括公司规模、公司年龄、名义税率这 3 个公司基本属性变量，以及总资产收益率、经营净现金流、固定资产比例、杠杆率这 4 个公司财务变量。名义税率直接决定了公司税负大小，而公司规模、年龄以及财务特征与盈利状况、融资约束、现金流情况紧密相关，是影响公司避税动机最为基本的因素，众多公司避税研究也将上述变量作为控制变量加入模型，故我们认为这些变量属于公司避税的基本预测特征。

接下来，我们从公司内部出发，将内部能够影响其避税决策的特征分成三类，即避税工具特征、公司治理特征、其他内部特征。我们再从公司外部出发，将外部能够影响避税决策的特征也分为三类，即宏观层面特征、行业层面特征、外部微观特征。

总体来说，本文将公司避税行为的预测特征分为七个类别：一是基本预测特征，共 7 个；二是避税工具特征，共 12 个；三是公司治理特征，共 29 个；四是其他内部特征，共 2 个；五是宏观层面特征，共 15 个；六是行业层面特征，共 3 个；七是微观层面特征，共 3 个。变量的具体定义和计算方式见附表 A-2。

接下来是公司内外部这六类变量的选取依据。

1. 公司内部特征

（1）避税工具特征。避税工具的供给和成本是决定避税程度的重要因素^[1]，当避税工具更多、成本更低时，公司更有可能避税。我们选取了捐赠支出、是否有企业捐赠金额数据、员工人数、员工薪酬、应计盈余、可操纵应计盈余、关联交易、是否有关联交易数据、研发支出、是否有研发支出数据、是否有外币业务、是否有注册地转移情况这 12 个变量。选取捐赠支出、是否有企业捐赠金额数据是由于企业慈善捐赠能够享受税收优惠，可能被用于避税。选取员工人数、员工薪酬指标是由于员工工资可以税前扣除，存在“就业税盾”^[2,3]，且部分企业可能通过虚列员工、虚发工资进行避税。选取应计盈余、可操纵应计盈余是因为这两个指标都是经典的盈余管理指标，而盈余管理也是公司避税的主要工具。选取关联交易、是否有关联交易数据是由于企业可能会通过关联交易进行避税^[4,5]。选取研发支出、是否有研发支出数据是由于我国存在很多研发创新的税收优惠，利用相关的税收优惠政策企业可以达到避税的目的。选取是否有外币业务是因为企业可能将利润转移到国外避税地实现避税^[6]。选取是否有注册地转移情况是因为企业可能利用地区间的税收差异，通过将注册地迁移至有税率优惠的地区进行避税^[7]。

（2）公司治理特征。大量文献还显示公司治理与避税行为息息相关，代理问题是影响企业避税行为的重要因素^[8,9]，因此我们加入了企业性质、股权结构、董事会治理、高管治理、监事会治理、代理成本这些公司治理特征。

在企业性质方面，我们选择了是否国企和是否家族企业这 2 个变量。选择是否国企是因为国有企业肩负着帮助政府实现经济社会目标的重要责任，缴税是其履行责任的重要任务，故国有企业的避税动机更小，相关研究也表明国有企业避税程度更低^[10-12]。选择是否家族企业是因为家族企业所有者更担心避税可能造成的非税成本，比如非家族成员股东的估值降低、监管部门的惩罚、家族名誉的损失，这导致家族企业避税动机更弱^[13]。

在股权结构方面，我们选择了是否发行 B 股 H 股、股权制衡度、第一大股东持股比例、前十大股东持股比例、股权集中度、机构投资者持股比例这 6 个变量。选择是否发行 B 股 H 股是因为大量文献^[14]显示境外投资者可以帮助改善公司治理，从而影响企业避税行为。选择股权制衡度、第一大股东持股比例、前十大股东持股比例、股权集中度是因为这些指标都是经典的股权结构指标，对公司治理有重要影响。选择机构投资者持股比例是因为现有文献发现机构投资者可能会鼓励公司采用更加激进的避税手段来提升绩效^[15-17]。

在董事会治理方面，我们选择了董事会规模、董事长是否兼任总经理、独立董事比例、董事薪酬、董事持股比例、女性董事比例、审计委员会独立性这 7 个变量。选择董事会规模、董事长是否兼任总经理、

独立董事比例是因为它们属于最基本的董事会特征，与公司治理息息相关，且 Armstrong 等^[8]发现独立董事比例与避税呈现非线性关系。选择董事薪酬、董事持股比例的原因在于这些指标是主要的董事会激励指标，在更强的激励下董事会更关注公司价值提升，从而影响避税行为。选择女性董事比例是因为以往文献发现女性更加风险规避^[18]、更注重社会责任感^[19]，因此女性董事多的企业可能更少进行避税。王雄^[20]发现审计委员会可以识别避税风险、减少代理冲突、帮助企业更加理性地认识避税的成本和收益，从而抑制企业避税，故我们加入了审计委员会独立性指标。

在高管治理方面，我们选择了高管薪酬、高管持股比例、CEO 薪酬、CEO 持股比例、CEO 任期、CEO 性别、CEO 年龄、CEO 海外经历、CEO 金融财务背景、CEO 学术背景共计 10 个变量。选择高管薪酬、高管持股比例、CEO 薪酬、CEO 持股比例是因为这些指标体现了企业对高管、CEO 的激励程度，高管和 CEO 获得的激励越强，代理问题越小，从而影响公司的避税行为。我们选择 CEO 任期、CEO 性别、CEO 年龄、CEO 海外经历、CEO 金融财务背景、CEO 学术背景是因为 Dyreng 等^[21]发现高管个体效应对公司避税有影响，因此，本文加入 CEO 这一至关重要的高管个人特征。

监事会治理方面，我们选择了监事会规模、监事薪酬、监事持股比例这些反映监事会的规模和激励水平的变量。

此外，我们还加入了代理成本这一反映代理问题严重程度的指标。

(3) 其他内部特征。除了避税工具和公司治理这两类内部特征外，Huang 等^[22]、曹越等^[23]等文献发现，客户集中度高的公司更需要持有现金，避税动机更大，故我们还控制了客户集中度和是否有客户集中度数据这两个其他公司内部特征。

2. 公司外部特征

(1) 宏观层面特征。在宏观层面，我们选取了税收征管强度、是否实施“金税三期”、最低工资、财政盈余、财政支出增长压力、经济增长、政府规模、经济政策不确定性、金融发展程度以及市场化程度相关指标，共计 15 个变量。选取税收征管强度、该省份是否实施“金税三期”工程是由于大量文献发现税收征管强度是影响企业避税行为的重要宏观因素^[24-28]，而且以“金税三期”为代表的大数据税收征管技术可以提高税收征管力度、抑制避税行为^[29]。选取各省份的月最低工资标准指标是因为刘行和赵晓阳^[30]发现最低工资标准的提高增加了企业劳动力成本和生产调整难度，导致企业避税需求增强。选取财政盈余、财政支出增长压力、经济增长、政府规模这四个指标是因为它们反映了地方政府的财政压力^[27,31,32]，而财政压力增大会导致税收征管力度增强、避税越少^[33]。选取经济政策不确定性指标是由于 Dang 等^[34]发现经济政策不确定性会导致政府财政压力增大、税收征管强度提高，且经济政策不确定性导致企业外部融资难度提高，使企业有更强的动机通过避税来拓展内源融资^[35]。选取金融发展程度是因为以往文献发现企业会进行激进的税务筹划以缓解融资约束^[36,37]，而更加发达的金融市场能够缓解企业融资约束，特别是民营企业的融资约束，减少避税动机^[38]。选取市场化程度相关指标是因为地区的市场化程度也会影响当地的金融发展程度、税收执法力度等方面，从而影响企业避税行为。

(2) 行业层面特征。在行业层面，我们选择了行业集中度、行业内企业数量、行业成长性 3 个变量。选择行业集中度、行业内企业数量是因为这两个指标均反映了行业竞争度，以往文献认为产品市场竞争越激烈，企业避税动机更强^[39]，故行业竞争度与公司避税程度呈正相关。此外，曹越和赵婷^[40]发现行业成长性越高，公司避税程度越高，故加入行业成长性指标。

(3) 外部微观特征。在微观层面，我们选择了审计师质量、分析师关注、媒体关注 3 个变量。选取审计师质量是因为一方面审计师是重要的外部监督机制，高质量的审计师可以提高公司治理水平，而公司治理水平又会影 响企业避税行为，另一方面，审计师也可能利用自身专业优势帮助企业避税^[41]。选取分析师关注、媒体关注同样是因为它们是重要的外部监督机制，与企业避税行为联系紧密。一方面，分析师和媒体都是企业避税行为的揭发者^[42-45]，因此分析师和媒体对企业的关注程度越高，企业避税程度可能越低，但另一方面，分析师和媒体的关注也会给管理层带来业绩压力，导致他们更有动机避税。

附录 2：算法选择、算法介绍、特征重要性与作用模式分析的基本思想

1. 算法选择依据

文章采用了两种主流的决策树算法——梯度提升回归树(Gradient Boosting Regression Tree, 简称 GBRT)算法和随机森林(Random Forest, 简称 RF)算法, 用于预测。选用上述两种算法有以下几点原因: 首先, 传统机器学习方法, 特别是梯度提升回归算法, 在许多情境下的表现不亚于更复杂化的模型^[46]。由于金融和财务数据集通常具有较高的结构性, 预测特征的识别和提取也更为容易, 因此采用传统机器学习方法可以取得良好效果, 并且从算力资源的角度来看, 更为高效。其次, 丰富的实践经验表明, 基于决策树的机器学习算法在有监督的机器学习问题中表现出色^[47,48]。这些算法在高维特征变量集和存在复杂非线性关系的情况下, 不仅能够提供更高的预测准确性, 还显现出更强的泛化能力, 尤其适用于处理公司避税问题中众多预测特征之间的复杂关系。其三, 与深度神经网络等复杂的机器学习模型相比, 决策树类模型具备一定的解释能力, 而在财务预测领域, 模型的解释性至关重要, 有助于理解预测结果, 提高模型可信度。最后, 为增强研究可比性, 同时避免采用单一决策模型的偶然性, 我们选择了梯度提升回归树和随机森林这两个主流且最具代表性的集成决策树模型对公司避税行为进行探究。

2. 算法介绍与参数说明

梯度提升回归树算法和随机森林算法均为成熟的机器学习算法, 下面将简单介绍两种算法, 更详细的算法实现步骤可参考相关文献^[49-51]。

(1) 梯度提升回归树

梯度提升回归树算法的基本思想: 从初始训练集中得到一个回归树作为基回归树, 计算当前预测误差, 然后在当前预测误差的基础上, 沿着损失函数梯度下降的方向训练新的基回归树, 迭代多次, 最后将多个基回归树加权结合成最后的函数。

算法具体步骤如下: 首先, 设定初始函数 $f_0(x) = \arg \min_c \sum_{i=1}^N L(y_i, c)$, $L(\cdot, \cdot)$ 为损失函数, $f_0(x)$ 即为使得损失函数极小化的常数值。在预测避税程度时, 由于避税程度为连续型数值, 我们选取常用的平方损失函数, 然后循环以下迭代过程($b = 1, \dots, B$):

(i): 计算损失函数的负梯度: $z_{i,b} = -\frac{\partial L(y_i, f(x))}{\partial f(x)} \Big|_{f(x)=f_{b-1}(x)}$ 。

(ii): 利用 $z_{i,b}$ 作为残差的近似值, 对 $z_{i,b}$ 拟合一个新的回归树 $T_b(x) = E(z|x)$ 。

(iii): 选择使得误差最小的梯度下降幅度 $\rho = \arg \min_{\rho} \sum_{i=1}^N L(y_i, \hat{f}_{b-1}(x) + \rho T_b(x))$ 。

(iv): 经过本轮迭代后新的预测函数: $f_b(x) = f_{b-1}(x) + \nu \rho T_b(x)$, 这里 ν 为学习率, 也被称为收缩参数。

迭代后的最终模型 $f_B(x)$ 即为我们的预测模型。

我们使用 R 软件的 gbm 工具包对梯度提升回归树进行拟合, 在拟合过程需要考虑三个关键参数: 回归树的交互深度, 学习率和回归树的数量。

交互深度指的是每个基回归树的分叉次数, 即树的深度。较高的交互深度意味着模型能够考虑变量之间更为复杂的交互关系, 但也会显著增加模型的复杂度和计算成本。Hastie 等^[49]发现在大多数问题中, 最佳的交互深度都较低, 因此我们将交互深度设为 5, 并将交互深度为 2、4、6 作为稳健性检验。

学习率指的是新回归树在模型中的权重, 也称为步长。较小的学习率意味着每个新回归树对模型的贡献较小, 模型收敛速度较慢, 但可以提高模型的稳定性。参照通常做法, 我们将学习率设定为 0.001, 并选择 0.0005、0.002 作为稳健性检验。

回归树的数量, 也被称为基学习器的数量。增加回归树的数量可以提高模型的复杂性和拟合能力, 但也可能导致过拟合, 通常情况下, 我们需要通过交叉验证来确定最优的树的数量。根据设备计算能力, 我们将树的数量最大值设定为 10000, 并通过 5 折交叉验证确定最优的树的数量。在交叉验证过程中, 我们仅使用训练集样本, 以确保模型的泛化能力。

（2）随机森林

随机森林的基本思想：通过构建多个决策树，并通过集成它们的预测结果来提高整体的预测准确性和稳定性。算法具体步骤如下：首先，从全体样本中运用随机抽取样本 b ，运用样本 b 构建基回归树 T_b 。这里，对基回归树 T_b 的每个节点，随机森林算法先从所有特征中随机抽取 m 个，再从这 m 个特征中选出最优特征进行划分。按照上述步骤进行 B 次之后（Bootstrap 方法）得到 B 个基回归树，再对每棵树的结果取均值即为随机森林的预测值， $\hat{f}_{rf}^B(x) = \frac{1}{B} \sum_{b=1}^B T_b(x)$ 。

我们使用 R 语言的 randomForest 工具包对随机森林进行拟合，关键参数包括树的数量和特征个数。一般来说，树的数量越多，预测性能越好，因此我们在能力范围内选择了尽可能大的值，2000 棵，并使用 1000 和 3000 棵进行稳健性检验。特征个数的选择则没有特定规则，在随机森林方法中，回归问题一般将特征个数 (m) 设定为所有特征个数的 1/3。虽然 randomForest 工具包不能够直接设置交互深度，但是可以通过设置模型的最大节点数和最小节点数来控制。一般来说，在数据量少或者特征量少的情况下，可以不限最大节点数，决策树会长到最大。本文使用的特征数量和样本量并不多，因此我们不限特征个数，使用 randomForest 包的默认值，总特征个数的三分之一，并用一系列其他取值进行稳健性检验。

3. 特征重要性与作用模式分析的基本思想

为理解模型的预测结果和经济机理，我们通过特征重要性和部分依赖图进行可解释性分析，接下来我们将介绍梯度提升回归树和随机森林算法下评价特征重要性的方法、计算方式以及部分依赖图的基本思想。

（1）特征重要性

特征重要性分析用于了解模型中各个特征对预测结果贡献程度的方法，是识别和理解关键特征以及关键特征影响程度的重点。

在梯度提升回归树算法中，相对重要性是评价特征预测能力的常用方法^[52]。其计算方式为：对于每棵回归树，评估特征作为节点时引起的损失函数（通常选用均方误差）减小幅度之和，然后，将所有回归树中损失函数减小值之和的均值作为该特征重要性的衡量指标。在此衡量标准下，一个特征的重要性越高、预测能力越强，那么它导致的损失函数下降幅度越大，相对重要性指标也相应越高。

在随机森林算法中，均方误差增加幅度（increase in MSE）是评价特征预测能力的常用方法。其计算方式为：分别对模型中每一个特征进行随机赋值，同时保持其他特征不变，计算该操作导致的均方误差上升幅度，如果特征越重要，均方误差上升的幅度越大。

在正文中我们采用相对重要性、均方误差增加幅度对特征重要性进行评价。由于 GBRT 中相对重要性指标加起来是 100%，而 RF 中均方误差增加幅度加总不一定是 100%，为了便于比较，我们将每个 RF 模型中所有特征的均方误差增加幅度总和标准化为 100%，再按比例折算各个特征的均方误差增加幅度。然后根据重要性排序识别关键避税特征。

（2）部分依赖图

部分依赖图可用于可视化特征与预测目标的关系，基本思想是通过对特征变量的取值范围进行均匀选取，观察模型预测结果的平均表现，以揭示解单个或多个特征与预测结果的联系。

具体地，假设我们需要考察的特征向量为 X_S ，其余特征为 X_C ，则 X_S 的部分依赖函数为 $f_S(X_S) = E_{X_C} f(X_S, X_C) = \int f(X_S, X_C) dP(X_C)$ ，其中 $P(X_C)$ 为 X_C 的概率密度函数。该函数的估计值通过模型在给定 X_S 的情况下求预测值的平均值，即 $\bar{f}_S(X_S) = \frac{1}{N} \sum_{i=1}^N f(X_S, x_{ic})$ ，其中 N 为训练集观测值个数， $\{x_{1c}, x_{2c}, \dots, x_{Nc}\}$ 为 X_C 在训练集中的观测值，根据 $\bar{f}_S(X_S)$ 画出的图即为部分依赖图。

附录 3：样本分布、变量定义与描述性统计

本文选用 2008 至 2019 年我国 A 股上市公司作为研究样本，数据处理分为三步：首先，我们剔除了金融业及总资产为负等特殊样本，剔除后剩余观测值为 29,316 个；随后，我们根据避税程度变量进行筛选，排除了税前利润为负或缺失、实际税率大于 1 或者小于 0 以及账面-税收差异缺失的样本，剔除税前利润为负的样本是因为这些公司无需缴纳企业所得税，剔除后剩余观测值为 18,826 个；最后，我们删除了重要避税特征缺失数据，最终得到 2,596 家公司，共计 13,113 个观测值。由于本文采用的损失函数主要为均方误差，这种统计方法对极端值比较敏感，因此，我们对连续变量进行了 1%水平的缩尾处理，不进行缩尾处理时结果仍然一致。

附表 A-1 样本分布

Appendix Table A-1 Sample Distribution		
年度	观测值个数	占比 (%)
2008 年	565	4.31
2009 年	744	5.67
2010 年	863	6.58
2011 年	1,162	8.86
2012 年	1,393	10.62
2013 年	1,358	10.36
2014 年	1,351	10.3
2015 年	1,411	10.76
2016 年	1,112	8.48
2017 年	1,028	7.84
2018 年	1,033	7.88
2019 年	1,093	8.34
总计	13,113	100

附表 A-1 展示了样本分布情况，可以看到，在 2008 年至 2015 年间，样本公司数量明显增加，与我国上市公司数量整体增长趋势相吻合，在 2016 年至 2018 年间，样本公司数量有所下降，这主要是由于这一时期亏损公司数量增加所致。

附表 A-2 展示了文章所使用变量的详细定义和计算方式。

附表 A-2 变量定义及计算方式

Appendix Table A-2 Variable Definition and Calculation Details				
变量类型		变量名称	变量符号	变量定义
避税程度指标		名义税率与实际税率之差	<i>TME</i>	名义税率-实际税率，再减去行业均值进行行业调整。其中，实际税率=当期所得税费用/税前利润
		1 减去实际税率	<i>NETR</i>	1-实际税率，再减去行业均值进行行业调整
		账面-税收差异	<i>BTD</i>	（税前利润-应纳税所得额）/期末资产总额，再减去行业均值进行行业调整
基本特征	基本属性	公司规模	<i>Size</i>	期末总资产的自然对数
		公司年龄	<i>Fage</i>	企业成立年限加一的自然对数
		名义税率	<i>Rate</i>	企业名义所得税税率
	财务指标	总资产收益率	<i>Roa</i>	息税前利润/总资产
		经营净现金流	<i>OCF</i>	公司经营活动净现金流/平均资产总额
		固定资产比例	<i>PPE</i>	固定资产/总资产
		杠杆率	<i>Lev</i>	总负债/总资产
年度		年度虚拟变量	<i>Year2008-Year2019</i>	年度虚拟变量，当年为 1，否则为 0
公司内	避税工	捐赠支出	<i>Donation</i>	企业捐赠金额/税前利润，缺失值取 0
		是否有企业捐赠金额数据	<i>Donation_Dummy</i>	数据不缺失取 1，缺失取 0
		员工人数	<i>Emp Size</i>	员工人数（单位：万人）

部 特 征	具	员工薪酬	<i>Emp_Pay</i>	平均员工薪酬的自然对数，平均员工薪酬=（员工薪酬-高管薪酬）/（员工人数-高管人数）
		应计盈余	<i>Tacc</i>	总应计盈余/总资产，总应计盈余=经营利润-经营活动净现金流
		可操纵应计盈余	<i>Absda</i>	可操纵应计盈余的绝对值 ^①
		关联交易	<i>Trade</i>	关联购销商品交易总额/总资产，缺失值取 0
		是否有关联交易数据	<i>Trade_Dummy</i>	数据不缺失取 1，缺失取 0
		研发支出	<i>RD</i>	研发支出/总资产，缺失值取 0
		是否有研发支出数据	<i>RD_Dummy</i>	数据不缺失取 1，缺失取 0
		是否有外币业务	<i>Foreign</i>	当年存在外币业务取 1，否则取 0
		是否有注册地转移情况	<i>Registran</i>	当年公司出现注册地转移取 1，否则取 0
	公司 治 理	是否国企	<i>State_Own</i>	是为 1，否为 0
		是否家族企业	<i>Fam_Own</i>	是为 1，否为 0
		是否发行 B 股 H 股	<i>BH_Share</i>	是为 1，否为 0
		股权制衡度	<i>Top1toTop25</i>	第一大股东持股数/第二到第五大股东持股数
		第一大股东持股比例	<i>Top1</i>	第一大股东持股数/总股数
		前十大股东持股比例	<i>Top10</i>	前十大股东持股数/总股数
		股权集中度	<i>Top3_SQ</i>	前三大股东持股比例的平方和
		机构投资者持股比例	<i>Inst_Own</i>	机构投资者持股数/总股数
		董事会规模	<i>Board_Size</i>	董事会人数的自然对数
		董事长是否兼任总经理	<i>Duality</i>	是为 1，否为 0
		独立董事比例	<i>Board_Ind</i>	独立董事数量/总董事数量
		董事薪酬	<i>Board_Pay</i>	公司董事年薪总额（单位：元）加一的自然对数
		董事持股比例	<i>Board_Own</i>	董事持股数/总股数
		女性董事比例	<i>Board_Gender</i>	女性董事数量/总董事数量
		审计委员会独立性	<i>Audit_Ind</i>	审计委员会独立董事数量加一的自然对数
		高管薪酬	<i>Exec_Pay</i>	高级管理人员年薪总额（单位：元）加一的自然对数
		高管持股比例	<i>Exec_Own</i>	高级管理人员持股数/总股数
		CEO 薪酬	<i>CEO_Pay</i>	CEO 年薪总额（单位：元）加一的自然对数
		CEO 持股比例	<i>CEO_Own</i>	CEO 持股数/总股数
		CEO 任期	<i>CEO_Tenure</i>	按月计算的 CEO 任职时间加一的自然对数
		CEO 性别	<i>CEO_Gender</i>	男性 CEO 取 1，女性 CEO 取 0
		CEO 年龄	<i>CEO_Age</i>	CEO 年龄的自然对数
		CEO 海外经历	<i>CEO_Aboard</i>	有海外经历取 1，否则取 0
		CEO 金融财务背景	<i>CEO_Career</i>	有金融、财务背景取 1，否则取 0
		CEO 学术背景	<i>CEO_Resea</i>	有学术背景取 1，否则取 0
		监事会规模	<i>Sup_Size</i>	监事会总人数的自然对数
		监事薪酬	<i>Sup_Pay</i>	监事年薪总额（单位：元）加一的自然对数
		监事持股比例	<i>Sup_Own</i>	监事持股数/总股数
		代理成本	<i>AC</i>	管理费用/主营业务收入
	其他	客户集中度	<i>CCHHI</i>	公司前五大客户各自销售占比的平方和，缺失值取 0
		是否有客户集中度数据	<i>CCHHI_Dummy</i>	数据不缺失取 1，缺失取 0
公 司 外 部 特	宏 观 层 面	税收征管强度	<i>TE</i>	公司注册地所在省份的税收征管强度，即地区真实的税收收入和预期可获取的税收收入之比 ^② 。
		是否实施“金税三期”	<i>Goldentax</i>	该省份在该年度实施了金税三期，取 1，否则取 0
		最低工资	<i>Minwage</i>	公司注册地所在省份该年度月最低工资，单位：千元
		财政盈余	<i>Gov_Fissurp</i>	（地方财政收入-地方财政支出）/地方财政收入

① 可操作性应计盈余根据修正 Jones 的模型^[53]进行估计，运用分年度、分行业的面板数据使用 OLS 方法对模型 $TA_t/A_{t-1} = \alpha_1 \times (1/A_{t-1}) + \alpha_2 \times (\Delta REV_t/A_{t-1}) + \alpha_3 \times (PE_t/A_{t-1}) + \varepsilon_t$ 进行估计，其中， TA_t 为第 t 期总应计盈余， A_{t-1} 为第 $t-1$ 期期末总资产， ΔREV_t 为第 t 期与第 $t-1$ 期主营业务收入的变化额， PE_t 为第 t 期期末固定资产原值，将上述模型得到的估计系数代入模型 $NDA_t = \alpha_1 \times (1/A_{t-1}) + \alpha_2 \times (\Delta REV_t - \Delta REC_t/A_{t-1}) + \alpha_3 \times (PPE_t/A_{t-1}) + \varepsilon_t$ ，其中， ΔREC_t 为第 t 期与第 $t-1$ 期应收账款的变化额，计算出经过第 $t-1$ 期期末资产调整后的非操纵性应计盈余项 NDA_t ，最终根据公式 $DA_t = TA_t/A_{t-1} - NDA_t$ 得到可操纵应计盈余 DA_t ，取绝对值即为本文的可操作性应计盈余指标 $Absda$ 。

② 参考江轩宇^[54]的研究，税收征管强度（TE）的计算公式如下： $TE = \frac{T_{it}/GDP_{it}}{T_{it}/GDP_{it,EST}}$ ，其中， T_{it}/GDP_{it} 表示各地区当年末的本地税收收入与国内生产总值的比值， $T_{it}/GDP_{it,EST}$ 表示根据下列模型估算的各地区预期可获取的税收收入。 $T_{it}/GDP_{it} = \alpha_0 + \alpha_1 \times IND1_{it}/GDP_{it} + \alpha_2 \times IND2_{it}/GDP_{it} + \alpha_3 \times OPEN_{it}/GDP_{it} + \varepsilon_{it}$ 。模型中， $IND1_{it}$ 表示各地区当年末的第一产业产值， $IND2_{it}$ 表示各地区当年末的第二产业产值， $OPEN_{it}$ 表示各地区当年末的进出口总额。

征		财政支出增长压力	Gov_Fispres	该年度地方财政支出/上一年地方财政支出
		经济增长	Gov_GDP	该省名义 GDP 增长率
		政府规模	Gov_Size	各省政府公职人员数量（单位：万人）的自然对数
		经济政策不确定性	EPU	中国经济政策不确定性月度指数平均值的自然对数 [®]
		金融发展程度	Findep	省份银行网点数量的自然对数
		市场化水平	MKT	市场化指数中市场化总评分
		政府和市场关系市场化水平	GM	市场化指数中政府与市场的关系评分
		非国有经济发展情况	NS	市场化指数中非国有经济的发展评分
		产品市场发展情况	CM	市场化指数中产品市场的发育程度评分
		要素市场发展情况	FM	市场化指数中要素市场的发育程度评分
		中介市场和法治环境情况	MIOLaw	市场化指数中市场中介组织的发育和法律制度环境评分
	行业 层面	行业集中度	Ind_HHI	公司所处行业的赫芬达尔指数。行业按照证监会 2012 年分类进行划分，其中制造业取二级分类，其余为一级分类
		行业内企业数量	Ind_Size	行业中企业数量的自然对数，行业划分同上
		行业成长性	Ind_Growth	所属行业该年度公司托宾 Q 值的中位数
外部 微观	审计师质量	Big4	是四大审计师事务所取 1，否则取 0	
	分析师关注	Analyst	该年度跟踪的分析师团队数加一的自然对数	
	媒体关注	Media	该年度公司相关新闻数量加一的自然对数	

附表 A-3 展示了衡量公司避税程度的 3 个变量，7 个基本预测特征，64 个公司内外部预测特征的描述性统计结果。从表中我们可以观察到，样本公司的平均名义税率 (*Rate*) 为 19%，总资产收益率均值为 10%。样本总资产收益均值偏高是因为我们在数据处理过程中剔除了税前利润为负的样本。

附表 A-3 描述性统计

Appendix Table A-3 Descriptive Statistics of Sample

Variable	Obs.	Mean	S.D.	Variable	Obs.	Mean	S.D.
<i>TME</i>	13,113	0.00	0.11	<i>Exec_Pay</i>	13,113	14.91	0.81
<i>NETR</i>	13,113	1.00	0.11	<i>Exec_Own</i>	13,113	0.06	0.13
<i>BTD</i>	13,113	0.00	0.02	<i>CEO_Pay</i>	13,113	13.32	0.81
<i>Size</i>	13,113	22.32	1.29	<i>CEO_Own</i>	13,113	0.05	0.11
<i>Fage</i>	13,113	2.75	0.37	<i>CEO_Tenure</i>	13,113	3.54	1.02
<i>Rate</i>	13,113	0.19	0.05	<i>CEO_Gender</i>	13,113	0.94	0.23
<i>Roa</i>	13,113	0.10	0.05	<i>CEO_Age</i>	13,113	3.88	0.13
<i>OCF</i>	13,113	0.06	0.08	<i>CEO_Aboard</i>	13,113	0.07	0.26
<i>PPE</i>	13,113	0.22	0.16	<i>CEO_Career</i>	13,113	0.17	0.37
<i>Lev</i>	13,113	0.43	0.20	<i>CEO_Resea</i>	13,113	0.18	0.38
<i>Donation</i>	13,113	0.01	0.02	<i>Sup_Size</i>	13,113	1.25	0.25
<i>Donation_Dummy</i>	13,113	0.80	0.40	<i>Sup_Pay</i>	13,113	13.00	1.64
<i>Emp_Size</i>	13,113	0.61	1.09	<i>Sup_Own</i>	13,113	0.00	0.01
<i>Emp_Pay</i>	13,113	11.41	0.56	<i>AC</i>	13,113	0.09	0.06
<i>Tacc</i>	13,113	0.01	0.08	<i>CCHHI</i>	13,113	0.03	0.07
<i>Absda</i>	13,113	0.06	0.06	<i>CCHHI_Dummy</i>	13,113	0.69	0.46
<i>Trade</i>	13,113	0.04	0.11	<i>TE</i>	13,113	1.01	0.19
<i>Trade_Dummy</i>	13,113	0.69	0.46	<i>Goldentax</i>	13,113	0.29	0.45
<i>RD</i>	13,113	0.02	0.02	<i>Minwage</i>	13,113	1.47	0.42
<i>RD_Dummy</i>	13,113	0.75	0.44	<i>Gov_Fissurp</i>	13,113	-0.48	0.55
<i>Foreign</i>	13,113	0.53	0.50	<i>Gov_Fispres</i>	13,113	1.14	0.08
<i>Registran</i>	13,113	0.05	0.21	<i>Gov_GDP</i>	13,113	0.11	0.05
<i>State_Own</i>	13,113	0.39	0.49	<i>Gov_Size</i>	13,113	4.09	0.53
<i>Fam_Own</i>	13,113	0.37	0.48	<i>EPU</i>	13,113	4.93	0.42
<i>BH_Share</i>	13,113	0.06	0.23	<i>Findep</i>	13,113	9.04	0.59
<i>Top1toTop25</i>	13,113	3.98	5.67	<i>MKT</i>	13,113	8.02	1.84

[®] 该指标根据中国香港《南华早报》每日新闻的关键词频信息统计得来，我们以这一月度指数平均值的自然对数衡量该年度中国经济政策变化情况，数据可以从 <http://www.policyuncertainty.com/> 网站获取。

<i>Top1</i>	13,113	0.36	0.15
<i>Top10</i>	13,113	0.59	0.15
<i>Top3_SQ</i>	13,113	0.17	0.12
<i>Inst_Own</i>	13,113	0.48	0.24
<i>Board_Size</i>	13,113	2.16	0.20
<i>Duality</i>	13,113	0.24	0.43
<i>Board_Ind</i>	13,113	0.37	0.05
<i>Board_Pay</i>	13,113	14.21	1.74
<i>Board_Own</i>	13,113	0.12	0.18
<i>Board_Gender</i>	13,113	0.13	0.12
<i>Audit_Ind</i>	13,113	0.42	0.57

<i>GM</i>	13,113	7.08	1.44
<i>NS</i>	13,113	8.68	1.94
<i>CM</i>	13,113	8.22	1.34
<i>FM</i>	13,113	7.35	3.12
<i>MIOLaw</i>	13,113	10.75	6.54
<i>Ind_HHI</i>	13,113	0.10	0.09
<i>Ind_Size</i>	13,113	4.48	0.81
<i>Ind_Growth</i>	13,113	3.02	4.63
<i>Big4</i>	13,113	0.06	0.24
<i>Analyst</i>	13,113	2.08	0.89
<i>Media</i>	13,113	4.02	0.72

附录 4：MSE、MAE 评价指标的预测结果

附表 A-4 展示了使用 OLS、梯度提升回归树（GBRT）算法、随机森林（RF）算法，各模型在 MSE、MAE 评价指标下的预测能力，可以看到与正文表 2 各模型在 R^2_{os} 评价指标下的模型预测能力结果相比，不论使用何种评价指标，模型预测能力的结果均高度一致。

正文中，我们使用 R^2_{os} 作为主要评价指标进行分析。使用 R^2_{os} 具有两点优势，一是更为直观地反映模型预测能力的提升程度，二是以百分比为单位，更清晰地呈现出模型预测能力的变化。

附表 A-4 MSE、MAE 评价指标下模型预测能力结果

Appendix Table A-4 Predictive Power of Models under MSE and MAE

附表 A-4.1 用 TME 作为公司避税程度衡量指标						
评价指标	MSE			MAE		
预测模型	OLS	GBRT	RF	OLS	GBRT	RF
	(4)	(5)	(6)	(7)	(8)	(9)
基准模型	0.01025	0.00970	0.00985	0.0690	0.0666	0.0673
<i>加入公司内部特征</i>						
基准模型+避税工具特征	<u>0.00978</u>	0.00926	<u>0.00930</u>	<u>0.0671</u>	0.0650	0.0654
基准模型+公司治理特征	0.01032	0.00975	0.00982	0.0695	0.0666	0.0673
基准模型+其他内部特征	0.01024	0.00967	0.00988	0.0689	0.0665	0.0675
内部模型	0.00981	0.00929	0.00934	0.0674	0.0650	0.0655
<i>加入公司外部特征</i>						
基准模型+宏观层面特征	0.01024	0.00976	0.00993	0.0692	0.0667	0.0678
基准模型+行业层面特征	0.01022	0.00965	0.00976	0.0688	0.0661	0.0667
基准模型+外部微观特征	0.01023	0.00969	0.00980	0.0688	0.0664	0.0670
外部模型	0.01021	0.00968	0.00972	0.0688	0.0663	0.0667
<i>加入全部公司内外部特征</i>						
全模型	0.00982	0.00928	0.00931	0.0675	0.0649	<u>0.0652</u>
附表 A-4.2 用 TME 作为公司避税程度衡量指标						
评价指标	MSE			MAE		
预测模型	(4)	(5)	(6)	(7)	(8)	(9)
基准模型	0.01032	0.00977	0.00991	0.0694	0.0675	0.0681
<i>加入公司内部特征</i>						
基准模型+避税工具特征	<u>0.00991</u>	0.00945	0.00948	<u>0.0682</u>	0.0667	0.0669
基准模型+公司治理特征	0.01035	0.00986	0.00993	0.0698	0.0677	0.0682
基准模型+其他内部特征	0.01032	0.00975	0.00992	0.0694	0.0675	0.0683
内部模型	0.00992	0.00950	0.00951	0.0684	0.0669	0.0671
<i>加入公司外部特征</i>						
基准模型+宏观层面特征	0.01030	0.00982	0.00998	0.0694	0.0677	0.0687
基准模型+行业层面特征	0.01034	0.00970	0.00981	0.0696	0.0670	0.0676
基准模型+外部微观特征	0.01031	0.00978	0.00987	0.0693	0.0675	0.0680
外部模型	0.01032	0.00973	0.00977	0.0696	0.0672	0.0676
<i>加入全部公司内外部特征</i>						
全模型	0.00993	0.00942	<u>0.00945</u>	0.0685	0.0664	<u>0.0665</u>

注：与正文相同，我们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，我们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，我们比较不同算法的预测效果，最佳结果以粗体显示。后续预测结果表格沿用相同标注方式。由于 MSE 指标评价下保留四位小数无法区分最佳模型，故改为保留五位小数。

附录 5：各类模型的特征重要性数值和排序

附表 A-5 加入各子类特征的模型中重要性排名前十的特征

Table A-5 Top 10 Most Important Features under Sub-Classification

模型 排序	基准模型 +避税工具特征	基准模型 +公司治理特征	基准模型 +其他内部特征	基准模型 +宏观层面特征	基准模型 +行业层面特征	基准模型 +外部微观特征
1	<i>Roa</i>	<i>Roa</i>	<i>Roa</i>	<i>Roa</i>	<i>Roa</i>	<i>Roa</i>
2	<i>Donation</i>	<i>PPE</i>	<i>PPE</i>	<i>PPE</i>	<i>PPE</i>	<i>PPE</i>
3	<i>Rate</i>	<i>Lev</i>	<i>Lev</i>	<i>Lev</i>	<i>Lev</i>	<i>Lev</i>
4	<i>Tacc</i>	<i>Rate</i>	<i>Size</i>	<i>Rate</i>	<i>Rate</i>	<i>Rate</i>
5	<i>Lev</i>	<i>Board_Pay</i>	<i>Rate</i>	<i>Size</i>	<i>Size</i>	<i>Size</i>
6	<i>OCF</i>	<i>Top1toTop25</i>	<i>OCF</i>	<i>OCF</i>	<i>Ind_Size</i>	<i>Media</i>
7	<i>PPE</i>	<i>Size</i>	<i>CCHHI</i>	<i>MIOLaw</i>	<i>Ind_HHI</i>	<i>OCF</i>
8	<i>Emp_Size</i>	<i>AC</i>	<i>Fage</i>	<i>Fage</i>	<i>OCF</i>	<i>Fage</i>
9	<i>Size</i>	<i>Inst_Own</i>	<i>CCHHI_Dummy</i>	<i>NS</i>	<i>Ind_Growth</i>	<i>Analyst</i>
10	<i>RD</i>	<i>CEO_Pay</i>	<i>Year2018</i>	<i>FM</i>	<i>Fage</i>	<i>Big4</i>

注：Year2018 为 2018 年的虚拟变量；本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，下同。

附表 A-5 展示了分别加入六个子类别特征后，位于前十的预测特征排序情况。可以看到，在分别加入子类别特征后，在避税工具特征中，捐赠支出（*Donation*）重要性最高，此外，应计盈余（*Tacc*）、员工人数（*Emp_Size*）、研发支出（*RD*）也位列前十，与全模型的结果一致；在公司治理特征中，董事薪酬（*Board_Pay*）的重要性最高，此外，股权制衡度（*Top1toTop25*）、代理成本（*AC*）、机构投资者持股比例（*Inst_Own*）、CEO 薪酬（*CEO_Pay*）也位列前十，这说明在公司治理特征中，高管薪酬、股权结构、代理成本相关特征在预测公司避税时更有效；在其他内部特征中，客户集中度（*CCHHI*）的重要性最高；在宏观层面特征中，中介市场和法治环境情况（*MIOLaw*）的重要性最高，此外，非国有经济发展情况（*NS*）、要素市场发展情况（*FM*）也位列前十，这说明在宏观特征中，公司所处地区的市场化、法治化程度在预测公司避税时更有效；在行业层面特征中，行业内企业数量（*Ind_Size*）重要性最高，行业集中度（*Ind_HHI*）的重要性紧随其后，这再次说明行业的规模和竞争度是预测公司避税最为有效的行业层面特征；在外部微观特征中，媒体关注度（*Media*）的重要性最高，这说明在外部监督机制中，媒体关注对公司避税程度的预测能力最强。

附表 A-6 展示了各类模型在预测公司避税行为时，重要性均值排名前十的特征，附表 A-7 进一步展示了全模型中所有特征的重要性排序及具体的特征重要性数值。

附表 A-6 加入各类特征模型的特征重要性均值（前十名）

Appendix Table A-6 Average Feature Importance when Adding Each Category of Features (Top 10)

附表 A-6.1 全模型、内部模型、外部模型特征重要性均值（单位：%）						
模型	全模型		内部模型		外部模型	
排序	特征	重要性均值	特征	重要性均值	特征	重要性均值
1	<i>Roa</i>	14.79	<i>Roa</i>	16.11	<i>Roa</i>	22.16
2	<i>Donation</i>	7.08	<i>Donation</i>	7.78	<i>Lev</i>	8.05
3	<i>Rate</i>	4.47	<i>Lev</i>	4.58	<i>PPE</i>	7.66
4	<i>Lev</i>	4.00	<i>OCF</i>	4.51	<i>Rate</i>	6.39
5	<i>OCF</i>	3.71	<i>Rate</i>	4.42	<i>Size</i>	3.89
6	<i>Tacc</i>	3.53	<i>Tacc</i>	4.40	<i>Ind_Size</i>	3.63
7	<i>Emp_Size</i>	3.07	<i>PPE</i>	3.67	<i>Ind_HHI</i>	3.22
8	<i>PPE</i>	2.90	<i>Emp_Size</i>	3.49	<i>MIOLaw</i>	3.01
9	<i>RD</i>	2.63	<i>RD</i>	3.03	<i>OCF</i>	2.96
10	<i>Board_Pay</i>	1.93	<i>Board_Pay</i>	2.69	<i>Media</i>	2.79
附表 A-6.2 基准模型分别加入各类公司内部特征后的特征重要性均值（单位：%）						
模型	基准模型 +避税工具特征		基准模型 +公司治理特征		基准模型 +其他内部特征	
排序	特征	重要性均值	特征	重要性均值	特征	重要性均值

1	<i>Roa</i>	20.16	<i>Roa</i>	21.28	<i>Roa</i>	32.43
2	<i>Donation</i>	9.82	<i>PPE</i>	7.42	<i>PPE</i>	14.98
3	<i>Rate</i>	7.20	<i>Lev</i>	7.19	<i>Lev</i>	13.47
4	<i>Tacc</i>	7.11	<i>Rate</i>	5.73	<i>Size</i>	10.11
5	<i>Lev</i>	7.09	<i>Board_Pay</i>	3.96	<i>Rate</i>	9.98
6	<i>OCF</i>	7.08	<i>Top1toTop25</i>	3.54	<i>OCF</i>	6.89
7	<i>PPE</i>	6.34	<i>Size</i>	3.25	<i>CCHHI</i>	5.76
8	<i>Emp_Size</i>	6.25	<i>AC</i>	2.99	<i>Fage</i>	5.69
9	<i>Size</i>	5.65	<i>Inst_Own</i>	2.95	<i>CCHHI_Dummy</i>	0.88
10	<i>RD</i>	5.31	<i>CEO_Pay</i>	2.83	<i>Year2018</i>	0.66

附表 A-6.3 基准模型分别加入各类公司外部特征后的特征重要性均值（单位：%）

模型	基准模型 +宏观层面特征		基准模型 +行业层面特征		基准模型 +外部微观特征	
排序	因素	重要性均值	因素	重要性均值	因素	重要性均值
1	<i>Roa</i>	24.41	<i>Roa</i>	28.87	<i>Roa</i>	29.74
2	<i>PPE</i>	9.38	<i>PPE</i>	12.34	<i>PPE</i>	13.45
3	<i>Lev</i>	8.85	<i>Lev</i>	12.04	<i>Lev</i>	12.44
4	<i>Rate</i>	6.57	<i>Rate</i>	9.11	<i>Rate</i>	9.05
5	<i>Size</i>	5.38	<i>Size</i>	7.88	<i>Size</i>	8.95
6	<i>OCF</i>	3.85	<i>Ind_Size</i>	6.72	<i>Media</i>	6.01
7	<i>MIOLaw</i>	3.54	<i>Ind_HHI</i>	5.87	<i>OCF</i>	5.94
8	<i>Fage</i>	3.37	<i>OCF</i>	5.52	<i>Fage</i>	5.21
9	<i>NS</i>	3.23	<i>Ind_Growth</i>	4.60	<i>Analyst</i>	4.60
10	<i>FM</i>	3.18	<i>Fage</i>	4.36	<i>Big4</i>	1.92

注：本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。

附表 A-7 全模型所有特征重要性排序及特征重要性数值

Appendix Table A-7 Feature Importance of the Full Model

附表 A-7.1 GBRT 算法下全模型特征重要性排序（单位：%）

排序	两个避税指标平均	特征重要性均值	以 <i>TME</i> 为避税指标	特征重要性数值	以 <i>NETR</i> 为避税指标	特征重要性数值
1	<i>Roa</i>	23.44	<i>Roa</i>	22.73	<i>Roa</i>	24.14
2	<i>Donation</i>	10.96	<i>Donation</i>	10.71	<i>Donation</i>	11.21
3	<i>OCF</i>	5.27	<i>Rate</i>	6.16	<i>OCF</i>	5.27
4	<i>Lev</i>	4.91	<i>OCF</i>	5.26	<i>Lev</i>	5.23
5	<i>Tacc</i>	4.77	<i>Tacc</i>	4.91	<i>Tacc</i>	4.62
6	<i>Rate</i>	4.22	<i>Lev</i>	4.59	<i>Emp_Size</i>	3.94
7	<i>Emp_Size</i>	3.55	<i>RD</i>	3.91	<i>PPE</i>	3.36
8	<i>PPE</i>	3.11	<i>Emp_Size</i>	3.16	<i>Rate</i>	2.28
9	<i>RD</i>	2.52	<i>PPE</i>	2.85	<i>Board_Pay</i>	2.14
10	<i>Board_Pay</i>	2.05	<i>Board_Pay</i>	1.95	<i>Top1toTop25</i>	1.8
11	<i>Top1toTop25</i>	1.8	<i>Top1toTop25</i>	1.8	<i>Analyst</i>	1.56
12	<i>Trade</i>	1.43	<i>Trade</i>	1.58	<i>Size</i>	1.45
13	<i>Size</i>	1.43	<i>Size</i>	1.4	<i>CEO_Pay</i>	1.41
14	<i>Analyst</i>	1.36	<i>Top10</i>	1.32	<i>Ind_Growth</i>	1.37
15	<i>Top10</i>	1.28	<i>Sup_Pay</i>	1.26	<i>CEO_Age</i>	1.36
16	<i>CEO_Age</i>	1.23	<i>AC</i>	1.19	<i>Ind_HHI</i>	1.32
17	<i>CEO_Pay</i>	1.22	<i>Analyst</i>	1.16	<i>Trade</i>	1.28
18	<i>Sup_Pay</i>	1.2	<i>Ind_Size</i>	1.1	<i>Top10</i>	1.23
19	<i>AC</i>	1.18	<i>CEO_Age</i>	1.1	<i>Ind_Size</i>	1.19
20	<i>Ind_Size</i>	1.15	<i>CEO_Pay</i>	1.02	<i>AC</i>	1.16
21	<i>Ind_HHI</i>	1.14	<i>Absda</i>	1.02	<i>Absda</i>	1.14
22	<i>Absda</i>	1.08	<i>Exec_Pay</i>	1	<i>Sup_Pay</i>	1.13
23	<i>Inst_Own</i>	1.04	<i>Inst_Own</i>	0.98	<i>RD</i>	1.12
24	<i>Top1</i>	1	<i>Ind_HHI</i>	0.95	<i>Inst_Own</i>	1.09
25	<i>Exec_Pay</i>	0.98	<i>Top1</i>	0.94	<i>Top1</i>	1.06
26	<i>Ind_Growth</i>	0.89	<i>Top3_SQ</i>	0.87	<i>Exec_Pay</i>	0.95
27	<i>CCHHI</i>	0.87	<i>CCHHI</i>	0.85	<i>CCHHI</i>	0.89
28	<i>Top3_SQ</i>	0.87	<i>MIOLaw</i>	0.84	<i>Top3_SQ</i>	0.87

29	<i>MIOLaw</i>	0.83	<i>CEO_Tenure</i>	0.76	<i>MIOLaw</i>	0.82
30	<i>NS</i>	0.75	<i>Media</i>	0.75	<i>NS</i>	0.78
31	<i>GM</i>	0.75	<i>GM</i>	0.72	<i>GM</i>	0.77
32	<i>CEO_Tenure</i>	0.74	<i>NS</i>	0.72	<i>Emp_Pay</i>	0.76
33	<i>Media</i>	0.71	<i>Fage</i>	0.66	<i>CEO_Tenure</i>	0.72
34	<i>Emp_Pay</i>	0.69	<i>FM</i>	0.64	<i>MKT</i>	0.69
35	<i>Fage</i>	0.67	<i>Emp_Pay</i>	0.61	<i>Fage</i>	0.68
36	<i>FM</i>	0.63	<i>Gov_GDP</i>	0.58	<i>Media</i>	0.66
37	<i>Findep</i>	0.6	<i>Findep</i>	0.56	<i>Findep</i>	0.64
38	<i>MKT</i>	0.58	<i>TE</i>	0.55	<i>FM</i>	0.62
39	<i>Gov_GDP</i>	0.58	<i>Board_Gender</i>	0.54	<i>Exec_Own</i>	0.61
40	<i>TE</i>	0.58	<i>Gov_Fissurp</i>	0.53	<i>TE</i>	0.6
41	<i>Exec_Own</i>	0.57	<i>Exec_Own</i>	0.53	<i>Gov_Fissurp</i>	0.59
42	<i>Gov_Fissurp</i>	0.56	<i>Gov_Size</i>	0.51	<i>Gov_GDP</i>	0.57
43	<i>Gov_Size</i>	0.53	<i>Minwage</i>	0.47	<i>Gov_Size</i>	0.54
44	<i>Board_Gender</i>	0.51	<i>MKT</i>	0.47	<i>Minwage</i>	0.51
45	<i>Minwage</i>	0.49	<i>CM</i>	0.43	<i>Board_Gender</i>	0.48
46	<i>Board_Own</i>	0.46	<i>Board_Own</i>	0.43	<i>Board_Own</i>	0.48
47	<i>CM</i>	0.43	<i>Sup_Size</i>	0.42	<i>CM</i>	0.42
48	<i>Gov_Fispres</i>	0.41	<i>Ind_Growth</i>	0.41	<i>Gov_Fispres</i>	0.41
49	<i>Sup_Size</i>	0.4	<i>Gov_Fispres</i>	0.4	<i>Sup_Own</i>	0.39
50	<i>Sup_Own</i>	0.32	<i>Board_Size</i>	0.27	<i>Sup_Size</i>	0.38
51	<i>Board_Size</i>	0.24	<i>Sup_Own</i>	0.24	<i>Board_Ind</i>	0.24
52	<i>Board_Ind</i>	0.23	<i>Board_Ind</i>	0.22	<i>Board_Size</i>	0.2
53	<i>CEO_Own</i>	0.18	<i>Foreign</i>	0.19	<i>CEO_Own</i>	0.19
54	<i>EPU</i>	0.16	<i>CEO_Own</i>	0.17	<i>EPU</i>	0.14
55	<i>Audit_Ind</i>	0.13	<i>EPU</i>	0.17	<i>Audit_Ind</i>	0.12
56	<i>Foreign</i>	0.13	<i>Audit_Ind</i>	0.13	<i>State_Own</i>	0.11
57	<i>State_Own</i>	0.1	<i>State_Own</i>	0.08	<i>Foreign</i>	0.06
58	<i>BH_Share</i>	0.05	<i>BH_Share</i>	0.05	<i>BH_Share</i>	0.05
59	<i>Fam_Own</i>	0.03	<i>Goldentax</i>	0.02	<i>Fam_Own</i>	0.03
60	<i>CEO_Career</i>	0.02	<i>CEO_Career</i>	0.02	<i>CEO_Career</i>	0.02
61	<i>Donation_Dummy</i>	0.02	<i>Fam_Own</i>	0.02	<i>Donation_Dummy</i>	0.02
62	<i>Duality</i>	0.02	<i>Donation_Dummy</i>	0.02	<i>Duality</i>	0.02
63	<i>CEO_Aboard</i>	0.01	<i>Year2013</i>	0.01	<i>Year2013</i>	0.01
64	<i>CEO_Gender</i>	0.01	<i>CEO_Gender</i>	0.01	<i>Year2016</i>	0.01
65	<i>CEO_Resea</i>	0.01	<i>CEO_Resea</i>	0.01	<i>Year2011</i>	0.01
66	<i>Goldentax</i>	0.01	<i>Duality</i>	0.01	<i>Year2015</i>	0.01
67	<i>Year2017</i>	0.01	<i>CEO_Aboard</i>	0.01	<i>CEO_Aboard</i>	0.01
68	<i>Year2011</i>	0.01	<i>Year2018</i>	0.01	<i>Registran</i>	0.01
69	<i>Year2013</i>	0.01	<i>Year2015</i>	0.01	<i>CEO_Gender</i>	0.01
70	<i>Year2015</i>	0.01	<i>Year2011</i>	0.01	<i>CEO_Resea</i>	0.01
71	<i>Big4</i>	0.01	<i>RD_Dummy</i>	0.01	<i>Year2017</i>	0.01
72	<i>RD_Dummy</i>	0.01	<i>Year2017</i>	0.01	<i>Big4</i>	0.01
73	<i>Registran</i>	0.01	<i>Registran</i>	0	<i>Year2008</i>	0
74	<i>Year2018</i>	0.01	<i>Year2019</i>	0	<i>Goldentax</i>	0
75	<i>Year2016</i>	0.01	<i>Year2012</i>	0	<i>Trade_Dummy</i>	0
76	<i>CCHHI_Dummy</i>	0	<i>Year2009</i>	0	<i>Year2018</i>	0
77	<i>Trade_Dummy</i>	0	<i>Year2016</i>	0	<i>Year2014</i>	0
78	<i>Year2008</i>	0	<i>Year2008</i>	0	<i>Year2019</i>	0
79	<i>Year2019</i>	0	<i>Year2010</i>	0	<i>Year2012</i>	0
80	<i>Year2009</i>	0	<i>Big4</i>	0	<i>Year2009</i>	0
81	<i>Year2010</i>	0	<i>Year2014</i>	0	<i>Year2010</i>	0
82	<i>Year2012</i>	0	<i>Trade_Dummy</i>	0	<i>RD_Dummy</i>	0
83	<i>Year2014</i>	0	<i>CCHHI_Dummy</i>	0	<i>CCHHI_Dummy</i>	0

附表 A7.2 RF 算法下全模型特征重要性排序（单位：%）

排序	两个避税指标平均	特征重要性均值	以 <i>TME</i> 为避税指标	特征重要性数值	以 <i>NETR</i> 为避税指标	特征重要性数值
1	<i>Roa</i>	6.15	<i>Rate</i>	6.38	<i>Roa</i>	6.18

2	Rate	4.72	Roa	6.12	Donation	3.32
3	Donation	3.2	RD	3.21	Lev	3.18
4	Lev	3.09	Donation	3.07	Rate	3.06
5	RD	2.74	Lev	3	PPE	2.77
6	PPE	2.7	PPE	2.64	Emp_Size	2.64
7	Emp_Size	2.58	Emp_Size	2.53	Tacc	2.31
8	Tacc	2.3	Tacc	2.3	RD	2.26
9	Size	2.18	Size	2.2	OCF	2.18
10	OCF	2.15	OCF	2.11	Size	2.16
11	Ind_Size	2.02	Ind_Size	2.05	Ind_Size	1.98
12	AC	1.85	AC	1.9	Board_Pay	1.91
13	Board_Pay	1.82	Trade	1.81	CEO_Pay	1.91
14	Exec_Pay	1.74	Board_Pay	1.73	Exec_Pay	1.86
15	CEO_Pay	1.74	Top1toTop25	1.72	AC	1.8
16	Top1toTop25	1.72	Sup_Pay	1.67	Top1	1.76
17	Gov_Size	1.67	Exec_Pay	1.63	Emp_Pay	1.76
18	Inst_Own	1.67	Findep	1.62	Inst_Own	1.74
19	Ind_HHI	1.64	Gov_Size	1.62	Top3_SQ	1.72
20	Findep	1.63	Ind_HHI	1.62	Gov_Size	1.72
21	MIOLaw	1.62	NS	1.61	Top1toTop25	1.72
22	Emp_Pay	1.62	Inst_Own	1.59	FM	1.7
23	Sup_Pay	1.62	CEO_Pay	1.58	MIOLaw	1.67
24	NS	1.61	MIOLaw	1.57	Ind_HHI	1.67
25	Trade	1.59	Top10	1.55	Findep	1.64
26	Top3_SQ	1.58	Fage	1.55	Ind_Growth	1.62
27	FM	1.55	Emp_Pay	1.49	NS	1.61
28	Top1	1.52	Top3_SQ	1.44	MKT	1.6
29	Top10	1.51	CM	1.43	Media	1.59
30	Fage	1.49	FM	1.4	Minwage	1.57
31	Media	1.48	Minwage	1.38	Sup_Pay	1.56
32	Minwage	1.48	Media	1.38	Gov_Fissurp	1.52
33	MKT	1.45	Board_Own	1.35	Analyst	1.47
34	CM	1.44	Gov_Fissurp	1.32	Top10	1.47
35	Gov_Fissurp	1.42	MKT	1.3	CM	1.45
36	Ind_Growth	1.41	GM	1.28	Fage	1.43
37	Analyst	1.36	Top1	1.28	Trade	1.37
38	Board_Own	1.33	Analyst	1.26	GM	1.32
39	GM	1.3	CCHHI	1.25	Board_Own	1.32
40	Exec_Own	1.23	Exec_Own	1.22	CEO_Age	1.27
41	CEO_Age	1.19	Ind_Growth	1.2	Exec_Own	1.23
42	CCHHI	1.19	CEO_Age	1.11	Sup_Own	1.16
43	Sup_Own	1.13	Sup_Own	1.1	CCHHI	1.13
44	Absda	1.06	TE	1.05	Absda	1.12
45	TE	1.06	Absda	1.01	TE	1.07
46	Sup_Size	1	CEO_Own	0.98	Sup_Size	1.03
47	Board_Size	0.98	Sup_Size	0.96	Board_Size	1.01
48	CEO_Own	0.96	Board_Size	0.95	CEO_Own	0.95
49	Board_Gender	0.86	Foreign	0.9	Board_Gender	0.89
50	Foreign	0.79	Board_Gender	0.83	State_Own	0.8
51	Board_Ind	0.75	RD_Dummy	0.77	Board_Ind	0.75
52	State_Own	0.71	Board_Ind	0.74	EPU	0.7
53	EPU	0.68	EPU	0.65	BH_Share	0.7
54	Big4	0.67	Gov_GDP	0.65	Big4	0.69
55	Gov_GDP	0.67	Big4	0.65	Gov_GDP	0.68
56	RD_Dummy	0.65	State_Own	0.62	Foreign	0.68
57	BH_Share	0.63	BH_Share	0.55	Fam_Own	0.53
58	Fam_Own	0.52	Audit_Ind	0.54	RD_Dummy	0.52
59	Audit_Ind	0.49	Fam_Own	0.51	CEO_Resea	0.5

60	<i>CEO_Aboard</i>	0.49	<i>Gov_Fispres</i>	0.51	<i>Duality</i>	0.48
61	<i>Gov_Fispres</i>	0.47	<i>CEO_Aboard</i>	0.51	<i>CEO_Aboard</i>	0.47
62	<i>CEO_Tenure</i>	0.46	<i>Trade_Dummy</i>	0.46	<i>CEO_Tenure</i>	0.46
63	<i>Duality</i>	0.45	<i>CEO_Tenure</i>	0.45	<i>Audit_Ind</i>	0.44
64	<i>CEO_Resea</i>	0.44	<i>Duality</i>	0.43	<i>Gov_Fispres</i>	0.43
65	<i>Trade_Dummy</i>	0.44	<i>Goldentax</i>	0.39	<i>Trade_Dummy</i>	0.42
66	<i>Goldentax</i>	0.38	<i>CEO_Resea</i>	0.39	<i>Goldentax</i>	0.37
67	<i>CEO_Career</i>	0.29	<i>CCHHI_Dummy</i>	0.35	<i>Year2019</i>	0.31
68	<i>CCHHI_Dummy</i>	0.28	<i>CEO_Career</i>	0.34	<i>CEO_Career</i>	0.23
69	<i>Year2019</i>	0.27	<i>Year2019</i>	0.23	<i>Year2015</i>	0.23
70	<i>CEO_Gender</i>	0.19	<i>CEO_Gender</i>	0.19	<i>CCHHI_Dummy</i>	0.22
71	<i>Year2015</i>	0.18	<i>Donation_Dummy</i>	0.18	<i>CEO_Gender</i>	0.18
72	<i>Year2009</i>	0.16	<i>Year2009</i>	0.15	<i>Year2008</i>	0.17
73	<i>Donation_Dummy</i>	0.16	<i>Registran</i>	0.14	<i>Year2009</i>	0.16
74	<i>Year2018</i>	0.13	<i>Year2015</i>	0.12	<i>Year2018</i>	0.14
75	<i>Registran</i>	0.13	<i>Year2018</i>	0.12	<i>Donation_Dummy</i>	0.13
76	<i>Year2010</i>	0.11	<i>Year2010</i>	0.12	<i>Registran</i>	0.12
77	<i>Year2008</i>	0.09	<i>Year2017</i>	0.04	<i>Year2010</i>	0.1
78	<i>Year2016</i>	0.05	<i>Year2008</i>	0.02	<i>Year2016</i>	0.09
79	<i>Year2011</i>	0.03	<i>Year2016</i>	0.01	<i>Year2011</i>	0.08
80	<i>Year2012</i>	0.01	<i>Year2013</i>	0.01	<i>Year2012</i>	0.02
81	<i>Year2013</i>	0	<i>Year2012</i>	-0.01	<i>Year2013</i>	-0.01
82	<i>Year2017</i>	-0.02	<i>Year2011</i>	-0.03	<i>Year2017</i>	-0.09
83	<i>Year2014</i>	-0.07	<i>Year2014</i>	-0.05	<i>Year2014</i>	-0.1

附录 6：关键特征对公司避税程度的作用模式验证

本文按照 Shrestha 等^[55]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，将机器学习算法所识别出的模式与实证研究方法相结合进行检验。

附表 A-8 至附表 A-11 展示了我们对捐赠支出、应计盈余、员工人数、研发支出与避税程度之间作用模式的检验结果，检验结果与部分依赖图展示的模式基本吻合，进一步印证了机器学习算法在识别避税行为模式方面的有效性和准确性，同样也说明基于机器学习的理论构建范式是有效的。

附表 A-8 捐赠支出对公司避税程度的回归结果

Appendix Table A-8 Regression of *Donation* on Corporate Tax Avoidance

避税衡量指标	<i>TME</i>	<i>NETR</i>
	(1)	(2)
<i>Donation</i>	-1.005*** (-7.901)	-0.977*** (-7.645)
公司固定效应	控制	控制
年度固定效应	控制	控制
样本量	4232	4232
R^2	0.416	0.440

注：***、**、*分别表示在 1%、5%和 10%的水平下显著。括号内为在公司层面进行聚类处理（cluster）的 t 值，样本量少于真实训练集样本数量的原因是，包含了单一观测值（Singleton group）的公司可能会低估标准差，因此回归中去除了只含有一年观测值的公司，下同。

附表 A-9 应计盈余对公司避税程度的回归结果

Appendix Table A-9 Regression of *Tacc* on Corporate Tax Avoidance

避税衡量指标	<i>TME</i>			<i>NETR</i>		
	(1)	(2)	(3)	(4)	(5)	(6)
<i>Tacc</i> 取值范围	(-∞,-0.05)	[-0.05,0.05]	(0.05,+∞)	(-∞,-0.05)	[-0.05,0.05]	(0.05,+∞)
<i>Tacc</i>	-0.053 (-0.289)	0.248** (2.498)	0.039 (0.742)	-0.023 (-0.133)	0.225** (2.289)	0.045 (0.817)
公司固定效应	控制	控制	控制	控制	控制	控制
年度固定效应	控制	控制	控制	控制	控制	控制
样本量	540	2242	911	540	2242	911
R^2	0.572	0.463	0.529	0.600	0.486	0.529

附表 A-10 员工人数对公司避税程度的回归结果

Appendix Table A-10 Regression of *Emp_Size* on Corporate Tax Avoidance

避税衡量指标	<i>TME</i>		<i>NETR</i>	
	(1)	(2)	(3)	(4)
<i>Emp_Size</i> 取值范围	(0,0.04]	(0.04,+∞)	(0,0.04]	(0.04,+∞)
<i>Emp_Size</i>	2.047 (0.824)	-0.008 (-1.208)	1.441 (0.602)	-0.009 (-1.373)
公司固定效应	控制	控制	控制	控制
年度固定效应	控制	控制	控制	控制
样本量	156	4031	156	4031
R^2	0.565	0.398	0.623	0.422

注：由于在分样本情况下样本量较少，假设检验的功效（Power）较低，因此结果缺乏显著性。鉴于此，我们将主要集中员工人数（*Emp_Size*）的回归系数是否符合预期。

附表 A-11 研发支出对公司避税程度的回归结果

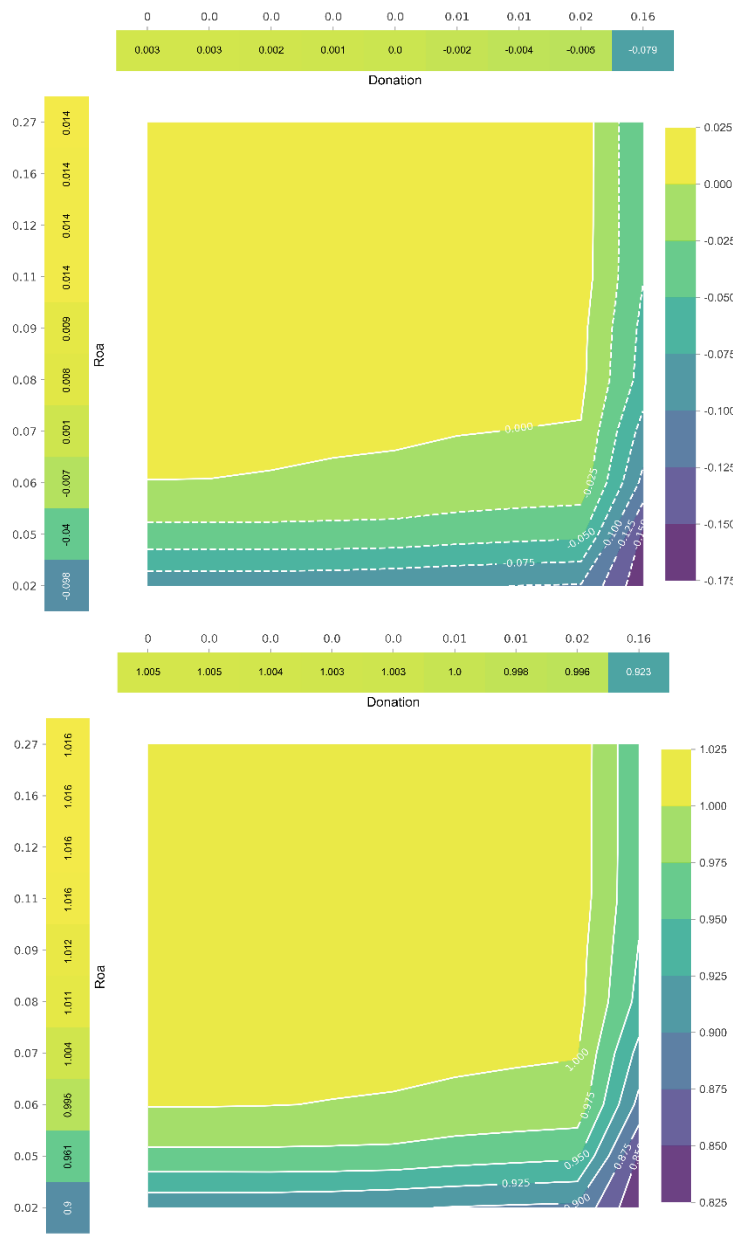
Appendix Table A-11 Regression of *RD* on Corporate Tax Avoidance

避税衡量指标	<i>TME</i>			<i>NETR</i>		
	(1)	(2)	(3)	(4)	(5)	(6)
<i>RD</i> 取值范围	[0,0.02)	[0.02,0.04]	(0.04,+∞)	[0,0.02)	[0.02,0.04]	(0.04,+∞)
<i>RD</i>	-0.067 (-0.099)	1.793** (2.205)	0.533 (0.881)	0.046 (0.071)	1.842** (2.302)	0.455 (0.691)
公司固定效应	控制	控制	控制	控制	控制	控制
年度固定效应	控制	控制	控制	控制	控制	控制
样本量	2538	1021	440	2538	1021	440
R^2	0.424	0.427	0.499	0.442	0.441	0.508

注：当 *RD* 取值范围在[0,0.02)且避税程度指标为 *TME* 时，*RD* 系数为负可能是由于研发支出缺失的样本导致。若控制 *RD_Dummy*，*RD* 的系数均为正。

附录 7：基于随机森林算法的双变量联合部分依赖图

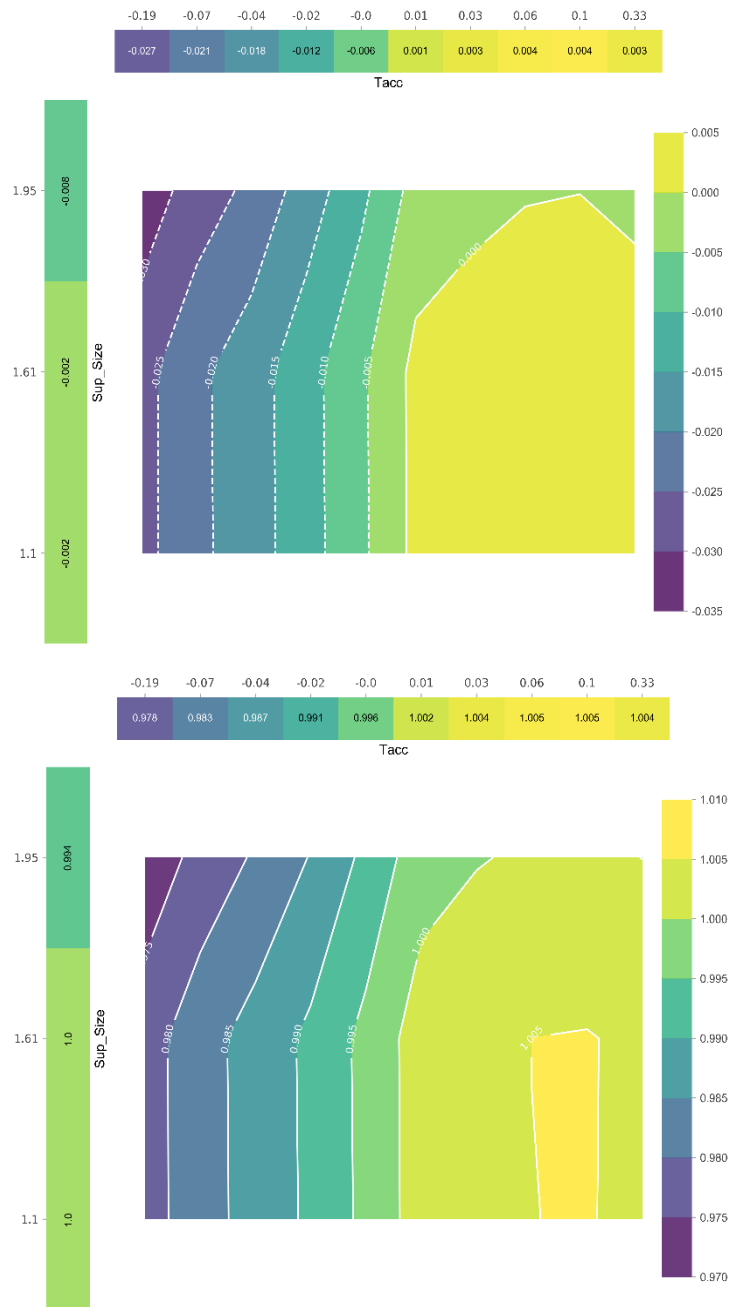
这里，由于 R 软件中绘制随机森林的部分依赖图耗时较长，我们采用 python 软件绘制随机森林算法下双变量的联合部分依赖图。



附图 A-1 捐赠支出和总资产收益率的部分依赖图

Appendix Figure A-1 Partial Dependence Plots of *Donation* and *Roa*

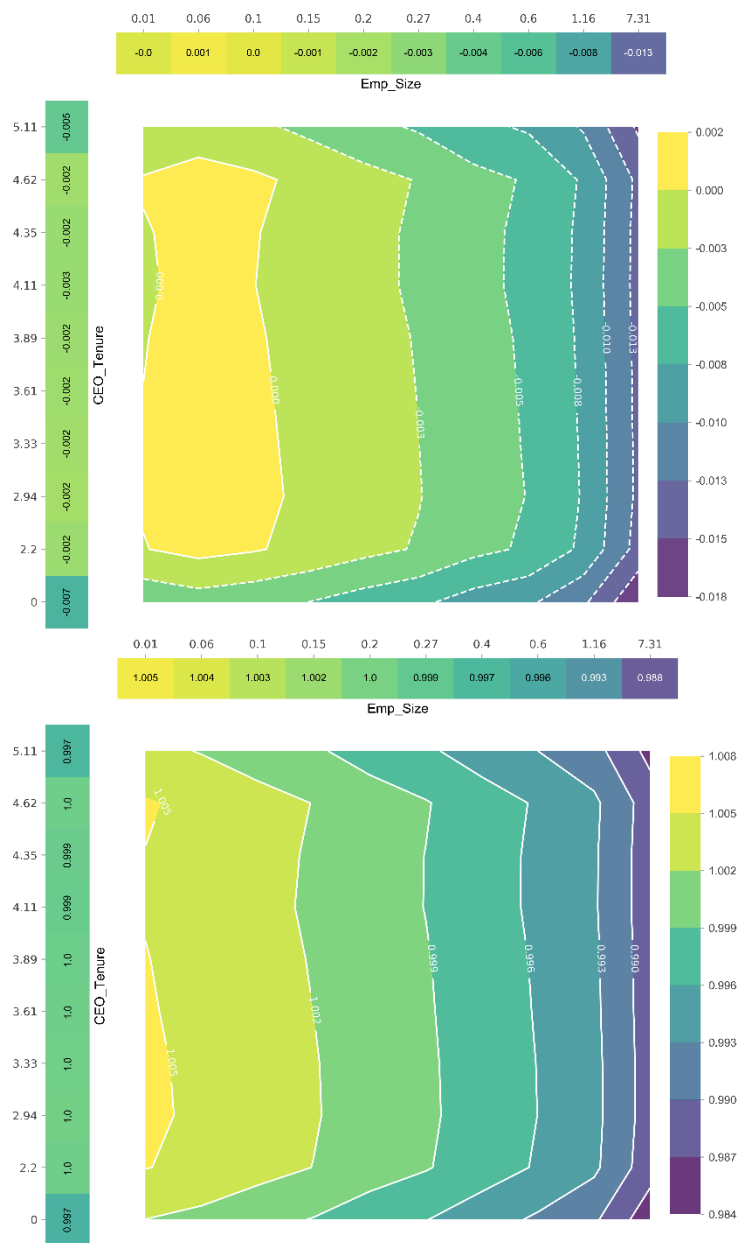
注：X 轴代表企业捐赠规模，Y 轴总资产收益率，右侧色卡代表公司避税程度，颜色越浅，避税程度越高，上图中公司避税程度采用 *TME*，下图采用 *NETR* 衡量。



附图 A-2 应计盈余和监事会规模的部分依赖图

Appendix Figure A-2 Partial Dependence Plots of *Tacc* and *Sup_Size*

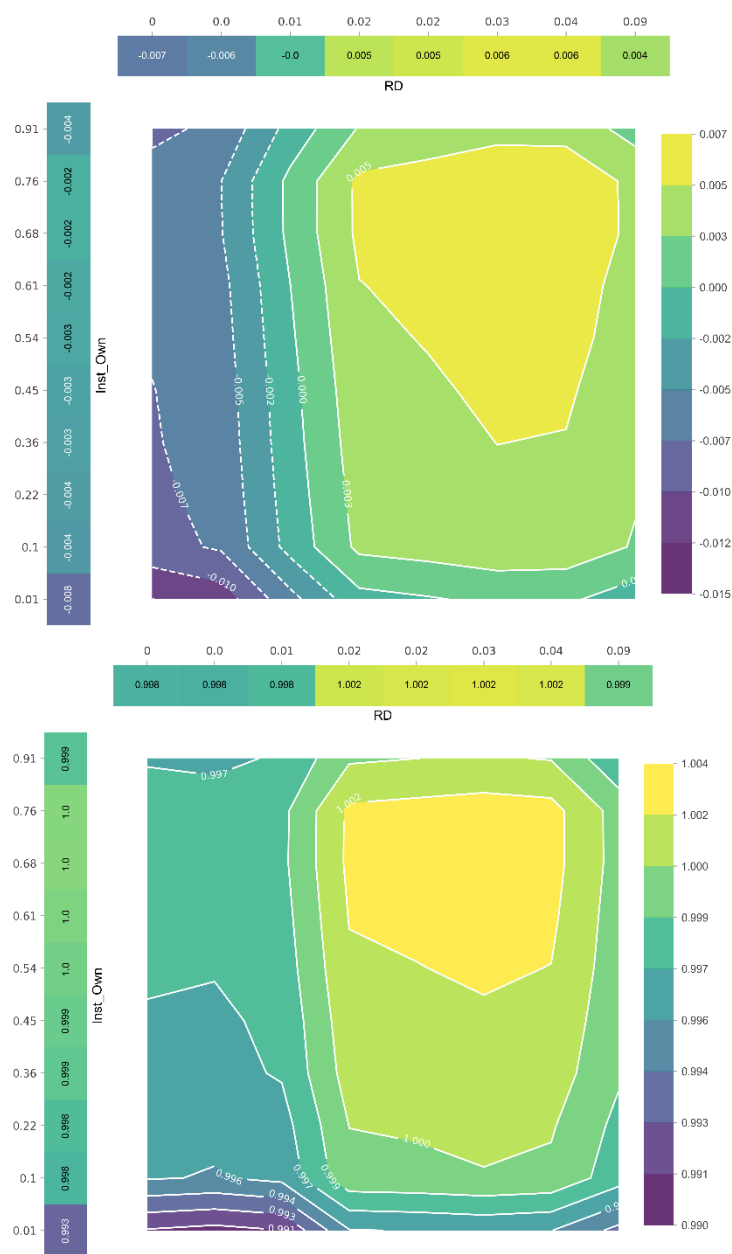
注：X 轴代表应计盈余，Y 轴代表监事会规模，右侧色卡代表公司避税程度，颜色越浅，避税程度越高，上图中公司避税程度采用 *TME*，下图采用 *NETR* 衡量。



附图 A-3 员工人数和 CEO 任期的部分依赖图

Appendix Figure A-3 Partial Dependence Plots of *Emp_Size* and *CEO_Tenure*

注：X 轴代表员工人数，Y 轴代表 CEO 任期，右侧色卡代表公司避税程度，颜色越浅，避税程度越高，上图中公司避税程度采用 *TME*，下图采用 *NETR* 衡量。



附图 A-4 研发支出和机构投资者持股比例的部分依赖图

Appendix Figure A-4 Partial Dependence Plots of RD and $Inst_Own$

注：X 轴代表研发支出，Y 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，上图中公司避税程度采用 TME ，下图采用 $NETR$ 衡量。

附录 8：分类排序评价体系下的预测能力

我们还重点关注了财务预测模型在公司税务稽查情境中的性能。税收征管强调“无风险不打扰、低风险预提醒、中高风险严监控”，这为评价模型实际稽查效果提供了明确指导，即在最小化稽查成本、减少对合规公司的干扰的同时，尽可能覆盖更多的高避税企业。具体而言，我们关注所有高避税公司被正确识别的比例（即召回率，Recall）和识别为高避税公司的实际准确性（即精确率，Precision），揭示财务预测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。

为计算召回率和准确率，我们将回归问题转换为分类问题，按照避税程度由高到低对样本进行排序，并且选择前 10%和 20%分别作为阈值。如果公司的避税程度在对应年度内位列前 10%或 20%，我们将其定义为高避税公司，否则为低避税公司。选择 10%、20%作为阈值是因为真实税务稽查中通常按照 10%-20%对公司进行抽查^④。

这里，我们还参考了 Bao 等^[56]的方法，将回归问题转为排序问题，并引入基于第 K 个位置计算的归一化折损累积收益指标，即 NDCG@K。通过引入这一排序评价指标，我们可以限制评价范围在测试集中具有最高避税水平的前 K 个观测值，从而有效解决高误报率，即被错误识别为高避税公司的数量（False Positive, FP）远高于真正高避税公司数量（True Positive, TP），对公司避税识别问题带来的负面影响。同样地，我们选择测试集样本量的 10%和 20%作为 K 的取值。

评价指标的计算过程如下：首先，为保持一致，我们采用了基于前 K 个样本的修正后的召回率和精确率，其具体定义如式（1）-（2）所示。其中，TP 代表在测试集中，预测避税程度和真实避税程度均位列前 K 个的样本数量，FN（False Negative）是指测试集中，预测避税程度未进入前 K 个，但真实避税程度位列前 K 个的样本数量，FP 是测试集中预测避税程度位于前 K 个，但真实避税程度未进入前 K 个的样本数量。

$$Recall = \frac{TP}{TP + FN} \quad (1)$$

$$Precision = \frac{TP}{TP + FP} \quad (2)$$

关于 NDCG@K 指标：我们首先按照 GBRT、RF 算法预测的公司避税程度由高到底排序，计算 DCG@K 值，计算公式见（3），对于每个公司-年度观察值，如果其真实避税程度位于该年度的前 K 位， rel_i 取值为 1，否则取 0。随后，计算归一化处理后的 NDCG@K，计算公式见（4），这里， $ideal\ DCG@K$ 代表所有真实避税企业都恰好位于前 K 个位置时的 DCG@K 值。NDCG@K 的取值范围在 0 到 1 之间，其数值越高表示模型性能越好。

$$DCG@K = \sum_{i=1}^K \frac{(2^{rel_i} - 1)}{\log_2(i + 1)} \quad (3)$$

$$NDCG@K = \frac{DCG@K}{ideal\ DCG@K} \quad (4)$$

分类排序评价体系下的预测能力结果见附表 A-12 所示。可以看到，在界定高避税公司的门槛设为前 20%时，无论选择何种算法、何种避税衡量指标、何种评价指标，加入避税工具特征均能够提升预测效果；当高避税公司的划分标准为前 10%时，使用 TME 作为避税衡量指标，同样能观察到加入避税工具特征对预测结果的增强，这主要是因为 TME 考虑了公司名义税率的差异，使得基于 TME 的划分更为准确，进而增强了模型的预测能力，而相比之下，采用 NETR 指标划分高避税和低避税公司时，模型的预测能力普遍较低。此外，采用 NETR 指标划分高避税和低避税公司时，行业层面指标表现出很强的预测能力，这可能是由于同行业公司受相同税收政策影响，其实际税率更加接近，因此，行业层面指标对于 NETR 的预测能力

④ 我们研究了各地税务部门税务稽查随机抽查工作方案，发现税务部门一般每年按照 10%或 20%的比例抽查重点稽查对象企业账目。例如：河北省税务局 2022 年内部联合随机抽查工作计划显示对于重点稽查对象的抽查比例为 10%，详情见：http://hebei.chinatax.gov.cn/hbswxgk/gkml/202201/t20220113_3024540.html；宁夏区税务局 2022 年税务稽查随机抽查工作方案显示对于重点稽查对象的抽查比例为 20%，详情见：http://ningxia.chinatax.gov.cn/art/2022/6/29/art_10940_270526.html。因此，我们选取 10%和 20%作为划分标准。考虑到上市公司通常是当地的大型企业，是重要的税收来源，很可能成为重点稽查对象，因此，我们根据重点稽查对象的抽查比例来划分高避税公司和低避税公司。在稳健性检验预测高避税公司与低避税公司一节中，我们采取相同的划分标准。

附表 A-12 分类评价体系下的预测能力结果

Appendix Table A-12 Recall Precision NDCG@K of Predictive Power of Models

附表 A-12.1 以当年度公司避税程度前 10% 为分界, 划分高避税与低避税公司的对比结果												
避税衡量指标	TME						NETR					
机器学习算法	GBRT			RF			GBRT			RF		
评价指标	Recall	Precision	NDCG@K	Recall	Precision	NDCG@K	Recall	Precision	NDCG@K	Recall	Precision	NDCG@K
基准模型	27.22	28.38	0.2870	25.41	27.18	0.2842	8.49	9.18	0.0874	10.33	11.57	0.1359
加入公司内部特征												
基准模型+避税工具模型	27.65	29.62	0.3103	26.90	29.45	0.3164	7.72	8.31	0.0767	12.24	13.39	0.1374
基准模型+公司治理特征	24.25	25.55	0.2634	24.61	25.80	0.2954	8.79	9.26	0.0870	10.24	11.21	0.1107
基准模型+其他内部特征	27.63	28.99	0.2905	25.14	27.20	0.2804	8.88	9.55	0.1000	11.15	12.36	0.1445
内部模型	27.17	29.32	0.3028	26.27	28.32	0.3020	5.73	6.12	0.0617	11.80	12.88	0.1228
加入公司外部特征												
基准模型+宏观层面特征	24.36	26.32	0.2635	23.13	25.75	0.2732	9.18	9.72	0.0951	10.47	11.74	0.1158
基准模型+行业层面特征	25.85	27.85	0.3036	28.38	30.81	0.3204	11.03	12.71	<u>0.1439</u>	15.71	17.24	0.1803
基准模型+外部微观特征	26.28	27.88	0.2800	24.12	26.55	0.2833	9.31	10.06	0.0918	10.03	11.39	0.1264
外部模型	24.89	26.93	0.3003	25.20	28.02	0.2958	<u>12.11</u>	<u>13.89</u>	0.1343	14.13	15.44	0.1579
加入全部公司内外部特征												
全模型	27.02	29.18	0.3033	27.38	29.55	0.3113	8.91	9.98	0.0971	12.41	13.64	0.1452
附表 A-12.2 以当年度公司避税程度前 20% 为分界, 划分高避税与低避税公司的对比结果												
避税衡量指标	TME						NETR					
机器学习算法	GBRT			RF			GBRT			RF		
评价指标	Recall	Precision	NDCG@K	Recall	Precision	NDCG@K	Recall	Precision	NDCG@K	Recall	Precision	NDCG@K
基准模型	27.17	45.81	0.4078	25.99	44.18	0.3946	12.72	22.56	0.2266	15.12	26.71	0.2601
加入公司内部特征												
基准模型+避税工具模型	<u>28.17</u>	47.19	0.4224	28.36	48.30	0.4123	13.54	23.94	0.2290	16.75	29.80	0.2628
基准模型+公司治理特征	25.63	42.11	0.4004	24.63	41.55	0.3828	13.59	23.83	0.2383	14.12	24.97	0.2355
基准模型+其他内部特征	27.17	45.95	0.4070	25.53	43.76	0.3870	13.22	23.41	0.2382	15.47	27.35	0.2639
内部模型	28.14	47.77	0.4114	27.51	46.97	0.4021	11.49	20.00	0.2266	15.61	27.41	0.2583
加入公司外部特征												
基准模型+宏观层面特征	25.99	42.85	0.4123	25.94	44.77	0.3776	12.16	21.30	0.2228	14.00	24.92	0.2304
基准模型+行业层面特征	27.10	46.20	0.4120	26.80	45.82	0.3957	<u>16.62</u>	<u>29.27</u>	<u>0.2770</u>	20.47	35.80	0.3120
基准模型+外部微观特征	26.90	45.02	0.4051	26.27	44.62	0.3956	13.50	23.89	0.2279	15.45	27.28	0.2587
外部模型	26.66	44.62	0.4217	27.12	46.53	0.3892	16.21	28.40	0.2589	18.78	32.53	0.2935
加入全部公司内外部特征												
全模型	27.88	47.23	0.4182	28.54	48.74	<u>0.4140</u>	13.64	23.71	0.2540	19.11	33.57	0.2950

较强。

为进一步反映公司避税预测模型的效果，我们将其与相似的财务预测任务进行了简单比较。Höglund^[57]的研究关注公司拖欠税款行为，其通过 100 次遗传算法预测得到的模型平均准确率在 71.60%~72.40%之间。Abedin 等^[58]研究了在公司拖欠税款行为预测中特征变换对多种机器学习算法的影响，经过特征转换后，其模型准确率大约在 61.33%~73.18%之间。相比之下，本文的公司避税行为预测模型表现更好，无论使用何种评价指标、模型或算法，其预测准确率均超过 74%，位于 74.06%~86.92%之间。

另外，Bao 等^[59]采用了与本文相同的分类评价指标体系来预测财务欺诈行为，其召回率为 4.88%，精确率为 4.48%，NDCG@K 为 0.049。与之比对，本文的预测模型在召回率、精确率、NDCG@K 等方面均更为出色，具体而言，本文的预测模型，其召回率介于 24.15%~28.73%之间，精确率介于 25.58%~49.42%之间，NDCG@K 大致在 0.28~0.42 之间。

附录 9：使用公司预测特征水平值对未来期避税行为预测

预测公司未来期避税行为也是一个十分重要的话题^[59]，具有重大实际意义。本文研究结果显示避税工具特征在预测公司当期避税行为中至关重要。那么，避税工具特征能否预测未来期避税程度？为回答该问题，我们运用第 T-1 期至第 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。其中，第 T-1 期预测特征的变化量等于第 T-1 期期初值减去第 T-1 期期末值，第 T-2 期预测特征的变化量等于第 T-2 期期初值减去第 T-1 期期末值，依次类推。选择 3 年时限来研究我国公司未来期避税行为，是基于我国税法关于结转年限的相关规定^⑤，确保契合我国企业的实际运营环境和避税行为模式。

附表 A-13 展示了在随机种子设定为 1 时，以 R^2_{oos} 作为评价指标，采用 OLS、GBRT、RF 对公司未来期避税行为预测的结果。我们观察到：随着时间跨度的增加，预测特征水平值和预测特征变化量对公司未来期避税行为的预测能力呈现先增强后减弱的趋势，并在使用第 T-1 期至第 T-2 期预测特征水平值和相应的预测特征变化量时效果最好。这表明，随时间推移，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，特别是较远期（如第 T-3 期）。这一趋势符合直觉，因为公司的避税策略本质上是动态且复杂的，过去较远期（如第 T-3 期）的信息已不足以准确捕捉公司未来期避税行为。第 T-2 期表现更好的原因可能在于，使用较短的时间跨度（如第 T-1 期）时，预测特征变化量可能过于灵敏，容易受到短期波动的影响。此外，在考虑预测特征变化量时，样本数量显著减少也对模型的预测能力产生了直接影响^⑥。由于训练集规模的缩减会降低模型内树之间的多样性，进而影响 RF 算法发掘数据潜在复杂模式的能力，故 RF 算法受样本减少影响尤为严重。鉴于此，我们着重分析基于第 T-1 期、第 T-1 期至第 T-2 期 OLS 与 GBRT 算法在预测公司未来期避税行为的效果。

在运用 OLS 算法预测避税程度指标 TME 、 $NETR$ 时，无论是基于第 T-1 期或第 T-1 期至第 T-2 期的预测数据（附表 A-13 列（1）、（4）所示），基准模型在加入避税工具特征之后均能显著提升 R^2_{oos} ，增幅为 3.26~3.98 个百分点。而当加入公司治理特征、其他内部特征、宏观经济特征、行业层面特征和外部微观特征时， R^2_{oos} 的增幅不明显或有所下降。而将 43 个内部特征和 21 个外部特征分别加入基准模型，形成内部模型和外部模型时，其预测能力也弱于仅加入避税工具特征。

运用 GBRT 预测避税程度指标 TME 、 $NETR$ 时，基于第 T-1 期、第 T-1 期至第 T-2 期的预测结果显示（附表 A-13 列（2）、（5）所示），在基准模型加入避税工具特征后， R^2_{oos} 平均增幅最大，达到 0.84 个百分点。说明避税工具特征同样是预测公司未来期避税行为的关键。

进一步，通过比较 GBRT 与 OLS 预测结果，我们发现在第 T-1 期、第 T-1 期至第 T-2 期预测特征，运用 GBRT 预测 TME 时，各模型提升效果在 3.83~7.40 个百分点，预测 $NETR$ 时，各模型提升效果介于 3.25~9.28 个百分点之间。同样表明，使用非线性算法在识别公司未来期避税能力具有优势，并且过去预测特征与公司未来期避税行为也极可能是非线性、复杂的关系。

⑤ 例如，《企业所得税法》第九条规定：“企业发生的公益性捐赠支出，在年度利润总额 12% 以内的部分，准予在计算应纳税所得额时扣除；超过年度利润总额 12% 的部分，准予结转以后三年内在计算应纳税所得额时扣除。”《企业所得税法》第十八条规定：“企业纳税年度发生的亏损，准予向以后年度结转，用以后年度的所得弥补，但结转年限最长不得超过五年。”Dyreng 等^[60]研究公司避税问题时选择 5 年和 10 年为时限，主要是因为发达国家，结转弥补期限通常是无期限的。

⑥ 由于 RF 算法不提供缺失值处理功能，我们在使用预测特征水平值及其变化量进行预测时，剔除了含有缺失值的样本。处理后最终得到：第 T-1 期、第 T-1 期至第 T-2 期、第 T-1 期至第 T-3 期预测特征水平值及其变化量样本分别为 8725、5949、4090 个，相较于原始样本数量明显减少。

附表 A-13 加入特征变化量的未来期避税行为预测结果（单位：%）

Appendix Table A-13 R^2_{OOS} of Predictive Power in Future Corporate Tax Avoidance with Feature Dynamics

附表 A-13.1 用 <i>TME</i> 作为公司避税程度衡量指标									
	第 T-1 期			第 T-1 期至第 T-2 期			第 T-1 期至第 T-3 期		
预测模型	OLS	GBRT	RF	OLS	GBRT	RF	OLS	GBRT	RF
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
基准模型	9.23	15.49	12.66	11.13	18.10	13.45	11.56	15.54	10.61
加入公司内部特征									
基准模型+避税工具特征	<u>13.21</u>	17.04	<u>14.08</u>	<u>14.39</u>	19.11	<u>14.93</u>	<u>12.19</u>	15.09	<u>12.10</u>
基准模型+公司治理特征	6.90	13.28	10.38	7.40	14.56	9.39	4.14	11.78	7.85
基准模型+其他内部特征	9.24	15.46	12.02	11.26	17.78	12.41	11.69	15.45	9.88
内部模型	11.69	15.97	12.82	11.54	17.34	13.01	6.57	14.08	10.87
加入公司外部特征									
基准模型+宏观层面特征	8.59	14.46	9.53	9.09	16.17	10.87	8.65	12.46	8.26
基准模型+行业层面特征	9.28	15.44	11.85	11.07	18.24	12.92	11.24	15.69	10.93
基准模型+外部微观特征	9.69	15.40	12.69	11.68	17.70	12.91	11.58	14.44	10.28
外部模型	8.99	14.60	10.57	9.53	16.50	11.50	8.12	12.68	9.20
加入全部公司内外部特征									
全模型	10.93	15.66	12.63	9.59	16.99	13.23	3.21	13.10	10.85
附表 A-13.2 用 <i>NETR</i> 作为公司避税程度衡量指标									
	第 T-1 期			第 T-1 期至第 T-2 期			第 T-1 期至第 T-3 期		
预测模型	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
基准模型	8.92	15.33	12.88	10.44	17.96	13.85	10.58	16.12	12.79
加入公司内部特征									
基准模型+避税工具特征	<u>12.90</u>	16.15	<u>14.20</u>	<u>13.76</u>	17.94	<u>14.85</u>	<u>11.44</u>	13.75	12.24
基准模型+公司治理特征	7.29	13.41	10.85	7.87	14.63	10.01	4.67	12.83	10.30
基准模型+其他内部特征	8.99	15.38	12.76	10.49	17.53	13.46	10.83	16.04	12.32
内部模型	11.78	15.18	12.92	11.54	15.83	13.30	6.58	13.01	10.98
加入公司外部特征									
基准模型+宏观层面特征	8.62	14.95	10.66	8.83	16.46	12.43	8.00	13.72	10.33
基准模型+行业层面特征	8.53	15.92	13.06	10.02	19.30	14.75	9.79	17.56	<u>13.70</u>
基准模型+外部微观特征	9.29	15.03	12.80	10.84	17.45	13.31	10.49	15.51	11.95
外部模型	8.50	15.33	11.74	8.75	17.93	13.02	6.74	15.36	11.25
加入全部公司内外部特征									
全模型	11.23	15.49	13.01	9.92	16.64	13.44	3.45	13.50	11.26

同时，我们也尝试了仅使用第 T-1 期至第 T-3 期预测特征水平值对第 T 期公司避税程度进行预测，结果如附表 A-14 所示。同样可以看到，预测特征对公司未来期避税行为的预测能力逐渐减弱。即使用第 T-1 期、第 T-2 期、第 T-3 期特征对第 T 期公司避税程度的预测能力逐渐下降，且相较于使用第 T-3 期预测特征，使用第 T-1 期、第 T-2 期预测特征时，模型在识别公司未来期避税行为方面表现相对更好。并且与额外引入预测特征变化量（附表 A-13）时相比，仅考虑预测特征过去期的水平值时模型对公司未来期避税行为的预测能力整体较弱，这表明，在预测公司未来期避税行为时需要考虑避税特征的动态演变，而不仅仅是当期水平，以更准确地捕捉公司未来期避税行为的趋势。

鉴于 RF 算法受缺失值影响严重，且较远期数据的预测能力较弱，所以我们主要关注基于第 T-1 期、第 T-2 期数据，OLS 与 GBRT 算法在预测公司未来期避税行为的效果。在运用 OLS、GBRT 算法，我们再次发现，无论使用 *TME* 还是 *NETR* 来衡量避税程度，加入避税工具特征均能够显著提升模型的预测能力，这进一步验证了避税工具是识别公司未来期避税行为的关键特征。

具体而言，在运用 OLS 算法预测避税程度指标 *TME*、*NETR* 时，无论是基于第 T-1 期或第 T-2 期的预测数据（附表 A-14 列（1）、（4）所示），基准模型在加入避税工具特征之后均能显著提升

R^2_{oos} ，增幅为 1.14~1.75 个百分点。相比之下，当加入公司治理特征、其他内部特征、宏观经济特征、行业层面特征和外部微观特征时， R^2_{oos} 的增幅不明显，甚至有所下降，即使将 43 个内部特征和 21 个外部特征分别加入基准模型形成内部模型和外部模型时，其预测能力也未能超越仅加入避税工具特征的模型；运用 GBRT 预测避税程度指标 TME 、 $NETR$ 时，我们也观察到类似现象，在基准模型加入避税工具特征后， R^2_{oos} 平均增幅最大，达到 0.34 个百分点。并且，通过比较 GBRT 与 OLS 预测结果，我们发现 GBRT 在预测 TME 和 $NETR$ 时，各模型提升效果更为显著，分别提升 0.66~2.33 个百分点和 0.50~2.83 个百分点。这表明，使用非线性算法在识别公司未来期避税能力具有优势，并且过去预测特征与公司未来期避税行为也极可能是非线性、复杂的关系。

附表 A-14 未来期避税行为预测结果（单位：%）

Appendix Table A-14 R^2_{oos} of Predictive Power of in Future Corporate Tax Avoidance

附表 A-14.1 用 TME 作为公司避税程度衡量指标									
	T-1 期			T-2 期			T-3 期		
预测模型	OLS	GBRT	RF	OLS	GBRT	RF	OLS	GBRT	RF
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
基准模型	3.51	4.98	1.51	1.94	3.39	0.45	2.44	2.88	0.19
加入公司内部特征									
基准模型+避税工具特征	<u>4.91</u>	5.57	3.35	<u>3.08</u>	3.85	1.46	2.28	2.86	0.31
基准模型+公司治理特征	1.97	4.30	2.26	0.33	2.06	0.29	0.67	1.86	0.78
基准模型+其他内部特征	3.52	4.79	1.05	2.05	3.31	0.06	<u>2.58</u>	2.95	0.12
内部模型	3.78	5.35	3.82	1.76	3.41	<u>1.79</u>	0.71	2.24	1.15
加入公司外部特征									
基准模型+宏观层面特征	3.18	4.82	-1.34	1.95	3.33	-0.91	2.29	2.28	-1.20
基准模型+行业层面特征	3.76	5.17	2.54	2.30	3.04	0.81	2.46	2.51	-0.51
基准模型+外部微观特征	3.44	4.93	2.33	1.72	3.46	1.43	2.05	2.74	0.18
外部模型	3.30	4.78	1.73	2.01	3.09	0.08	1.82	2.23	-0.01
加入全部公司内外部特征									
全模型	3.48	5.40	<u>3.93</u>	1.60	3.44	1.98	-0.32	1.72	<u>1.48</u>
附表 A-14.2 用 $NETR$ 作为公司避税程度衡量指标									
	T-1 期			T-2 期			T-3 期		
预测模型	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
基准模型	4.72	6.89	4.00	2.87	4.86	2.56	2.61	3.37	1.58
加入公司内部特征									
基准模型+避税工具特征	<u>6.47</u>	7.11	5.66	<u>4.28</u>	4.95	3.44	2.83	3.52	1.93
基准模型+公司治理特征	3.98	6.30	4.42	2.26	3.87	2.37	1.66	2.22	1.34
基准模型+其他内部特征	4.79	6.78	3.68	2.85	4.74	2.35	2.63	3.32	1.74
内部模型	6.02	6.87	6.15	3.80	4.30	3.52	2.07	2.79	2.32
加入公司外部特征									
基准模型+宏观层面特征	4.78	6.87	1.58	3.36	<u>5.22</u>	1.38	2.88	2.78	-0.53
基准模型+行业层面特征	4.52	7.35	5.34	2.87	4.71	3.33	2.38	3.16	1.66
基准模型+外部微观特征	4.68	6.87	4.60	2.76	4.86	3.12	2.18	3.28	1.37
外部模型	4.48	7.12	4.09	3.12	5.15	2.58	2.09	2.67	1.14
加入全部公司内外部特征									
全模型	5.95	7.41	<u>6.25</u>	4.03	4.78	<u>3.69</u>	1.55	2.21	<u>2.33</u>

附录 10：稳健性检验结果

1. 更换模型参数

由于机器学习算法需要事先指定参数，为避免参数选择产生的影响，我们更换参数重新预测。

对于 GBRT，我们将学习率设为 0.0005、0.001、0.002，将交互深度设为 2、4、6，重新进行预测，结果如表 A-15 所示。可以看到，更换参数之后，样本外 R 方与主要结果非常接近。

附表 A-15 更换 GBRT 算法参数的样本外 R 方 (R^2_{os})

Appendix Table A-15 R^2_{os} of Using Alternative GBRT Parameters											
附表 A-15.1 TME 作为避税程度的衡量指标 (单位: %)											
参数组合		基准模型	基准模型+避税工具特征	基准模型+公司治理特征	基准模型+其他内部特征	内部模型	基准模型+宏观层面特征	基准模型+行业层面特征	基准模型+外部微观特征	外部模型	全模型
学习率	交互深度										
0.0005	2	12.64	<u>15.21</u>	12.19	12.68	15.08	12.42	12.67	12.61	12.48	15.08
	4	13.33	<u>16.41</u>	12.87	13.46	16.21	12.99	13.60	13.28	13.31	16.28
	6	13.36	16.68	12.99	13.52	16.59	13.03	13.71	13.33	13.46	<u>16.73</u>
0.001	2	13.27	<u>16.52</u>	12.62	13.46	16.09	12.82	13.53	13.19	13.12	16.08
	4	13.23	<u>16.96</u>	12.76	13.49	16.68	12.74	13.66	13.25	13.32	16.75
	6	12.90	<u>16.98</u>	12.53	13.19	16.82	12.39	13.51	13.09	13.10	16.96
0.002	2	13.13	<u>17.00</u>	12.18	13.32	16.38	12.44	13.39	13.05	12.80	16.33
	4	12.81	<u>16.94</u>	11.67	12.74	16.70	11.89	12.93	13.19	12.40	16.67
	6	12.70	<u>16.81</u>	11.30	12.43	16.59	11.55	12.82	13.03	11.77	16.65
附表 A-15.2 NETR 作为避税程度的衡量指标 (单位: %)											
参数组合		基准模型	基准模型+避税工具特征	基准模型+公司治理特征	基准模型+其他内部特征	内部模型	基准模型+宏观层面特征	基准模型+行业层面特征	基准模型+外部微观特征	外部模型	全模型
学习率	交互深度										
0.0005	2	12.70	<u>14.53</u>	12.11	12.77	14.29	12.43	12.75	12.66	12.55	14.39
	4	13.49	15.60	12.69	13.57	15.29	13.18	13.79	13.38	13.58	<u>15.61</u>
	6	13.57	15.91	12.81	13.67	15.63	13.37	14.04	13.49	13.88	<u>16.09</u>
0.001	2	13.60	<u>15.86</u>	12.73	13.74	15.27	13.13	13.97	13.45	13.63	15.65
	4	13.77	16.44	12.88	13.92	15.97	13.40	14.32	13.66	14.12	<u>16.56</u>
	6	13.56	16.60	12.78	13.69	16.18	13.23	14.23	13.59	14.06	<u>16.85</u>
0.002	2	13.72	<u>16.60</u>	12.62	13.92	15.90	13.12	14.28	13.52	13.73	16.44
	4	13.60	16.65	12.33	13.34	16.35	13.01	14.02	13.58	13.65	<u>17.03</u>
	6	13.44	16.68	12.04	13.30	16.45	12.73	13.94	13.54	13.36	<u>17.10</u>

对于 RF，我们将树的个数 (n) 设为 1000、2000、3000，将内部模型的变量个数 (m) 设为 7、11、18、22，将外部模型的变量个数设为 5、8、12、15，将全模型的变量个数设为 8、17、31、40，重新进行预测^⑦，结果如附表 A-16 所示。可以看到，更换参数之后，结果仍然一致。

附表 A-16 更换 RF 算法参数的样本外 R 方 (R^2_{os})

Appendix Table A-16 R^2_{os} of Using Alternative RF Parameters												
附表 A-16.1 内部模型 (单位: %)												
参数组合	n=3000				n=2000				n=1000			
	7	11	18	22	7	11	18	22	7	11	18	22

⑦ 这里我们选取变量个数 m 的规则是在 0 和总变量个数二分之一之间随机选取。这里由于加入 6 个子类特征时，特征总数较少， m 的取值范围较小，因此我们只修改树的数量进行预测，结果仍然一致，为了简洁，在此不展示该结果。

<i>TME</i>	15.60	16.30	16.21	16.25	15.50	16.26	16.29	16.27	15.46	16.26	16.31	16.19
<i>NETR</i>	15.07	15.77	15.95	15.78	15.02	15.74	16.02	15.83	14.95	15.79	15.90	15.87
附表 A-16.2 外部模型（单位：%）												
参数组合	n=3000				n=2000				n=1000			
	5	8	12	15	5	8	12	15	5	8	12	15
<i>TME</i>	11.38	12.56	12.86	12.86	11.34	12.51	12.94	12.85	11.32	12.49	12.96	12.79
<i>NETR</i>	12.17	13.21	13.70	13.70	12.08	13.18	13.69	13.70	12.09	13.12	13.58	13.66
附表 A-16.3 全模型（单位：%）												
参数组合	n=3000				n=2000				n=1000			
	8	17	31	40	8	17	31	40	8	17	31	40
<i>TME</i>	15.05	16.52	16.49	16.19	15.03	16.53	16.50	16.15	14.88	16.53	16.47	16.17
<i>NETR</i>	15.10	16.33	16.50	16.29	15.02	16.37	16.50	16.29	15.02	16.24	16.55	16.10

2. 更换样本

主要结果中，我们选随机取了样本中三分之二的公司作为训练集，三分之一公司作为测试集。为避免样本对结果产生影响，我们重新随机抽取三分之二的公司作为新训练集、另外三分之一公司作为新测试集进行预测，结果见附表 A-17。可以看到，更换训练集和测试集后，内部模型、全模型预测能力明显仍然好于外部模型、基准模型，且 GBRT、RF 的预测能力仍明显好于 OLS，与主要结果一致。

附表 A-17 更换训练集和测试划分方式后的样本外 R 方 (R^2_{oos})

Appendix Table A-17 R^2_{oos} of Changing Training Set and Test Set										
附表 A-17.1 用 <i>TME</i> 作为公司避税程度衡量指标（单位：%）										
	基准模型	基准模型+避税工具特征	基准模型+公司治理特征	基准模型+其他内部特征	内部模型	基准模型+宏观层面特征	基准模型+行业层面特征	基准模型+外部微观特征	外部模型	全模型
算法	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
OLS	8.91	<u>11.76</u>	7.53	8.76	10.52	8.86	9.19	8.90	9.16	10.48
GBRT	13.02	17.09	12.53	13.27	17.06	12.94	13.63	13.28	13.76	17.17
RF	11.70	15.73	10.71	11.67	15.49	11.39	12.52	13.12	13.69	<u>15.85</u>
附表 A-17.2 用 <i>NETR</i> 作为公司避税程度衡量指标（单位：%）										
	基准模型	基准模型+避税工具特征	基准模型+公司治理特征	基准模型+其他内部特征	内部模型	基准模型+宏观层面特征	基准模型+行业层面特征	基准模型+外部微观特征	外部模型	全模型
算法	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
OLS	7.41	<u>9.83</u>	6.08	7.25	8.48	7.54	7.67	7.46	7.95	8.88
GBRT	12.14	15.90	11.25	12.31	15.60	12.50	13.05	12.31	13.46	16.14
RF	10.71	14.65	10.15	11.06	14.25	11.39	12.52	13.12	13.42	<u>14.96</u>

3. 更换算法

为确保结果稳健，我们还选用了 LASSO（Least Absolute Shrinkage and Selection Operator）和极度梯度提升回归树（eXtreme Gradient Boosting，简称 XGBoost）这两个机器学习算法重新进行拟合。

LASSO 回归与 OLS 回归类似，也是线性回归算法。其区别在于，OLS 以均方误差作为损失函数，选取能够最小化均方误差的参数作为估计值，这样会导致在样本特征很多、样本数量相对较少时出现过拟合问题。而 LASSO 回归通过对损失函数添加惩罚项 $\|\beta\|_1$ ，防止模型过拟合，其损失函数为 $\sum (y_i - \beta^T x_i)^2 + \lambda \|\beta\|_1$ ，其中 λ 为正则化参数， $\lambda > 0$ ， i 为观测个体， β 为线性模型系数。若 λ

取 0，该模型则退化成 OLS 模型；若将惩罚项 $\|\beta\|_1$ （L1 正则化项）换为 $\|\beta\|_2^2$ （L2 正则化项），该模型则变为岭回归模型。LASSO 回归的优势在于， β 的估计值中会出现更多的零向量，有助于我们在众多的、存在高度相关性的指标中选出重要的指标。

XGBoost 是梯度提升回归树算法的一个优化算法，它与 GBRT 的主要区别在于 XGBoost 在优化过程中对于损失函数进行了二阶泰勒展开，而 GBRT 沿着负梯度的方向进行优化时只运用了一阶梯度信息，而且 XGBoost 能够实现并行计算。具体算法见 Chen 和 Guestin^[61]。这里参数的选取与 GBRT 一致。

LASSO 和 XGBoost 的预测结果如附表 A-18 所示。可以看到，在更换算法后，加入避税工具因素对于 R^2_{OOS} 的提升作用仍然最为显著，LASSO 回归的表现与 OLS 表现近似，XGBoost 算法的表现与 GBRT 算法表现非常接近，显著好于 OLS 和 LASSO。这些结果再次说明非线性的机器学习算法在预测能力上要优于线性算法。

附表 A-18 更换算法后的样本外 R 方 (R^2_{OOS})

Appendix Table A-18 R^2_{OOS} of Changing Machine Learning Algorithm

附表 A-18.1 用 <i>TME</i> 作为公司避税程度衡量指标（单位：%）										
	基准模型	基准模型+避税工具特征	基准模型+公司治理特征	基准模型+其他内部特征	内部模型	基准模型+宏观层面特征	基准模型+行业层面特征	基准模型+外部微观特征	外部模型	全模型
算法	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
LASSO	7.99	<u>12.28</u>	7.95	8.00	12.09	8.05	8.18	8.18	8.42	8.13
XGBoost	12.72	16.67	11.17	12.93	16.80	11.84	13.15	12.90	12.34	16.89
附表 A-18.2 用 <i>NETR</i> 作为公司避税程度衡量指标（单位：%）										
	基准模型	基准模型+避税工具特征	基准模型+公司治理特征	基准模型+其他内部特征	内部模型	基准模型+宏观层面特征	基准模型+行业层面特征	基准模型+外部微观特征	外部模型	全模型
算法	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
LASSO	8.42	<u>12.04</u>	8.58	8.38	11.91	8.50	8.48	8.29	8.49	11.66
XGBoost	13.47	16.19	12.08	13.49	15.95	12.52	13.75	13.41	13.45	16.84

4. 更换公司避税程度衡量指标

由于避税程度的衡量方式很多，为了避免避税衡量方式不同对结果的影响，我们借鉴以往研究，使用行业调整的账面-税收差异 (*BTD*) 作为另一公司避税程度衡量方式，重新进行预测。账面-税收差异等于（税前利润－应纳税所得额）/期末资产总额，其中，应纳税所得额等于（所得税费用－递延所得税费用）/名义税率。较高的账面-税收差异表明公司避税程度较大。

预测结果如附表 A-19 所示。可以看到，在更换了公司避税衡量方式之后，加入避税工具因素对于 R^2_{OOS} 的提升作用仍然最为显著，且 GBRT 的预测能力也明显好于 OLS，RF 在特征数量多且特征重要性高的模型中，其预测能力也优于 OLS，与主要结果一致。

附表 A-19 运用 *BTD* 作为避税衡量指标的样本外 R 方 (R^2_{OOS})

Appendix Table -19 R^2_{OOS} of Using *BTD* as a Measure of Tax Avoidance

	基准模型	基准模型+避税工具特征	基准模型+公司治理特征	基准模型+其他内部特征	内部模型	基准模型+宏观层面特征	基准模型+行业层面特征	基准模型+外部微观特征	外部模型	全模型
算法	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)

OLS	7.48	<u>9.76</u>	6.42	7.54	8.96	7.14	7.80	7.67	7.68	8.89
GBRT	8.21	<u>12.76</u>	8.23	8.48	12.58	7.76	8.93	8.95	8.77	12.56
RF	6.41	12.27	7.70	6.76	<u>12.29</u>	5.51	7.41	7.61	7.85	12.14

5. 预测高避税公司与低避税公司

税收征管需要根据税收风险进行精准监管。据国家税务总局发布的《推进税务稽查随机抽查实施方案》指出，定向抽查对象名单要按照稽查对象的税收风险等级等特定条件进行抽取。因此，在实践中识别公司避税所处的水平等级往往比预测避税程度大小更为重要。因此我们将样本分为高避税公司与低避税公司，运用梯度提升决策树（GBDT）和随机森林算法预测公司是否为高避税公司。

为了保证结果的稳健可信，我们通过4种方式定义高避税公司：（1）以每年 *TME* 处在前 10% 为高避税公司；（2）以每年 *TME* 处在前 20% 为高避税公司；（3）以每年 *NETR* 处在前 10% 为高避税公司；（4）以每年 *NETR* 处在前 20% 为高避税公司。这里，由于预测公司是否为高避税公司是二分类问题，GBDT 算法中我们使用指数损失函数进行优化，其余参数设定与上文一致。由于存在样本不平衡问题，我们使用 AUC（area under the curve）值来判断分类模型的预测能力。

AUC（Area under the Curve）值是二分类问题中常用的评价指标，AUC 值由受试者工作特征曲线（Receiver Operating Characteristic Curve，简称 ROC 曲线）得出，ROC 曲线是以真正例率（True Positive Rate，即所有被模型判定为正例的样本中，真正为正例个体的占比）为 Y 轴，以假正例率（False Positive Rate，即所有实际为负例的样本中，被模型判断为正例的个体占比）为 X 轴的二维曲线。ROC 曲线与 X 轴围成的面积即为 AUC 值，AUC 值介于 0 与 1 之间，AUC 值越大，说明预测能力越强。特别地，AUC 值为 0.5 时，说明模型的预测能力与随机猜测一样；小于 0.5 时，说明模型的预测能力不如随机猜测；在 0.5 至 1 之间时，说明模型优于随机猜测；为 1 时，说明模型能够完美地进行分类。关于 ROC 曲线和 AUC 值的详细理论与解释可以参考周志华^[62]。AUC 值的优点在于其能够全面地考察模型在处理分类问题时的表现，且不会受样本不平衡问题的干扰。

附表 A-20 展示了分类预测的结果。可以看到，无论使用何种高避税公司的定义，加入避税工具特征都能够提升 AUC 值，这也再次说明避税工具特征对公司避税有预测能力。但该结果也有两点值得注意之处：第一，与附录 8 得出的结论类似，采用根据 *NETR* 指标来划分高避税和低避税公司时，模型的预测能力普遍较低，并且行业层面指标表现出很强的预测能力。第二，按照 10% 进行划分时模型预测能力更好，这可能是因为偷逃税等高风险避税行为与合理的税务筹划有很大差异，而实际中进行高风险避税行为的公司比例很低，因此按照 20% 进行划分会将很多实际从事合理税收筹划的公司也被误认为是高避税公司，导致模型效果不佳。

附表 A-20 预测是否为高避税公司的 AUC 值

Appendix Table A-20 The AUC value of Predicting High- or Low- Tax-Avoidance Firms										
	基准模型	基准模型+避税工具特征	基准模型+公司治理特征	基准模型+其他内部特征	内部模型	基准模型+宏观层面特征	基准模型+行业层面特征	基准模型+外部微观特征	外部模型	全模型
算法	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
以每年 <i>TME</i> 处在前 10% 为高避税公司，其余为低避税公司										
GBDT	0.7388	<u>0.7507</u>	0.7360	0.7410	0.7480	0.7287	0.7387	0.7465	0.7375	0.7454
RF	0.5009	0.7478	0.7393	0.7400	0.7433	0.7106	<u>0.7513</u>	0.7380	0.7226	0.7276
以每年 <i>TME</i> 处在前 20% 为高避税公司，其余为低避税公司										
GBDT	0.6807	<u>0.7052</u>	0.6819	0.6831	0.7014	0.6791	0.6884	0.6829	0.6869	0.6993

RF	0.5209	<u>0.6940</u>	0.6702	0.6711	0.6796	0.6609	0.6909	0.6720	0.6719	0.6718
以每年NETR处在前10%为高避税公司, 其余为低避税公司										
GBDT	0.5811	0.6041	0.5765	0.5779	0.5957	0.5658	0.6269	0.5967	0.6224	<u>0.6307</u>
RF	0.5000	0.6134	0.5854	0.5846	0.5836	0.5691	<u>0.6559</u>	0.6043	0.6127	0.5902
以每年NETR处在前20%为高避税公司, 其余为低避税公司										
GBDT	0.5737	0.5922	0.5459	0.5706	0.5748	0.5690	0.6278	0.5763	0.6214	<u>0.6260</u>
RF	0.5002	0.5832	0.5380	0.5505	0.5514	0.5448	<u>0.6309</u>	0.5662	0.6021	0.5846

6. 采用“减法”方式构造预测模型

在主要结果中, 我们以只包括 7 个基本预测特征和 12 个年度虚拟变量的预测模型作为基准模型, 通过分别加入公司内外部 6 类特征, 即“加法”方式, 考察模型预测能力的提升幅度来判断每一类特征的预测能力。为确保预测结果稳健, 我们以全模型为基准, 分别减去公司内外部 6 类特征, 采用“减法”方式, 通过考察模型预测能力的下降幅度对每一类特征的预测能力进行判断。如果减去某一类特征后, 模型预测能力有大幅下降, 那么说明这一类特征对于公司避税程度有较强的预测能力。

预测结果如附表 A-21 所示。可以看到, 在更换预测模型构造方式之后, 减去避税工具特征和公司内部特征造成的预测能力下降最为明显, 且 GBRT 的预测能力也明显好于 OLS, 与主要结果一致。

附表 A-21 采用“减法”方式构造预测模型的样本外 R 方 (R^2_{oos})

Appendix Table A-21 R^2_{oos} of Using "Subtraction" Approach								
附表 A-21.1 用 TME 作为公司避税程度衡量指标 (单位: %)								
	OLS 下模型的 R^2_{oos}	GBRT 下模型的 R^2_{oos}	RF 下模型的 R^2_{oos}	OLS 下相比全模型 R^2_{oos} 的下降	GBRT 下相比全模型 R^2_{oos} 的下降	GBRT 相比 OLS R^2_{oos} 的提升	RF 下相比全模型 R^2_{oos} 的下降	RF 相比 OLS R^2_{oos} 的提升
预测模型	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
全模型	11.96	16.88	16.58	—	—	4.92	—	4.62
减去公司内部特征								
全模型—避税工具特征	<u>7.98</u>	13.26	12.97	3.98	3.63	5.27	3.61	4.99
全模型—公司治理特征	12.18	17.04	16.86	-0.22	-0.15	4.85	-0.28	4.68
全模型—其他内部特征	11.96	16.82	16.57	0.01	0.07	4.86	0.02	4.61
全模型—所有公司内部特征	8.48	13.21	12.91	3.48	3.67	4.73	3.67	4.43
减去公司外部特征								
全模型—宏观层面特征	12.09	16.99	16.66	-0.13	-0.10	4.89	-0.08	4.57
全模型—行业层面特征	11.94	16.70	16.45	0.02	0.18	4.76	0.13	4.51
全模型—外部微观特征	11.98	16.84	16.52	-0.02	0.04	4.87	0.07	4.54
全模型—所有公司外部特征	12.11	16.75	16.31	-0.15	0.14	4.63	0.27	4.20
减去全部公司内外部特征								
基准模型	8.19	<u>13.04</u>	<u>11.73</u>	3.77	3.84	4.85	4.85	3.54
附表 A-21.2 用 NETR 作为公司避税程度衡量指标 (单位: %)								
	OLS 下模型的 R^2_{oos}	GBRT 下模型的 R^2_{oos}	RF 下模型的 R^2_{oos}	OLS 算法下相比全模型 R^2_{oos} 的下降	GBRT 算法下相比全模型 R^2_{oos} 的下降	GBRT 算法相比 OLS 算法 R^2_{oos} 的提升	RF 算法下相比全模型 R^2_{oos} 的下降	RF 算法相比 OLS 算法 R^2_{oos} 的提升
预测模型	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
全模型	12.31	16.78	16.52	—	—	4.47	—	4.22
减去公司内部特征								
全模型—避税工具特征	<u>8.62</u>	13.75	13.17	3.69	3.03	5.13	3.35	4.55
全模型—公司治理特征	12.35	17.13	16.92	-0.04	-0.35	4.78	-0.40	4.57

全模型—其他内部特征	12.31	16.71	16.28	-0.01	0.07	4.39	-0.24	3.97
全模型—所有公司内部特征	8.87	14.08	13.64	3.44	2.70	5.21	2.88	4.78
减去公司外部特征								
全模型—宏观层面特征	12.21	16.73	16.53	0.10	0.05	4.52	0.01	4.32
全模型—行业层面特征	12.38	16.24	16.17	-0.08	0.54	3.86	0.35	3.79
全模型—外部微观特征	12.36	16.62	16.30	-0.05	0.15	4.27	0.22	3.95
全模型—所有公司外部特征	12.34	16.09	16.00	-0.03	0.69	3.75	0.52	3.66
减去全部公司内外部特征								
基准模型	8.81	13.70	12.49	3.50	3.08	4.89	4.03	3.68

7. 以三年期进行滚动预测

在主要结果中，我们运用三分之二的公司数据作为训练集建立模型预测另外三分之一的公司，在截面上外推预测。那么，利用历史数据作为训练集建立模型而用未来数据作为预测集、进行时间上外推预测时，结果是否仍然成立呢？为回答该问题，我们借鉴 Gu 等^[63]的做法，进行了三年期滚动预测，即以 T-3 至 T-1 期的数据作为训练集，以 T 期的数据作为测试集，分别预测 2011 至 2019 年的避税程度。

附表 A-22 展示了各个模型下样本外 R 方的年度均值，可以看到，该预测结果与主要结果的趋势保持一致，运用机器学习模型进行时间上外推预测效果也较好。

附表 A-22 三年期滚动预测的样本外 R 方 (R^2_{os}) 均值

Appendix Table A-22 The average R^2_{os} of three-year rolling prediction										
附表 A-22.1 用 TME 作为公司避税程度衡量指标（单位：%）										
	基准模型	基准模型+避税工具特征	基准模型+公司治理特征	基准模型+其他内部特征	内部模型	基准模型+宏观层面特征	基准模型+行业层面特征	基准模型+外部微观特征	外部模型	全模型
算法	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
OLS	9.39	<u>13.36</u>	8.96	9.38	12.77	7.81	9.47	8.28	5.37	9.95
GBRT	14.57	20.21	15.47	14.89	20.82	14.90	15.09	14.71	15.51	<u>21.08</u>
RF	16.08	21.90	19.49	16.73	<u>22.95</u>	17.40	17.04	16.29	17.91	22.74
附表 A-22.2 用 NETR 作为公司避税程度衡量指标（单位：%）										
	基准模型	基准模型+避税工具特征	基准模型+公司治理特征	基准模型+其他内部特征	内部模型	基准模型+宏观层面特征	基准模型+行业层面特征	基准模型+外部微观特征	外部模型	全模型
算法	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
OLS	8.13	<u>11.48</u>	7.84	8.06	10.96	6.74	7.52	6.65	4.19	8.41
GBRT	14.06	18.73	15.23	14.35	19.64	14.44	14.52	14.24	15.07	<u>19.97</u>
RF	15.52	20.53	19.03	16.07	<u>21.74</u>	17.04	16.07	15.54	17.11	21.60

8. 考虑预测变量存在缺失的情况

主要结果中，为方便对比机器学习与线性模型的预测效果，我们使用了不包含缺失值的数据集，但现实中预测变量可能存在大量数据缺失情况，那么机器学习模型能否在数据缺失时也取得较好效果呢？

GBRT 和 XGBoost 算法均能够自动处理预测变量缺失的情况。生成回归树时，若遇到某个节点上的预测变量有缺失值，GBRT 算法会生成一个新节点，将有缺失值的样本单独分类；XGBoost 则将有缺失值的样本直接按照指定方向分类^[61]。附表 A-23 展示了考虑预测变量存在缺失值的预测结果，可以看到，该预测结果与主要结果的趋势保持一致，而且即使考虑缺失值，GBRT 和 XGBoost

的预测能力也好于正文表 2 中 OLS，说明 GBRT 和 XGBoost 算法能够较好地处理预测变量缺失的问题。

附表 A-23 包含预测变量缺失样本进行预测的样本外 R 方 (R^2_{oss})

Appendix Table A-23 The R^2_{oss} of Prediction that Including Missing Predictors

附表 A-23.1 用 <i>TME</i> 作为公司避税程度衡量指标 (单位: %)										
	基准模型	基准模型+避税工具特征	基准模型+公司治理特征	基准模型+其他内部特征	内部模型	基准模型+宏观层面特征	基准模型+行业层面特征	基准模型+外部微观特征	外部模型	全模型
算法	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
GBRT	11.18	12.65	12.09	11.33	<u>13.48</u>	10.43	11.16	11.20	10.67	13.29
XGBoost	9.98	12.32	11.54	11.47	13.67	10.51	10.36	11.63	10.29	13.86

附表 A-23.2 用 <i>NETR</i> 作为公司避税程度衡量指标 (单位: %)										
	基准模型	基准模型+避税工具特征	基准模型+公司治理特征	基准模型+其他内部特征	内部模型	基准模型+宏观层面特征	基准模型+行业层面特征	基准模型+外部微观特征	外部模型	全模型
算法	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
GBRT	9.98	11.20	10.79	10.04	11.81	9.25	9.88	9.97	9.49	<u>11.79</u>
XGBoost	9.12	11.09	10.85	8.85	11.99	8.47	9.29	9.68	8.79	12.29

9. 考虑机器学习算法的随机性

GBRT 和 RF 算法具有一定的随机性，为排除主要结果是随机性所致，我们将种子数取值设为 2 到 101，对主要模型分别进行 100 次预测，这 100 次预测的样本外 R 方的均值和标准差以及 MSE、MAE 的均值如附表 A-24 所示。在 MSE 指标评价下，由于保留四位小数点无法区分最佳模型，我们改为了保留五位小数点。可以看到，对于每个模型来说，无论采用何种评价指标，100 次预测均值与表 1、附表 A-4 都非常接近，而且 100 次预测 R^2_{oss} 标准差均在 0.1% 以下，这说明主要结果并未受到随机性的影响，GBRT 和 RF 预测能力稳定。

附表 A-24 100 次模型预测能力结果

Appendix Table A-24 The Predictive Power of Models of 100 Iterations

附表 A-24.1 用 <i>TME</i> 作为公司避税程度衡量指标								
评价指标	R^2_{oss} (单位: %)				MSE		MAE	
机器学习算法	GBRT		RF		GBRT	RF	GBRT	RF
	均值	标准差	均值	标准差	均值		均值	
基准模型	13.21	0.0793	11.70	0.0765	0.00970	0.00985	0.0666	0.0673
加入公司内部特征								
基准模型+避税工具模型	16.99	0.0429	<u>16.72</u>	0.0758	0.00926	<u>0.00930</u>	0.0650	0.0654
基准模型+公司治理特征	12.64	0.0368	11.84	0.0942	0.00975	0.00982	0.0666	0.0673
基准模型+其他内部特征	13.36	0.0482	11.59	0.0911	0.00967	0.00988	0.0665	0.0675
内部模型	16.78	0.0422	16.22	0.0747	0.00929	0.00934	0.0650	0.0655
加入公司外部特征								
基准模型+宏观层面特征	12.60	0.0494	11.11	0.0769	0.00976	0.00993	0.0667	0.0678
基准模型+行业层面特征	13.65	0.0559	12.57	0.0753	0.00965	0.00976	0.0661	0.0667
基准模型+外部微观特征	13.21	0.0777	12.22	0.0730	0.00969	0.00980	0.0664	0.0670

外部模型	13.20	0.0445	12.92	0.0845	0.00968	0.00972	0.0663	0.0667
加入全部公司内外部特征								
全模型	16.89	0.0510	16.52	0.0778	0.00928	0.00931	0.0649	<u>0.0652</u>
附表 A-24.2 用 <i>NETR</i> 作为公司避税程度衡量指标								
评价指标	R^2_{oos} (单位: %)				MSE		MAE	
机器学习算法	GBRT		RF		GBRT	RF	GBRT	RF
	均值	标准差	均值	标准差	均值		均值	
基准模型	13.69	0.0383	12.46	0.0829	0.00977	0.00991	0.0675	0.0681
加入公司内部特征								
基准模型+避税工具模型	16.51	0.0428	16.27	0.0724	0.00945	0.00948	0.0667	0.0669
基准模型+公司治理特征	12.87	0.0466	12.28	0.0893	0.00986	0.00993	0.0677	0.0682
基准模型+其他内部特征	13.80	0.0417	12.26	0.0734	0.00975	0.00992	0.0675	0.0683
内部模型	16.12	0.0544	15.83	0.0891	0.00950	0.00951	0.0669	0.0671
加入公司外部特征								
基准模型+宏观层面特征	13.26	0.0421	11.84	0.0799	0.00982	0.00998	0.0677	0.0687
基准模型+行业层面特征	14.32	0.0377	13.43	0.0839	0.00970	0.00981	0.0670	0.0676
基准模型+外部微观特征	13.59	0.0431	12.77	0.0792	0.00978	0.00987	0.0675	0.0680
外部模型	14.07	0.0518	13.63	0.0976	0.00973	0.00977	0.0672	0.0676
加入全部公司内外部特征								
全模型	16.72	0.0518	<u>16.46</u>	0.0826	0.00942	0.00945	0.0664	<u>0.0665</u>

参考文献

- [1] Slemrod J. The economics of corporate tax selfishness [J]. *National Tax Journal*, 2004, 57: 877-899.
- [2] Kesselman J R, Williamson S H, Berndt E R. Tax credits for employment rather than investment [J]. *American Economic Review*, 1977, 67(3): 339-349.
- [3] 王跃堂, 王国俊, 彭 洋. 控制权性质影响税收敏感性吗?——基于企业劳动力需求的检验 [J]. *经济研究*, 2012, 47(4): 52-63.
- Wang Yuetang, Wang Guojun, Peng Yang. Does ownership nature affect tax sensitiveness?——Evidence based on firm's labor demand [J]. *Economic Research Journal*, 2012, 47(4): 52-63. (in Chinese)
- [4] Lo A W, Wong R M, Firth M. Tax, financial reporting, and tunneling incentives for income shifting: An empirical analysis of the transfer pricing behavior of Chinese-listed companies [J]. *Journal of the American Taxation Association*, 2010, 32(2): 1-26
- [5] 魏志华, 赵悦如, 吴育辉. “双刃剑”的哪一面: 关联交易如何影响公司价值 [J]. *世界经济*, 2017, 40(1): 142-167.
- Wei Zhihua, Zhao Yueru, Wu Yuhui. Which side of the double-edged sword: How related party transaction affect firm value [J]. *The Journal of World Economy*, 2017, 40(1): 142-167. (in Chinese)
- [6] 黄炳艺, 林嘉伟, 王艳艳. 资本弱化税制与外资企业税收规避行为研究 [J]. *管理科学学报*, 2020, 23(4): 38-54.
- Huang Bingyi, Ling Jianwei, Wang Yanyan. Thin capitalization rules and tax avoidance of foreign-funded enterprises [J]. *Journal of Management Sciences in China*, 2020, 23(4): 38-54. (in Chinese)
- [7] Tang T Y. A review of tax avoidance in China [J]. *China Journal of Accounting Research*, 2020, 13(4): 327-338.
- [8] 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.
- [9] Desai M A, Dharmapala D. Corporate tax avoidance and high-powered incentives [J]. *Journal of Financial Economics*, 2006, 79(1): 145-179.
- [10] 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.
- [11] 吴联生. 国有股权、税收优惠与公司税负 [J]. *经济研究*, 2009, 44(10): 109-120.
- Wu Liansheng. State ownership, preferential tax, and corporate tax burdens [J]. *Economic Research Journal*, 2009, 44(10): 109-120. (in Chinese)
- [12] 王跃堂, 王亮亮, 彭 洋. 产权性质、债务税盾与资本结构 [J]. *经济研究*, 2010, 45(9): 122-136.
- Wang Yuetang, Wang Liangliang, Peng Yang. Ownership nature of ultimate controller, debt-related tax shields and capital structure [J]. *Economic Research Journal*, 2010, 45(9): 122-136. (in Chinese)
- [13] Chen S, Chen X, Cheng Q, et al. Are family firms more tax aggressive than non-family firms? [J]. *Journal of Financial Economics*, 2010, 95(1): 41-61.
- [14] 李春涛, 刘贝贝, 周 鹏, 等. 它山之石: QFII 与上市公司信息披露 [J]. *金融研究*, 2018, (12): 138-156.
- Li Chuntao, Liu Beibei, Zhou Peng, et al. QFII and corporate information disclosure [J]. *Journal of Financial Research*, 2018, (12): 138-156. (in Chinese)
- [15] 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.
- [16] Cheng C S A, Huang H H, Li Y, Stanfield J. The effect of hedge fund activism on corporate tax avoidance [J]. *The Accounting Review*, 2012, 87(5): 1493-1526.
- [17] Khan M, Srinivasan S, Tan L. Institutional ownership and corporate tax avoidance: New evidence [J]. *The Accounting Review*, 2017, 92(2): 101-122.

- [18] 祝继高, 叶康涛, 严冬. 女性董事的风险规避与企业投资行为研究——基于金融危机的视角 [J]. 财贸经济, 2012, (4): 50-58.
Zhu Jigao, Ye Kangtao, Yan Dong. Risk-avoidance of women directors and firm investment——From the perspective of financial crisis [J]. Finance & Trade Economics, 2012, (4): 50-58. (in Chinese)
- [19] Williams R. J. Women on corporate boards of directors and their influence on corporate philanthropy [J]. Journal of Business Ethics, 2003, 42(1): 1-10.
- [20] 王雄. 审计委员会能抑制企业避税吗? [J]. 财经问题研究, 2018, (3): 109-115.
Wang Xiong. Can audit committee reduce corporate tax avoidance? [J]. Research on Financial and Economic Issues, 2018, (3): 109-115. (in Chinese)
- [21] 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.
- [22] Huang H H, Lobo G J, Wang C, Xie H. Customer concentration and corporate tax avoidance [J]. Journal of Banking and Finance, 2016, 72: 184-200.
- [23] 曹越, 孙丽, 醋卫华. 客户集中度、内部控制质量与公司税收规避 [J]. 审计研究, 2018, (1): 120-128.
Cao Yue, Sun Li, Cuo Weihua. Customer concentration, internal control quality and tax avoidance [J]. Auditing Research, 2018, (1): 120-128. (in Chinese)
- [24] 陈德球, 陈运森, 董志勇. 政策不确定性、税收征管强度与企业税收规避 [J]. 管理世界, 2016, (5): 151-163.
Chen Deqiu, Chen Yunsen, Dong Zhiyong. Economic policy uncertainty, tax enforcement, and corporate tax avoidance [J]. Management World, 2016, (5): 151-163. (in Chinese)
- [25] Hoopes J L, Mescall D, Pittman J A. Do IRS audits deter corporate tax avoidance? [J]. The Accounting Review, 2012, 87(5): 1603-1639.
- [26] 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 Ziyang, Tian Binbin. Tax Competition, tax Enforcement and tax Avoidance [J]. Economic Research Journal, 2013, 48(9): 99-111. (in Chinese)
- [27] 马光荣, 李力行. 政府规模、地方治理与企业逃税 [J]. 世界经济, 2012, 35(6): 93-114.
Ma Guangrong, Li Lixing. Government size, local governance and corporate tax avoidance [J]. The Journal of World Economy, 2012, 35(6): 93-114. (in Chinese)
- [28] Slemrod J. Tax compliance and enforcement [J]. Journal of Economic Literature, 2019, 57(4): 904-954.
- [29] Wilson R. An examination of corporate tax shelter participants [J]. The Accounting Review, 2009, 84 (3): 969-999.
- [30] 刘行, 赵晓阳. 最低工资标准的上涨是否会加剧企业避税? [J]. 经济研究, 2019, 54(10): 121-135.
Liu Hang, Zhao Xiaoyang. Does increasing the minimum wage induce tax avoidance? [J]. Economic Research Journal, 2019, 54(10): 121-135. (in Chinese)
- [31] 于文超, 殷华, 梁平汉. 税收征管、财政压力与企业融资约束 [J]. 中国工业经济, 2018, (1): 100-118.
Yu Wenchao, Yin Hua, Liang Pinghan. Tax enforcement activities, fiscal pressure and firms' financial constraints [J]. China Industrial Economics, 2018, (1): 100-118. (in Chinese)
- [32] 刘思义, 翁若宇, 杨道广. 自然灾害、财政压力与企业避税——基于台风灾害的实证分析 [J]. 会计研究, 2018, (3): 34-41.
Liu Siyi, Weng Ruoyu, Yang Guangdao. Natural disaster, fiscal pressure, and tax avoidance: An empirical study based on typhoon disaster [J]. Accounting Research, 2018, (3): 34-41. (in Chinese)
- [33] 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(08): 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(08): 32-50. (in Chinese)
- [34] Dang D, Fang H, He M. Economic policy uncertainty, tax quotas and corporate tax burden: Evidence from China [J]. *China Economic Review*, 2019, 56: 101303.
- [35] 赵 萌, 叶 莉. 经济政策不确定性、融资约束与企业税收规避行为 [J]. *中央财经大学学报*, 2020, (2): 67-78.
- Zhao Meng, Ye Li. Economic policy uncertainty, financial constraints and firm tax avoidance [J]. *Journal of Central University of Finance & Economics*, 2020, (2): 67-78. (in Chinese)
- [36] Edwards A, Schwab C, Shevlin T. Financial constraints and cash tax savings [J]. *The Accounting Review*, 2016, 91(3): 859-881.
- [37] 王亮亮. 金融危机冲击、融资约束与公司避税 [J]. *南开管理评论*, 2016, 19(01): 155-168.
- Wang Liangliang. External financial shocks, financial constraints, and corporate tax avoidance [J]. *Nankai Business Review*, 2016, 19(01): 155-168. (in Chinese)
- [38] 刘 行, 叶康涛. 金融发展、产权与企业税负 [J]. *管理世界*, 2014, (3): 41-52.
- Liu Hang, Ye Kangtao. Financial development, ownership and corporate tax burdens [J]. *Management World*, 2014, (3): 41-52. (in Chinese)
- [39] Cai H, Liu Q. Competition and corporate tax avoidance: Evidence from Chinese industrial firms [J]. *Economics Journal*, 2009, 119(537): 764-795.
- [40] 曹 越, 赵 婷. 行业成长性是否降低了公司税负——来自中国上市公司的证据 [J]. *中南财经政法大学学报*, 2018, (2): 14-24+158.
- Cao Yue, Zhao Ting. Whether industry growth reduces corporate tax burden: Evidence from Chinese listed companies [J]. *Journal of Zhongnan University of Economics and Law*, 2018, (2): 14-24+158. (in Chinese)
- [41] McGuire S T, Omer T C, Wang D. Tax avoidance: Does tax-specific industry expertise make a difference? [J]. *The Accounting Review*, 2012, 87(3): 975-1003.
- [42] Chen S, Schuchard K, Stomberg B. Media coverage of corporate taxes [J]. *The Accounting Review*, 2019, 94(5): 83-116.
- [43] 刘笑霞, 李明辉. 媒体负面报道、分析师跟踪与税收激进度 [J]. *会计研究*, 2018, (9): 64-71.
- Liu Xiaoxia, Li Minghui. Negative media coverage, analyst following and tax aggressiveness [J]. *Accounting Research*, 2018, (9): 64-71. (in Chinese)
- [44] Allen A, Francis B B, Wu Q, et al. Analyst coverage and corporate tax aggressiveness [J]. *Journal of Banking and Finance*, 2016, 73: 84-98.
- [45] Chen T, Lin C. Does information asymmetry affect corporate tax aggressiveness? [J]. *Journal of Financial and Quantitative Analysis*, 2017, 52(5): 2053-2081.
- [46] Elsayed S, Thyssens D, Rashed A, et al. Do we really need deep learning models for time series forecasting? [J]. *arXiv preprint arXiv:2101.02118*, 2021.
- [47] Bojer C S, Meldgaard J P. Kaggle forecasting competitions: An overlooked learning opportunity [J]. *International Journal of Forecasting*, 2021, 37(2): 587-603.
- [48] Januschowski T, Wang Y, Torkkola K, et al. Forecasting with trees [J]. *International Journal of Forecasting*, 2022, 38(4): 1473-1481.
- [49] Hastie T, Tibshirani R, Friedman J. *The Elements of Statistical Learning* [M]. New York: Springer, 2009.
- [50] Freund Y, Schapire R E. A decision-theoretic generalization of on-line learning and an application to boosting [J]. *Journal of Computer and System Sciences*, 1997, 55(1): 119-139.
- [51] Breiman L. Random forests [J]. *Machine learning*, 2001, 45: 5-32.
- [52] Friedman J H. Greedy function approximation: a gradient boosting machine [J]. *Annals of statistics*, 2001: 1189-

1232.

- [53] Dechow P M, Sloan R G, Sweeney A P. Detecting earnings management [J]. The Accounting Review, 1995, 70(2): 193-225.
- [54] 江轩宇. 税收征管、税收激进与股价崩盘风险[J]. 南开管理评论, 2013, 16(5): 152-160.
Jian Xuanyu. Tax enforcement, tax aggressiveness and stock price crash risk [J]. Nankai Business Review, 2013, 16(5): 152-160. (in Chinese)
- [55] 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.
- [56] Bao Y, Ke B, Li B, et al. Detecting accounting fraud in publicly traded US firms using a machine learning approach [J]. Journal of Accounting Research, 2020: 58(1): 199-235.
- [57] Höglund H, Tax payment default prediction using genetic algorithm-based variable selection[J]. Expert Systems with Applications, 2017, 88: 368-375.
- [58] Abedin M Z, Chi G, Uddin M M, et al. Tax default prediction using feature transformation-based machine learning [J]. IEEE Access, 2020, 9: 19864-19881.
- [59] Hanlon M, Heitzman S. A review of tax research [J]. Journal of accounting and Economics, 2010, 50(2-3): 127-178.
- [60] Dyreng S D, Hanlon M, Maydew E L. Long-run corporate tax avoidance [J]. The accounting review, 2008, 83(1): 61-82.
- [61] Chen T, Guestrin C. Xgboost: A scalable tree boosting system [R/OL]. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016. Available at: arXiv:1603.02754.
- [62] 周志华: 《机器学习》, 《清华大学出版社》, 2016 年(2022 年重印)。
Zhou Zhihua, Machine Learning, Tsinghua University Press, 2016 (reprinted in 2022).
- [63] Gu S, Kelly B, Xiu D. Empirical asset pricing via machine learning [J]. Review of Financial Studies, 2020, 33(5): 2223-2273.