厦門大學

本 科 毕 业 论 文 （设 计）

（主修）

基于logistic模型和随机森林模型的财务造假因素的研究

Research on Factors of Financial Falsification Based on Logistic Model and Random Forest Model

姓　　名：　林振炜

学　　号：　15220162202239

学　　院：　经济学院

专　　业：　经济统计

年　　级：　2016级

校内指导教师：　冯峥晖　副教授

二〇一九年十二月二十四日

厦门大学本科学位论文诚信承诺书

本人呈交的学位论文是在导师指导下独立完成的研究成果。本人在论文写作中参考其他个人或集体已经发表的研究成果，均在文中以适当方式明确标明，并符合相关法律规范及《厦门大学本科毕业论文（设计）规范》。

该学位论文为（经济学院 属性数据分析）课题（组）的研究成果，获得（经济学院 属性数据分析）课题（组）经费或实验室的资助，在（实验室）实验室完成。（请在以上括号内填写课题或课题组负责人或实验室名称，未有此项声明内容的，可以不作特别声明。）

另外，本人承诺辅修专业毕业论文（设计）（如有）的内容与主修专业不存在相同与相近情况。

学生声明（签名）：

年　　月　　日

致谢

致谢语应以简短的文字对课题研究与论文撰写过程中曾直接给予帮助的人员(例如指导教师、答疑教师及其他人员)表示自己的谢意。

作为毕业论文提交时，应注意事项：致谢内容用小四号宋体。根据2016年2月施行的《厦门大学本科毕业论文（设计）规范》，致谢被放在论文起首。致谢结构一般分为三个部分：1，回顾；2，感谢； 3，承担责任以及献辞。第一部分可以简述论文写作的经历，所面对的挑战以及你如何应对。第二部分具体感谢在论文过程中给与你帮助的人。第三部分指出你将为自己的论文承担责任，如果你希望将此论文献给谁，可以在最后指出。致谢内容请亲自撰写，使其具备你个人的特色。抄袭任何模板内容是极其懒惰、没有意义、不负责任和错误的行为。

摘要

由于我国证券市场发展时间相对较短，相关监管政策不健全等原因，上市公司财务造假仍呈现高发态势。为实现对财务欺诈行为进行准确预测，本文基于2018年上市公司的财务数据，构建了logistic模型和随机森林模型作为预测模型。logistic模型研究发现，中国的上市公司往往会选择通过对营业成本进行改动以对财务报表进行粉饰；在偿债能力越强的公司，其财务主管领取薪酬会更加减少财务造假的可能。随机森林模型的研究发现，利润总额增长率和可持续增长率是判别一个公司是否产生财务造假重要指标。

本文的研究为中国上市公司的财务造假识别提供了新的角度与方法，建立的模型有利于资源配置优化。为审计师审计财务造假提供了新的参考依据。

关键词：财务造假；logisitc模型；随机森林模型；营业成本；偿债能力；利润总额增长率；可持续增长率

Abstract

Due to the relatively short development time of China's securities market and the inadequacy of relevant regulatory policies, financial fraud in listed companies is still on the rise. In order to predict financial fraud accurately, this paper builds a logistic model and a RandomForest model based on the financial data of listed companies in 2018. The logistic model research found that Chinese listed companies often choose to decorate their financial statements by changing operating costs. In companies with stronger solvency, their financial officers will be more likely to receive financial remuneration when they receive compensation. The study of the Random orest model found that the total profit growth rate and sustainable growth rate are important indicators for judging whether a company has financial fraud.

The research in this paper provides a new perspective and method for the identification of financial fraud in Chinese listed companies. The model established is conducive to the optimization of resource allocation. It provides a new reference for auditors to audit financial fraud.

Keywords: financial fraud; logisitc model; RandomForest model; operating cost; debt service ability; total profit growth rate; sustainable growth rate

目 录

Table of Contents

第一章　引言

中国的股票市场自1989年成立以来历经40年时间。在此期间，上市公司数量和总体规模增长迅速，为我国经济发展注入了强劲动力。但是由于我国证券市场发展时间相对较短，相关监管政策不健全等原因，上市公司财务造假仍呈现高发态势，损害了投资者的切身利益，影响了中国证券市场的健康稳定发展。因此，如何对上市公司财务造假问题进行精准识别和有效预警就成为了监管层、机构投资者和个人投资者共同关注的问题。

财务造假一直是与财务报告相伴相生的一个问题，也是困扰监管层、金融业界和投资者多年的难题。某些企业出于粉饰公司业绩、提升投资者信心或逃避监管等目的，通过虚增交易、隐瞒相关信息、篡改财务数据等手段进行财务造假。常见手段包括特殊交易的不当核算（如对债权、债务重组以及关联交易的错误核算）、错用会计准则以及虚构交易、资产负债造假等等。这种行为不仅影响了投资者对公司情况的正确判断，还挑战了法律监管的权威。财务造假的“爆雷效应”还容易导致重大金融风险。因此，如何对上市公司财务造假问题进行精准识别和有效预警成为了各方面共同关注的问题。

回顾国外证券市场的发展史，我们可以发现，财务造假的先例可以追溯至英国南海公司的财务造假案，这起案件也被认为是非官方审计的开端。近几十年来，即使在监管体系较为完善的发达国家资本市场，上市公司财务造假案件也时有发生。其中，美国Enron公司[1]财务造假案影响较为巨大。

与发达国家相比，我国的证券业务开展时间较晚，部分上市公司抱有侥幸心理，各种违规现象时有发生。从违规类型来看，其中相当一部分属于财务造假。其中不仅有蓝田股份等典型案例，也有北大荒、雅百特等近年来新出现的案例。

对上市公司财务造假案例进行分析后，有研究者认为，只要公司的内部控制失效，公司管理层凌驾于内控体系之上，任何公司都有发生财务舞弊的可能，只是财务舞弊的手段和隐蔽性各有不同。

本文尝试构建一个公司财务造假风险识别模型，为财务造假的识别提供技术支持和一定辅

助。从资本市场及其参与者角度来看，本文的研究有以下几方面的现实意义。

首先，促进上市公司规范经营。加强对上市公司财务造假的识别，将提升上市公司的违法成本，可以促进上市公司和规范经营和真实披露，通过规范经营来获取经营业绩，而非通过财务造假来获得有吸引力的财务数据。

第二，提高资本市场效率。上市公司财务造假可能导致资本市场信息不真实，降低资本市场效率。因此加强对上市公司财务造假的识别有助于真实准确反映企业经营状况，提升资本市场效率，促进资本市场健康稳定发展。

第三，强化监管能力。以往的财务造假识别往往基于审计人员的主观判断，主要来自于审计人员对三张财务报表之间的逻辑关系进行推断分析，找出矛盾点和可疑点，从而对财务造假行为进行识别。但是这种识别方法也存在着一些不足，如耗时长、成本高、错误率和遗漏率高、过度依赖个人经验、缺乏统一标准等问题。因此，引入效率和准确率更高的算法和设计对于财务造假的精准识别具有重要意义。本文对上市公司财务造假识别的研究，将有助于监管部门加强对上市公司财务造假的预警和监测，提升违法成本，促使上市公司如实反映经营业绩。

第四，保护上市公司和投资者利益。在市场参与者层面，研究财务造假识别对维护上市公司和投资者利益都具有重要意义。财务数据是投资者进行投资决策的主要依据之一，篡改后的财务数据往往使得公司显得更有发展前景，对于证券投资者行为会产生误导，最终导致投资损失。对于上市公司而言，如果财务造假行为得不到有效遏制，就会在市场中引起“劣币驱逐良币”效应，损害上市公司利益。因此加强对上市公司财务造假识别的研究实质上有助于保护市场参与者的切身利益。

第二章　文献综述

2.1　财务造假原因综述

关于财务造假的成因，国外研究者提出了三角理论、GONE成因理论等。Cresscy(1973)提出著名了财务造假三角理论，企业进行财务造假的诱因主要由三个方面，即压力、机会和自我接受。压力即代表了目前公司遇到问题的紧迫性，如利润不达标即将暂停上市、公司收入降低影响股价、公司现金流紧张等，在这种情况下，公司进行财务造假就有了行为上的动机。

Bologna et al.(1993)提出GONE[①]成因理论，也被称作“四因子理论”．该理论认为，公司财务造假的原因主要可以分为贪婪、机会、需要和暴露四个因子组成。此理论把个体的主观因素归为贪婪和需要，阐述了造假者的动机；机会和暴露更多是指财务造假的客观条件，即内控机制不完善带来的造假机会和造假行为被发现和惩罚的可能性。这几类因素从不同的层面影响公司的决策行为，最终造成了财务造假行为的产生。企业董事会或管理层有不良动机，有粉饰公司经营状况现实需要，且公司客观条件可以进行财务造假，且事后不易被发现，那么企业就可能对公司的经营情况进行虚假的说明。

在GONE理论提出后，后续的研究不断对其进行完善。如财务造假因子理论把GONE理论中的成因分为主观和客观，主观条件是指个人的道德品质、造假动机等，客观条件是指公司内控严格程度、被发现的可能性以及惩罚力度等。两类因素共同作用，决定了一个公司是否会进行财务造假。

在国外理论研究较为完善的基础上，近年来国内学者对上市公司财务造假也进行了研究。如梅丹 等(2014)以2006年至2015年因财务造假受到监管机构处罚的上市公司为样本进行研究，发现公司财务造假的成因可以是为了成功进行IPO或取得增发资格、对投资者隐瞒经营状况、防止出现ST、操纵股价等。

[①]GONE由四个单词：greed,opportunity,need, exposure首字母组成

2.2 财务造假识别研究综述

财务造假的识别研究一直与财务造假本身的研究是相伴相生的，数十年来，国内外研究者已经对此问题进行了深入的研究和讨论，取得了一系列研究成果。相关研究主要分为三个阶段。

第一阶段是案例分析和特征总结，研究者主要对上市公司财务造假的典型案例进行分析比较，归纳总结这些公司及其财务数据多一些特征和矛盾点，研究变量和研究方法较为单一。Pavlović et al.(2019)将Benford定律[②]应用于财务造假的研究，利用塞尔维亚工商局2008年-2013年的数据，说明财务数据中，若第二个数字明显偏离Benford定律，那么有可能是公司对财务数据进行了改动。

第二阶段是建模分析，研究者开始由定性研究转为定量研究，通过建立数学模型对公司财务数据进行分析研判，从中寻找财务造假的痕迹。如Md Nasir et al.(2018)使用SCM(Securities Commissions of Malaysia)和Bursa Malaysia数据库对马来西亚上市公司的2001年到2008年的财务造假行为建立时间序列预测模型，进行实证分析。研究结果表明，在财务造假发生的前四年，盈余管理往往是非常有效的；生产成本在财务造假公司与非财务造假的公司的显著不同往往只发生在财务造假发生的前两年；Malaysia的上市公司往往喜欢通过操纵应计费用项目来进行财务造假等;Gao et al.(2017)通过实证研究表明财务造假的过程中往往伴随着外部董事的异常换手率，而且在此期间的换手率还同时与上市板块，会议频繁程度，以及财务高管的数量有关，研究还发现财务问题较多的公司其公司高管的离职率也会越高。在国内，Liao et al.(2019)实证研究表明，中国的企业社会责任可能是减少公司欺诈发生的有效途径。

第三阶段是大数据运用。随着信息技术的发展研究者开始使用数据挖掘技术对公司财务造假问题进行研究，取得了一定的研究成果。Tseng et al.(2005)使用logistic模型研究财务报表风险甄别，得出logistic模型优于判别分析模型；Chen(2017)建立SVM模型检测财务造假，结果表明SVM模型的预测效果较好；Abbasi et al.(2012)使用在包含数千家合法和欺诈性公司的数据上进行了一系列实验，得出其使用的meta-learning(元学习)框架在对于财务欺诈具有较好的效果。结果表明，框架的每个组成部分均对其整体有效性做出了重要贡献，额外的实验证明了元学

[②]由统计学家Benford于1938年提出，得出人们在处理的数据中，1-9作为首位数出现的概率是不同的

习框架相对于最新的欺诈检测方法的有效性。此外，该框架会生成与每个预测相关的置信度分数，这可以前所未有地促进财务欺诈检测性能，并可以作为有用的决策辅助工具。国内使用数据挖掘算法研究财务造假的研究发展仍较为缓慢，卢涛(2013)应用logistic模型、判别分析模型、神经网络模型等对123家上市公司财务造假样本与非财务造假样本进行分类预测,得出logisitc模型的预测效果较好，李明(2015)应用主成分分析方法和CART算法建立财务风险预测模型,邵朝等(2017)通过线性组合构造混合核函数，建立基于混合核学习的支持向量机财务欺诈预测模型，与单核的支持向量机模型相比，其模型的鲁棒性与识别精度都有所提高。

2.3 研究评述

国外财务报表造假的甄别的研究方式相对于国内更加成熟，但是建模的样本数据具有特定的西方国家市场经济环境，并没有具体结合中国的国情，存在一定的局限性，模型指标的构建缺乏通用性，因此模型的应用推广能力不足，在解决我国实际问题中预测效果并不理想。

机器学习算法的预测能力非常强，但是在解释性角度有一定的不足。

鉴于此，本文在国内外研究的基础上，对财务报表数据进行统计分析，从财务盈利能力、运营能力、资产管理效率等角度，并引入一些用于判断财务造假的账目数据。本文的主要目的在于探究中国上市财务造假最容易发生在哪些账目，以及如何尽量提高准确率以达到更好的财务监管的目的。

第三章　样本选取和变量选择

3.1　数据来源

本文从2018年中国上市公司财务年报和相关违规信息数据中提取样本，数据主要来源于国泰安（CSMAR）数据库，所使用的财务造假公司信息主要从国泰安数据库的公司研究违规信息数据中提取，构建模型所用的自变量指标数据主要来自于国泰安数据库的公司研究系列，并使用Wind数据对其进行了补充。

3.2　样本选取

国泰安金融数据库对上市公司违规性质的分类共有16类，具体分类有

表 3.1: 国泰安金融数据库企业违规分类表

违规编码	违规行为	违规编码	违规行为
P2501	虚构利润	P2509	擅自改变资金用途
P2502	虚列资产	P2510	占用公司资产
P2503	虚假记载(误导性陈述)	P2511	内幕交易
P2504	推迟披露	P2512	违规买卖股票
P2505	重大遗漏	P2513	操纵股价
P2506	披露不实(其它)	P2514	违规担保
P2507	欺诈上市	P2515	一般会计处理不当
P2508	出资违规	P2599	其他

在表3.1中本文直接把虚构利润和虚列资产的违规行为全部归为财务造假[15]。对于其他的几种违规行为，需要结合具体违规行为进行判断，最终得到了22个造假样本。

再结合各上市公司的年度审计报告中审计师给出的意见，包括标准无保留意见、保留意见、否定意见、无法发表意见、无保留意见加事项段、保留意见加事项段，本文对给出保留意见、否定意见、保留意见加事项段的事项进行审阅，最终认定有极强的造假嫌疑的样本共120个，无造假嫌疑的样本有3088个。

3.3 变量选择

3.3.1 初步变量选择

本文参照GONE成因理论，再结合公司财务造假的手段等因素。从公司的财务数据、治理结构、造假动力与造假表现四个方面进行构建变量。

在财务数据的分类构建中，本文参照了国泰安数据库——公司研究的分类方法，从公司偿债能力、盈利能力、经营能力、发展能力4个方面构建财务数据指标，其中偿债能力选取了流动比率、速动比率、资产负债率、经营活动产生的的现金流量净额/负债和，经营能力选取了存货与收入比、流动资产与收入比、总资产周转率，盈利能力选取了营业毛利率，发展能力选取了资本保值增值率、利润总额增长率、可持续增长率。

在治理结构的分类构建中，本文股权集中度以及监管层的持股比例进行描述。

在造假动力的指标构建中，由于高层管理人员的报酬与财务业绩或公司股票的市场表现挂钩[16]本文选取了财务类高管是否在上市公司领取薪酬这个二值变量。

在造假表现方面，由于上市公司财务造假多围绕提升业绩展开，造假手段多样，主要有虚增收入、减少成本费用等[17]，故本文选取了管理费用增长率、营业总成本增长率、销售费用增长率、坏账准备计提比例的变化率、2018年坏账准备比例这5个变量，最终得到初始变量选取情况表 3.2中

3.3.2 数据预处理与数据描述

由于本文选取的的变量皆为比率变量，故排除了上市公司的体量大小对结果的影响。对于bd_ratio，该数据来自于Wind,其不仅报告了2018年的坏账比例，而且还报告了该坏账持续的时间，从财务知识可以知道，坏账持续时间越长则其收回的可能性越低，即成为坏账的可能性越高。因此本文对来自不同年份的坏账进行了加权处理，对于年份越久远的坏账给予更高的权重。

在处理缺失值方面，对于除get_paid和bd_ratio即财务类高管是否领取薪酬和坏账计提比例其他数据选择用2017年的数据对其进行补足，如果2017年的数据没有，则用该行业的平均水平

来代替[①]；对于get_paid，如果没有找到当前财务主管是否领取薪酬，则选取其有报告的最高的管理层是否领取薪酬来代替；对于bd_ratio的缺失值，本文选择将其为0代替，即2018年的坏账比例相对2017年来说明没有发生变化。

观察如表3.3所示，发现流动比率、速动比率、存货与收入比、流动资产与收入比、资本保值增值率、利润总额增长率、管理费用增长率、营业总成本增长率、销售费用增长率的最大值比上四分位数都大出10倍左右，出于对数据真实性考虑以及希望排除某些数据录入错误造成的影响，对这些异常值进行处理，Nyitrai et al.(2019)的研究表明在财务比率的异常值处理方面，winsorize[②]处理具有非常好的效果。因此，本文采用了winsorize方法，将上述变量超出95%分位数的值用其95%分位数进行替代。

3.3.3 多重共线性分析

由于财务报表数据本身所具有的内部逻辑性，本文初选出的指标之间也必然存在着某种程度的自相关性，这对模型估计的参数准确性会产生影响。因此，在代入模型之前，我们首先要分析多重共线性，把部分自相关性较高的变量剔除。

首先对所有变量进行VIF检验[③]最终得到如表3.4所示,其中流动比率，速动比率、资本保值增值率、可持续增长率的VIF值大于10，可以认为其具有较强的多重共线性，最终去掉了流动比率与资本保值增值率这两个变量，得到的最终变量都通过了vif检验。而且根据两两之间的线性关系图如5.1所示(详见附录)，两两之间的关系也是比较弱的。至此，我们的初步变量选择完成。

[①]依据证监会2012版行业分类标准

[②]是一种处理离群值的方法，在公司金融、财务管理等微观领域应用非常广泛，将超出变量特定百分位范围的数值替换为其特定百分位数值的方法

[③]VIF(方差扩大(膨胀)因子法)是通过考察给定的解释变量被方程中其他所有解释变量所解释的程度，以此来判断是否存在多重共线性的一种方法。

表 3.2: 初始变量选取情况表

指标类型	指标代码	指标名称	指标计算方法
偿债能力	liquidity_ratio	流动比率	流动资产/流动负债
	quick_ratio	速动比率	（流动资产-存货）/流动负债
	assets_liabilities	资产负债率	总负债/总资产
	cash_liabilities	经营活动产生的现金流量净额/负债合计	经营活动产生的现金流量净额/总负债
经营能力	inventory_income	存货与收入比	存货/总收入
	assets_income_ratio	流动资产与收入比	流动资产/总收入
	asset_turnover	总资产周转率	收入净额/平均资产总额
盈利能力	operating_margin	营业毛利率	营业毛利额/主营业务收入
发展能力	capital_preservation	资本保值增值率	期末所有者权益/期初所有者权益
	profit_growth	利润总额增长率	（本期利润总额-上期利润总额）/上期利润总额
	sustain_grow_rate	可持续增长率	资产收益率*收益留存率/（1-净资产收益率*收益留存率）
治理结构	TopTenHoldersRate	股权集中度	前十名股东持股比例之和(%)
	reg_shareholding	监管层持股比例	监管层持股数/总股数
动力指标	get_paid	是否领取薪酬	财务类高管是否在上市公司领取薪酬 1=在上市公司领取薪酬，2=未在上市公司领取薪酬
财务表现	management_rate	管理费用增长率	（管理费用本年本期金额— 管理费用上年同期金额）/(管理费用上年同期金额)
	operating_cost	营业总成本增长率	（营业总成本本年本期金额— 营业总成本上年同期金额）/（营业总成本上年同期金额）
	sales_expense_rate	销售费用增长率	（销售费用本年本期金额— 销售费用上年同期金额）/（销售费用上年同期金额）
	bd_ratio	2018年坏账准备比例	(2018年坏账准备计提数/2018年应收账款数) -(2017年坏账准备计提数/2017年应收账款数)
	bad_debt18	坏账准备计提比例的变化率	（2018年坏账准备比例- 2017年坏账准备比例）/2017年坏账准备比例

表 3.3: 描述性统计

Statistic	N	Mean	St. Dev.	Min	Pctl(25)	Pctl(75)	Max
liquidity_ratio	3,208	2.370	2.673	0.123	1.180	2.580	54.507
quick_ratio	3,208	1.902	2.414	0.081	0.799	2.032	41.266
assets_liabilities	3,208	0.421	0.197	0.017	0.265	0.561	0.993
cash_liabilities	3,208	0.176	0.348	−2.213	0.022	0.259	3.770
inventory_income	3,208	0.404	1.837	0.000	0.099	0.327	83.662
assets_income_ratio	3,208	6.468	251.986	0.068	0.696	1.576	14,236.340
asset_turnover	3,208	0.648	0.531	0.0001	0.357	0.800	9.663
operating_margin	3,208	0.307	0.183	−0.728	0.181	0.397	0.990
capital_preservation	3,208	1.155	1.682	0.022	1.010	1.124	69.264
profit_growth	3,208	−0.399	10.785	−213.431	−0.358	0.331	371.127
sustain_grow_rate	3,208	0.040	0.141	−0.977	0.018	0.089	1.592
TopTenHoldersRate	3,208	60.209	14.726	10.890	51.227	69.755	100.970
bad_debt18	3,208	53.899	12.819	0	50	61.5	91
bd_ratio	3,208	0.012	0.256	−1	0	0	12
management_rate	3,208	0.298	2.346	−0.737	0.049	0.309	125.092
sales_expense_rate	3,208	1.723	73.196	−1.000	−0.001	0.315	4,133.569
Ownership	3,208	1.860	0.809	1	1	2	8
operating_cost	3,208	0.418	8.991	−0.829	0.038	0.299	504.738
reg_shareholding	3,208	0.115	0.178	0.000	0.000	0.196	0.823

表 3.4: VIF检验

变量	VIF
流动比率	83.047
速动比率	81.285
资产负债率	1.825
经营活动产生的现金流量净额/负债合计	1.454
存货与收入比	1.315
流动资产与收入比	1.014
总资产周转率	1.131
营业毛利率	1.330
资本保值增值率	10.679
利润总额增长率	1.407
可持续增长率	12.630
股权集中度	1.456
监管层持股比例	1.1279
是否领取薪酬	1.0539
管理费用增长率	1.130
营业总成本增长率	1.677
销售费用增长率	1.041
坏账准备计提比例的变化率	1.038
2018年坏账准备比例	1.083

第四章　实证分析

4.1　数据平衡与样本重建

初步我们得到有极强的造假嫌疑的样本共120个，无造假嫌疑的样本有3088个，二者比例为1：25左右，存在较为严重的样本不平衡问题。对于这个问题的解决主要通过欠采样和过采样的方法，考虑到欠采样将导致样本的数据急剧下降，因此本文采用的过采样的方法进行解决。

本文采用SMOTE方法[19]进行数据过采样，最终得到造假样本440条，非造假样本528条，二者比例接近1:1。

4.2　Logistic模型实证分析

4.2.1　Logistic模型

Logistic 回归模型是传统经典的分类方法，在因变量取值为二分类时进行回归分析，基本思路是通过 Logistic 非线性变换，利用极大似然估计的方法，通过Newton-Raphson方法进行迭代求解，求出系数的估计值，建立一个概率拟合函数。它是一个参数线性的判别类，用于解释一个二元变量与一个或多个度量自变量之间的关系[20]。

本文通过 Logistic 回归模型将自变量的取值代入到概率拟合函数，在二元值中通过预测因变量的概率取值，假设预测概率大于 0.5，则财务舞弊预测发生，预测概率小于 0.5，则财务造假预测不发生。概率取值为度量尺度，从而判断样本的所属类别进行二分类。

4.2.2　Logistic模型的构建与实证分析

首先本文先对前文提到的变量共17个[①]，进行建模得到一个完整的全模型，而后，根据AIC原则，采用向前和向后两种变量选择方法得到了相同的结果，再删除掉不显著的无关变量，最终

[①]排除了两个共线性变量liquidity_ratio与capital_preservation。

得到了如表4.1中的方程(1)，该方程的变量选择涉及到了我们所构建指标的各个方面，其中财务表现，发展能力，治理结构，动力指标，偿债能力这几个方面都有变量入选。

表4.1汇报了基于先前初步选择变量的结果。从方程(1)可以看出，在财务造假的表现方面，operationg_cost显著为负的，即营业总成本增长，对于财务造假的预测具有明显的负效应。换句话说，中国的上市公司在财务造假的过程中，往往会选择通过即营业总成本减少以达到虚增利润的效果,而较少会通过计提坏账准备，减少销售费用和减少管理费用的方式；在发展能力方面sustain_grow_rate是显著为负的，而且其对最终财务造假的预测影响较大，说明一个发展前景较好的公司具有较低的财务造假的可能，这与之前的理论分析相契合；

在方程(1)中财务类高管是否领取薪酬对于财务造假的预测是没有显著影响的，然而作为动力指标，不领取薪酬可能会导致更强的财务嫌疑，因为其报酬极有可能与公司的财务表现相挂钩，其主观动力会更强。

为探究在不同的财务指标下，财务类高管领取与不领取薪酬对财务造假是否存在不同影响，得到了回归方程(2)加入了get_paid与quick_ratio的交叉项，可以知道，财务类高管领取薪酬确实对财务造假预测有一定的影响，且其参数显著为负，表明财务类高管没有在上市公司领取薪酬，将会使降低财务报表造假预测的概率，即认为财务类高管没有领取薪酬的上市公司财务造假的可能会小于那些财务类高管领取薪酬的上市公司，并且该影响将会随着quick_ratio的上升而弱化，当$quick_ratio > 1.318$,该影响将会由负转正。速动比率衡量的是短期偿债能力，从衡量长期偿债能力的assets_liabilities即资产负债率来看，结论也是类似的。速动比率是一个正向指标[②]，资产负债率是一个负向指标。因此，同样对比方程(1)与方程(3)可以发现在不同资产负债率的公司，财务类高管是否领取薪酬对上市公司是否财务造假具有显著的影响，甚至随着资产负债率的不断上升，其影响会由正转负。因此这就说明了观点：在不同偿债能力的公司，财务类高管是否领取薪酬对上市公司是否财务造假具有显著的影响是不同的，偿债能力越强的公司，其财务类高管领取薪酬对财务造假的预测具有更强的负向影响，即认为财务类高管领取薪酬会减少财务造假的可能，而且这种影响在偿债能力强的公司将会更加显著。

[②]正向指标指认为指标值越大，偿债能力越强，负向指标与之相反

表 4.1：回归结果

	Dependent variable:		
	ViolationTypeID		
	(1)	(2)	(3)
get_paid2	0.484	−1.566**	2.863**
	(0.407)	(0.774)	(1.250)
quick_ratio	0.177***	0.160***	0.176***
	(0.061)	(0.059)	(0.060)
assets_liabilities	1.421**	1.525**	1.571**
	(0.632)	(0.630)	(0.637)
cash_liabilities	−2.343***	−2.371***	−2.400***
	(0.475)	(0.480)	(0.479)
sustain_grow_rate	−7.302***	−7.329***	−7.321***
	(0.709)	(0.712)	(0.711)
TopTenHoldersRate	−0.012**	−0.011*	−0.011*
	(0.006)	(0.006)	(0.006)
operating_cost	−0.361*	−0.333*	−0.319*
	(0.208)	(0.188)	(0.187)
get_paid2:quick_ratio		1.188***	
		(0.438)	
get_paid2:assets_liabilities			−5.151**
			(2.517)
Constant	−0.249	−0.309	−0.353
	(0.512)	(0.510)	(0.514)
Observations	968	968	968
Hosmer-Lemeshow X2	47.229***	48.862***	47.947***
Note:			*p<0.1; **p<0.05; ***p<0.01

4.2.3 Logistic模型评价

本文通过变量选择得到了模型 (4.1)其方程为：

$$\begin{aligned}\log it(\hat{\pi}) = &-0.309 + 1.566 get\_paid + 0.160 quick\_ratio + 1.525 assets\_liabilities \\ &- 2.371 cash\_liabilities - 7.329 sustain\_grow\_rate - 0.011 TopTenHoldersRate \\ &- 0.333 operating\_cost + 1.188 get\_paid2 \times quick\_ratio\end{aligned} \tag{4.1}$$

其中：$\log it(\hat{\pi}) = ln(\frac{\hat{\pi}}{1-\hat{\pi}})$

对上述模型，本文选择了常用的auc指标与混淆矩阵[③]对其预测能力进行评判。根据模型可得到如图 4.1所示的ROC曲线图，其AUC值达到了0.8834，具有较好的识别财务造假的能力。为使得到的结果更加具有稳健性，本文接着采用了十折交叉检验[④]的方式对样本进行切割为训练集和测试集，用训练集拟合模型参数，并用训练集进行检验预测能力，并将得到的混淆矩阵十次的结果取平均，得到如表4.2所示，其specificity的均值为0.911，sensitivity的均值为0.702，十次得到的auc取均值为0.886。综上，是可以认为我们建立得到的模型是具有较强的预测能力的。

表 4.2: Confusion matrix

Fraud In Reality	Fraud In Prediction	
	T	F
T	40.0	3.9
F	15.5	36.6

4.3 随机森林模型分析

为探究本文选择的变量在使用其他预测模型时仍然有较好的效果，本文选择了树模型对本文选择的变量进行预测，树模型包括决策树，随机森林，提升树等，其中随机森林(Randomforest)，提升树作为一种组合分类器算法，在大样本、高维度特征和异常值数据上仍然能够保持较好的预测准确率，最终通过准确率的比较，本文选择了随机森林模型作为最终模型，在这项预测中，随机森林模型在多种树模型中具有明显的优点。

[③]其中π_0为样本比例0.545

[④]即将所有的数据随机切割成十份,依次选择其中九份作为训练集，剩下一份样本为测试集

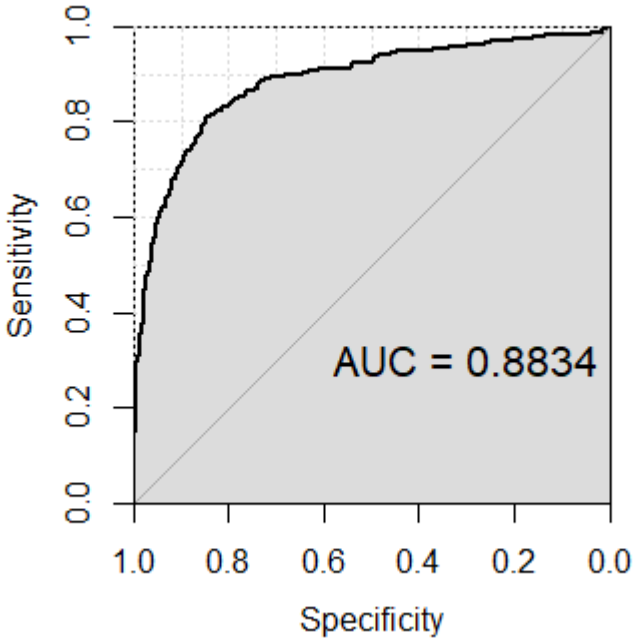

图 4.1: logistic ROC

4.3.1 随机森林模型

随机森林模型通过对树作相关处理，实现对袋装法树的改进。在随机森林中需对自助抽样训练集建立一系列决策树，但是在建立决策树时，每考虑树上的一个分裂点，都要从全部的p个预测变量中选出一个包含m个预测变量的随机样本作为候选变量。随机森林算法可以用于连续变量的预测，也可以用于分类型变量的预测，对于本文的二值变量的预测同样具有较好的效果[21]。随机森林模型的决策树分类器的构建过程如下：

(1)对原始数据集有放回的随机抽样，引导生成与原始实例集相同数量的样本，组成一定规模的随机子集，随机子集的规模对应随机森林的规模;

(2)在每个随机子集的生成过程中，可以证明，约有37%的样本不会被采用，称为OOB(out of bagging)。并且，利用OOB进评估模型的训练损失，是一种无偏估计[22]。

(3)随机子集对应决策数根节点。在每个节点随机抽样样本特征，最后到达限制层数或者当每个叶节点是纯数据的时候，停止分裂。

4.3.2 随机森林模型的选择与构建

本文采用前文提到的十折交叉验证来对进行模型的选择，通过对树的棵树进行调参，并采用十折交叉验证来计算其平均的准确度，最终得到三种模型的准确率的对比如图4.2。从图中可以到随机森林算法的误差率是显著小于Bagging和Adaboost的，故最终选择了随机森林算法，找到其中使误差最小的树的棵数为505。

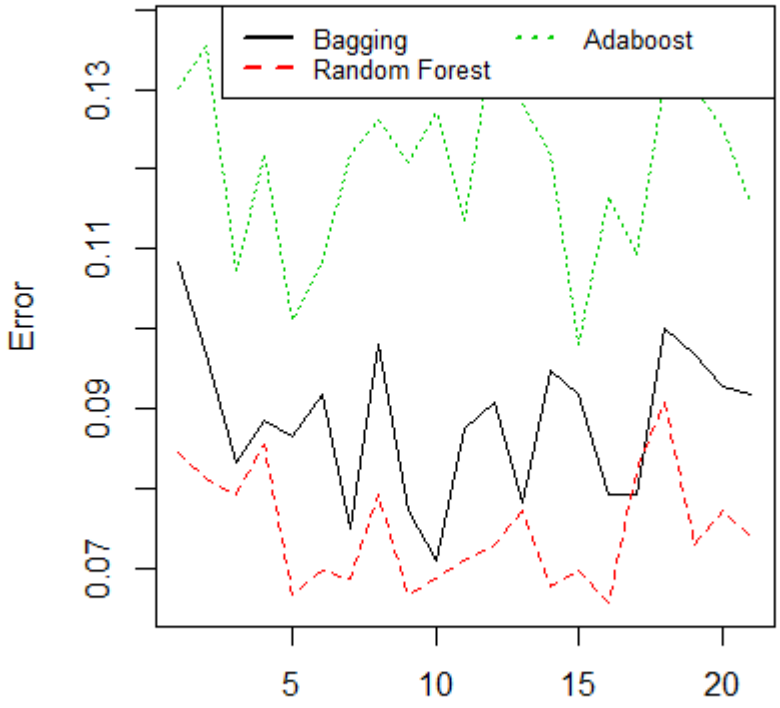

图 4.2: 模型平均误差图

4.3.3 随机森林模型的评价

使用准确率最高的参数即$ntree = 505, mtry = 4$建立随机森林模型可以得到其准确率达93.4%。

对于上述模型，本文同样选择了常用的auc指标与混淆矩阵对其预测能力进行评判。根据模型可以得到如图4.3所示的ROC曲线，其AUC值达到了0.976，相对于logistic模型来说是显著的提升。同样使用十折交叉检验的方式对样本进行切割，并得到十次的平均混淆矩阵，得到如表表4.3所示，其specificity的均值为0.900,sensitivity的均值为0.914,相对于logistic模型来说，其specificity的值没有明显的下降，但是sensitivity却有明显的提高。

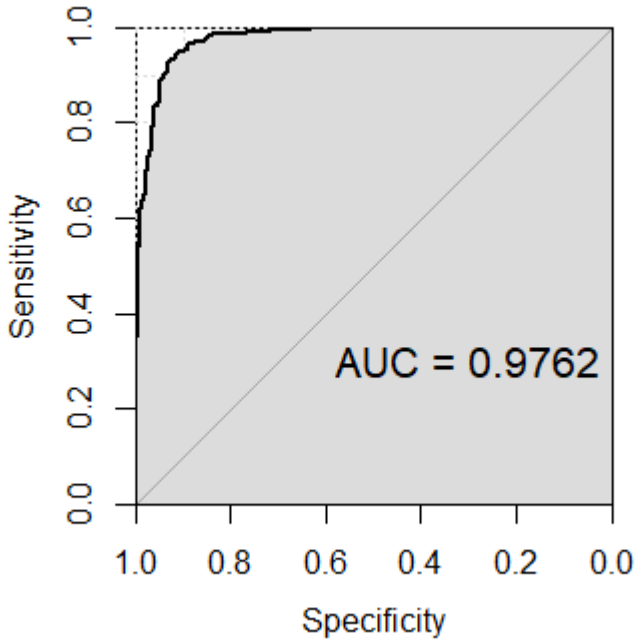

图 4.3: RandomForest ROC

表 4.3: Confusion matrix

Fraud In Reality	Fraud In Prediction	
	T	F
T	38.0	4.2
F	4.6	49.2

对于随机森林模型为什么可以显著提高sensitivity的原因除了该模型可以解决非线性分类等模型本身的优势之外，其预测变量的选择与logistic模型选择的变量不同也是重要的一方面，如

图4.4所示，可以看到在随机森林模型判别标准中利润总额增长率和可持续增长率是判别模型中最重要的变量，其中利润总额增长率在logistic模型甚至没有出现在最终选择的模型中。

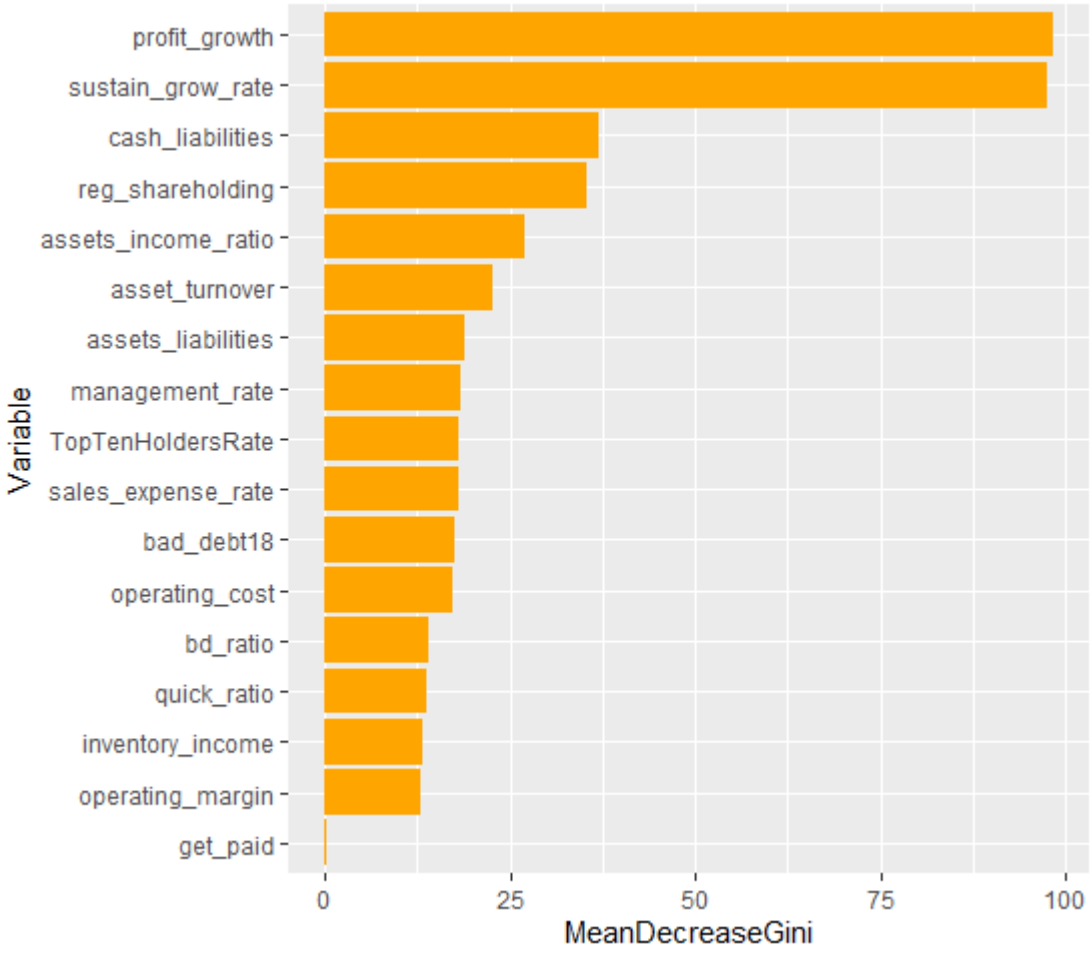

图 4.4: 变量重要性排序

第五章　结论与局限性

本文以2018年上市公司财务报表结合违规内容以及会计师审计意见的数据，考察了当前财务造假的财务表现，建造一个财务造假的预测模型。

5.1　结论

logistic模型研究结果显示：中国的上市公司往往会选择通过营业成本进行改动来达到自己美化财务报表的目的；发展前景较好的公司往往具有较低的财务造假的可能；在偿债能力越强的公司，其财务主管领取薪酬会更加减少财务造假的可能。从对logistic模型的评价中我们可以发现，logistic模型对财务造假的预测效果总体来说并不是非常的好，尤其是甄别造假公司的确造假的准确率(sensitivity)并不是特别高。鉴于此，logistic模型对于统计解释有着较大的意义，但是对财务造假模型的预测仍然需要进一步的改进。

随机森林模型研究结果显示：机器学习算法对于财务造假的分类问题十分有效。其不论是在specificity和sensitivity两方面都表现的非常好。且该模型对于变量的选择也与logistic模型具有十分大的区别，认为利润总额增长率和可持续增长率是判别一个公司是否产生财务造假的最重要的两个指标。

5.2　局限性

会计账目中存在一些如人力资本、知识产权及专利技术等无形资产难以用货币衡量。会计报表遵循谨慎性原则，并没有包括未来企业价值和报表外的信息，如企业的社会责任的衡量。这些问题可能会影响控制变量的计算结果，进而影响模型分类的准确率与解释性。

参考文献

[1] ULICK J. 2002: Year of the scandal - dec. 17, 2002[EB/OL]. [2019-12-24]. https://money.cnn.com/2002/12/17/news/review_scandals/index.htm.

[2] CRESSCY D. Other people’s money, a study in the social psychology of embezzlement montclair[J]. NJ Patterson smith, 1973.

[3] BOLOGNA J, LINDQUIST R J, WELLS J T. The accountant’s handbook of fraud and commercial crime[M]. [S.l.]: Wiley New York, NY, 1993.

[4] 梅丹, 徐颖. 我国制造业上市公司内部控制信息披露问题与对策[J]. 财政监督, 2014(29): 16–21.

[5] PAVLOVIć V, KNEžEVIć G, JOKSIMOVIć M, et al. Fraud detection in financial statements applying benford’s law with monte carlo simulation: volume 69[EB/OL]. 217–239(2019-06)[2019-12-21]. https://www.akademiai.com/doi/10.1556/032.2019.69.2.4.

[6] MD NASIR N A B, ALI M J, RAZZAQUE R M, et al. Real earnings management and financial statement fraud: evidence from malaysia: volume 26[Z]. 2018: 508–526[2019-12-21]. DOI: 10.1108/IJAIM-03-2017-0039.

[7] GAO Y, KIM J B, TSANG D, et al. Go before the whistle blows: an empirical analysis of director turnover and financial fraud: volume 22[EB/OL]. 320–360(2017-03)[2019-12-21]. http://link.springer.com/10.1007/s11142-016-9381-z.

[8] LIAO L, CHEN G, ZHENG D. Corporate social responsibility and financial fraud: evidence from china[EB/OL]. acfi.12572(2019-11-19)[2019-12-21]. https://onlinelibrary.wiley.com/doi/abs/10.1111/acfi.12572.

[9] TSENG F M, LIN L. A quadratic interval logit model for forecasting bankruptcy: volume 33 [EB/OL]. 85–91(2005-02)[2019-12-04]. https://linkinghub.elsevier.com/retrieve/pii/S0305048304000623. DOI: 10.1016/j.omega.2004.04.002.

[10] CHEN Y J. Enhancement of fraud detection for narratives in annual reports[Z]. [S.l.: s.n.], 2017: 14.

[11] Abbasi, Albrecht, Vance, et al. MetaFraud: A meta-learning framework for detecting financial fraud: volume 36[Z]. 2012: 1293[2019-12-21]. DOI: 10.2307/41703508.

[12] 卢涛. 我国上市公司财务报告舞弊行为识别及其监管研究[D]. 大连: 东北财经大学, 2013.

[13] 李明. 基于cart树的上市公司财务风险预测研究[D]. 武汉: 武汉科技大学, 2015.

[14] 邵朝, 林路路, 周谋. 混合核svm财务欺诈识别[J]. 西安邮电大学学报, 2017(2).

[15] 徐延. 基于数据挖掘的公司财务造假识别模型研究[D]. 南京: 南京大学, 2019.

[16] 靳超. 财务报表舞弊的影响因子研究[D]. 山东: 山东财经大学, 2013.

[17] 詹红雁. 上市公司财务造假常用的手段、识别及防范建议[J]. 商业会计, No.632(8): 101–103.

[18] NYITRAI T, VIRáG M. The effects of handling outliers on the performance of bankruptcy prediction models: volume 67[Z]. 2019: 34–42[2019-12-04]. DOI: 10.1016/j.seps.2018.08.004.

[19] CHAWLA N V, BOWYER K W, HALL L O, et al. Smote: Synthetic minority over-sampling technique[J]. Journal of Artificial Intelligence Research, 2011, 16(1): 321–357.

[20] AGRESTI A. Wiley series in probability and mathematical statistics: An introduction to categorical data analysis[M]. 2nd ed ed. [S.l.]: Wiley-Interscience, 2007.

[21] JAMES G, WITTEN D, HASTIE T, et al. Springer texts in statistics: volume 103 an introduction to statistical learning[M]. Springer New York, 2013[2019-12-19]. DOI: 10.1007/978-1-4614-7138-7.

[22] BREIMAN L. Random forests[J]. Machine Learning, 2001, 45(1): 5–32.

[23] FOREMAN R D. A logistic analysis of bankruptcy within the US local telecommunications industry[Z]. [S.l.: s.n.], 2003: 32.

[24] HESSING D, ELFFERS H, ROBBEN H, et al. Needy or greedy? the social psychology of individuals who fraudulently claim unemployment benefits 1[J]. Journal of Applied Social Psychology, 1993, 23(3): 226–243.

附录A

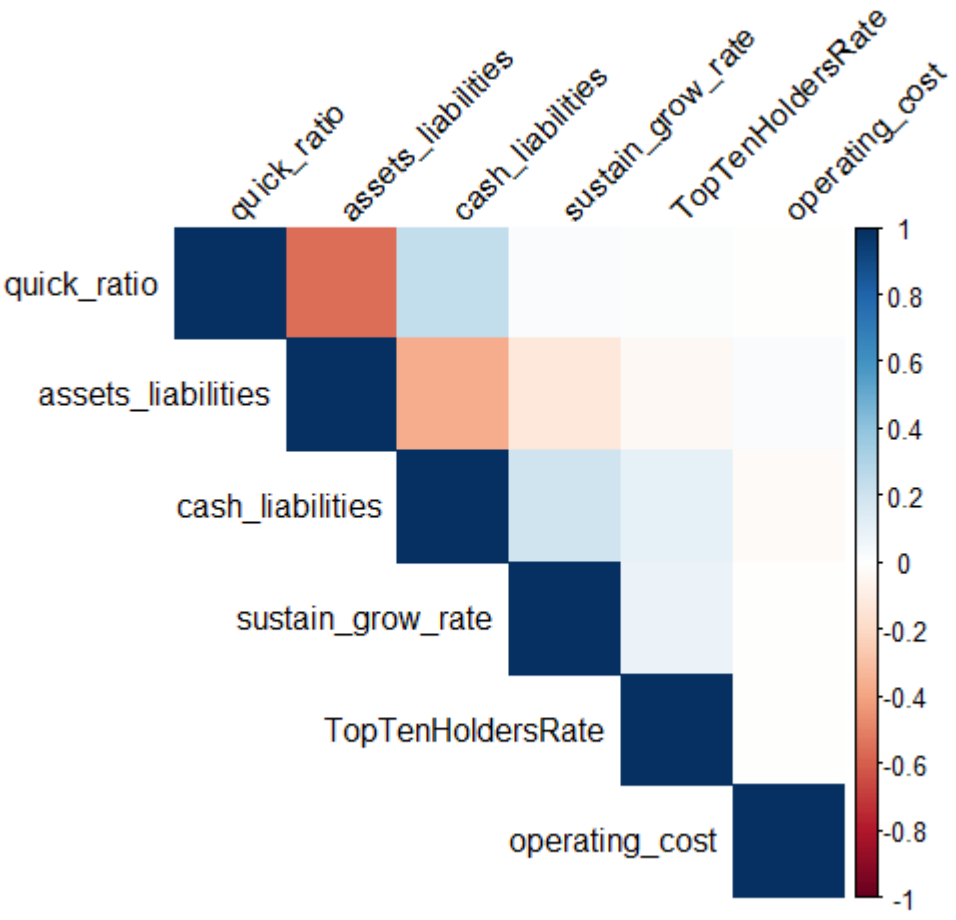

图 5.1：相关系数图