《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. 附录一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描述性统计情况。

2. 附录二展示了各类模型的特征重要性数值和排序。

3. 附录三展示了关键特征与公司避税程度的作用模式验证内容。

4. 附录四展示了基于随机森林算法的双变量联合部分依赖图。

5. 附录五展示了使用公司预测特征水平值对未来期避税行为预测的结果。

6. 附录六展示了本文稳健性检验的结果。

目录

附录 1：样本分布和预测变量描述性统计

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　描述性统计

Appendix Table A-2　Descriptive Statistics of Sample

Variable	Obs.	Mean	S.D.	Variable	Obs.	Mean	S.D.
TME	13,113	0.00	0.11	*Exec_Pay*	13,113	14.91	0.81
NETR	13,113	1.00	0.11	*Exec_Own*	13,113	0.06	0.13
BTD	13,113	0.00	0.02	*CEO_Pay*	13,113	13.32	0.81
Size	13,113	22.32	1.29	*CEO_Own*	13,113	0.05	0.11
Fage	13,113	2.75	0.37	*CEO_Tenure*	13,113	3.54	1.02
Rate	13,113	0.19	0.05	*CEO_Gender*	13,113	0.94	0.23
Roa	13,113	0.10	0.05	*CEO_Age*	13,113	3.88	0.13
OCF	13,113	0.06	0.08	*CEO_Aboard*	13,113	0.07	0.26
PPE	13,113	0.22	0.16	*CEO_Career*	13,113	0.17	0.37
Lev	13,113	0.43	0.20	*CEO_Resea*	13,113	0.18	0.38
Donation	13,113	0.01	0.02	*Sup_Size*	13,113	1.25	0.25
Donation_Dummy	13,113	0.80	0.40	*Sup_Pay*	13,113	13.00	1.64
Emp_Size	13,113	0.61	1.09	*Sup_Own*	13,113	0.00	0.01
Emp_Pay	13,113	11.41	0.56	*AC*	13,113	0.09	0.06

Tacc	13,113	0.01	0.08	*CCHHI*	13,113	0.03	0.07
Absda	13,113	0.06	0.06	*CCHHI_Dummy*	13,113	0.69	0.46
Trade	13,113	0.04	0.11	*TE*	13,113	1.01	0.19
Trade_Dummy	13,113	0.69	0.46	*Goldentax*	13,113	0.29	0.45
RD	13,113	0.02	0.02	*Minwage*	13,113	1.47	0.42
RD_Dummy	13,113	0.75	0.44	*Gov_Fissurp*	13,113	-0.48	0.55
Foreign	13,113	0.53	0.50	*Gov_Fispres*	13,113	1.14	0.08
Registran	13,113	0.05	0.21	*Gov_GDP*	13,113	0.11	0.05
State_Own	13,113	0.39	0.49	*Gov_Size*	13,113	4.09	0.53
Fam_Own	13,113	0.37	0.48	*EPU*	13,113	4.93	0.42
BH_Share	13,113	0.06	0.23	*Findep*	13,113	9.04	0.59
Top1toTop25	13,113	3.98	5.67	*MKT*	13,113	8.02	1.84
Top1	13,113	0.36	0.15	*GM*	13,113	7.08	1.44
Top10	13,113	0.59	0.15	*NS*	13,113	8.68	1.94
Top3_SQ	13,113	0.17	0.12	*CM*	13,113	8.22	1.34
Inst_Own	13,113	0.48	0.24	*FM*	13,113	7.35	3.12
Board_Size	13,113	2.16	0.20	*MIOLaw*	13,113	10.75	6.54
Duality	13,113	0.24	0.43	*Ind_HHI*	13,113	0.10	0.09
Board_Ind	13,113	0.37	0.05	*Ind_Size*	13,113	4.48	0.81
Board_Pay	13,113	14.21	1.74	*Ind_Growth*	13,113	3.02	4.63
Board_Own	13,113	0.12	0.18	*Big4*	13,113	0.06	0.24
Board_Gender	13,113	0.13	0.12	*Analyst*	13,113	2.08	0.89
Audit_Ind	13,113	0.42	0.57	*Media*	13,113	4.02	0.72

附表 A-2 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。

附录 2：各类模型的特征重要性数值和排序

附表 A-3　加入各子类特征的模型中重要性排名前十的特征

Table A-3　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	*Roa*	*Roa*	*Roa*	*Roa*	*Roa*	*Roa*
2	***Donation***	*PPE*	*PPE*	*PPE*	*PPE*	*PPE*
3	*Rate*	*Lev*	*Lev*	*Lev*	*Lev*	*Lev*
4	***Tacc***	*Rate*	*Size*	*Rate*	*Rate*	*Rate*
5	*Lev*	***Board_Pay***	*Rate*	*Size*	*Size*	*Size*
6	*OCF*	***Top1toTop25***	*OCF*	*OCF*	***Ind_Size***	***Media***
7	*PPE*	*Size*	***CCHHI***	***MIOLaw***	***Ind_HHI***	*OCF*
8	***Emp_Size***	***AC***	*Fage*	*Fage*	*OCF*	*Fage*
9	*Size*	***Inst_Own***	***CCHHI_Dummy***	***NS***	***Ind_Growth***	***Analyst***
10	***RD***	***CEO_Pay***	*Year2018*	***FM***	*Fage*	***Big4***

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-3 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-4 展示了各类模型在预测公司避税行为时，排名前十的特征重要性均值信息，附表 A-5 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-4　加入各类特征模型的特征重要性均值（前十名）

Appendix Table A-4　Average Feature Importance when Adding Each Category of Features (Top 10)

附表 A-4.1　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	*Roa*	14.79	*Roa*	16.11	*Roa*	22.16
2	***Donation***	7.08	***Donation***	7.78	*Lev*	8.05
3	*Rate*	4.47	*Lev*	4.58	*PPE*	7.66
4	*Lev*	4.00	*OCF*	4.51	*Rate*	6.39
5	*OCF*	3.71	*Rate*	4.42	*Size*	3.89
6	***Tacc***	3.53	***Tacc***	4.40	***Ind_Size***	3.63
7	***Emp_Size***	3.07	*PPE*	3.67	***Ind_HHI***	3.22
8	*PPE*	2.90	***Emp_Size***	3.49	***MIOLaw***	3.01
9	***RD***	2.63	***RD***	3.03	*OCF*	2.96
10	***Board_Pay***	1.93	***Board_Pay***	2.69	***Media***	2.79
附表 A-4.2　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	*Roa*	20.16	*Roa*	21.28	*Roa*	32.43
2	***Donation***	9.82	*PPE*	7.42	*PPE*	14.98
3	*Rate*	7.20	*Lev*	7.19	*Lev*	13.47
4	***Tacc***	7.11	*Rate*	5.73	*Size*	10.11
5	*Lev*	7.09	***Board_Pay***	3.96	*Rate*	9.98
6	*OCF*	7.08	***Top1toTop25***	3.54	*OCF*	6.89
7	*PPE*	6.34	*Size*	3.25	***CCHHI***	5.76
8	***Emp_Size***	6.25	***AC***	2.99	*Fage*	5.69
9	*Size*	5.65	***Inst_Own***	2.95	***CCHHI_Dummy***	0.88
10	***RD***	5.31	***CEO_Pay***	2.83	*Year2018*	0.66

附表 A-4.3　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	*Roa*	24.41	*Roa*	28.87	*Roa*	29.74
2	*PPE*	9.38	*PPE*	12.34	*PPE*	13.45
3	*Lev*	8.85	*Lev*	12.04	*Lev*	12.44
4	*Rate*	6.57	*Rate*	9.11	*Rate*	9.05
5	*Size*	5.38	*Size*	7.88	*Size*	8.95
6	*OCF*	3.85	***Ind_Size***	6.72	***Media***	6.01
7	***MIOLaw***	3.54	***Ind_HHI***	5.87	*OCF*	5.94
8	*Fage*	3.37	*OCF*	5.52	*Fage*	5.21
9	***NS***	3.23	***Ind_Growth***	4.60	***Analyst***	4.60
10	***FM***	3.18	*Fage*	4.36	***Big4***	1.92

附表 A-5　全模型所有特征重要性排序及特征重要性数值

Appendix Table A-5 Feature Importance of the Full Model

附表 A-5.1　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 *TME* 为避税指标	特征重要性数值	以 *NETR* 为避税指标	特征重要性数值
1	*Roa*	23.44	*Roa*	22.73	*Roa*	24.14
2	*Donation*	10.96	*Donation*	10.71	*Donation*	11.21
3	*OCF*	5.27	*Rate*	6.16	*OCF*	5.27
4	*Lev*	4.91	*OCF*	5.26	*Lev*	5.23
5	*Tacc*	4.77	*Tacc*	4.91	*Tacc*	4.62
6	*Rate*	4.22	*Lev*	4.59	*Emp_Size*	3.94
7	*Emp_Size*	3.55	*RD*	3.91	*PPE*	3.36
8	*PPE*	3.11	*Emp_Size*	3.16	*Rate*	2.28
9	*RD*	2.52	*PPE*	2.85	*Board_Pay*	2.14
10	*Board_Pay*	2.05	*Board_Pay*	1.95	*Top1toTop25*	1.8
11	*Top1toTop25*	1.8	*Top1toTop25*	1.8	*Analyst*	1.56
12	*Trade*	1.43	*Trade*	1.58	*Size*	1.45
13	*Size*	1.43	*Size*	1.4	*CEO_Pay*	1.41
14	*Analyst*	1.36	*Top10*	1.32	*Ind_Growth*	1.37
15	*Top10*	1.28	*Sup_Pay*	1.26	*CEO_Age*	1.36
16	*CEO_Age*	1.23	*AC*	1.19	*Ind_HHI*	1.32
17	*CEO_Pay*	1.22	*Analyst*	1.16	*Trade*	1.28
18	*Sup_Pay*	1.2	*Ind_Size*	1.1	*Top10*	1.23
19	*AC*	1.18	*CEO_Age*	1.1	*Ind_Size*	1.19
20	*Ind_Size*	1.15	*CEO_Pay*	1.02	*AC*	1.16
21	*Ind_HHI*	1.14	*Absda*	1.02	*Absda*	1.14
22	*Absda*	1.08	*Exec_Pay*	1	*Sup_Pay*	1.13
23	*Inst_Own*	1.04	*Inst_Own*	0.98	*RD*	1.12
24	*Top1*	1	*Ind_HHI*	0.95	*Inst_Own*	1.09

25	*Exec_Pay*	0.98	*Top1*	0.94	*Top1*	1.06
26	*Ind_Growth*	0.89	*Top3_SQ*	0.87	*Exec_Pay*	0.95
27	*CCHHI*	0.87	*CCHHI*	0.85	*CCHHI*	0.89
28	*Top3_SQ*	0.87	*MIOLaw*	0.84	*Top3_SQ*	0.87
29	*MIOLaw*	0.83	*CEO_Tenure*	0.76	*MIOLaw*	0.82
30	*NS*	0.75	*Media*	0.75	*NS*	0.78
31	*GM*	0.75	*GM*	0.72	*GM*	0.77
32	*CEO_Tenure*	0.74	*NS*	0.72	*Emp_Pay*	0.76
33	*Media*	0.71	*Fage*	0.66	*CEO_Tenure*	0.72
34	*Emp_Pay*	0.69	*FM*	0.64	*MKT*	0.69
35	*Fage*	0.67	*Emp_Pay*	0.61	*Fage*	0.68
36	*FM*	0.63	*Gov_GDP*	0.58	*Media*	0.66
37	*Findep*	0.6	*Findep*	0.56	*Findep*	0.64
38	*MKT*	0.58	*TE*	0.55	*FM*	0.62
39	*Gov_GDP*	0.58	*Board_Gender*	0.54	*Exec_Own*	0.61
40	*TE*	0.58	*Gov_Fissurp*	0.53	*TE*	0.6
41	*Exec_Own*	0.57	*Exec_Own*	0.53	*Gov_Fissurp*	0.59
42	*Gov_Fissurp*	0.56	*Gov_Size*	0.51	*Gov_GDP*	0.57
43	*Gov_Size*	0.53	*Minwage*	0.47	*Gov_Size*	0.54
44	*Board_Gender*	0.51	*MKT*	0.47	*Minwage*	0.51
45	*Minwage*	0.49	*CM*	0.43	*Board_Gender*	0.48
46	*Board_Own*	0.46	*Board_Own*	0.43	*Board_Own*	0.48
47	*CM*	0.43	*Sup_Size*	0.42	*CM*	0.42
48	*Gov_Fispres*	0.41	*Ind_Growth*	0.41	*Gov_Fispres*	0.41
49	*Sup_Size*	0.4	*Gov_Fispres*	0.4	*Sup_Own*	0.39
50	*Sup_Own*	0.32	*Board_Size*	0.27	*Sup_Size*	0.38
51	*Board_Size*	0.24	*Sup_Own*	0.24	*Board_Ind*	0.24
52	*Board_Ind*	0.23	*Board_Ind*	0.22	*Board_Size*	0.2
53	*CEO_Own*	0.18	*Foreign*	0.19	*CEO_Own*	0.19
54	*EPU*	0.16	*CEO_Own*	0.17	*EPU*	0.14
55	*Audit_Ind*	0.13	*EPU*	0.17	*Audit_Ind*	0.12
56	*Foreign*	0.13	*Audit_Ind*	0.13	*State_Own*	0.11
57	*State_Own*	0.1	*State_Own*	0.08	*Foreign*	0.06
58	*BH_Share*	0.05	*BH_Share*	0.05	*BH_Share*	0.05
59	*Fam_Own*	0.03	*Goldentax*	0.02	*Fam_Own*	0.03
60	*CEO_Career*	0.02	*CEO_Career*	0.02	*CEO_Career*	0.02
61	*Donation_Dummy*	0.02	*Fam_Own*	0.02	*Donation_Dummy*	0.02
62	*Duality*	0.02	*Donation_Dummy*	0.02	*Duality*	0.02
63	*CEO_Aboard*	0.01	*Year2013*	0.01	*Year2013*	0.01
64	*CEO_Gender*	0.01	*CEO_Gender*	0.01	*Year2016*	0.01
65	*CEO_Resea*	0.01	*CEO_Resea*	0.01	*Year2011*	0.01
66	*Goldentax*	0.01	*Duality*	0.01	*Year2015*	0.01
67	*Year2017*	0.01	*CEO_Aboard*	0.01	*CEO_Aboard*	0.01
68	*Year2011*	0.01	*Year2018*	0.01	*Registran*	0.01
69	*Year2013*	0.01	*Year2015*	0.01	*CEO_Gender*	0.01
70	*Year2015*	0.01	*Year2011*	0.01	*CEO_Resea*	0.01
71	*Big4*	0.01	*RD_Dummy*	0.01	*Year2017*	0.01
72	*RD_Dummy*	0.01	*Year2017*	0.01	*Big4*	0.01
73	*Registran*	0.01	*Registran*	0	*Year2008*	0
74	*Year2018*	0.01	*Year2019*	0	*Goldentax*	0
75	*Year2016*	0.01	*Year2012*	0	*Trade_Dummy*	0
76	*CCHHI_Dummy*	0	*Year2009*	0	*Year2018*	0
77	*Trade_Dummy*	0	*Year2016*	0	*Year2014*	0
78	*Year2008*	0	*Year2008*	0	*Year2019*	0
79	*Year2019*	0	*Year2010*	0	*Year2012*	0
80	*Year2009*	0	*Big4*	0	*Year2009*	0
81	*Year2010*	0	*Year2014*	0	*Year2010*	0
82	*Year2012*	0	*Trade_Dummy*	0	*RD_Dummy*	0

83	*Year2014*	0	*CCHHI_Dummy*	0	*CCHHI_Dummy*	0
附表 A5.2　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 *TME* 为避税指标	特征重要性数值	以 *NETR* 为避税指标	特征重要性数值
1	*Roa*	6.15	*Rate*	6.38	*Roa*	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	*CEO_Aboard*	0.49	*Gov_Fispres*	0.51	*Duality*	0.48
61	*Gov_Fispres*	0.47	*CEO_Aboard*	0.51	*CEO_Aboard*	0.47
62	*CEO_Tenure*	0.46	*Trade_Dummy*	0.46	*CEO_Tenure*	0.46
63	*Duality*	0.45	*CEO_Tenure*	0.45	*Audit_Ind*	0.44
64	*CEO_Resea*	0.44	*Duality*	0.43	*Gov_Fispres*	0.43
65	*Trade_Dummy*	0.44	*Goldentax*	0.39	*Trade_Dummy*	0.42
66	*Goldentax*	0.38	*CEO_Resea*	0.39	*Goldentax*	0.37
67	*CEO_Career*	0.29	*CCHHI_Dummy*	0.35	*Year2019*	0.31
68	*CCHHI_Dummy*	0.28	*CEO_Career*	0.34	*CEO_Career*	0.23
69	*Year2019*	0.27	*Year2019*	0.23	*Year2015*	0.23
70	*CEO_Gender*	0.19	*CEO_Gender*	0.19	*CCHHI_Dummy*	0.22
71	*Year2015*	0.18	*Donation_Dummy*	0.18	*CEO_Gender*	0.18
72	*Year2009*	0.16	*Year2009*	0.15	*Year2008*	0.17
73	*Donation_Dummy*	0.16	*Registran*	0.14	*Year2009*	0.16
74	*Year2018*	0.13	*Year2015*	0.12	*Year2018*	0.14
75	*Registran*	0.13	*Year2018*	0.12	*Donation_Dummy*	0.13
76	*Year2010*	0.11	*Year2010*	0.12	*Registran*	0.12
77	*Year2008*	0.09	*Year2017*	0.04	*Year2010*	0.1
78	*Year2016*	0.05	*Year2008*	0.02	*Year2016*	0.09
79	*Year2011*	0.03	*Year2016*	0.01	*Year2011*	0.08
80	*Year2012*	0.01	*Year2013*	0.01	*Year2012*	0.02
81	*Year2013*	0	*Year2012*	-0.01	*Year2013*	-0.01
82	*Year2017*	-0.02	*Year2011*	-0.03	*Year2017*	-0.09
83	*Year2014*	-0.07	*Year2014*	-0.05	*Year2014*	-0.1

附录 3：关键特征与公司避税程度的作用模式验证

本文按照 Shrestha 等[1]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-6 至附表 A-10 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-6　捐赠支出对公司避税程度的回归结果

Appendix Table A-6　Regression of *Donation* on Corporate Tax Avoidance

避税衡量指标	*TME*	*NETR*
	（1）	（2）
Donation	-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[2]，因此回归中去除了只含有一年观测值的公司，下同。

附表 A-7　应计盈余对公司避税程度的回归结果

Appendix Table A-7　Regression of *Tacc* on Corporate Tax Avoidance

避税衡量指标	*TME*			*NETR*		
	（1）	（2）	（3）	（4）	（5）	（6）
Tacc 取值范围	(-∞,-0.05)	[-0.05,0.05]	(0.05,+∞)	(-∞,-0.05)	[-0.05,0.05]	(0.05,+∞)
Tacc	-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-8　员工人数对公司避税程度的回归结果

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

避税衡量指标	*TME*		*NETR*	
	（1）	（2）	（3）	（4）
Emp_Size 取值范围	(0,0.04]	(0.04,+∞)	(0,0.04]	(0.04,+∞)
Emp_Size	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-9　研发支出对公司避税程度的回归结果

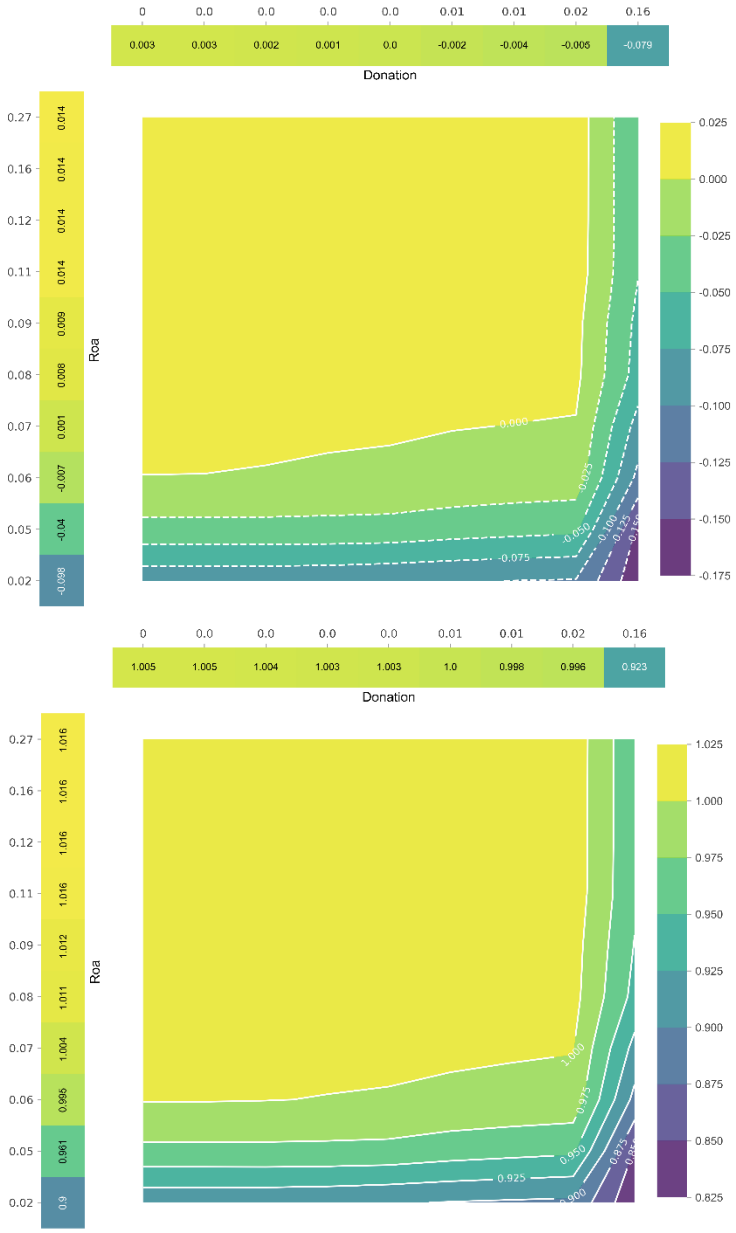
Appendix Table A-9　Regression of *RD* on Corporate Tax Avoidance

避税衡量指标	*TME*			*NETR*		
	（1）	（2）	（3）	（4）	（5）	（6）
RD 取值范围	[0,0.02)	[0.02,0.04]	(0.04,+∞)	[0,0.02)	[0.02,0.04]	(0.04,+∞)
RD	-0.067	1.793**	0.533	0.046	1.842**	0.455
	(-0.099)	(2.205)	(0.881)	(0.071)	(2.302)	(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 4：基于随机森林算法的双变量联合部分依赖图

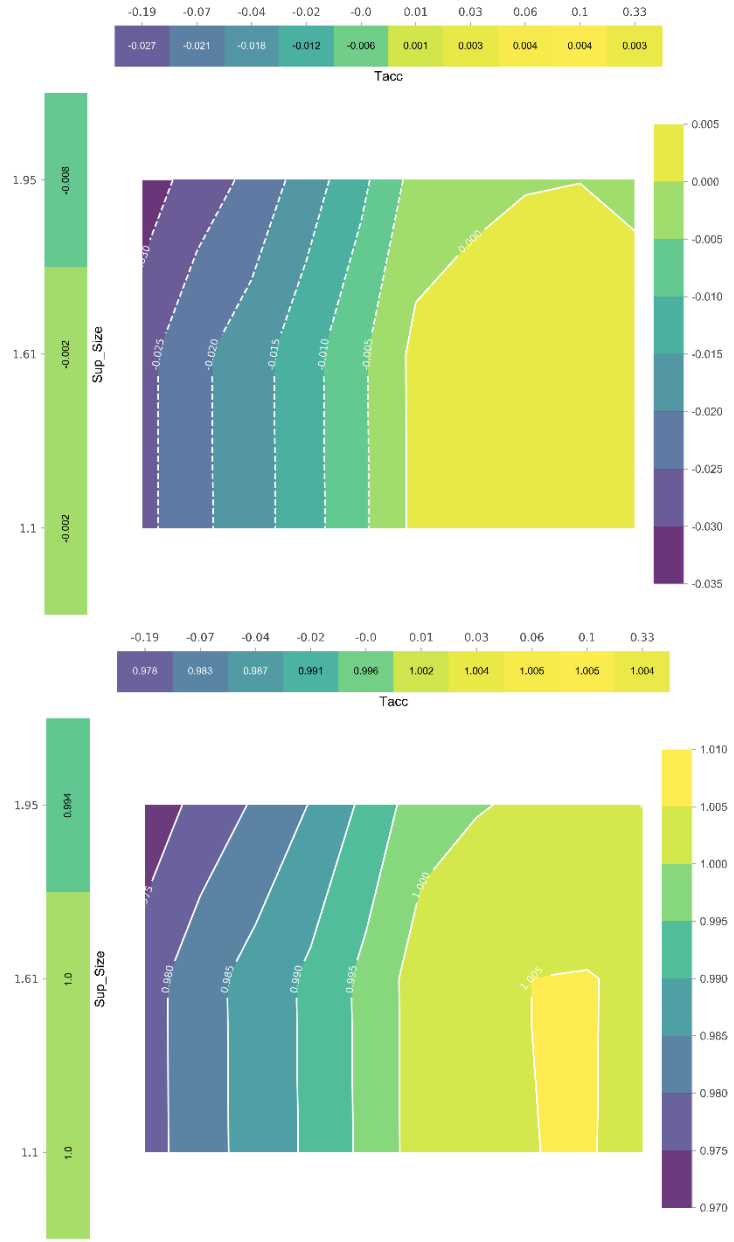
这里，由于 R 中画随机森林的部分依赖图耗时较长，我们采用 python 绘制随机森林算法下双变量的联合部分依赖图。



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*

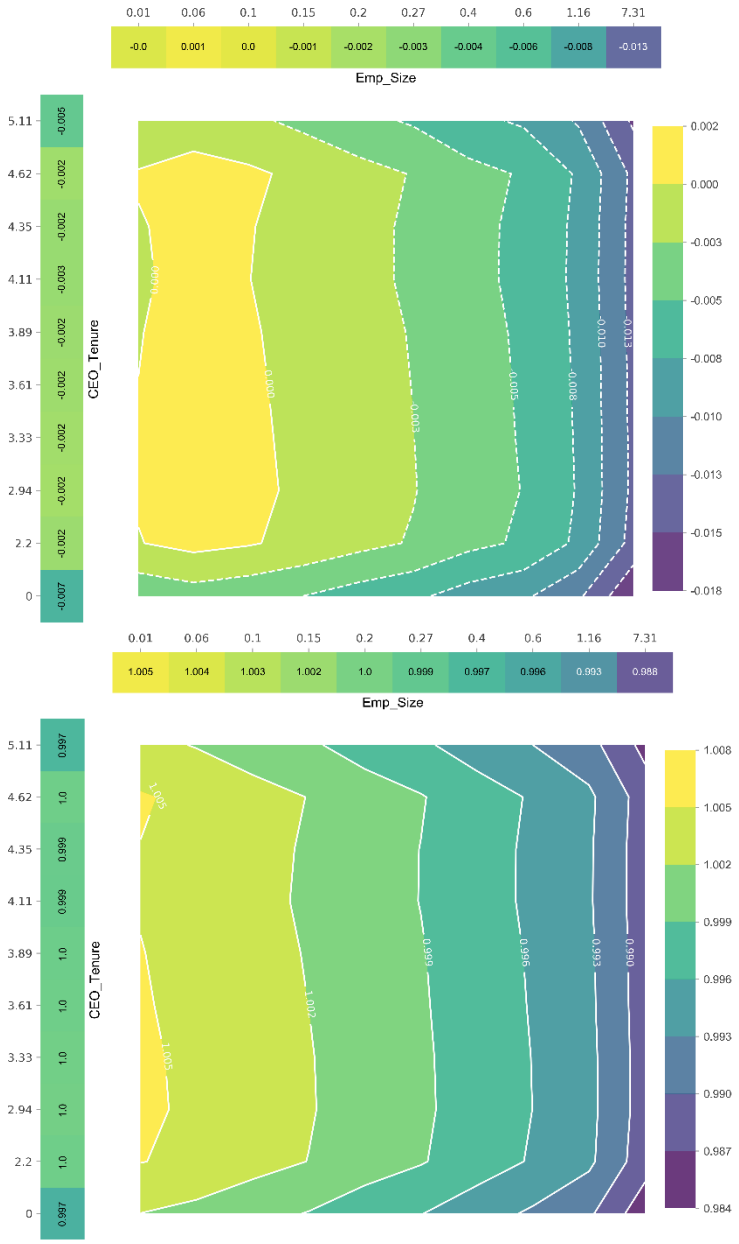
注：X 轴代表企业捐赠规模，Y 轴总资产收益率，右侧色卡代表公司避税程度，颜色越浅，避税程度越高，上图中公司避税程度采用 *TME*，下图采用 *NETR* 衡量。



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*

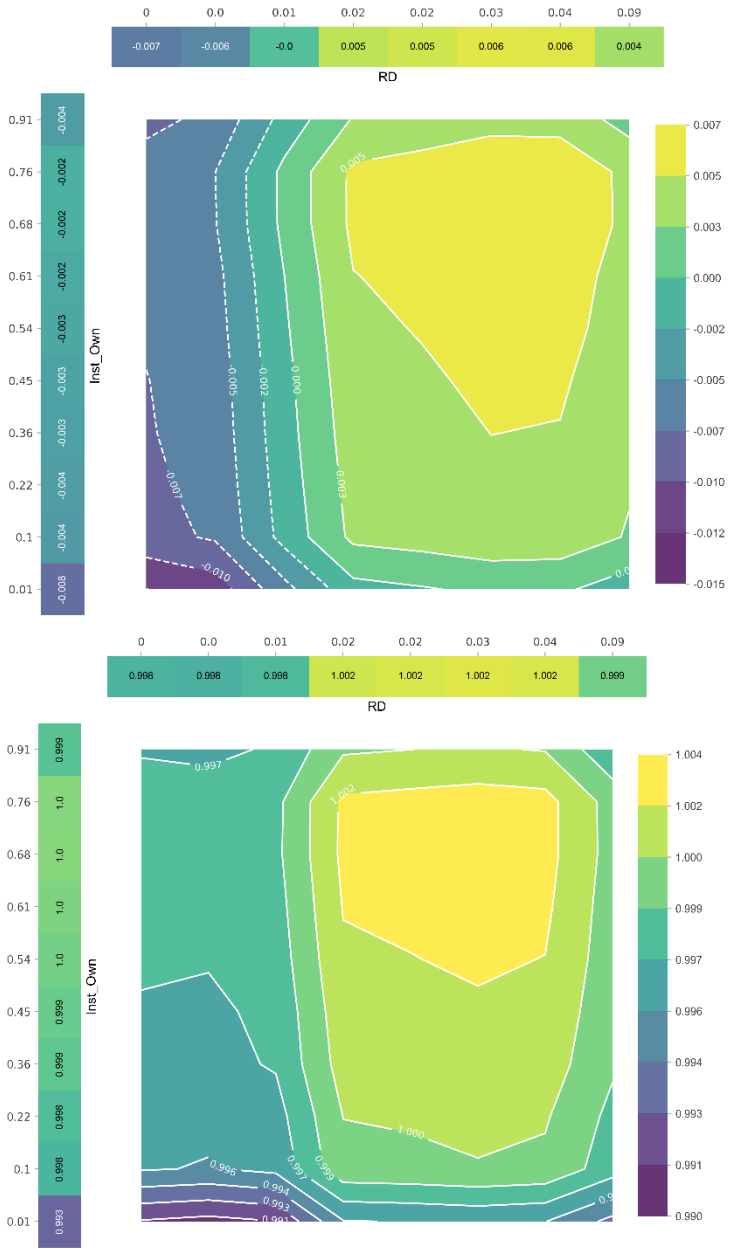
注：X 轴代表应计盈余，Y 轴监事会规模，右侧色卡代表公司避税程度，颜色越浅，避税程度越高，上图中公司避税程度采用 *TME*，下图采用 *NETR* 衡量。



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，左图中公司避税程度采用 *TME*，右图采用 *NETR* 衡量。

附录 5：使用公司预测特征水平值对未来期避税行为预测

在正文部分，我们运用第 T-1 期至第 T-3 期预测特征水平值和变化量，对第 T 期公司避税程度进行预测，旨在识别公司未来期避税行为的预测特征。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-10 所示。

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（正文表 6）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（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* 时和 *NETR*，各模型提升效果更为显著，分别达到 0.66~2.33 个百分点和 0.50~2.83 个百分点之间。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-10　未来期避税行为预测结果（单位：%）

Appendix Table A-10　R^2_{oos} of Predictive Power of in Future Corporate Tax Avoidance

附表 A-10.1　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	4.91	**5.57**	3.35	3.08	**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	2.58	**2.95**	0.12
内部模型	3.78	**5.35**	3.82	1.76	**3.41**	1.79	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**	3.93	1.60	**3.44**	1.98	-0.32	**1.72**	1.48
附表 A-10.2　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	6.47	**7.11**	5.66	4.28	**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	**5.22**	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**	6.25	4.03	**4.78**	3.69	1.55	**2.21**	2.33

注：这里我们采取两种突出标注方式，从两个维度进行比较分析：1.在相同的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.在相同的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的标注方式。

附录 6：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-11 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-11　更换 GBRT 算法参数的样本外 R 方（R^2_{oos}）

Appendix Table A-10　R^2_{oos} of Using Alternative GBRT Parameters

附表 A-11.1　*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	15.21	12.19	12.68	15.08	12.42	12.67	12.61	12.48	15.08
	4	13.33	16.41	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	16.73
0.001	2	13.27	16.52	12.62	13.46	16.09	12.82	13.53	13.19	13.12	16.08
	4	13.23	16.96	12.76	13.49	16.68	12.74	13.66	13.25	13.32	16.75
	6	12.90	16.98	12.53	13.19	16.82	12.39	13.51	13.09	13.10	16.96
0.002	2	13.13	17.00	12.18	13.32	16.38	12.44	13.39	13.05	12.80	16.33
	4	12.81	16.94	11.67	12.74	16.70	11.89	12.93	13.19	12.40	16.67
	6	12.70	16.81	11.30	12.43	16.59	11.55	12.82	13.03	11.77	16.65

附表 A-11.2　*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	14.53	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	15.61
	6	13.57	15.91	12.81	13.67	15.63	13.37	14.04	13.49	13.88	16.09
0.001	2	13.60	15.86	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	16.56
	6	13.56	16.60	12.78	13.69	16.18	13.23	14.23	13.59	14.06	16.85
0.002	2	13.72	16.60	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	17.03
	6	13.44	16.68	12.04	13.30	16.45	12.73	13.94	13.54	13.36	17.10

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[①]，结果如附表 A-12 所示。可以看到，更换参数之后，结果仍然一致。

附表 A-12　更换 RF 算法参数的样本外 R 方（R^2_{oos}）

Appendix Table A-12　R^2_{oos} of Using Alternative RF Parameters

附表 A-12.1　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

① 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。

TME	15.60	16.30	16.21	16.25	15.50	16.26	16.29	16.27	15.46	16.26	16.31	16.19
NETR	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-12.2　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
TME	11.38	12.56	12.86	12.86	11.34	12.51	12.94	12.85	11.32	12.49	12.96	12.79
NETR	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-12.3　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
TME	15.05	16.52	16.49	16.19	15.03	16.53	16.50	16.15	14.88	16.53	16.47	16.17
NETR	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-13。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-13　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-13　R^2_{oos} of Changing Training Set and Test Set

附表 A-13.1　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	8.91	11.76	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	15.85
附表 A-13.2　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	7.41	9.83	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	14.96

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[3]。这里参数的选取与 GBRT 一致。

LASSO 和 XGBoost 的预测结果如附表 A-14 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好于线性算法。

附表 A-14　更换算法后的样本外 R 方（R^2_{oos}）

Appendix Table A-14　R^2_{oos} of Changing Machine Learning Algorithm

附表 A-14.1　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)
LASSO	7.99	12.28	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-14.2　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)
LASSO	8.42	12.04	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。

预测结果如附表 A-15 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-15　运用 *BTD* 作为避税衡量指标的样本外 R 方（R^2_{oos}）

Appendix Table -15　R^2_{oos} of Using *BTD* as a Measure of Tax Avoidance

	基准模型	基准模型+避税工具	基准模型+公司治理	基准模型+其他内部	内部模型	基准模型+宏观层面	基准模型+行业层面	基准模型+外部微观	外部模型	全模型

		特征	特征	特征		特征	特征	特征		
算法	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
OLS	7.48	9.76	6.42	7.54	8.96	7.14	7.80	7.67	7.68	8.89
GBRT	**8.21**	**12.76**	**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	12.29	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[4]。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-16 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 5.1 节得出的结论类似，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-16　预测是否为高避税公司的 AUC 值

Appendix Table A-16　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 TME 处在前 10%为高避税公司，其余为低避税公司										
GBDT	**0.7388**	**0.7507**	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	**0.7513**	0.7380	0.7226	0.7276
以每年 TME 处在前 20%为高避税公司，其余为低避税公司										
GBDT	**0.6807**	**0.7052**	**0.6819**	**0.6831**	**0.7014**	**0.6791**	0.6884	**0.6829**	**0.6869**	**0.6993**
RF	0.5209	0.6940	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**	**0.6307**
RF	0.5000	**0.6134**	**0.5854**	**0.5846**	0.5836	**0.5691**	**0.6559**	**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**	**0.6260**
RF	0.5002	0.5832	0.5380	0.5505	0.5514	0.5448	**0.6309**	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-17 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-17　采用“减法”方式构造预测模型的样本外 R 方（R^2_{oos}）

Appendix Table A-17　R^2_{oos} of Using "Subtraction" Approach

附表 A-17.1　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	7.98	**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	**13.04**	11.73	3.77	3.84	4.85	4.85	3.54
附表 A-17.2　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	8.62	**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[5]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-18 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-18　三年期滚动预测的样本外 R 方（R^2_{oos}）均值

Appendix Table A-18　The average R^2_{oos} of three-year rolling prediction

附表 A-18.1　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	13.36	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	21.08
RF	**16.08**	**21.90**	**19.49**	**16.73**	**22.95**	**17.40**	**17.04**	**16.29**	**17.91**	**22.74**
附表 A-18.2　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	11.48	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	19.97
RF	**15.52**	**20.53**	**19.03**	**16.07**	**21.74**	**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[3]。附表 A-19 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 的预测能力也好于表 4 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-19　包含预测变量缺失样本进行预测的样本外 R 方（R^2_{oos}）

Appendix Table A-19　The R^2_{oos} of Prediction that Including Missing Predictors

附表 A-19.1　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）
GBRT	**11.18**	**12.65**	**12.09**	11.33	13.48	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-19.2　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）
GBRT	**9.98**	**11.20**	10.79	**10.04**	11.81	**9.25**	**9.88**	**9.97**	**9.49**	11.79
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-20 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 次预测均值与正文表 2 都非常接近，而且 100 次预测 R^2_{oos} 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-20　100 次模型预测能力结果

Appendix Table A-20　The Predictive Power of Models of 100 Iterations

附表 A-20.1　用 *TME* 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	**13.21**	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	16.72	0.0758	**0.00926**	0.00930	**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**	0.0652

附表 A-20.2　用 *NETR* 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	16.46	0.0826	**0.00942**	0.00945	**0.0664**	0.0665

参考文献

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

[2] Correia S. Singletons, cluster-robust standard errors and fixed effects: A bad mix[R]. Technical Note, Duke University, 2015.

[3] 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.

[4] 周志华. 机器学习[M]. 北京：清华大学出版社，2016.
Zhou Zhihua. Machine Learning[M]. Beijing: Tsinghua University Press, 2016.

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