基于随机森林算法的上市公司财务报告审计预测

王　琼

江苏信息职业技术学院，中国·江苏　无锡　214153

【摘　要】针对传统财务报告审计方法存在的不足，本文提出了利用随机森林算法对上市公司财务报告进行审计预测的新思路。通过构建随机森林模型，并结合实际上市公司的财务报告数据，实现了对财务报告审计意见的自动分类和预测。实验结果表明，基于随机森林算法的财务报告审计预测模型在审计预测任务中展现出良好的性能。本文的研究不仅为财务报告审计提供了一种新的有效工具，也为资本市场的健康稳定发展提供了有力支持。同时，本文还探讨了该方法的改进方向，为后续研究提供了有益的参考。

【关键词】随机森林算法；财务报告；审计预测；机器学习

1　引言

财务报告审计作为确保财务报告质量的重要环节，其预测和评估的准确性直接影响到资本市场的公平性和稳定性。传统的财务报告审计依赖于审计师的专业知识与经验，然而，面对日益庞大的数据量和复杂多变的会计信息，人工审计面临着效率低下和主观判断差异等局限性。因此，探索一种更为高效、准确的财务报告审计预测方法，成为当前审计领域亟待解决的问题。而随着机器学习和数据挖掘技术的发展，特别是集成学习算法，为提高审计效率和准确性提供了新的解决方案，有助于揭示财务数据中潜在的风险，为监管者、投资者以及公司管理层提供更为科学的决策支持。

2　上市公司财务报告审计情况整体分析

财务报告审计是一个系统的检查过程，由专业的审计人员（通常是注册会计师）进行。这个过程遵循国际或国内认可的审计标准，以及其他相关的法律和职业规范。审计的主要目标是对财务报告的真实性、准确性、完整性和合规性给出专业意见。如果审计师出具了标准的审计意见，则意味着对该公司的财务报表在财务状况、经营成果、现金流量、重大事项等方面给予了肯定。标准审计意见对于企业的健康运营和持续发展至关重要，它不仅为企业赢得了市场的信任，还为其未来的融资和扩张提供了坚实的基础。如果标准审计意见是由四大会计师事务所（即普华永道、德勤、安永和毕马威）中的任何一家出具的，这通常被视为财务报表质量和企业信誉的高度认可，能为企业在全球市场上的发展提供有力支持。

本文从国泰安数据库（CSMAR）里选取了我国上市公司2010~2022年财务报告相关数据来分析近12年财务报告审计情况，在删除标准审计意见缺失的样本后，最终得到数据43030条。通过分析发现，12年来被出具标准审计意见公司占比在92.86%~96.57%之间浮动，比较稳定，历年均值为95.30%；未被出具标准审计意见公司占比均值为4.70%。由四大会计师事务所出具标准审计意见公司占比在6.12%~7.58%之间，平均值为6.68%。可见上市公司出具标准审计意见占比高且波动浮动不大，整体质量良好；由四大出具标准审计意见占比较少，但自2019年来有上涨趋势。总体而言，审计质量较高，能为后续的模型构建做强有力的数据支撑。

3　随机森林算法在审计预测中的应用

3.1随机森林算法基本原理

随机森林是一种集成学习算法，由多个决策树组合而成，其核心思想是通过集成多个弱预测模型来形成一个强预测模型。具体来说，随机森林在构建每个决策树时会随机选择一部分数据作为训练集，并且从全部的特征中随机选取一部分特征用于节点分裂，[1]这种两次随机性的引入使得模型具有较强的泛化能力和较低的过拟合风险。在预测阶段，随机森林通过对所有决策树的预测结果进行投票或取均值（分类问题为投票，回归问题则取均值）来确定最终的输出。[2]

3.2随机森林算法在审计中的适用性

随机森林算法具有高维数据处理能力、抗过拟合能力、准确性与鲁棒性、特征重要性评估等特征。这些特征使其在处理复杂的财务数据时表现出色。如随机森林算法能够有效处理财务报告审计涉及的高维数据；它的强大随机性能够防止过拟合，确保审计预测的可靠性；通过组合多个弱分类器可以提高审计预测模型的精确度；随机森林能够

评估各个财务指标对审计结果的影响，从而指导重点审计领域。审计意见模型预测结果，不仅能为审计机构识别财务报告舞弊提供参考，也能帮助投资者更好地规避投资风险，从而维护资本市场稳定发展。

4 基于随机森林算法的审计预测模型构建

黄志刚(2020)等人在研究中提出能否将机器学习算法应用到财务报告舞弊识别中以及何者最优的问题，并将支持向量机、随机森林、神经网络应用到财务舞弊识别中，经过实证发现随机森林表现最佳。[3]本文沿用他们的做法，从众多算法中选择随机森林算法应用于财务报告审计预测中。

4.1数据收集与预处理

仍结合上述2010~2022年上市公司数据进行分析，为消除离群值的影响，保证结果的稳健性，所有的连续变量都进行缩尾处理后余38939条数据。运行代码df=df.dropna(axis=0)去掉空值后剩余有效数据24857条。

4.2特征提取与选择

根据财务报告的特点和审计需求，提取3个反映偿债能力，4个反映盈利能力，2个反映营运能力和2个反映发展能力的指标作为相关的特征指标（共计11个指标，具体指标名称见表2）为自变量，选取字段“标准审计意见”为因变量（1表示出具，0表示未出具，1的总数为24181，占比97.28%），以构建有效的随机森林模型。（见表1）

从表1中可以看出，有效统计数为24857条，数据最分散的指标是“净利润增长率”和“流动比率”，表明部分企业存在较大优势，且存在两级分化现象。数据较为集中的是“资产收益率”和“现金流比率”，表明各个企业在这个指标上水平相近。

4.3模型训练

在python中利用上述提取的特征指标和标签数据，将所有样本中 70%(17399条)的数据作为训练集，剩余 30%(7458条)的数据作为测试集。训练随机森林模型，并通过调整模型参数和优化算法，提高模型的预测性能。

4.4审计预测

将训练好的随机森林模型应用于实际财务报告的审计预测中，通过对财务报告的关键指标进行分析和评估，对是否出具标准审计意见进行预测。预测有效性主要通过准确度(accuracy)、精确度(precision)、召回率(recall)和 F1值（F1-score）来判断。具体公式如下：

accuracy = (+TN)/(+ +FN+TN) (公式 4.1)

precision = /(+) (公式 4.2)

recall = /(+) (公式 4.3)

F1-score = 2 precision recall/(precision +

表1：主要变量的描述性统计

能力	变量	N	均值	标准差	最小值	最大值	分位数		
							25%	50%	75%
偿债能力	资产负债率	24857	0.44	0.20	0.03	0.91	0.28	0.44	0.59
	流动比率	24857	2.34	2.62	0.27	35.50	1.12	1.59	2.50
	现金流比率	24857	0.04	0.07	-0.22	0.27	0.01	0.04	0.08
盈利能力	净资产收益率	24857	0.06	0.13	-0.93	0.44	0.03	0.07	0.12
	销售毛利率	24857	0.29	0.17	-0.06	0.87	0.16	0.26	0.38
	销售净利率	24857	0.06	0.17	-1.54	0.55	0.02	0.07	0.13
	资产收益率	24857	0.04	0.06	-0.37	0.25	0.01	0.04	0.07
营运能力	应收账款占比	24857	0.12	0.10	0.00	0.51	0.04	0.10	0.18
	总资产周转率	24857	0.66	0.45	0.06	3.09	0.37	0.55	0.81
发展能力	总资产增长率	24857	0.19	0.39	-0.38	5.12	0.02	0.10	0.23
	净利润增长率	24857	-0.37	3.70	-36.53	15.50	-0.43	0.05	0.37

recall）（公式 4.4）

TP是指模型预测为正类别实际上也是正类别的情况。FP是指模型预测为正类别但实际上是负类别的情况。FN是指模型预测为负类别但实际上是正类别的情况。TN指模型预测为负类别实际上也是负类别的情况。[4]

4.5结果反馈

本文选用了七家公司财务报告进行随机森林算法预测，审计预测结果为array([1, 1, 0, 1, 1, 1, 1]，准确度为0.973183159023867。这表明7家公司数据中有1家不应该出具标准的审计意见，预测准确度约为97.32%。更详尽的结果如下表所示：

表2：随机森林预测结果

	precision	recall	F1-score	support
0	0.83	0.09	0.17	216
1	0.91	1.00	0.99	7242
accuracy			0.97	7458
macro avg	0.90	0.55	0.58	7458
weighted avg	0.97	0.97	0.96	7458

数据显示，Accuracy为0.97，意味着预测正确的正例与预测正确的反例的和占总样本的比例为97%；Precision加权平均值为0.97，说明预测可信度高；Recall加权平均值为0.97，说明模型在识别正例方面表现良好，能够有效地捕捉到大部分的真实正例；F1-score加权平均值为0.96，意味着模型在精确度和召回率之间有很好的平衡。Support为每个特征或每个类别在训练数据集中出现的次数，正好是训练集的数据数7458。总的来说，这个随机森林模型预测有效，它在区分正类和负类时表现好，并且不会过于偏向任何一方。

5 结论与展望

本文基于随机森林算法，对2010~2022年上市公司财务数据构建了审计预测模型。研究结果表明，随机森林算法不仅在理论上具备优越的性能，而且在实践中证明其在财务报告审计领域具有的应用价值，能够提高审计工作的效率和准确性，降低人为因素的干扰，提高审计效率和准确性，有助于保障资本市场的健康稳定发展。随着技术的不断进步和数据量的增加，随机森林有望在未来的审计工作中发挥更加关键的作用。

然而，本文仍存在以下的局限性：（1）在特征指标的选取上，只在 CSMAR 数据库中抽取有限的关键财务指标；（2）没有分析非财务指标对审计的影响，比如上市公司的客户议价能力、公司是否存在董事长和总经理两职合一现象、公司内部控制制度、审计主体的专业性和独立性等[5]；（3）对模型也没有做进一步的优化，也没有跟其它算法进行比较。今后还可以在以下几个方面进行改进：（1）收集更多财务指标数据，用科学的方法对财务指标重要性进行排名，选用更关键的财务指标进行模型构建；（2）引入更多非财务数据非结构化数据，以更全面地反映企业的经营状况和管理机制；（3）一方面通过优化算法参数和模型结构，进一步提高预测精度和稳定性；另一方面，考虑将该方法与其他机器学习算法进行融合，以发挥各自的优势，提高整体预测性能。

参考文献：

[1]张成刚.基于深度神经网络的短期电力负荷概率预测[D].武汉大学，2021.[中国优秀硕士学位论文全文数据库].

[2]陈若冰,陈萱等.基于机器学习的海洋浮标寿命及轨迹预测[J].海洋通报.2021,(03):25-36.

[3]张志敏.基于随机森林的上市公司财务报告舞弊智能识别[D].山西财经大学，2021.[中国优秀硕士学位论文全文数据库].

[4]徐嘉琦.基于社交文本和深度学习的抑郁症分析研究[D].哈尔滨理工大学，2022.[中国优秀硕士学位论文全文数据库].

[5]任婷.上市公司治理结构对审计费用的影响研究[D].西南大学，2015.[中国优秀硕士学位论文全文数据库].

作者简介：

王琼（1983-），女，湖南醴陵人，江苏信息职业技术学院商学院会计教研室，讲师，硕士，研究方向：会计、数据分析。