分类号_____ 密级_____
U D C _____ 编号 10736

西北師範大學

硕士学位论文

（专业学位）

基于深度学习的上市公司财务欺诈识别模型构建研究

——以ST康美为例

研 究 生 姓 名：龙 瑶
指导教师姓名、职称：赵雪梅 教授
实践指导教师姓名、职称：蔡恒备 高级会计师

专 业 学 位 类 别：会计硕士
专 业 学 位 领 域：会计
专 项 计 划：

二〇二三年五月

Research on the construction of financial fraud identification model of Listed companies based on deep learning

—— Take ST Kangmei as an example

A Thesis Submitted to

Northwest Normal University

in partial fulfillment of the requirement

for the degree of

Master of Professional Accounting

by：

Long Yao

Supervisor : Professor Zhao Xuemei

May, 2023

西北师范大学学位论文原创性声明

本人郑重声明：所呈交的学位论文是本人在导师的指导下独立进行研究工作所取得的成果。除文中已经注明引用的内容外，本论文不含任何其他个人或集体已经发表或撰写过的作品成果。对本文的研究做出重要贡献的个人和集体，均已在文中以明确方式标明。因本学位论文引起的法律后果完全由本人承担。

学位论文作者签名：龙瑶

导师签名：

签字日期：2023年5月25日

西北师范大学学位论文版权使用授权书

本学位论文作者完全了解西北师范大学有关保留、使用学位论文的规定，有权保留并向国家有关部门或机构送交论文的纸质版和电子版，允许论文被查阅和借阅。本人授权西北师范大学可以将学位论文的全部或部分内容编入有关数据库进行检索，可以采用影印、缩印或扫描等复制手段保存、汇编学位论文，可以公开学位论文的全部或部分内容。

（保密的学位论文在解密后适用本授权书）

学位论文作者签名：尤瑶

签字日期：2023年5月25日

摘要

近年来，随着资本市场和经济的快速发展，上市公司越来越多且规模越来越大，与此同时，伴随出现的财务欺诈问题也屡见不鲜，情节越来越恶劣，后果越来越严重。从 2017 年开始的数据整理可以看出证监会发出的行政处罚通知涉及财务造假行为的处罚与日俱增，涉及到的上市公司被确认为财务欺诈公司的也越来越多。通过整理和分析可以看出，财务欺诈识别问题广泛存在并且还没有得到很好的解决。

本文用 SMOTE 采样数据训练 DT、RF、ET、XGBoost、KNN、LR、LGBM、AdaBoost 八种机器学习算法,通过单模型构建评分和融合模型效果选出最优的三种算法 RF、ET、XGBoost 去求对上市公司财务数据造假有较大影响的特征。在 AUC 指标最优的情况下，将 RF、ET、XGB 这三种算法模型的特征选择结果进行综合。计算出特征重要性权重值，筛选出特征权重值排名前 31 的作为对上市公司财务数据造假有较大影响的特征因子。对我国 2015 年以来上市公司的财务数据和证监会等各项处罚文件进行整理和搜集，选定了 5125 个沪深 A 股上市公司财务数据为数据集，其中制造业行业 3336 家，其他行业 1789 家，计划以数据集中 3336 家制造业企业中的 2730 个公司作为训练集，剩下的 606 个为测试集；以数据集中 1789 个其他行业中的 1464 个企业作为训练集，其余 325 个为测试集。

最后本文确定以曾是医药制造业龙头的代表性企业 ST 康美作为研究对象。其曾披露了一则称有近三百亿“会计差错”的更正公告，这是一次百亿“白马股”的财务欺诈，属于恶意利用前期会计差错更正，并以此为手段试图掩饰财务欺诈这一事实，这是管理层有意且违法的操作，造成震动资本市场的巨大损失，说明财务欺诈问题不只是存在于监管可能有疏漏的中小企业，在大企业中也同样普遍。将 ST 康美代入融合模型进行验证，模型成功识别出 ST 康美的造假年度。用同样的方法检测了金亚科技、尔康制药、联建光电三个企业，均成功识别出各企业造假年度，模型验证为有效。通过建立本文的模型，也得知影响因子的选择对于模型建立十分关键。

关键词：Bagging＋DCRN；机器学习；模型融合；财务欺诈

Abstract

In the past several years, owing to the swift progress of the capital market and the economy, the number of listed companies has increased significantly, and many of them have also experienced substantial growth in size and scale. At the same time, the accompanying financial fraud problems also appear from occasional to common, the circumstances are more and more bad, and the consequences are more and more serious. From the data collation starting in 2017, it can be seen that the administrative punishment notice issued by the CSRC involves more and more penalties for financial fraud,and more and more listed companies involved have been identified as financial fraud companies. Through sorting and analysis, it can be seen that the problem of financial fraud identification exists widely and has not been well solved.

In this paper, SMOTE sampling data is used to train DT, RF, ET, XGBoost, KNN, LR, LGBM and AdaBoost to build single model scoring and fusion model effect, and select the best three algorithms RF, ET and XGBoost to find the characteristics that have a great influence on financial data fraud in publicly listed companies. In the case of optimal AUC index, the feature selection results of the three algorithm models of RF, ET and XGB are integrated. The feature importance weight values were computed, and the top 31 features were selected as financial indicators of listed companies based on their relevance to fraud characteristics.For financial data of listed companies since 2015 and CSRC punishment documents, 5125 A-share listed companies were selected as the data set, including 3336 companies in manufacturing industry and 1789 companies in other industries, 2730 companies in 3336 manufacturing enterprises in the data set as training sets, the remaining 606 as test sets; 1464 enterprises in 1789 other industries as training sets, and the remaining 325 as test sets.

Finally, this paper selects ST Kangmei, a representative company in the leading pharmaceutical manufacturing industry, as the research subject. Its disclosed a said nearly three hundred billion “accounting error” correction announcement, this is a billions “white” financial fraud, belongs to the malicious use of early accounting error correction, and this means to hide the financial fraud the fact, this is the management intentionally and illegal operation, shock the capital market, financial fraud problem is not only widespread in regulation may have omissions of small and medium-sized

enterprises, is also common in large enterprises. ST Kangmei was replaced into the fusion model for validation and the model was successfully identified ST Kang mei fraud year. With the same method, we tested the three enterprises of Jinya Technology, Erkang Pharmaceutical and Lianjian Optoelectronics, and all successfully identified the fraud year of each enterprise, and the model verification was effective. Through establishing the model in this paper, we also know that the choice of influence factor is very critical for the model establishment.

Key words: Bagging + DCRN; machine learning; model fusion; financial fraud

目录

第 1 章 绪论

1.1 研究背景

对于上市公司来说，财务欺诈行为的问题普遍存在，不单单在我们国家频发，对于很多市场经济发展成熟于我国的国家，不论大小企业，都或多或少存在财务欺诈问题。财务欺诈行为所引发的后果对政府、企业本身到投资者都有很多危害，最后甚至会导致各种社会和经济问题发生。所以，财务欺诈如何有效被识别出来并及时采取相应的防范以及应对措施一直在学术界的研究中占重要地位。

在对数据进行了整理后，可知证监会在 2017 年一共发出了 109 份行政处罚通知，其查处的重点主要是针对财务造假等四大领域。如果只针对数量来比较，确实比 2016 年少了，但从性质上来说，有关财务造假行为的处罚有 29 份，占处罚总量的 26.6%，远远高于上一年；2018 年有中兵红箭、恒顺众昇、众和股份等 11 个上市公司被确认为财务欺诈公司；2020 年 3 月 24 日中国证监会在官网发布《证监会严厉打击上市公司财务造假》文章，称 2019 年以来已累计对 22 家上市公司财务造假行为立案调查，对 18 起典型案件做出行政处罚，向公安机关移送财务造假涉嫌犯罪案件 6 起。2021 年 4 月 16 日证监会通报上市公司财务造假案件办理情况称 2020 年以来，证监会共办理该类案件 59 起，占办理信息披露类案件的 23%，向公安机关移送相关涉嫌犯罪案件 21 起。2021 年 1 月 1 日至 12 月 31 日，共有 32 家上市公司因财务造假被证监会或证监局实施行政处罚。其中深主板 17 家，沪主板 9 家，创业板 6 家。通过以上数据可以看出，财务欺诈识别问题广泛存在并且还没有得到解决。

在 2018 年收到中国证监会《调查通知书》，并在 2019 年变更股票名称的“ST 康美”，此前备受投资者追捧，因受到重重质疑后竟披露了一则称有近三百亿“会计差错”的更正公告。在 2010 年上市的康得新，自上市后在七年内市值翻了近 30 倍，对其进行审计的会计师事务所已经连续多年出具了标准无保留意见的审计报告，直到康得新财务欺诈事实浮出水面，在债券违约事件曝光后，瑞华会计师事务所才作出了“无法表示意见”的审计报告，而其股票名称也最终变更为“*ST 康得”，直到最终退市。而对资本市场震动如此之大的此类百亿甚至千亿的“白马股”财务欺诈竟蒙蔽了监管机构和投资者许久，造成巨额损失，说明财务欺诈问题不只是存在于中小企业，在大企业中也同样普遍。

如今应用较多的财务欺诈识别手段有检查以及延伸检查、实施审计程序、分

析性程序、研究变量之间关系的线性模型、概率统计模型等传统方法。无论从曾经的乐视网、蓝田股份事件，还是近年的康得新、康美药业、德豪润达的财务造假案例来看，财务造假行为都越来越隐蔽，注册会计师越来越难以设计合理审计程序进行财务造假的识别。本文认为应对越来越难以识别的财务欺诈行为，传统的审计方法和程序不够有效。

综上，可以看出现有的财务欺诈识别技术和手段还远远不够，并且当前对各公司的审计都太依赖审计师和监管者，而审计师和监管者不管从主观还是客观上来说都可能存在无意的疏漏或者有意的错误，这就导致了财务欺诈的发生一直广泛存在甚至越来越多，而对财务欺诈的识别也由于人为处理而受到精力、资源等的制约，从发生到被识别的时间滞后；同时，由于欺诈方式越来越层出不穷，越来越难快速且精准地进行人为识别。因此，通过机器学习建模来高效识别财务欺诈显得越发迫切。

1.2 研究意义

上市公司因财务欺诈可能造成经济波动、资本市场混乱，更会直接给投资者和债权人带来或多或少的损失。因此，能够快速有效地对财务欺诈进行检测是社会各级人员翘首以盼的。当将财务欺诈检测分为传统意义上的审计与运用大数据识别模型这两种方法时，通过对比可以看到传统方法将耗费巨大的人力，为了确保审计工作的顺利进行，通常需要专业的审计人员进行审核。这些人员拥有财务知识，可以对企业的财务数据报表进行审查，效率低、准确性差都是对有效识别财务欺诈行为的制约。在大数据技术被广泛应用的今天，越来越有可能在机器学习方法的基础上建立更好的财务欺诈识别模型。

基于以上背景，本文在一系列上市公司财务以及非财务的数据挖掘后选取国内医药制造业代表性企业 ST 康美作为模型验证对象，通过案例研究法、文献研究法、模型构建法等对其各项指标进行了分析和考量，以期为我国医药制造业企业的财务欺诈识别提供一定的参考，向各行各业上市公司敲响警钟，进一步保障资本市场的稳定，发挥市场监督积极的作用，完善我国股权激励机制和体系。本文构建的模型在财务欺诈行为识别方面能够帮助政府监管部门以及审计机构更加快速准确地发现有异常迹象的公司，从而能够在监管过程中更加高效且更有针对性，该模型的应用还有助于投资者更清晰全面地认识上市公司，从而更好地保护他们的利益。

我国对于财务欺诈识别系统案例研究时间较短，数量有限，研究仍处于进一步发展的阶段。基于以上考虑，对于财务欺诈识别模型的深入研究，具有一定的

理论和实践价值。

1.2.1 理论价值

丰富了财务欺诈问题的研究角度。财务舞弊以及财务欺诈问题的研究已经较深入，但较少有学者从财务欺诈识别模型的角度深度剖析该问题。因此本文通过对财务欺诈识别模型的建立和对所选定的案例公司检测并验证，丰富了财务欺诈识别系统问题的研究角度。

1.2.2 实践价值

我国证券市场还不够成熟，监管环境也日益复杂，欺诈行为越发难以识别，监管部门面临着诸多问题，注册会计师对财务欺诈行为的识别难度也越来越大。

本文主要立足于证券市场迫切待解决的财务欺诈问题，运用了大数据技术，通过机器学习，建立财务欺诈识别模型，选定案例企业进行模型验证，以期为财务欺诈识别研究提供更为有效准确的识别方法和系统，从制度角度加强约束力，防止挪用和侵占资金，同时还能给利益相关者带来帮助，给其他类似上市公司一些启示，为企业敲响警钟，对投资者作出正确的投资决策提供帮助和指导，为维持资本市场的稳定出一份力，有一定的实践价值。

1.3 文献综述

1.3.1 财务欺诈动因相关研究

西方学术界探讨了导致财务欺诈行为实施的动因，对于这个问题的研究历史悠久，成果丰硕。这主要得益于发达、成熟的证券市场。AlbreCht(1995)[1]结合实践提出了“欺诈三角理论”。压力、机会和借口三个因素息息相关、密不可分组成了“欺诈三角理论”的三大核心点，这三个因素在企业正常经营运作中导致财务欺诈的诞生。该理论认为上市公司财务欺诈需要在以下主要条件同时存在时才会发生：一个是压力，管理层来自工作的压力和企业来自经济的压力都包含在其中。对于企业发生财务欺诈行为来说，压力是极大的行为动机；另一个是机会，企业存在一些机会，使财务欺诈行为的实施处于在隐蔽中进行；最后一个是借口，欺诈者可以通过编造虚假的理由或辩解来掩盖其欺诈行为，让其看起来合理或正当。是由 Bologna 等人(1993)[2]提出 GONE 理论，该理论认为，贪婪(Greed)、机会(Opportunity)、需要(Need)、暴露(Exposure)组成了财务欺诈行为被实施的成因，这四个因素共同影响财务欺诈的发生概率。该理论认为在贪婪因子的作用下的，只要需求和机会并存，并且财务欺诈行为被认为能够隐蔽进行，那么毋庸置疑，财务欺诈的行为就会发生。对于人的行为，有“贪婪”和“需要”两大因子，而对于组织环境来说，有“机会”和“暴露”两大因子。虞宛静(2022)[3]以瑞幸咖啡为例，从舞弊风险因子理论的两个角度，即个别风险因子和一般风险因子，从道德品质、商业模式、公司治理、跨境监管等分析瑞幸咖啡财务舞弊的动因。宋子豪(2013)[4]指出我国上市公司利用壳资源的现象已泛滥。其中，财务欺诈主要受地方政府的宽容和公司高管的贪婪所驱动。地方政府尽管已经掌握了这些企业大量负面的信息，但往往受到政绩考核的影响，会鼓励或甚至协助当地企业采取欺诈性行为，如虚假包装和上市等。这些欺诈行为旨在误导投资者和监管机构。政府官员往往需要在一定时间内完成一系列任务和指标，以证明自己的工作表现和能力。这些任务和指标通常与民生、经济、环保等方面的问题有关，因此政府官员需要通过各种手段来完成这些目标。此外，高管由于想要满足自身利益，且其拥有获得外部投资者无法获取的特权信息和职业晋升机会。便可能利用信息优势和权威性，可操纵向公众披露的信息，不断筹集资金并消耗公司资源。张敦力、王沁文(2022)[5] 通过对康美药业财务造假案件的全过程进行回顾。通过结合现有法律法规、实践经验以及学术理论对独立董事勤勉尽责的认定和评价标准，探究其独立董事在该事件中的履职情况，从而揭示导致我国监管失效的根源。Yangetal(2017) [6]通过研究发现，中国企业的股权结构、双重领导职位、外部审计以及监管机构的要

求可能导致财务诈骗。刘为岩(2021)[7]认为商业模式有缺陷，如产品结构和盈利重点等；公司治理的风险，如内部治理体系和股权结构集中度等；上市地与经营地的地域差异导致的信息不对称以及经营管理的不同政策问题，都是财务欺诈的动因。范秋如(2014) [8]认为财务欺诈的动因可以归结为以下四个。第一，基于宏观视角，为了适应环境改变，公司可能需要进行战略调整。第二，高管由于金融激励的影响，推动他们通过再融资和限售股解禁等手段追逐利益，但企业缺乏自我激励机制，可能会致使企业为达利益需求而采取冒险行为；第三是我国 IPO 标准相对严格，企业为了上市开始实施财务欺诈行为；最后，监管机制薄弱使得违法成本不够高。

屈文洲(2007)[9]提出大股东滥用权力和不当行为、公司治理不良会增加财务欺诈的风险，资产被掏空、内幕交易和盈余管理增加了财务欺诈的可能性。同样的，汪昌云，孙艳梅(2010)[10]在公司治理研究领域中研究者通过分类代理冲突，以研究公司治理中的欺诈行为。刘惠萍(2011)[11]认为尽管拟上市企业的保荐人可以获得可观的收益，但他们所承担的责任不足以弥补其所获得的收益，这导致了保荐人和企业之间的勾结，从而促成了财务欺诈的发生。李琳(2022)[12]基于 2008 年以来 13 年间我国沪深两市 A 股上市公司数据，探究了财务舞弊和税收在强度方面的表现这两类典型机会主义行为之间的关系，以及外部监督机制对这两类行为的作用。深度研究与分析之后发现，如果企业面临现金流不足的问题，那么其在同期采取税收激进决策的动机就会大大增强。韩成(2013)[13]指出，监管机构在审核体系上存在问题。由于需要处理的待上市企业数量众多，每家企业又有大量文件需要审核，监管机构的审核体系尚不完善。加之监管机构人力和物力不足，监管常常出现疏漏，给某些公司留下钻空子的机会。

牛羿恒(2020)[14]先以风险因子理论和博弈论为基础，研究我国上市公司的财务欺诈。在他看来，导致公司财务舞弊的因素主要包括管理层的私欲追求、企图规避证券市场的惩罚以及对违法行为的隐瞒。接下来，建立了一个模型，从三个不同的角度来分析研究哪些外部因素对财务欺诈行为有影响。徐凡卿、章之旺(2021)[15]在研究国内舞弊动因现状时以国外主流观点为基础。然而他们发现我们国家不但过度借鉴国外的动因理论，自身对于财务欺诈问题也分析得不够透彻。因此，其剖析了个人、组织和社会几个层面，来对上市公司存在的财务欺诈行为的多元动因进行分析。杨振宇(2020)[16]旨在介绍上市公司财务舞弊的现状，并深入研究造成难以杜绝该现象的原因。特别关注导致舞弊行为的内部动因，如内部控制结构缺陷和信息不透明，以及外部动因，如审计机构职业道德缺失、法律法规不够完善和处罚力度不足等。通过对这些因素的分析，揭示财务舞弊问题的根源，

并提出对策以促进上市公司财务诚信与规范发展。赵昱(2021)[17]认为有许多原因导致财务造假，其中包括了利益驱动、监管机构不力、审计质量不高、股权结构不完善、内部控制意识不足以及高管和会计人员缺乏职业道德等。刘艺璐(2021)[18]提出为了追求一些不合法的利益，一家公司可能会向其会计人员发布指示，生成虚假的财务数据，旨在欺骗利益相关者，让公司得以生存和增长。财务欺诈的目的都是为了利益。罗韵轩、陈卷逸(2021) [19]认为当公司提出改变主营业务方向时，这种转型给公司带来了经营上的压力和挑战，同时也面临了满足投资者预期的压力。会计师事务所未能履行职责，为公司进行财务欺诈创造了机会。此外，公司内部控制和信息披露方面的缺陷也为企业实施财务欺诈提供了借口。这些都是导致公司财务欺诈的重要动因。

1.3.2 财务欺诈手段相关研究

Beasleyetal(2000) [20]进行了一项关于 1980 年代末至 1990 年代技术、医疗保健和金融服务行业的财务报表欺诈研究后发现，不同行业的欺诈活动存在显著差异。技术公司常常从事收入欺诈，而金融服务机构更倾向于资产挪用和欺诈。这些欺诈手段的不同选择，是由于各行业特点的不同所致。黄月菡、陈庆杰(2021)[21]提出，上市公司经常使用不诚实的手段欺骗投资者和债权人，包括夸大其资产价值、隐藏债务、虚增收入以及通过关联方交易等进行欺诈。刘启亮等(2023)[22]近年来，财务欺诈的类型并未发生显著变化，但欺诈手段越来越隐蔽。虚增收入或虚减费用以实现虚构利润，仍是财务欺诈的主要手段。此外，虚列资产、隐瞒关联方交易、重大遗漏也是常见的手段。募集资金滥用、欺诈上市、隐瞒亏损、虚增资本公积等类型也屡见不鲜。这些欺诈行为严重损害了公司声誉，甚至威胁到企业的生存与发展。黄世忠、叶钦华等(2020)[23]经对 2010 至 2019 年样本公司的财务欺诈类型和会计科目进行分析，发现财务欺诈行为主要针对利润表进行篡改和操纵。在财务欺诈中，收入欺诈是最常见的，占比高达 68.14%，成为欺诈行为的主要手段。费用和成本欺诈以及资产负债表上的货币资金欺诈按照占比也是财务欺诈前几类型之一。这一分析表明，财务欺诈手段越来越隐蔽，投资者应审慎评估企业财务状况，以避免遭受欺诈行为的损害。

我国学者王晶(2012)[24]认为财务报表欺诈行为不可避免地会在公司报告中显露端倪，因为公司经常通过混淆来掩盖实际的财务表现。会计学的基本原则是财务报表编制的核心，同时也是发现财务报表欺诈的理论基础。通过贯彻会计学的基本原则，可以减少公司在财务报表中掩盖真实业务状况的机会，进而更容易发现潜在的财务报表欺诈行为。吴晓迪(2011)[25]揭露了财务欺诈的方法会根据政策的

退出和废止而不断演变。现如今，财务欺诈手段层出不穷，对收入和费用进行调整、虚增利润、以及隐瞒重要信息等手段。针对这些手段，增强社会的道德水平和诚信意识、加强执法监管、提高违法成本等对策被提出。徐丽萍(2013)[26]认为财务舞弊是企业面临的严重问题之一。为了预防和发现财务舞弊，需要了解财务舞弊的征兆。财务舞弊的征兆可以从多个方面体现出来。其中之一是财务状况不佳，包括但不限于经营亏损、现金流量不足等；另一个方面是高负债的资本结构，如债务比率和利息保障倍数的异常高值；此外，当公司的财务比率出现异于寻常的变化时，这可能意味着公司存在操纵财务数据的风险，例如应收账款周转率、存货周转率和毛利率等比率数值明显偏离历史数据或行业平均水平。因此，企业需要定期进行财务数据分析，及时发现和解决可能存在的财务舞弊问题。钱玉、徐立文(2014)[27]分析了 2007 年以来被违规处罚的上市公司，总结为：上市公司在那时实施的财务欺诈手段大多在新的会计准则实施后出现，包括虚增各项资产和收入等。暗示着财务欺诈者在不断寻找新的漏洞，以逃避监管和审计的监督。这也提示我们，审计和监管机构需要时刻关注会计准则的变化，并不断改进监管方式，以更好地发现和防范财务欺诈行为。Reurink A(2018)[28]提出，金融欺诈已经成为一个广泛记录在学术文献中的经验。它描述了在金融市场背景下出现的各种欺诈行为，以及之前研究已经确定的这类行为的普遍性和后果，以及促进欺诈的经济和市场结构。为了更好地讨论，该文将财务欺诈分为三类：虚假财务披露、财务骗局和财务不当销售。何欣惟(2016)[29]认为近年来，财务报告中涌现出新型的手段来实现错报，如摊销固定资产或其他长期资产时，采用不同的摊销方法进行计算，例如年数平均法和双倍余额递减法。在财务报表中，将采用年数平均法计算的资产价值与采用双倍余额递减法计算的资产价值进行比较，并选择较高的价值。反复执行上述步骤，以达到虚增利润的目的；以及在应付账款和其他应付款中虚报少数款项或将一部分应付账款计入其他应付款，从而降低负债总额，增加资产负债表的净资产，或将已收到的部分其他应收款计入未收款项，或将其他应收款计入其他流动资产，从而降低负债总额，增加资产负债表的净资产等。同时，对信息进行虚假披露的企业数量也呈上升趋势。这些行为的主要目的是企图通过欺诈手段获得不当收益。曾汝林(2011)[30]提出在会计和财务领域，财务欺诈手段可以从不同的角度进行归类，其中包括“假账真做”和“真账假做”。前者指企业虚构业务并但在记录时却采用正确的记录方法，而后者则指在业务真实的基础上，企业粉饰报表。王淑玲(2012)[31]研究了财务欺诈行为最盛行的年份，得出了有的上市公司通过故意隐瞒而达到欺诈目的。郑贤龙(2013)[32]分析了实际案例，讨论了财务欺诈手段公司在首次公开发行过程中是如何实施的，这些手段包括：不公开公司

内部制度和流程的问题，伪造研究报告等。刘天敏(2015)[33]提出随着经济全球化和电子信息技术的快速发展，市场经济行为变得更加复杂和多样，财务舞弊手法也不断创新，呈现出新的特点。这些特点主要包括以下三个方面：一是舞弊手法越来越借助高科技手段；二是舞弊手法更加隐蔽；三是舞弊手法更加多样化。而张彤(2019)[34]根据会计和财务方面的知识，对当前最新的财务欺诈手段进行了总结：调整会计核算方式、操纵盈利水平、虚报收入和费用等。

陈澜(2019)[35]对财务造假方式进行总结，包括收入和费用两类。收入造假主要包括虚增收入、提前透支未来收入、推迟确认相关收入等方式；费用造假主要包括将企业支出进行转移、延迟对费用的确认、设立资产减值准备等方式。曹越和伍中信(2015)[36]通过造假三角理论将财务欺诈识别模型应用于农业类上市企业。研究发现，我国农业类上市公司的财务欺诈主要出现在增发配股过程中。

1.3.3 财务欺诈识别模型相关研究

王言、周绍妮、石凯(2021)[37]构建了对国有上市公司并购风险有影响的指标评价体系，结合机器学习中的算法，利用 Python 语言构建出能够对财务欺诈行为进行预警的模型，用于评估、监控、测算和控制并购中存在的风险，在做了对比之后，以多元线性回归模型为落脚点对并购事件中存在的风险进行进一步研究。钱苹、罗玫(2015)[38]以 1994-2011 年沪深两市财务造假上市公司为样本，在对企业财务报表等数据进行分析和比对后得到了特征指标，根据其与财务造假或盈余质量问题之间的关系来进行筛选和策略性地组合，并结合欧美资本市场经常出现的研究变量，最终建立了在中国市场中最具有适用性的模型用来进行财务欺诈预警。通过变量筛选，对影响中国上市公司实施财务欺诈行为的因素进行鉴别。所建立的模型易于理解，整合了更广泛的变量，能够更深入地揭示潜在的欺诈风险因素，从而提高了其检测和预防针对中国的欺诈计划的能力。对于国内造假企业的识别优于国外的模型。Abbasi A 等人(2012)[39]为了强化财务欺诈检测，采用设计科学方法开发了 MetaFraud 元学习框架。在测试平台上进行的实验结果表明，MetaFraud 的每个组件都对整体有效性都有显著贡献。此外，还发现 MetaFraud 比最先进的金融欺诈检测方法更加有效，且生成置信度分数有助于提高检测性能并作为决策辅助。杨贵军等人(2022)[40]为了解决上市公司粉饰财务数据对财务风险预警模型准确性的影响问题。通过 Benford 律和 Myer 指数两种方法，构建了 BM-BP 模型，基于对 2000 年以来的数据研究和分析，表明实验结果：该模型预测准确率更高、误判率更低、稳定性良好，同时还采用决策树算法提高了模型的预测准确性。阮素梅,杜旭东等人(2022)[41]介绍了中文信息在智能财务风险识别中的作用，构建了能

够更好地理解文本的情感色彩和作者态度的指标体系，开发了一种基于机器学习的智能系统，可用于识别财务风险，从而提高风险管理的效率和准确性。经过一系列实验，最终得出结果：由于采用了更加综合和全面的数据来源，提高了风险识别的准确性和全面性，文本信息对于上市公司财务风险的识别起了很大作用。Ali A 等人(2022)[42]认为，财务欺诈已成为当前公司和组织面临的普遍威胁。然而，传统的技术难以精确检测欺诈交易，因此需要引入更加智能的检测方法。基于机器学习的方法能够通过分析大量的金融数据来智能地检测欺诈交易，成为新的解决方案。本文进行了文献综述，分析了 93 篇文章，总结了流行的机器学习技术、欺诈类型和评估指标。研究结果表明，支持向量机(SVM)和人工神经网络(ANN)是常用的欺诈检测算法。王昱、杨珊珊(2021)[43]提出了用于预测上市公司的财务困境的四个维度的模型，研究的结果显示：由于多维视角的存在，此模型效果优于一般的传统识别模型。Craja P 等(2020)[44]提出了一种检测报表舞弊的方法，该方法结合了公司年报中的财务比率信息和管理层评论信息。使用 HAN 算法从上市公司年报的管理讨论与分析（MD&A）部分中提取文本特征，以便更好地理解公司的经营情况和未来展望，将其与财务比率相结合，并对其进行分析和预测。结果表明，能够产生很好的分类结果。Kamal(2016)[45]使用 M-score 模型对马来西亚 17 家上市公司进行财务欺诈检测的研究发现，在财务欺诈行为被披露前，该模型就能识别出财务欺诈行为，正确率超过了 80%，这为该模型如何很好地应用于中国市场有很好的借鉴意义。

杨子晖等人(2022)[46]使用机器学习方法探究系统性风险指标在预测我国企业财务危机中的作用。最终得出：系统性风险具有可靠的预测准确性，基于因子分析构建的系统性风险指标与随机森林模型相结合预测效果更佳。此外，对于我国的大多数企业来说，随机森林模型和回归模型相结合的财务欺诈预警模型是更为有效的，同时对陷入财务状况恶化的情况进行了不同原因的分析。李锋锐(2022)[47]结合大数据理论和智能算法，研究提出了企业财务危机预警的方法和技术，构建了融合大数据指标的企业财务危机预警模型。实证检验表明，该预警体系可以有效提高基于财务和非财务数据的企业财务风险预警效果。Dechowetal(2011)[48]为确保数据的完整性和可靠性，建立了可以根据需求进行灵活扩展、添加数据的数据库系统。在财务数据中进行了各类公司的分析和比较，并深入探究实施了财务欺诈行为的公司与未实施财务欺诈行为的公司有什么区别。对市场相关变量等公开数据进行了多维讨论，并使用 Logistic 回归来检验效果。结果显示，最高的预测精度超过了 70%。我们采用的 M-score 模型中的指标延续了国外使用 logic 模型的思路，并可以作为补充指标。

袁先智等人(2022)[49]利用大数据框架对上市公司的财务欺诈风险进行特征刻画，并构建了财务欺诈风险模型。通过 MCMC 框架下的 Gibbs Sampling 方法进行随机抽样，成功提出了财务欺诈特征的有效提取方法，解决了在高维空间中进行数据分析和建模时存在的数据分布的不均匀性和过拟合等问题。最后将与财务欺诈行为相关度最高的因子提取出来，基于这些因子，我们能够更加清晰准确地看到财务欺诈风险。Huang D 等人(2018)[50]提出了一种名为 CoDetect 的全新财务欺诈检测框架，该框架能够同时利用网络信息和特征信息来检测财务欺诈，并能够检测与财务欺诈活动相关的特征模式。研究结果证明，CoDetect 对识别和打击财务欺诈是有效且高效的。向有涛等人(2022)[51]介绍了企业财务风险预测的重要性，并使用财务风险预测指标体系训练深度学习模型，构建了 MOPSO-CD-DNN 混合预测模型。实验证明该模型效果较好，并且适用于我国的企业，引入多目标优化算法可显著提升模型泛化能力。叶钦华等人(2022)[52]针对近年来中国的上市公司和在海外上市的中国公司发生了许多财务欺诈事件而引起学术领域和商业界的高度关注，提出了一个五维度的财务舞弊识别框架，包括财务税务、行业业务、公司治理、内部控制和数字特征维度。基于复式记账法等理论，每个维度与会计系统对企业交易和经济事件进行记录、分类、汇总、报告和分析的各个过程相对应。蔡志岳(2006)[53]在研究财务欺诈识别模型时，选取财务、公司治理等多个指标，研究被证监会因财务欺诈行为进行处罚的上市公司，采用混合 BP 神经网络和 Logistic 回归构建了一个预警模型，能够对信息披露进行识别和预警。结果显示，混合后的模型总体精度高，各类误判率较低，而预测的效果也优于单模型。需要注意的是，不能一味关注模型的精度和误判率，由于技术受限，难以进行深入探究和可视化操作，比起单模型，混合的模型结构复杂且繁琐。刘云菁等人(2022)[54]提出了一种包括但不限于深度学习、强化学习、迁移学习、集成学习等的能够对公司的财务欺诈行为进行预测的机器学习模型，将变量还在输入到模型之前的初始状态时输入财务比率指标和非财务指标，在处理非平衡样本问题时采取欠采样方法来平衡不同类别的样本数量，以确保机器学习算法的准确性和稳定性。使用 LightGBM 算法建立分类模型，这个模型在测试集上获得了最好的性能表现，有助于实时高效地识别舞弊并及时进行监管，提升资本市场治理效能，促进经济平稳运行。周卫华等人(2022)[55]本文探讨了如何运用大数据和机器学习方法分析和挖掘上市公司的财务和非财务数据，提出了一种名为 Xscore 的基于机器学习的财务舞弊预测模型。研究表明，相较于其他模型，该模型在我国上市公司财务舞弊预测方面表现更佳，能够显著提高预测准确性。

随着计算机技术趋于成熟，应用也越来越广泛，数字挖掘技术可以通过对海

量的财务数据进行分析和挖掘，寻找异常和不规则的模式，从而识别出潜在的欺诈行为和风险。如果识别方法传统和单一，可能会有较大的局限性，结果准确性也会降低，因此许多学者通过对比不同方法对财务欺诈甄别的准确率，选出精度更高的甄别模型。Al-Hashedi K G(2021)[56]修订了 2009 年至 2019 年的金融欺诈检测研究，对欺诈类型和数据挖掘技术进行了分类。通过审查 75 篇相关文章，发现 SVM 是最常用于金融欺诈检测的技术之一，而大部分数据挖掘技术都被用于银行和保险欺诈。该研究提供了学术界和行业的参考，重要数据挖掘技术信息也得以呈现，并列出了面临金融欺诈的国家列表。

Chen Y, Wu Z 等人(2022)[57]通过对 28 组财务报告原始数据的分析，比较了单分类机器学习算法和集成学习算法在中国上市公司欺诈检测方面的效果。结果表明，集成学习算法整体表现更好，其中堆叠学习算法效果最佳。这些结果为利用财务报告原始数据和集成学习算法进行快速欺诈检测提供了重要依据，并提出了基于叠加算法的识别模型，可为投资者、监管机构和管理层提供一种简单有效的识别方法，也可为其他欺诈场景的检测提供参考。Jarrod West(2016)[58]讨论了金融欺诈问题对各个领域的影响，新技术的应用对该问题产生了加剧的影响。传统的人工检测方法在大数据时代已经不再实用，金融机构转向使用自动化流程，并采用统计和计算方法进行欺诈检测。本文系统回顾了数据挖掘方法在金融欺诈检测领域的研究，尤其关注了基于计算智能技术的应用。Hajek 和 Henriques(2017)[59]研究发现，集成方法在欺诈检测中表现更佳，而贝叶斯网络在识别非欺诈性公司方面最优。此外，对公司年度财务报告进行负面词汇相对频率分析，可检测非欺诈性公司。建议使用更加具体的词汇和简洁的句子结构，以更清晰地传达研究结果。Xia H 等(2022)[60]研究发现，财务欺诈往往源于对商业利益的夸大，准确预测或发现财务欺诈对于企业管理和资本市场具有重要意义。虽然目前有多种计算机技术可供选择，但无监督学习算法是其中最为重要的一种。然而，大多数方法忽略了企业间的行为模式和同伴效应，这会限制其准确性和检测性能。为此，该研究提出了一种基于误差分布的财务欺诈检测方法，其表现良好。建议避免使用“指出”这类的描述，使用更加简练和明确的词汇和句子结构，同时确保论文中使用的术语和定义符合相关的会计和财务规定。

李林杰(2021)[61]的研究使用四种机器学习模型对我国制造业 A 股上市公司进行财务欺诈检测，AUC 评分也表现很好，实际操作中也能较为有效识别出实施了财务欺诈的公司。指标来源于基本财务数据的静态、个体和时间三个维度，并采用贝叶斯优化和网格搜索两种方法进行参数调优。最终，通过等权重 voting 法将调优后的模型进行融合。袁先智，周云鹏，严诚幸等(2022)[62]基于会计和财务、计

算机等领域的知识，采用吉布斯随机搜索算法，建立了一种针对公司财务欺诈风险的特征因子筛选方法。通过实证分析和数据样本验证测试，该方法可有效识别财务欺诈。孙玲莉、杨贵军等(2021)[63]将首位数字法则引入到到随机森林模型中，本研究构建了一系列数字中，以特定数字作为首位数字的概率分布，并提出了在首位数字法则基础上建立的的随机森林模型。该模型的预测结果与真实数据的相关性较高，说明模型能够捕捉到数据中的关键特征和变化趋势，随机森林模型也变得更加实用。华长生(2008)[64]在沪深证券市场选取的样本中，该研究从反映公司财务状况和能力的 26 个财务指标中初步筛选，通过相关性分析、T 检验和逐步判别分析等多元统计方法建立了可识别财务欺诈的模型，并进行了实证检验，结果良好。

张引弟(2021)[65]公司风险需要综合考虑多方面内容进行分析，财务欺诈不能通过单一指标反映，需要对财务报表进行分析检测异常指标或数据，以判断公司是否存在财务欺诈行为。范宇晨(2021)[66]本文探讨了使用 XGBoost 和 CatBoost 两种 boosting 方法对财务造假进行识别。通过 GridSearchCV 选择最优超参数，将其应用于两种方法上进行比较分析。结果表明，相比 XGBoost 方法，CatBoost 方法在准确率和性能方面更优，是一种更有效的财务造假识别方法。李天霞(2021)[67]研究利用 CiteSpace 软件分析 CNKI 核心期刊数据库中 185 篇关于“财务造假”的文献，评估相关研究的数量、作者、机构、热点和趋势等方面，以深入了解我国财务造假研究现状。研究结合多领域的知识，旨在提出更加精准、全面的研究观点和建议。

樊蕾，吴明远(2022)[68]分析了 185 篇“财务造假”为主题的文献显示：我国学者更倾向于独立研究财务造假，从组织管理、审计和实际案例等研究角度出发，研究人员正试图通过法律和监管手段来防范和打击企业财务造假行为，以维护市场的公平和透明。审计是对企业财务报表的全面检查和评估，能够有效发现和防范财务造假行为，保证财务报表的真实性和准确性。因此未来会更注重两者之间的关系以及影响。李爱华,王迪文(2022)[69]研究新开发了一种异常检测模型，此模型是一种综合多种数据指标和算法的融合方法，对于异常样本较为罕见或数量较少的情况，非平衡处理包括欠采样、过采样和结合采样等技术，可以通过改变样本的分布和权重，使得模型更加关注和学习异常样本的特征和规律，提高对异常样本的识别和分类准确率。经实验证明，模型各项评价指标优于未处理的结果，含有多种类型特征的检测结果更优。结论是非平衡处理有助于提升模型对异常样本的识别能力，将来自不同数据源、不同类型或格式的数据进行整合和合并，形成一个统一的数据集，这对财务造假的识别准确性很有帮助。

1.3.4 文献评述

对于财务欺诈的动因研究，相对而言国外的起步较早，我国开始研究的时间比较滞后，但是由于事件性质恶劣并且发生频率很高，研究劲头很足。基于国内外长期深入的研究，具体动因越来越细化和清楚，在不同的分类标准下划分了不同的动因，在不同的案例研究中能更有针对性地发现问题。每个国家处于不同的环境下，有不同的政策规定，从国情出发是最科学的，我国学者对此确实也做出了很大贡献，提出很多独特又能够适应我国发展的观点。在对财务欺诈手段的研究中，可以看出财务欺诈的手段越来越隐蔽和难以识别，这也进一步印证了对财务欺诈的识别系统和方法有越来越高的要求。在对财务欺诈的识别研究中，基于深度学习的研究较少，研究深度也不够，各种算法模型与财务欺诈识别研究的联系也不够紧密。

总的来说，我国对于财务欺诈识别系统案例研究时间较短，数量有限，研究仍处于进一步发展的阶段。本文通过对财务欺诈动因、手段和识别模型的文献进行研究，得到一些结论：在国内外财务欺诈及其识别的研究中，大部分学者都是基于传统的财务指标以及传统会计模型的构建，对财务欺诈的预测效果多数都较为滞后或是不够准确，也缺乏一定的创新性。尽管我国的证券市场相关制度和规范越来越完善，但财务欺诈事件依旧频发并且更难通过传统手段准确识别。

1.4 研究内容及方法

1.4.1 研究内容

第一部分，主要是针对论文的研究背景及意义展开，并通过回顾国内外财务欺诈识别的发展历程来引出研究思路。在进一步总结文献的基础上将梳理和整合后的内容进行概括性阐述，为本文后面的研究提供依据。

第二部分，阐述财务欺诈的相关概念以及公司财务欺诈的动因、手段、特征，介绍各种财务欺诈识别的模型，为下文模型的建立和分析ST康美财务欺诈做铺垫。

第三部分，基于 Bagging 和深度学习，选取数据集并对数据进行一系列探索、处理，对 DecisionTree、RandomForest、ExtraTrees、XGBoost、KNN、LogisticRegression、LightGBM、AdaBoost 八种算法进行分析以及评价，构建模型并对所选数据集进行训练集和测试集的划分，在构建了模型的基础上对模型参数调优并选取不同模型下影响财务欺诈的权重最大的前 31 个特征。在融合模型之后输出财务欺诈预测结果。

第四部分，在财务欺诈公司中选取 ST 康美作为验证模型的案例公司，针对 ST 康美进行案例分析和说明，介绍 ST 康美概况、股权结构、财务状况，并对 ST 康美财务欺诈案件采用传统与本文研究模型进行识别，对比二者差异，体现所研究模型的先进之处。

第五部分，有效识别财务欺诈行为的启示。介绍了影响因子的作用和选择、算法模型的特点和选择。

第六部分，研究结论和不足。先是对本文得出的研究结论进行了总结，接着指出本文研究由于技术和个人能力而存在的局限性。

1.4.2 研究方法

本文在一系列上市公司财务以及非财务的数据挖掘后选取国内医药制造业代表性企业 ST 康美作为分析对象，通过案例研究法、文献研究法、定量分析法对其各项指标进行了分析和考量，将财务欺诈识别问题与大数据手段进行交叉，并通过传统模型与 Bagging＋DCRN 模型的对比，体现出本文研究的识别模型的先进和科学之处。

1.4.2.1 案例研究法

本文在有效建立财务欺诈识别模型并进行分析后，选取了 ST 康美作为模型研

究和验证对象，并对传统识别方法和 Bagging 融合模型的效果进行对比分析并进行了总结。结合 ST 康美的概述和财务状况以及年报数据分析其财务欺诈行为，深入进行了 ST 康美的具体案例分析，以期让模型的准确性得到验证。

1.4.2.2 文献研究法

在查阅了大量相关资料后，对国内外学者的相关研究进行了分析，具体从财务欺诈识别系统的动因、手段以及识别模型来展开，了解到深度学习下的财务欺诈识别模型的发展以及研究现状，为本文后续的撰写打牢了基础。

1.4.2.3 模型构建法

本文对多种机器学习算法模型进行研究，对比各类算法模型的优劣后提出 Bagging 集成模型构建，并将构建的模型用于财务欺诈的研究。包括数据处理前的数据准备、特征处理、算法模型的评估与选择、模型及参数的调整和优化等步骤。并在这个动态的过程，不断地根据任务和数据的特点进行调整和改进。

1.5 研究思路及技术路线图

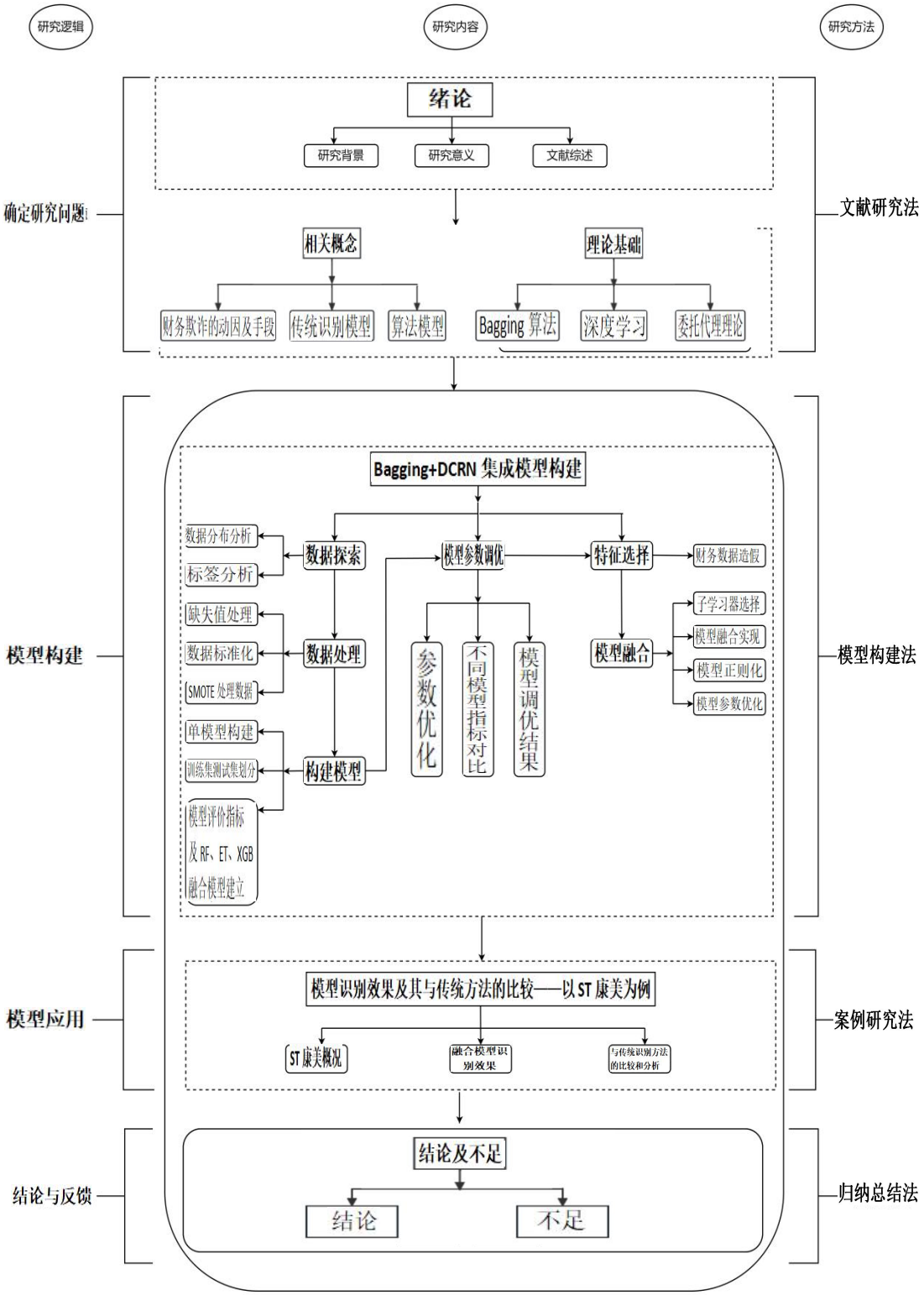
1.5.1 研究思路

首先，确定研究问题：基于当前环境下的研究背景，以国内外研究现状引出研究的方向并阐明研究意义，理清选题依据。

然后，进行模型模型构建：先爬取研究所需数据集，数据清洗后得到所需数据并对数据进行进一步的分析和处理。划分好训练集与测试集后开始构建模型，模型构建好后对模型参数进行调优，通过特征选择后进行最终的模型融合和调参并进行预测。用 SMOTE 采样数据训练 DT、RF、ET、XGBoost、KNN、LR、LGBM、AdaBoost 八种机器学习算法。选出最优的三种算法 RF、ET、XGB 去求数据集中公司财务数据造假有较大影响的特征。在 AUC 指标最优的情况下，将 RF，ET，XGB 这三种基于树模型的特征选择结果进行综合。计算出特征重要性权重值，分别选出不同训练模型中特征权重值排名前 31 的特征，作为上市公司财务数据造假有较大影响的特征因子。用深度学习模型代替了传统的树模型，以多层感知机，多层残差网络，Cross 网络作为子网络构建了(Deep-Cross Residual-NetWork，DCRN)网络模型。子网络完成特征的交叉组合，全连接层实现逻辑回归的二分类功能。其中多层残差网络通过短路操作解决梯度消失问题，Cross 网络通过运算来进一步增加特征之间的交互力度。此外，还引入了 Batch Normalize 层，起到了加速模型收敛，防止梯度消失与爆炸，缓解过拟合等作用。同时，舍弃了过采样的机制，依靠神经学习强大的学习能力对少量样本的特征进行捕捉。最后，在 DCRN 模型的基础上进行 Bagging 集成，提高模型的泛化能力，并缓解样本不均衡带来的影响。最终 Bagging+DCRN 集成模型在验证集上的 AUC 得分高于所有单独的树模型，可见验证集效果好，稳定性高。

最后，进行模型应用：在输出 flag 为“1”的结果中经内外部分析以及代表性分析后选择 ST 康美为研究目标，用传统分析方法以及其实施财务欺诈时间段会计师事务所对其出具的审计报告进行分析研究，与本文构建的模型进行对比分析，旨在突出本文研究方法的科学及优越之处。

1.5.2 技术路线图



第 2 章 相关概念与理论基础

2.1 公司财务欺诈的相关概述

2.1.1 公司财务欺诈的概念界定

制造虚假的经济业务，虚增资产或收益，减少负债或费用等，财务欺诈的手法多种多样。财务欺诈，是指在会计活动中，相关当事人为了追求个人利益，而采取违规、欺诈等手段来误导和隐瞒财务信息，从而达到避税等目的。以提前通过伪造凭证、假账真算、隐匿收入、账外设账、公款私汇等一系列准备和安排而有意地制造和提供虚假会计信息以欺骗财务报表使用者的行为。财务欺诈行为可能会涉及虚假记账、隐瞒重要信息、虚构交易等手段，导致企业财务状况失真、信誉受损、股价下跌、投资者利益受损等后果。美国注册会计师协会(AICPA)在SAS82《财务报表审计中对欺诈的考虑》把财务欺诈定义为“在财务报表中蓄意错报、漏报或泄露以欺骗财务报表使用者。”

2.1.2 公司财务欺诈的动因

2.1.2.1 GONE 理论

由于在市场交易中不同交易方之间拥有的信息水平不同，投资者会倾向于选择信任那些大股东，并且基于其大股东的身份和地位对公司的经营、盈利等各项状况进行更准确深入的了解。造假者为了降低成本会向大股东承诺更高的收益，从而诱发投资者的贪婪心理，造假者和投资者同样具有逐利动机，不管处于盈利还是亏损状态，具有贪婪心理的企业都会趋向于造假。造假因子分为外部和内部机会因子，企业在市场经济中扮演着重要角色，但其在经营过程中也要面对多重因素的挑战，如法律法规、政府监管、社会舆论等。内部因子与企业内部控制、股权结构等有关。上市能够提高企业信誉及形象，也会大大增加股东权益，因此大多企业都想达到上市条件，便产生了通过隐藏不利经营状况、虚构财务信息等来吸引更多的社会投资的需求。造假的暴露风险和暴露后受到的惩罚程度也是企业进行财务欺诈的动因，违法成本低、惩罚程度轻都造成了我国 A 股市场财务欺诈频发。

2.1.2.2 舞弊三角理论

舞弊三角理论认为：企业舞弊产生的原因是由压力、机会和借口三要素组成的。对于处于跨越式发展的企业来说，资金的需求压力比较大。被收购企业的亏损也会使得母公司的财务状况发生改变，从而导致净利润下降。对于行业整体发展迟缓，上市要求屡被驳回的企业，为了推动企业的发展，企业往往需要转向新风口进行投资及研发。在新产业的投资以及建厂等方面，需要大量的资金支持，而企业自有的资金有限，贷款过高也给企业的资金链带来了较大的压力。同时，企业退市的压力也不得不面对。根据《上海证券交易所退市公司重新上市实施办法（2020 年 12 月修订）》的规定，强制退市的上市公司即使满足了重新上市的条件，也只有达到对应要求的间隔期，才能向交易所申请重新上市。公司控股股东与实际控制人直接或者间接持有的公司股份，即使公司重新上市了，36 个月内不得转让、委托他人管理或者由公司回购。公司退市期间发行的新增股份，除已通过证券竞价交易等方式公开转让的股份之外，自重新上市之日起 12 个月内不得转让。一系列严格的规定使经营情况不理想的上市企业倍感退市的压力，若是触发强制退市流程，也会大大打击投资者们对企业的信心，由此引发的各种负面消息对企业也会产生不利影响。外部审计缺乏独立性等原因造就了企业财务欺诈的机会。在压力和机会面前，企业开始寻找借口来进行财务欺诈。

2.1.3 公司财务欺诈的手段

对于已经上市的企业来说，常用的财务欺诈手段有虚增资产和收入，包括伪造收入或在不满足收入确认准则的情况下确认收入、虚构收入等。虚假披露，推迟确认费用，虚减成本、费用，一些企业还会借法律法规中的漏洞来使用不当的会计政策或会计估计，利用关联交易来达成目的。

2.2 公司财务欺诈识别的传统模型

2.2.1 财务分析法

基于财务分析的财务欺诈识别方法主要是对企业过去一段时期的经营业绩及效益、盈利能力、偿债能力和发展潜力等的分析。而财务状况分析主要是对企业过去一定时期内的财务活动情况及其影响因素进行全面深入的考察，并通过一定程序确定未来一个时期内企业生产经营活动可能出现与会计信息有关的主要问题，以及这些问题在不同时期内的发展趋势，以揭示这些内容及变化趋势。财务报表

分析就是在对企业现有资产进行全面分析的基础上，以企业财务状况为依据或以货币资金、实物资产或劳务等资源为对象，对企业过去一定时期和未来一定时期内有关资产、负债、所有者权益等经济利益变动原因进行综合分析和判断。这种综合概括是对历史数据或资料进行科学分析后得出的。它是一种用现代科学方法对会计资料所提供信息进行加工处理后得到的综合反映，其目的不是为了揭示企业过去或未来经济利益变动情况下发生的信息，而是为了反映企业经营管理活动在一定时期内所取得成果和出现的问题。财务分析的方法一般都有不同的侧重点，有对各种比率指标进行计算和分析的方法，也有通过财务报表上各个项目的占比情况，来分析企业财务结构和变化趋势等。

2.2.2 现金流量分析法

追求现金流量最大化是企业的起点，在企业经济活动中，企业的现金流总是在资产负债表之后，并非不合理地流动，而是不断得到补充和发展，这就要求我们要加强现金流量管理。现实中，大多企业都会采用现金流量分析法来分析企业的利润质量。简单来说就是将企业在各项活动中产生的现金以及现金净流量分别与利润和投资收益进行比较分析，从而对企业的利润质量和净利润是否达到预定目标进行识别和判断。如通过对营业利润总额以及其他业务收入的分析，可以看出各项业务的占比，并分析企业的盈利主要是哪项业务所带来的。

2.2.3 应收款项和存货分析法

应收账款分析是对流动资产的分析，即对应收账款占主营业务收入的比重来分析企业应收账款带来收入的能力。在对应收账款周转率各期数进行比较后，来分析回款情况，并据此分析企业应收账款管理策略是否合理。还可以分析应收款客户的信用情况，各项规模和能力之间是否平衡等。

存货分析可以通过企业存货项目中的原材料、在产品、产成品等具体金额，分析所研究企业的市场占有率、产销量、对上下游企业控制力的强弱。主要是对存货周转率的分析，存货周转速度越快，反映存货的占用水平越低，流动性越强，同时其转换为现金等其他货币形式的速度就越快。当然，存货天数并不意味着越少就一定越好，存货也并非越多越好，应该与特定的经营条件下存货的最佳水平相比较，进而分析企业各项活动及其利润。

2.3 公司财务欺诈识别的算法模型

2.3.1 Boosting 算法

作为 Boosting 集成算法，XGBoost、LightGBM、Adaboost 三种常用算法有相似之处也有区别。

XGBoost 在监督学习时，是在建立的数据模型中输入相关特征完成的。为了能够构建成准确率较高的模型，此算法通过组合多个树模型并迭代生成新的树来完成。为了能够顺利预测出目标，往往通过定义目标函数来完成。

目标函数 L(x)由误差函数 F(x)和复杂度函数Ω(x)组成:

$$L(x) = F(x) + \Omega(x) \tag{2-1}$$

$$L(x) = \sum_i l\left(y_i, \$_{y_i}\right) + \gamma T + \frac{1}{2}\lambda\sum_{j=1}^{T}\|W_j\|^2 \tag{2-2}$$

其中 L 是在机器学习或优化算法中用于评估模型预测结果与实际结果之间差距的函数，其在定义域内可以求导，且在函数曲线上任意两点的连线所在区间内，函数值不大于这条连线上对应两点线性插值的函数，用集成学习方法，通过多次迭代不断增加基分类器（或基回归器）来构建一个更为复杂的集成模型。通过不断增加新的基分类器来提高模型的泛化性能，同时避免模型出现过拟合的情况。可得评价函数：

$$L_m(x) = \sum_{i=1} l\left[y_i, \$_{y_i}^{(m-1)} + f_m(x_i)\right] + \Omega(f_m) \tag{2-3}$$

这个算法通过贪婪地一步步优化目标函数来构建一个强大的集成分类器。为了更高效地优化目标函数，可以采用一些算法优化技巧，比如利用二阶泰勒展开：

$$f(x + \Delta x) = f(x) + f'(x)\Delta x + \frac{1}{2}f''(x)\Delta x^2 \tag{2-4}$$

令

$$g_i = \frac{\partial l\left(y_i, \$_y^{(t-1)}\right)}{\partial \$_y^{(t-1)}}, \hbar_i = \frac{\partial^2 l\left(y_i, \$_y^{(t-1)}\right)}{\partial\left(\$_y^{(t-1)}\right)^2} \tag{2-5}$$

最终的目标函数表达式在剔除常数项后如下所示：

$$L_m(x) = \sum_{i=1}^{n}\left[g_i + f_m(x_i) + \frac{1}{2}\hbar_i f_m^2(x_i)\right] + \Omega(f_m) \tag{2-6}$$

LightGBM 和 XGBoost 算法都是通过多轮迭代，不断拟合残差，每轮迭代都会拟合一颗新的决策树，而这颗新的决策树的构建方式就是利用损失函数对模型预测值的导数的相反数作为当前决策树的残差近似值，进而构建新的决策树来拟

合这个残差值。通过这样的方式，每轮迭代都可以不断地优化模型，提高模型的准确性，最终得到一个更加准确的集成模型。LightGBM 算法比传统的机器学习算法训练时间更短，训练过程更快速高效，在支持并行化学习、低内存使用的特点下，在大规模数据处理中较为可靠。

LightGBM 算法采用了基于叶子节点的生长策略。其中，leaf-wise 生长策略是 LightGBM 算法中一种常用的生长策略，它的特点是在每一轮迭代中，只选择当前最重要的那个节点进行生长，从而能够更快地找到最优解。不同于其他大多数决策树算法所采用的 level-wise 生长策略，此策略是将每一层的所有叶子节点一起考虑，在这些叶子节点中选择一个最优的节点进行分裂。而 LightGBM 算法采用的是 leaf-wise 生长策略，则是在当前的所有叶子节点中选择一个具有最大分裂增益的节点进行分裂，相对于其他算法，该算法可以提供更高的精度，但是如果使用的样本量较小，则该算法容易过度适应训练数据，从而在新的数据上表现不佳，即产生过拟合现象。

2.3.2 树模型

在对一系列数据进行分类的过程中，决策树算法是最常用的。决策树在计算机科学中已经有了很多应用，包括数据挖掘、机器学习和系统建模。它通常用来评估各种算法，包括回归模型、分类模型和预测模型。在很多情况下，决策树能提供比其他方法更好的预测结果。通过这种方法，可以根据给定的条件确定所需要的值。决策树算法优点是直接从数据集得到结论，缺点是容易产生过拟合，所以一般在分类任务中用决策树，比如使用 Bagging 算法。

决策树的构建过程就是找到最能区分样本的特征，将其作为树的节点，并以该特征的取值作为分支进行分类。即在决策树的构建过程中，通过分析样本中的各个特征，找到那个对分类有关键作用的特征，将其作为根节点。接着，对于每个根节点的子节点（即对应该特征取值的分支），将子数据集按照该特征的取值进行划分，然后针对每个子数据集，重复上述步骤，即找到子数据集中对分类起决定性作用次大的特征，并以此构建子树。递归地进行上述步骤，直到子数据集中所有数据都属于同一类别为止。这时，该叶子节点就代表了该类别。数据有不同的特征，决策树来基于这些差异进行分类就是其构造的本质。在分类时，决策树有三个根本，一是信息，二是熵，三是信息增益。

在进行特征选择时，主要的目的就是在决策树算法中，需要选择合适的分裂标准来确定当前节点的划分方式。由于决策树的特点包括容易过拟合，因此预剪枝或后剪枝两种剪枝技术通常都被需要。决策树学习算法的主要思想是将输入空

间划分成一些矩形区域，并在每个矩形区域内拟合出一个离散值，以此来逼近目标函数。在机器学习中，生成的决策树虽然具有很好的分类和预测能力，但由于其结构过于复杂，很难为人类理解和解释。因此可以将其转换为一组 If-Then 规则，使得决策树的输出可以被更好地解释。随机森林算法就是以决策树为基础，运用集成学习的思想，将分类能力最强的决策树识别出来后与其他分类能力强的树进行组合，得到一个性能良好的分类器。在分类问题中，分类器就是不同的决策树，在输入样本不变时，不同的树会有不同的分类结果。而随机森林是一种基于决策树集成的机器学习算法，它采用 Bagging 思想进行集成。作为一种大数据算法，随机森林将其得出的分类结果再汇总，最终得出不易过拟合的分类结果。相对其他模型来说，随机森林模型准确率较高，数据的部分缺失也不太影响模型判断结果。

Extra-Trees（Extremely randomized trees，极端随机树）算法与随机森林算法都是基于决策树的集成学习方法。在 Extra Trees 使用的全部样本中，特征是随机选取的，对决策树实现分叉的分叉值是在完全随机的情况下得到的。例如，以类别作为二叉树特征属性时，左右分支的选择是随机的，以数值作为其特征属性时，数字的选择是随机的，预测误差也是由于最佳分叉属性的随机选择产生的。因此 Extra Trees 算法的随机性较强。正是因为其随机性，与单一决策树相比，多颗决策树的组合可以通过集成学习的方式来实现，以提高预测效果。

2.3.3 回归模型

Logistic 回归分析是数据挖掘中常用的线性回归分析模型，可以用于经济预测领域。当回归分析中的被解释变量属于二分类或多分类型变量时通常采用 Logistic 回归分析来替代回归模型进行研究。在分析中，自变量是一组实数值，这些连续值的自变量通常会被分成几个离散的区间，形成一些划分点，以便于进行决策树的分裂。同时自变量也可以是分类的，通过对数据进行建模，Logistic 回归可以确定每个自变量对因变量的影响程度，以及这些自变量的权重。这些权重可以解释为每个自变量对预测因变量的贡献程度，即每个自变量的重要性。在财务欺诈识别中用 Logistic 回归分析模型构建时，通常是通过财务指标作为主要变量进行预测，虽然使用财务指标作为主要变量来预测企业的财务危机是可行的，但这种方法的局限性在于财务指标只能反映企业在某一特定时间点或一段时间内的财务状况。财务危机是一个复杂的概念，包括很多方面的因素，而财务指标只能提供有限的信息。因此，仅依靠财务指标可能无法完全反映企业的财务危机状况，还需要结合其他因素进行综合分析和判断。

KNN（K-Nearest Neighbor）法是一种分类算法，也就是 K 最邻近法。这种分

类算法的目标就是在特征空间中，要是最邻近的大多数或者说 K 个最相似可以看做是同一个类别，那么这个类别包括这个样本。KNN 分类算法在实现过程中，进行数据分析之前需要先收集、整理和清理数据并对数据进行预处理，接着要对测试样本点到其他每个样本点的距离进行计算。计算完成后，按距离来排序，再选择出距离最小的 K 个点。比较了各个点的类别后，寻找 K 个点中占比最高的一类，并将测试样本点归入那一类。

2.4 相关理论基础

2.4.1 Bagging 算法

Bagging 算法（Bootstrap aggregating），也就是装袋算法，在机器学习中，是一种利用多个学习器的组合来改善学习性能的一类算法。Bagging 是一种集成技术，具有一定的代表性。Bagging 算法在与其他算法结合时，能够在降低结果的方差时解决过拟合的问题，也能让结果更加稳定和准确。Bagging 算法通过构建实例来改善学习性能，这些实例是基于一个基学习器训练的。进行预测时，通常会生成多个估计器（estimator）并使用它们来得出最终的预测结果。将多个估计器的输出进行结合形成最终的预测结果。

Bagging 算法是一种集成学习方法，旨在通过构建多个弱学习器并将它们组合起来来提高学习算法的准确性。它可以用于分类或回归问题。具体来说，Bagging 算法通过从训练集中随机抽取一定数量的样本，构建多个训练集，并分别用这些训练集训练出多个弱学习器。然后将这些弱学习器组合起来，得到一个强学习器，用于预测新的样本。Bagging 算法的关键在于如何组合弱学习器。通常采用投票的方式，即对于分类问题，强学习器输出最多的类别作为预测结果；对于回归问题，强学习器输出各个弱学习器输出的平均值作为预测结果。其优点在于可以有效地减少模型的方差，从而提高模型的准确性。它还能够降低模型的过拟合风险，增加模型的泛化能力。

2.4.2 深度学习

深度学习(DL,Deep Learning)是机器学习(ML,Machine Learning)深度学习是一种机器学习的分支，其目标是通过多层神经网络等深度结构来建立模型，实现对数据的自动分析和学习，从而实现各种智能任务。与传统机器学习算法相比，深度学习可以自动发现更加抽象、高级的特征，具有更强的表征能力，能够处理更

加复杂的数据和任务。深度学习的引入使得机器学习更加接近于最初的目标——人工智能，因为深度学习算法可以通过自我学习和自我优化来不断提高模型的性能和准确度，从而实现更加智能化的任务和应用。深度学习已经在多个领域展现出强大的能力，这些应用场景中，深度学习模型可以自动地从大量的数据中学习并提取出特征，从而实现更加准确和高效的任务处理深度学习是学习样本数据的内在规律和表示层次。通过学习，可以增强对各种不同形式数据的理解能力。这项技术旨在使机器能够拥有人类的感知和认知能力。深度学习在处理语音和图像等复杂数据时具有更高的准确性和效率，这使得它成为当前最为流行的机器学习算法之一。

2.4.3 委托代理理论

由于所有权和经营权分离，委托代理理论应运而生。其认为：股东与管理者之间出现了委托关系，即股东将自己的资产委托给管理者去经营，委托代理关系中存在着信息不对称。由于存在的信息差异，又因为管理者对公司有一定的控制权，他们有可能会利用这种权力来获取更多的个人利益，而不顾及股东的利益，公司存在着代理问题。高管权力较大且较容易获取自身利益最大化。为了自己的利益，高管可能会做出损害公司整体利益的事情。从公司治理结构来看，若高管与股东有共同的目标——追求企业价值最大化，那么其在公司治理结构中拥有权力对企业价值最大化是有利的。

第 3 章 基于 Bagging 和深度学习的财务欺诈识别模型构建

3.1 数据探索

3.1.1 数据分布分析

对上证、深证 A 股财报数据进行初步的探索，并了解数据的架构和内容。数据包含了 5125 家上市企业的数据。按证监会 2012 版行业分类进行数据分组，制造业 3336 家，农林牧渔业 53 家，采矿业 87 家，电力、热力、燃气及水生产和供应业 134 家，建筑业 114 家，批发和零售业 206 家，交通运输、仓储和邮政业 119 家，住宿和餐饮业 10 家，信息传输、软件和信息技术服务业 426 家，金融业 131 家，房地产业 123 家，租赁和商务服务业 68 家，科学研究和技术服务业 104 家，水利、环境和公共设施管理业 95 家，居民服务、修理和其他服务业 1 家，教育业 12 家，卫生和社会工作 17 家，文化、体育和娱乐业 65 家，综合 24 家。以此标准划分，每家企业均有具体对应的行业，数据分布也更为科学。

3.1.2 标签分析

如果按照申银万国行业分类2021修订版进行数据分组，除了未划分行业的148家企业，其余企业来自 31 个不同行业，在最初的财务报表分析中，一共有 260 个特征，统计了正负标签分布情况后，可知与财务欺诈的公司相比，未发生财务欺诈的公司数量明显更多，因此正负样本分布极其不均衡。如果按照证监会 2012 版行业分类进行数据分组，制造业有 3336 家，占比较大且所有数据集中的企业均可以分到具体的行业中，数据分布也更加均衡。所以采用证监会 2012 版行业分类进行研究。在行业分类中对于数据缺失情况，将数据缺失率超过 30%的列直接删除，留下的特征再用于进一步的数据降维。具体操作如下：

```
# 删掉缺失率大于 30%的列
asset_miss_rate = asset.isnull().sum()/len(asset)
asset_miss_rate = asset_miss_rate[asset_miss_rate > 0.3]
profit_miss_rate = profit.isnull().sum()/len(profit)
profit_miss_rate = profit_miss_rate[profit_miss_rate>0.3]
asset = asset.drop(asset_miss_rate.index,axis=1)
profit = profit.drop(profit_miss_rate.index,axis=1)
```

```
asset = asset.drop(['证券代码','报表类型','是否发生差错更正'],axis=1)
profit = profit.drop(['证券代码','报表类型','是否发生差错更正'],axis=1)
```

以上操作过程先是将资产数据集中每个属性缺失值的数量除以该数据集的总行数，计算出每个属性的缺失比例，创建了包含所有属性的缺失比例的变量。筛选出缺失比例大于 30%的属性，保留这些属性的缺失比例数据并更新变量，删除了在变量中保存的资产数据集、利润数据集等缺失比例大于 30%的属性。

3.2 数据处理

3.2.1 缺失值的处理

因为在后面的问题求解过程中需要利用现有模型进行预测，采用机器学习相关的算法模型是不能让数据中存在缺失值的，而原始数据存在大量的缺失值，所以需要采用数据填充的方式来填补缺失值。

数值型特征采用插值填补法，遍历所有的特征，利用现有的数据进行多项式拟合，将存在缺失值的样本的其他有值的特征代入拟合多项式中算出拟合值作为填充数。

类别型特征采用众数填补法，找出每一个数值特征的众数，然后再将众数作为填充值填入缺失处，若存在一个特征有多个众数的情况则随机选取一个众数作为填充数。具体过程如下：

```
# 缺失值填充
imputer = KNNImputer(n_neighbors=4)#邻居样本求平均数
data.iloc[:,2:-1] = imputer.fit_transform(data.iloc[:,2:-1])
print(data.isnull().sum())
data.to_excel(r'C:\Users\Lenovo\Desktop\data.xlsx')
data = pd.read_excel(r'C:\Users\Lenovo\Desktop\data.xlsx')
```

以上操作在 Python 的常用库中主要用了 Pandas 和 Scikit-learn，创建了一个 KNNImputer 的实例，用于填充缺失值。KNNImputer 是一个来自于 Scikit-learn 库的类，它使用 KNN 算法来估计缺失值。n_neighbors 参数指定了用于估计缺失值的邻居的个数。接着使用 KNNImputer 实例来填充 data 数据集中从第三列到倒数第二列之间的缺失值。接着用 fit_transform()方法将缺失值替换为 KNN 算法的预测值。最后使用代码检查填充后的数据集中是否还有缺失值。输出结果中所有列的值都是 0，说明所有的缺失值都已经被填充了。

3.2.2 数据标准化

Z-Score 标准化是数据处理的一种常用方法。通过它能够将不同量级的数据转化为统一量度的 Z-Score 分值进行比较。因为上市公司的财务数据的范围很大，不同属性的量纲差别较大，因此需要使用标准化数据消除不同量纲带来的影响，让带入模型训练的数据更规整。本文用直接标准化方法将上市公司财务数据中所有非描述型属性缩放处理成均值为 0，方差为 1。

具体过程如下：

```
# 标准化
transfer = StandardScaler()
data.iloc[:,2:-1] = transfer.fit_transform(data.iloc[:,2:-1])
```

上述操作涉及到数据预处理中的标准化（Standardization）操作，其中 StandardScaler()是一个用于对数据进行标准化的类。具体来说，data.iloc[:,2:-1]表示对 data 数据中从第 2 列到倒数第 2 列的所有数据进行标准化操作。这个范围可以根据实际需求进行调整。fit_transform()方法是 StandardScaler()类的一个方法，用于计算并应用标准化操作。fit_transform()方法接收一个数据集作为输入，并返回一个新的经过标准化处理的数据集。在这个例子中，fit_transform()方法被应用于 data.iloc[:,2:-1]数据集，并将标准化后的数据集替换掉原来的数据。

总体来说，这段代码的作用是将 data 数据集中的一部分数据进行标准化处理，可以使得数据在不同的尺度和单位下具有可比性，消除数据中的偏差和噪声，使得数据更加稳定和可靠。这种处理可以帮助分析人员更好地理解数据的特征和模式，并且可以减少模型的误差和不确定性，提高模型的准确性和预测能力。

3.2.3 SMOTE 处理数据

财务欺诈属于风险检测问题，由于数据集中的数据分布是很不平衡的，即数据中的上市公司财务欺诈的数量远小于不造假的数量。定义财务欺诈为小类样本，不财务欺诈为大类样本。采用 AUC 为模型的评价指标，可以通过重采样减小类别不平衡带来的影响。重采样策略可进一步分为降低大类样本的“欠采样”以及增加小类样本的“过采样”。随机重复小类的有过拟合风险，但删除大类样本可能丢失有用信息。因为小类样本数量少，采用欠采样无法保证深度学习网络得到有效训练，最终可能影响网络的收敛速度，故采样过采样方法。

为了提升过采样的效果，使用 SMOTE 算法，用人工生成样本替代复制样本以降低过拟合风险。该算法使用在小类样本和它的同类近邻值间随机插值的方法产

生人工样本。依次将 SMOTE 采样的数据放到不同的算法模型中，AUC 分数最高的模型是 XGBoost，其次是 RandomForest，排名最后两位的算法是 LightGBM、AdaBoost，他们的效果并不好，因此在后续进行特征选择的过程中就不再使用这两种算法。

```
    # smote 过采样
sm = SMOTE(random_state=42, n_jobs=-1)
x, y = sm.fit_resample(x, y)
print(len(y))
print(y.sum())
```

首先用代码创建了一个 SMOTE 类对象，其中 random_state 和 n_jobs 参数分别指定了随机数种子和并行处理的数量。sm.fit_resample(x, y)方法是 SMOTE 类的一个方法，用于将输入数据集 x 和 y 进行过采样处理。其中 x 表示样本特征，y 表示样本标签，fit_resample()方法返回过采样后的特征和标签数据。接下来，通过 print(len(y))和 print(y.sum())分别输出了过采样后样本数量和正样本数量。其中，len(y) 表示样本数量，y.sum()表示正样本数量。

总体来说，这段代码的作用是使用 SMOTE 方法对输入数据进行过采样处理，从而解决数据不平衡问题。

3.3 构建模型

3.3.1 单模型构建

```
    # 建立模型
dt = DecisionTreeClassifier()
rf = RandomForestClassifier()
et = ExtraTreesClassifier()
xgb = XGBClassifier()
knn = KNeighborsClassifier()
lr = LogisticRegression()
lgb = lightgbm.LGBMClassifier()
ada = AdaBoostClassifier()
kf = KFold(n_splits = 5, shuffle=True, random_state=0)
model_dict                                                    =
```

```
{'DecisionTree':dt,'RandomForest':rf,'ExtraTrees':et,'XGBoost':xgb,'KNN':knn,'Logistic
Regression':lr,'LightGBM':lgb,'AdaBoost':ada}
for model_name in model_dict.keys():
    precision = 0
    recall = 0
    f1 = 0
    auc = 0
    for train_index, val_index in kf.split(data):
        estimator = model_dict[model_name]
        estimator.fit(x.loc[train_index],y.loc[train_index])
        y_pre = estimator.predict(x.loc[val_index])
        precision = precision + precision_score(y.loc[val_index],y_pre)
        recall = recall + recall_score(y.loc[val_index],y_pre)
        f1 = f1 + f1_score(y.loc[val_index],y_pre)
        auc = auc + roc_auc_score(y.loc[val_index],y_pre)
    print('{} Precision:'.format(model_name), precision/5)
    print('{} Recall:'.format(model_name), recall/5)
    print('{} F1-Score:'.format(model_name), f1/5)
    print('{} AUC:'.format(model_name), auc/5)
```

首先，初始化了几个分类模型，如 RandomForestClassifier、ExtraTreesClassifier、DecisionTreeClassifier、XGBClassifier 等，使用默认的超参数。然后，为了方便循环遍历所有模型并进行评估，创建了一个模型名称和对应初始化模型的字典。接下来，对数据进行交叉验证来评估模型的性能。最后，对于模型字典中的每个模型，循环通过交叉验证的训练集训练模型并在验证集上进行评估。性能指标是使用验证集的真实标签和模型预测标签计算的。计算了 5 个交叉验证折叠的所有性能指标的平均值，并为每个模型打印了它们。打印的性能指标是精确率、召回率、F1 分数和 AUC。

总的来说，这段代码使用交叉验证和多种性能指标评估多个分类模型，并输出每个模型的平均性能。

AUC（Area Under the ROC Curve）是一种二分类模型评估指标，它的作用是评估模型对样本的排序能力。AUC 评分的取值范围在 0.5 到 1 之间，其中 0.5 表示模型预测的准确率等同于随机预测的准确率，1 表示模型的预测完全正确。使用 AUC 评分可以比较不同模型的性能，越高的 AUC 分数意味着模型在分类问题中

具有更好的表现，具有更好的分类能力。AUC 评分也可以帮助选择最优的阈值，以在准确率和召回率之间获得最佳平衡点，从而使模型在不同的业务场景中获得最佳性能。

最后的结果如下表所示：

表 3-1 算法模型 AUC 评分

模型/评分	Precision	Recall	F1-Score	AUC
Decision Tree	66.3%	82.3%	78.4%	81.3%
Random Forest	78.2%	89.1%	87.3%	87.7%
Extra Trees	79.9%	90.1%	84.5%	86%
XGBoost	77.8%	88.2%	88.0%	88.0%
KNN	62.3%	76.1%	74.2%	75%
Logistic Regression	57.4%	75.7%	71.3%	69.5%
LightGBM	63.4%	71.6%	62.5%	62.7%
AdaBoost	56.8%	73.2%	61%	64.8%

从上表中可以看出，LightGBM 和 AdaBoost 算法模型的 AUC 指标评分远远低于其他算法模型，因此在此后续建模中就直接将这两种算法模型舍弃。

3.3.2 训练集、测试集划分

按照数据集中每一家企业所属行业进行划分，制造业 3336 家，其他行业 1789 家，将制造业有标签的观测行设为 2730 家，无标签的预测行设为 606 家；其他行业有标签的观测行设为 1464 家，无标签的预测行设为 325 行，划分好训练集和测试集，再进行 k 折交叉验证，让划分的效果更好。具体过程如下：

```
    # 制造业有标签的观测行
manu_labeled = np.arange(2730)
    # 制造业无标签的预测行
manu_unlabeled = np.arange(2730, 3336)
    # 其他行业有标签的观测行
other_labeled = np.arange(1464) + 3336
    # 其他行业无标签的预测行
other_unlabeled = np.arange(1464, 1789) + 3336
    # 将有标签的观测行和标签拼接在一起
labeled_indices = np.concatenate([manu_labeled, other_labeled])
```

```
labels = np.concatenate([np.zeros_like(manu_labeled), np.ones_like(other_labeled)])
    # 将数据集划分为训练集和测试集
train_indices, test_indices, train_labels, test_labels = train_test_split(
    labeled_indices, labels, test_size=0.2, random_state=42, stratify=labels)
    # 将制造业有标签的观测行和其他行业有标签的观测行分别划分为 k 折
k = 5
manu_labeled_folds = StratifiedKFold(n_splits=k, shuffle=True,
random_state=42).split(
    manu_labeled, np.zeros_like(manu_labeled))
other_labeled_folds = StratifiedKFold(n_splits=k, shuffle=True,
random_state=42).split(
    other_labeled, np.ones_like(other_labeled))
    # 将所有的折拼接在一起
folds = []
for i in range(k):
    manu_train_indices, manu_val_indices = next(manu_labeled_folds)
    other_train_indices, other_val_indices = next(other_labeled_folds)
    train_indices = np.concatenate([manu_train_indices, other_train_indices])
    val_indices = np.concatenate([manu_val_indices, other_val_indices])
    folds.append((train_indices, val_indices))
    # 最后将制造业无标签的预测行和其他行业无标签的预测行拼接在一起
unlabeled_indices = np.concatenate([manu_unlabeled, other_unlabeled])
```

上述代码先将所有的观测行和标签拼接在一起，并将其按一定比例划分为训练集和测试集。再将制造业有标签的观测行和其他行业有标签的观测行分别划分为 k 折，把所有的折拼接在一起。

3.4 模型参数调优及结果

3.4.1 参数优化

模型参数优化的目的是让模型输出的结果和实际观测数据之间的误差最小化，从而使得模型能够更好地拟合实际数据。为了达到这个目的，我们需要使用一些优化算法来调整模型的参数，使得模型在预测数据时具有更高的准确性和可靠性。优化算法的核心思想是通过不断地调整模型参数，使得目标函数的值不断减小，

从而达到优化模型的目的。这个过程需要不断地迭代和调整，直到模型的输出和实际观测数据之间的误差达到最小值为止。在模型调参中常用的方法有网格调参、贝叶斯调参。网格调参通过循环遍历，尝试每一种参数组合，返回最好的得分值的参数组合。网格搜索是一种通过枚举所有可能的参数组合来寻找最佳超参数的方法。具体来说，我们将需要调整的超参数放在一个网格中，每个超参数的可能取值形成一个维度，这样就可以组合出所有可能的参数组合。本文将使用网格搜素的方式进行参数调优。具体过程为：

```
    # 参数调优
models = [dt,rf,xgb,knn,lr]
dt_param =
{'max_depth':range(5,21),'min_samples_leaf':range(5,11),'min_samples_split':range(2,9)
}
rf_param =
{'n_estimators':np.arange(500,901,10),'max_depth':range(6,21),'min_samples_leaf':np.a
range(10,101,10),'min_samples_split':[2,3]}
xgb_param =
{'n_estimators':np.arange(700,901,10),'min_child_weight':range(1,6),'max_depth':range
(4,11),'gamma':np.arange(0.1,1,0.1),'subsample':np.arange(0.5,1,0.1),'Colsample
bytree':np.arange(0.5,1,0.1),'Reg_alpha':range(0,4),'Reg_lambda':np.arange(0.01,2,0.1),'
learning_rate':np.arange(0.001,0.2,0.001)}
knn_param = {'n neighbors':[2,3],'weights':['uniform','distance']}
lr_param = {'penalty':['l1','l2'],'c':[1,4],'max_iter':np.arange(10,800,10)}
params = [dt_param,rf_param,xgb_param,knn_param,lr_param]
for i in range(len(models)):
    grid = GridSearchCV(models[i],params[i],scoring='roc_auc',cv=kf)
    grid.fit(x,y)
    print(grid.best_params_)
    print(grid.best_score_)
```

在前文舍弃掉 LightGBM 和 AdaBoost 算法模型后，开始对剩下五种算法模型进行调参，用代码实现了针对五种不同机器学习模型的超参数调优，使用了 GridSearchCV。需要调优的模型包括决策树（dt）、随机森林（rf）、XGBoost（xgb）、K-最近邻（knn）和逻辑回归（lr）。为每个模型定义了相应的参数字典，分别是 dt_param、rf_param、xgb_param、knn_param 和 lr_param。使用 sklearn 中的

GridSearchCV 函数来穷举搜索所有可能的超参数组合，并找到每个模型的最佳超参数集。GridSearchCV 函数在 for 循环中调用，循环迭代每个模型，并使用定义的超参数使用训练数据（x，y）拟合模型。

3.4.2 参数优化结果

在参数优化后，打印出每个模型的最佳超参数和相应的最佳得分（由'roc_auc'指标衡量）。如下表所示：

表 3-2 参数调优后 AUC 评分结果

模型名称	AUC 初始值	调参后的 AUC 值
Decision Tree	81.3%	79.4%
Random Forest	87.7%	88.1%
Extra Trees	86.0%	87.3%
XGBoost	88.0%	84.8%
KNN	75.0%	74.4%
Logistic Regression	69.5%	66.9%

从评价指标 AUC 的结果来看，Random Forest、Extra Trees、XGBoost 模型的训练效果较好。而 Decision Tree、KNN、Logistic Regression 的模型训练效果不佳，故在模型融合时舍弃此三个算法。

3.5 模型融合

许多经典的传统机器学习模型比如逻辑回归具有一定局限性：表达能力不强，无法进行充分的特征交叉，往往需要经过费力繁琐的特征工程，比如特征选择，特征组合等。而且许多传统模型无法完成特征的充分交叉，因此不可避免地造成信息的损失。除了表达能力不强，会不可避免地造成有效信息损失之外，在仅仅利用单一特征而非交叉特征进行判断的情况下，有时不仅是信息损失的问题，甚至会得出错误的结论，比如著名的“辛普森悖论”。为了解决此类问题，利用多层神经网络，用深度学习模型代替传统机器学习模型。最后还将对 DCRN 网络进行 Bagging 集成，提高模型泛化能力。

3.5.1 子学习器选择

多层感知机是一种神经网络模型，它能够对输入特征向量中的每一个维度进行交叉组合，从而捕捉到更多的非线性特征和特征之间的组合信息。相比于传统的机器学习模型，多层感知机的表达能力更强，可以更好地适应复杂的数据分布和任务。简单来说，多层感知机能够更好地发现数据中的规律和模式，从而提高了模型的性能。先通过多层感知机完成特征的交叉与组合，然后通过 Batch Normalization 层进行数据批正则化，最后送往输出层（逻辑回归模型）完成二分类。

Residual 部分：多层感知机具备强大的非线性学习能力，但是当网络加深之后，往往存在过拟合现象，使得网络越深，在测试集上表现越差。此外，当网络加深之后，往往存在严重的梯度消失现象。即指在梯度方向传播过程中，越靠近输入端，梯度的幅度越小，参数收敛的越慢。

为了解决这两个问题，多层残差网络(Multi-Layer Residual Network)被提出。残差神经网络就是由残差单元(Residual Unit)组成的神经网络。

Cross 部分：增加特征之间的交互力度，使用多层交叉层(Cross Layer)对输入向量进行特征交叉。由多层交叉层组成的 Cross 网络进行特征的自动化交叉，在传统的机器学习方法中，需要手动提取和选择特征，这需要专业知识和经验，并且很容易受到主观因素的影响。而机器学习算法可以自动地从原始数据中学习特征，并找到最有效的特征组合来解决实际问题，避免了人工选择特征的主观性和局限性，从而可以更好地解决问题。

Batch Normalize 层：通过对系统参数搜索空间进行一定的限制或约束，可以增加系统的鲁棒性，使其更能够适应不同的环境和应对各种情况。这些约束可以是在参数范围、变量关系、约束条件等方面进行设置，以减少无效的搜索和排除不合理的方案，从而提高系统的性能和效率。通过对每一层的输入数据进行归一化处理，将其均值和方差进行标准化，使得数据分布更加稳定，减少了多层之间协调更新的问题，从而加速收敛，保证梯度，缓解过拟合等问题。

综上，用深度网络、多层残差网络、Cross 网络这三个子网络与 Batch Normalize 层设计 DCRN 网络。先通过三个子网络充分挖掘特征的相互关联，对特征进行自动化组合，然后三个子网络的输出进行拼接(Concatenate)，形成新的包括特征不同表示形式的特征向量。然后，经过 Batch Normalization 后送往全连接输出层采用逻辑回归（Logistic Regression）进行二分类。

3.5.2 模型融合实现

用 Bagging 集成学习来降低模型方差：假设有一个数据集，其中包含了所有的企业数据。为了进行训练或者评估模型，需要从这个数据集中随机抽取一部分数据来作为采样集，以此来代表整个数据集。首先从数据集中随机抽取一个样本，然后将该样本放入采样集中。与一般的抽样方法不同的是，在有放回抽样中，将该样本再放回到数据集中，使得下次采样时该样本仍有可能被选中。这样可以保证采样集中的样本数量与数据集中的样本数量相同，且可以对同一样本进行多次采样，增加了样本的多样性和代表性。有放回抽样是一种简单而有效的抽样方法，可以避免数据集中某些样本被过度选中或忽略，从而保证了采样集的代表性和可靠性。在实际应用中，有放回抽样通常与交叉验证、自助法等技术结合使用，可以有效提高模型的泛化能力和鲁棒性。这样，经过多次随机采样操作，得到多个采样集，基于各个采样集训练基学习器，再将这些基学习器进行结合。Bagging 通常对分类任务使用简单投票法，对回归任务使用简单平均法。若分类预测时出现两个类收到同样票数的情形，则最简单的做法是随机选择一个，也可进一步考察学习器投票的置信度来确定最优者。

Bagging 算法描述为输入：训练集 D={（x1,y1）,（x2,y2）,……,（xm,ym）};基学习算法ᔓ ;训练轮数 T;1:fort=1,2,……,Tdo2:ht=ᔓ (D,Dba)3:endfor 输出：

假定基学习器的计算复杂度为 O(m), 则 Bagging 的复杂度大致为 T(O(m)+O(s)),Bagging 可以用于多分类、回归等。通过对原始训练样本集进行 T 次 Bootstrap 采样得到 T 个与原始训练样本集大小相同的不同采样集，然后根据这 T 个不同的样本集训练 T 个不同的 DCRN 基分类器，最终通过投票表决法综合多个基分类器得到模型最终的预测输出。

```
    # 融合模型
x_train,x_val,y_train,y_val = train_test_split(x,y,test_size=0.2,random_state=99)
predict = []
for model in [rf,et,xgb]:
    model.fit(x_train,y_train)
    predict.append(list(model.predict(x_val)))
y_pre = np.array(predict).sum(axis=0)
y_pre = np.int64(y_pre>2)
print('Precision:', precision_score(y_val,y_pre))
print('Recall:', recall_score(y_val,y_pre))
print('F1-Score:', f1_score(y_val,y_pre))
```

print('AUC:', roc_auc_score(y_val,y_pre))

使用三种分类器（rf, et, xgb）来预测验证集的标签，并将这些分类器的预测结果进行了加权投票（通过计算每个分类器预测结果的和来实现）。然后使用这个加权投票结果进行评估，得到融合模型的 AUC 指标评分。

具体思路为：首先使用 train_test_split 函数将数据集划分为训练集和测试集，划分后得到训练集特征 x_train 和标签 y_train，测试集特征 x_val 和标签 y_val。其次使用循环遍历三个分类器（rf, et, xgb），并分别在训练集上拟合每个分类器，使用拟合好的分类器在验证集上进行预测，并将每个分类器的预测结果保存在一个列表中。再次将三个分类器的预测结果进行加权投票，将每个预测结果的值相加，得到总和 y_pre，最后使用 precision_score、recall_score、f1_score 和 roc_auc_score 函数计算加权投票结果的各指标评分值。

3.5.3 模型正则化减少过拟合

用 Dropout 提供的方法来减少过拟合：Dropout 能够训练和评估指数级数量的神经网络，其集成包括所有从基础网络中出去非输出单元后形成的网络。在 Bagging 的情况下，所有模型都是独立的。在 Dropout 的情况下，所有模型共享参数，其中每个模型继承父神经网络参数的不同子集。参数共享是一种在神经网络等深度学习模型中常用的技术，其基本思想是对于同一个模型的不同部分，使用相同的参数进行共享。这样一来，可以大大减少模型所需的参数数量，从而降低模型的存储和计算复杂度。例如，对于卷积神经网络（CNN）中的卷积层，可以共享卷积核（kernel）的权重，从而使得模型可以表示指数级数量的特征，同时在有限可用的内存下进行高效存储和计算。在 Dropout 的情况下，通常大部分模型都没有显式地被训练，因为通常父神经网络会很大，以至于在合理的时间内几乎不可能采样完所有的网络。取而代之的是，在单个步骤中训练一小部分的子网络，参数共享会使得剩余的子网络也能有好的参数设定。Dropout 与 Bagging 算法一样，每个子网络中遇到的训练集是又放回采样的原始训练集的一个子集。

因为训练的 Bagging+DCRN 模型表现能力强大甚至会过拟合，训练误差会随着时间的推移逐渐降低但验证集的误差会再次上升，所以引入提前终止(Early Stopping)。提前终止企图返回使验证集误差最低的参数设置，以获得验证集误差更低的模型。在每次验证集误差有所改善后，存储模型参数的副本。当训练算法终止时，返回这些参数。

3.5.4 融合模型评价

建立好 Bagging+DCRN 集成学习模型后，将数据代入进行测试，得到结果：Precision 为 96.1%，Recall 为 84.6%，F1-Score 为 89.9%，AUC 为 90.6%。可见，单个机器学习算法中表现最好的 XGBoost 模型评估指标值明显低于融合模型。即融合模型比所有单个算法模型的表现都要好。

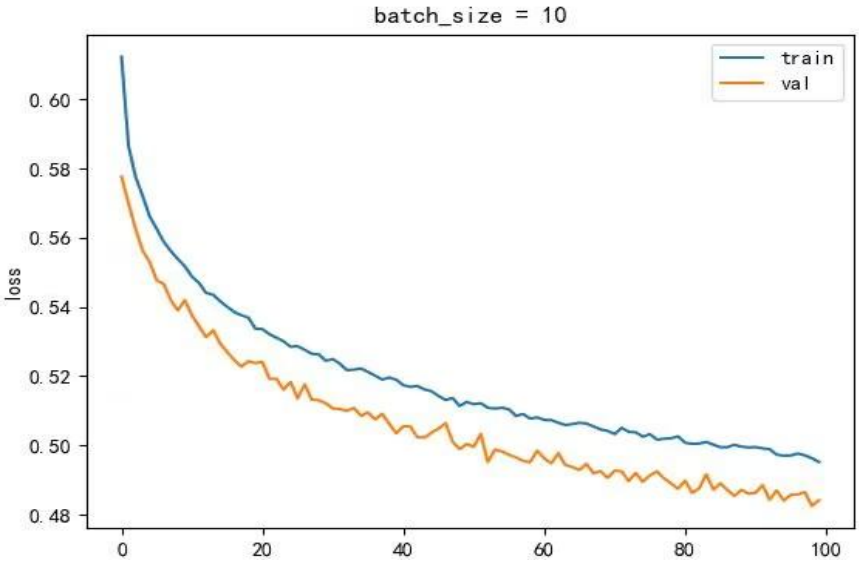


图 3-1 融合模型损失随训练轮次变化图

从图 3-1 融合模型损失随训练轮次变化图中可以看出在足够的训练轮次后，模型最终达到收敛状态，接近最优值。

第 4 章 模型识别效果及其与传统方法的比较——以 ST 康美为例

4.1 ST 康美概况

4.1.1 公司简介

康美药业股份有限公司(600518)是一家民营企业，成立于 1997 年，2001 年 3 月 19 日在上海证券交易所主板上市，曾经的 A 股上市医药名企之一，证券简称曾为“康美药业”，现称“ST 康美”。业务贯穿中药产业链的上、中、下游产业，渗透产业链各个环节。包括药材种植和交易，生产开发以及终端销售。

4.1.2 ST 康美股权结构

康美投资控股实业有限公司在2016、2017、2018三年分别持有ST康美31.35%、32.76%、32.91%的股份。同时，马兴田夫妇以 100%持股比例完全控制康美投资控股有限公司，因此马兴田夫妇就是其实际控制人。从表 4-1 和表 4-2 中可以看出，在此三年间，马兴田夫妇的合计持股比例分别为 35.64%、37.38%、37.31%，平均合计持股比例为 36.78%，在前十大股东所持的平均股份中，占 67.31%，且其他的前十大股东持股比例远远低于马兴田夫妇的持股比例。可见康美的股权架构较为集中。

表 4-1 马兴田夫妇持股比例

姓名/年份	2018 年	2017 年	2016 年	担任职务
马兴田	35.38%	35.45%	33.73%	董事长、总经理
许冬瑾	1.93%	1.93%	1.91%	副董事长、副总经理
合计	37.31%	7.38%	35.64%	

数据来源：东方财富网

表 4-2 康美药业前十大股东持股比例

股东姓名/年份	2018 年	2017 年	2016 年
康美实业投资控股有限公司	32.91%	32.76%	31.35%
五矿国际信托有限公司	4.57%	4.57%	4.53%
深圳市前海重明万方股权投资有限公司	3.29%	3.97%	3.31%
中国证券金融	2.99%	3.29%	1.92%
常州燕泽永惠投资中心	2.86%	3.31%	1.49%
天津市鲲鹏融创企业管理咨询有限公司	1.97%	1.97%	1.98%
许冬瑾	1.93%	1.93%	1.91%
普宁市金信典当行	1.83%	1.82%	1.79%
普宁市国际信息服务咨询有限公司	1.87%	1.87%	1.88%
陈树雄	1.59%	1.68%	1.41%
合计	54.81%	57.54%	51.57%

数据来源：东方财富网

4.1.3 ST 康美财务状况

表 4-3 ST 康美 2015-2021 年偿债能力分析

项目/年份	2015	2016	2017	2018	2019	2020	2021
流动比率	2.12	2.22	2.03	1.93	1.69	0.80	1.67
速动比率	1.42	1.59	0.66	0.74	0.56	0.55	1.03
现金比率	1.14	1.36	0.16	0.06	0.02	0.02	0.49

数据来源：国泰安数据库

看表 4-3 可知，ST 康美在 2017 年及以前流动比率都维持在 2.0 以上，2017 年调整了会计差错之后，从 2018 年开始流动比率低于 2.0；速动比率也从 1.5 左右下降到 1 以下；现金比率从大于 1 下降到 0.5 以下。由这些数据可以看出，企业偿债能力减弱，资本风险增加。

表 4-4 ST 康美 2015-2021 年营运能力分析

项目/年份	2015	2016	2017	2018	2019	2020	2021
应收账款周转率(%)	6.755	6.961	4.098	2.965	2.301	1.771	2.175
存货周转率(次)	1.509	1.354	0.451	0.336	0.304	0.282	0.783
总资产周转率(次)	0.548	0.466	0.293	0.247	0.166	0.111	0.168

数据来源：新浪财经 2015-2021 年报数据计算整理

从表 4-4 营运能力指标对比中可以看出，ST 康美从 2017 年调整会计差错之后应收账款周转率就直线下降，即其应收账款转为现金的平均次数变少，说明企业款项收回变慢或者企业坏账增加。存货周转率下降，说明企业的库存商品可能出现滞销，企业资金周转也可能出现困难。总资产周转率下降，说明企业销售能力变弱，资产投资的效益不好，资产利用效率变低。

表 4-5 ST 康美 2015-2021 年盈利能力分析

项目/年份	2015	2016	2017	2019	2020	2021
净资产收益率(加权)(%)	18.54	8.07	4.81	-21.2	-416.16	-120.39
总资产收益率(加权)(%)	8.35	7.18	3.57	-6.76	-63.52	32.08
毛利率（%）	28.34	29.9	38.63	13.2	3.43	16.04
净利率（%）	15.26	15.42	12.19	-40.67	-574.57	190.73
营业总收入(亿元)	180.7	216.4	175.8	114.5	54.12	41.53
归属净利润(亿元)	27.57	33.40	21.50	-46.61	-310.8	79.18

数据来源：新浪财经 2015-2021 年报数据计算整理

从表 4-5 盈利能力指标分析中可以看出，ST 康美加权净资产收益率、总资产收益率整体呈逐年下降的趋势，直到 2021 年才有所回升，说明企业经营状态变差，资本获取收益的能力变弱，投资带来的收益变低，企业的毛利率和净利率也在 2020 年及之前呈波动性下降，从 2017 年开始到 2020 年营业总收入、归属净利润也呈

下降趋势，说明企业运营效益越来越差。

表 4-6 ST 康美 2015-2021 年发展能力分析

项目/年份	2015	2016	2017	2018	2019	2020	2021
营业总收入同比增长(%)	13.28	19.79	-18.78	-2.92	-32.93	-52.72	-23.27
归属净利润同比增长(%)	20.60	21.17	-35.64	-82.58	-1344.53	-556.96	125.47

数据来源：新浪财经 2015-2021 年报数据计算整理

从表 4-6 营运能力指标分析中可以看出，ST 康美从 2017 年开始营业总收入同比增长、归属净利润同比增长都呈下降趋势，说明企业发展能力降低。

4.2 Bagging+DCRN 模型识别效果

4.2.1 数据处理

```
    # 标准化
transfer = StandardScaler()
data.iloc[:,3:-1] = transfer.fit_transform(data.iloc[:,3:-1])
    # 划分数据集
x = data.loc[(data['是否造假']==0)|(data['是否造假']==1),'货币资金':'其他收益']
y = data['是否造假'].dropna()
test = data.loc[data['证券代码']==600518,:]
x_test = data.loc[data['是否造假']!=data['是否造假'],'货币资金':'其他收益']
    # smote 过采样
sm = SMOTE(random_state=42, n_jobs=-1)
x, y = sm.fit_resample(x, y)
```

上述操作是对数据进行标准化。目的是使每个特征的均值为 0，标准差为 1。再将数据集划分为训练集和测试集，在 SMOTE 过采样和重采样后，完成了针对公司是否存在欺诈行为的机器学习分类模型的数据预处理和准备工作。

4.2.2 构建并融合模型

```
    # 建立模型
rf = RandomForestClassifier()
et = ExtraTreesClassifier()
xgb = XGBClassifier()
    # 融合模型
x_train,x_val,y_train,y_val = train_test_split(x,y,test_size=0.2,random_state=99)
predict = []
for model in [rf,et,xgb]:
    model.fit(x,y)
    predict.append(list(model.predict(data.iloc[:,3:-1])))
y_pre = np.array(predict).sum(axis=0)
y_pre = np.int64(y_pre>2)
data['预测是否造假'] = y_pre
data.to_excel(r'C:\Users\Lenovo\Desktop\result.xlsx',index=False)
```

最终，输出的 flag 标签如下：

表 4-7 2015-2019 年 ST 康美 flag 输出表

标签/年份	2015	2016	2017	2018	2019
Flag	0	1	1	1	1

从证监会的处罚来看，康美医药在 2016 年到 2019 年收到了中国证监会、广东证监局、上海证券交易所的处罚，有因虚构利润，虚假记载(误导性陈述)，重大遗漏，占用公司资产和其他行为。违反了《证券法》《上市公司信息披露管理办法》《关于规范上市公司与关联方资金往来及上市公司对外担保若干问题的通知》等多项相关规定。经检验，模型准确，且其效果与实际情况相符。

4.3 与传统方法识别方法的比较和分析

4.3.1 数据处理区别

传统的财务欺诈识别是完全基于企业的年报数据来进行分析和判断，在数据的选择上比较被动，且在处理指标筛选时需要对庞大的数据逐一对比和分析，处理数据效率低且容易对指标重要性分类错误，进而影响判断效果。

本文研究的基于深度学习的财务欺诈识别模型是通过大数据处理手段，将企业的所有报表数据整合在一起，用评价指标判断模型效果后直接进行所有数据权重的排序，避免了人工识别财务信息所导致的重要性指标识别的误差，保持了客观性。且在研究时引入处罚报告等非财务信息，使识别更为科学准确。

4.3.2 识别思路区别

传统的财务欺诈识别思路是根据企业的财务数据，通过具体的数据分析，发现并识别异常数据，对异常信息再进行进一步的拆解和分析，查找异常信息产生的原因，识别是否为财务欺诈行为。

本文研究的基于深度学习的财务欺诈识别思路是综合数据集中多年的所有财务信息，结合证监会处罚等非财务信息，筛选出最优模型所提供的对于财务欺诈行为来说影响大的一些特征数据，划分训练集和测试集并进行检验，最后从 flag 标签中清晰直观地看出各个企业在各个年份中是否有财务欺诈行为。

4.3.3 存在的风险区别

传统的财务欺诈识别主观性较强，可能会因个人对财务数据的理解产生认知偏差，从而导致识别结果不准确；人为进行数据分析也很容易发生错误或遗漏，影响识别结果；由于财务数据的复杂性，传统的人工识别效率低，可能存在识别结果滞后而带来新的欺诈行为的风险。

本文研究的财务欺诈识别模型建立在机器学习的基础上，克服了主观性误差，但由于各个企业所处行业的特征不同，在模型中数据的表现结果会有差异。并且有的企业为满足现实生产中的需要会发生一些非日常行为，本质上是没有进行财务欺诈，但由于数据异常也会被识别出来，因此也有误差风险。

4.3.4 识别效果区别

传统的财务欺诈识别效果依赖于审计人员的专业能力，同时也有可能受到外界压力等影响而导致偏差。而本文研究的基于深度学习的财务欺诈识别过程和结果都较为客观，用数据说话，不受外界干扰，结果较为可靠。

第 5 章 有效识别财务欺诈行为的启示

5.1 影响因子的作用及选择

5.1.1 影响因子的作用

影响因子让模型建立更有理有据，也让模型识别更准确。影响因子的选择越科学，模型建立才越有效，结果也越可靠。在财务欺诈的识别模型建立过程中，影响因子是财务欺诈行为识别的依据，能够根据特征值的变化和数据集里的大量数据找到特征规律，准确有效地识别财务欺诈行为。

5.1.2 影响因子的选择

影响因子对模型的表现至关重要，而由于行业特征的区别，在选择影响因子时如果一味追求所有类型的企业都完全适用，会存在一定的误差性。因此在选择影响因子时，要充分考虑特征因子在不同行业中的影响效果，可以通过在数据集中按行业分类来进行影响因子的选择。通过科学的方法，根据特征因子的特征权重值，判断出各个影响因子对于财务欺诈行为的影响作用大小，并据影响因子的权重高低来排序筛选，再进行进一步的财务欺诈行为识别。

具体过程为：

```
# 特征选择
feature = pd.DataFrame()
feature['feature_name'] = x.columns
total_importances = rf.feature_importances_ + et.feature_importances_ + xgb.feature_importances_
average_importances = total_importances / 3
for i in range(len(x.columns)):
    feature.loc[i,'feature_weight'] = total_importances[i] / 3;feature = feature.sort_values(by='feature_weight',ascending=False).reset_index(drop=True)
print(feature.head())
plt.figure()
plt.barh(feature.loc[:30,'feature_name'],feature.loc[:30,'feature_weight'])
plt.xlabel('特征权重值')
```

```
plt.ylabel('特征因子')
plt.show()
```

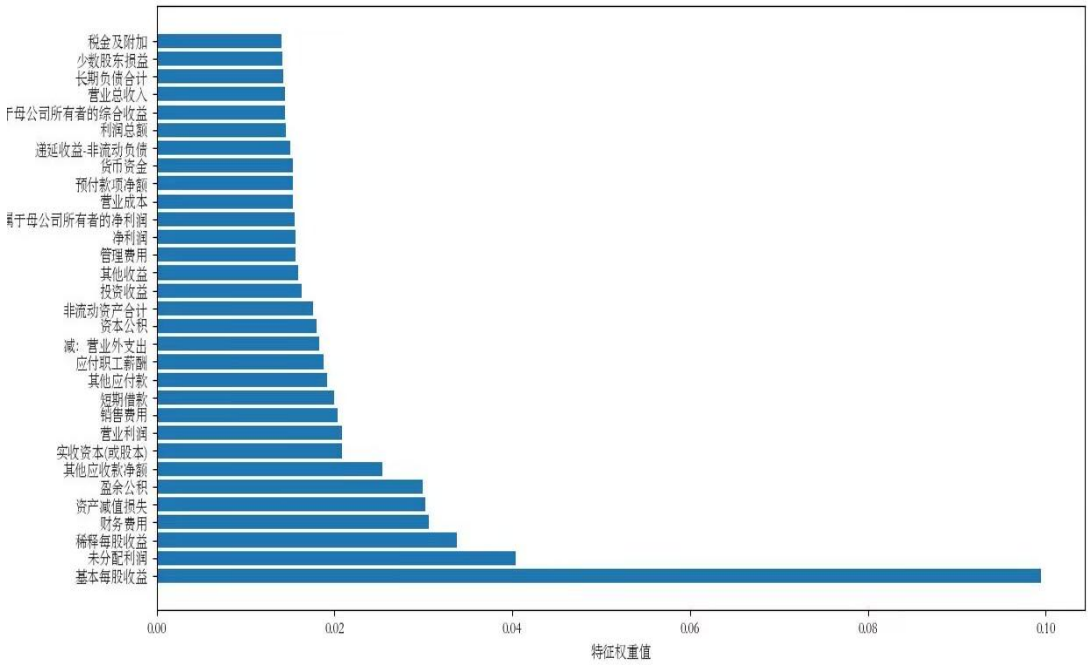


图 5-1 特征因子的选择

上述操作是特征因子选取的过程，最后生成一个机器学习模型中各个特征的重要性排名和可视化图表。具体操作是：首先，创建一个 feature 用来存储每个特征的名称和权重；然后，通过将三个算法的特征权重相加并取平均值，计算每个特征的平均权重值。接着，对于每个特征，通过除以三个算法的总特征权重的平均值，计算其相对权重值。最后，将特征按照权重值从大到小排序，并使用水平条形图可视化前 31 个特征的相对权重值，如图 5-1 所示。

5.2 算法模型的特点及选择

5.2.1 算法模型的特点

在本文的财务欺诈识别模型建立过程中，由于各个算法的区别，表现出不同的特点。决策树容易过拟合，一般来说需要剪枝，用预剪枝和后剪枝两种剪枝技术来缩小树结构规模、缓解过拟合。随机森林集成了所有的分类投票结果，由成百上千棵树组成，将投票次数最多的类别指定为最终的输出，是集成思想的体现。极端随机树使用所有的样本，特征是随机选取的，在某种程度上比随机森林得到

的结果更好。XGBoost 是将成百上千个树模型组合起来成为一个准确率很高的模型，此模型通过不断迭代生成新的树。KNN 是一种分类算法，其在定类决策上只依据最邻近的一个或者几个样本的类别来决定待分样本所属的类别。LogisticRegression 是一个二分类模型，可推广至多分类。LigthGBM 算法使用 leaf-wise 生长策略，每次在当前叶子节点中找出分裂增益最大的叶子节点进行分裂，而不是所有节点都进行分裂，这样可以提高精度，但在样本量较小时容易造成过拟合。AdaBoost 算法核心思想就是调整错误样本的权重，进而迭代升级。

5.2.2 算法模型的选择

在经典的机器学习算法中模型评估指标很多，本文用 AUC 的值来对六种机器学习算法模型进行评估。先通过对 SMOTE 采样后的六种算法模型进行第一次 AUC 评估，将表现较差的两种算法舍弃。再通过对每个单方法模型进行调参后的 AUC 评分来确定最终排名，筛选出对模型效果最好的几种方法后，将其建成一个优于所有单个机器算法的融合模型。

第 6 章 研究结论与不足

6.1 研究结论

本文研究财务欺诈识别模型，选取医药行业的代表性企业 ST 康美作为案例模型验证对象。首先，对财务欺诈的概念、动因、手段和特征进行整理和分析；然后，对几种有代表性的传统财务欺诈识别方法进行分析；接着，整理并介绍了几种算法模型并选取适当方法进行算法评估和择优，融合为新的模型并划分训练集和测试集进行训练；最后，以 ST 康美为例，带入模型进行验证分析。综上所述，本文研究结论主要有以下几方面：

首先，传统的财务欺诈识别方法有局限性。通过对财务分析法、现金流量分析法、应收款项与存货分析法的介绍，可以看出传统的财务欺诈识别方法由于过分依赖审计人员的专业分析能力和判断能力，同时受到主观性和外界压力的影响，识别结果可能存在偏差。以后文的 ST 康美为例，在传统识别方法中不难发现，在财务分析时虽然有异常信号，但由于医药行业的特点以及生物资产的特殊性，可能会提供便利条件，使得财务欺诈行为更容易发生。同时，这些行业特点也可能会为审计人员带来识别困难，因为这些行业特点可能会导致财务数据复杂多样，而且难以分析和比对，从而增加了发现欺诈行为的难度。因此，审计人员需要更加专业化和细致的审计方法来应对这些行业特点带来的挑战，因而很容易忽视潜在的舞弊风险。

然后，特征因子的选取有一定的方法。在财务欺诈识别模型建立过程中，特征因子的选取直接影响到识别结果的准确性。由于各个行业的各项资产、成本、研发投入等特征有所区别，表现在财务报表中的数据特征也就各不相同，如果单凭经验和主观判断来决定特征因子的选择是不科学的，因此在选取特征因子时应当基于行业特征对各项特征因子进行权重分析后再选择。本文提出先进行行业分类，再对特征因子进行权重计算，在此过程中按权重高低进行排序后再进一步根据需要做筛选。只有按照科学的特征因子选取方法，才能保证财务欺诈识别模型的效果。

最后，Bagging+DCRN 融合模型有一定的优越性。通过对各个机器学习算法的介绍和评分以及分析，可以看出单个算法模型对于财务欺诈识别的研究，由于各方法的思路不同，存在不同的识别效果。接下来，对融合模型的评估也验证了此观点，融合模型的评分高于所有的单个机器学习算法。因此，对于单个学习算法来说，融合模型有一定的优越性。同时，与传统识别模型对比可以看出，融合

模型克服了传统模型存在的主观性，并且可以通过对研究对象进行行业细分来进行更加准确的预测和判断，充分考虑到了各个行业的不同特征对财务信息产生的影响。因此，对于传统的财务欺诈识别模型来说，融合模型也有一定的优越性。

6.2 研究不足

本文在研究过程中，首先，因资料受限，企业的各项非财务信息难以完全获取，较偏重于财务数据支撑，对于识别模型中数据挖掘有一定的制约；其次，由于本人理论研究水平的限制，还不能对识别模型的效果进行广泛性的评价。综上，本文仍存在诸多不足之处。

参考文献

[1]Albrecht,W.S.,Wernz,G.W.&Williams,T.L.(1995).Fraud:Bring the Light to the Dark Side of Business[J].NewYork IrwinInc,22(3):15-52.

[2]Bologna J, Lindquist R J, Wells J T. The accountant's handbook of fraud and commercial crime[M]. New York, NY: Wiley, 1993.

[3]虞宛静. 基于风险因子理论的财务舞弊研究[J]. 上海商业, 2022.

[4]宋子豪.上市公司财务欺诈成因分析及防范[J].商,2013(13):124.

[5]张敦力, 王沁文. “包庇” 抑或 “蒙蔽”——由上市公司财务欺诈反观独立董事问责之困[J]. 财会月刊, 2022.

[6]Dan Yang,Hao Jiao,Roger Buckland. The determinants of financial fraud in Chinese firms: Does corporate governance as an institutional innovation matter?[J]. Technological Forecasting & Social Change,2017,125.

[7]刘为岩.瑞幸咖啡财务造假事件若干问题研究[J].商业会计,2021,No.698(02):39-42.

[8]郑贤龙.浅析 IPO 财务造假动因、手段及防范对策[J].商业会计,2013(17):92-93.

[9]屈文洲,蔡志岳.我国上市公司信息披露违规的动因实证研究[J].中国工业经济,2007(04):96-103.

[10]汪昌云,孙艳梅.代理冲突、公司治理和上市公司财务欺诈的研究[J].管理世界,2010(07):130-143+188.

[11]刘惠萍.我国创业板 IPO 保荐模式差异及其监管部门角色因应[J].改革,2011(05):115-119.

[12]李琳.上市公司“吹牛上税”了吗——来自财务舞弊与税收激进关系的证据[J].山西财经大学学报,2022,44(04):84-98.

[13]韩成.浅析 IPO 过程中会计造假的动因及防范措施[J].财务与会计,2013(05):52.

[14]牛羿恒.上市公司财务舞弊动因与治理研究——基于舞弊风险因子理论与博弈论共同视角[J].财会通讯,2020(12):99-103.

[15]徐凡卿,章之旺.上市公司财务舞弊多元动因研究[J].商业会计,2021(20):38-41.

[16]杨振宇.我国上市公司财务舞弊动因及治理对策研究[J].时代金融,2020(21):116-117.

[17]赵昱.从欣泰电气看企业财务造假[J].科技经济市场,2021.

[18]刘艺璐.上市公司财务欺诈识别与防范对策研究[J].全国流通经济,2021.

[19]罗韵轩,陈卷逸.基于舞弊三角理论的企业财务造假分析及思考——以康得新为例[J].商业会计,2021,No.717(21):70-72.

[20]Beasley M S, Carcello J V, Hermanson D R, et al. Fraudulent financial reporting: Consideration of industry traits and corporate governance mechanisms[J]. Accounting horizons, 2000, 14(4): 441-454.

[21]黄月菡,陈庆杰.上市公司财务舞弊手段及其审计研究[J].科技和产业,2021,21(10):197-202.

[22]刘启亮,邓瑶,陈惠霞,李洋洋,俞浩岚.上市公司财务舞弊的演变：1990～2022——基于典型个案的研究[J].财会月刊,2023,44(01):122-130.

[23]黄世忠,叶钦华,徐珊等.2010～2019年中国上市公司财务舞弊分析[J].财会月刊,2020,No.882(14):153-160.

[24]王晶.财务报表舞弊之手段分析[J].中国商贸,2012(8):83-84.

[25]吴晓迪.财务造假的手段剖析及防范措施[J].现代商业,2011(21):251-251.

[26]徐丽萍.上市公司财务报表舞弊手段分析[J].中外企业家,2013(27):159-160.

[27]钱玉,徐立文.中国上市公司财务报告舞弊手段及后果研究——基于新准则的实施[J].商场现代化,2014(26):179-182.

[28]Reurink A. Financial fraud: a literature review[J]. Journal of Economic Surveys, 2018, 32(5): 1292-1325.

[29]何欣惟. 财务舞弊的主要手段变化研究[D].云南财经大学,2015.

[30]曾汝林.浅析企业财务欺诈的主要手段及其防范[J].中国商论,2011(27):93-94.

[31]王淑玲.财务舞弊的手段与治理方法分析[J].现代商业,2012(10):242-243.

[32]郑贤龙.浅析IPO财务造假动因、手段及防范对策[J].商业会计,2013(17):92-93.

[33]刘天敏.上市公司财务舞弊手段有注册会计师的审计策略[D].2015.

[34]张彤.中国上市公司财务报告舞弊手段、成因及对策研究[J].财会学习,2019(12).

[35]陈澜. 创业板上市公司财务舞弊的透析与治理——基于金亚科技财务舞弊事件的分析[J]. 商业经济, 2019(03):170-173.

[36]曹越,吕亦梅,伍中信.其他综合收益的价值相关性及其原因——来自中国资本市场的经验证据[J][J]财贸研究,2015,26(06):132-141.

[37]王言,周绍妮,石凯.国有企业并购风险预警及其影响因素研究——基于数据挖掘和XGBoost算法的分析[J].大连理工大学学报(社会科学版),2021,42(03):46-57.

[38]钱苹,罗玫.中国上市公司财务造假预测模型[J].会计研究,2015(07):18-25+96.

[39]Abbasi A, Albrecht C, Vance A, et al. Metafraud: a meta-learning framework for detecting financial fraud[J]. Mis Quarterly, 2012: 1293-1327.

[40]杨贵军,杜飞,贾晓磊.基于首末位质量因子的BP神经网络财务风险预警模型[J].统计与决策,2022,38(03):166-171.

[41]阮素梅,杜旭东,李伟,陈旭.数据要素、中文信息与智能财务风险识别[J].经济问题,2022(01):107-113.

[42]Ali A, Abd Razak S, Othman S H, et al. Financial Fraud Detection Based on Machine Learning: A Systematic Literature Review[J]. Applied Sciences, 2022, 12(19): 9637.

[43]王昱,杨珊珊.考虑多维效率的上市公司财务困境预警研究[J].中国管理科学,2021,29(02):32-41.

[44]Craja P, Kim A, Lessmann S. Deep learning for detecting financial statement fraud[J].

Decision Support Systems, 2020, 139: 113421.

[45]Kamel,Mohamad EM,Sallehetal.(2016).Detecting financial statement fraud by Malaysian public listed companies[J].Pengurusan,46:23-32.

[46]杨子晖,张平淼,林师涵.系统性风险与企业财务危机预警——基于前沿机器学习的新视角[J].金融研究,2022(08):152-170.

[47]李锋锐.大数据时代企业财务风险预警系统构建与应用——评厦门大学出版社《大数据与企业财务危机预警》[J].价格理论与实践,2022(03):209.

[48]Dechow P M, Ge W, Larson C R, et al. Predicting material accounting misstatements[J]. Contemporary accounting research, 2011, 28(1): 17-82.

[49]Yuan G X, Zhou Y, Liu H, et al. The framework of CAFE credit risk assessment for financial markets in China[J]. Procedia Computer Science, 2022, 202: 33-46.

[50]Huang D, Mu D, Yang L, et al. CoDetect: Financial fraud detection with anomaly feature detection[J]. IEEE Access, 2018, 6: 19161-19174.

[51]向有涛,王明,曹琳.基于多目标深度学习模型的财务风险预测方法[J].统计与决策,2022,38(10):184-188.

[52]叶钦华,叶凡,黄世忠.财务舞弊识别框架构建——基于会计信息系统论及大数据视角[J].会计研究,2022(03):3-16.

[53]蔡志岳,吴世农.基于公司治理的信息披露舞弊预警研究[J].管理科学,2006,19(4).

[54]刘云菁,伍彬,张敏.上市公司财务舞弊识别模型设计及其应用研究——基于新兴机器学习算法[J].数量经济技术经济研究,2022,39(07):152-175.

[55]周卫华,翟晓风,谭皓威.基于 XGBoost 的上市公司财务舞弊预测模型研究[J].数量经济技术经济研究,2022,39(07):176-196.

[56]Al-Hashedi K G, Magalingam P. Financial fraud detection applying data mining techniques: A comprehensive review from 2009 to 2019[J]. Computer Science Review, 2021, 40: 100402.

[57]Chen Y, Wu Z. Financial Fraud Detection of Listed Companies in China: A Machine Learning Approach[J]. Sustainability, 2022, 15(1): 105.

[58]West J, Bhattacharya M. Intelligent financial fraud detection: a comprehensive review[J]. Computers & security, 2016, 57: 47-66.

[59]Hajek P, Henriques R. Mining corporate annual reports for intelligent detection of financial statement fraud – A comparative study of machine learning methods[J]. Knowledge-Based Systems, 2017, 128: 139-152.

[60]Xia H, Ma H, Cheng P. PE – EDD: An efficient peer – effect – based financial fraud detection approach in publicly traded China firms[J]. CAAI Transactions on Intelligence Technology, 2022, 7(3): 469-480.

[61]李林杰.机器学习在财务欺诈风险预警问题中的应用研究——基于制造业上市公司财务数据的实证分析[J].今日财富,2021.

[62]袁先智,周云鹏,严诚幸,等.财务欺诈风险特征筛选框架的建立和应用[J].中国管理科学,2022,30(3):43-54.

[63]孙玲莉,杨贵军,王禹童.基于 Benford 律的随机森林模型及其在财务风险预警的应用[J].数量经济技术经济研究,2021,38(09):159-177.

[64]华长生. 逐步判别分析模型在识别上市公司财务欺诈中的应用[J]. 当代财经, 2008 (12): 119-122.

[65]张引弟.瑞幸咖啡财务造假案例研究——基于财务分析的视角[J].全国流通经济,2021.

[66]范宇晨. 基于两类 Boosting 的财务造假识别方法对比[J]. Advances in Applied Mathematics, 2021, 10: 4227.

[67]李天霞.基于分析性复核的上市公司财务舞弊识别的典型案例研究[J].Finance,2021,11:312.

[68]樊蕾, 吴明远. 财务造假研究的动态与进展——基于 CNKI 核心合集的可视化分析[J]. Advances in Applied Mathematics, 2022, 11: 3395.

[69]李爱华,王迪文,续维佳,等.基于多数据源融合的创业板上市公司财务造假异常检测[J].数据分析与知识发现, 2022: 1.

致谢

儿时因为一个棒棒糖就无比开心的画面似乎还在眼前，很快我竟要结束我的漫漫求学之路而步入社会了。这时光，终是说不清是快还是慢。当我现在安静地坐在图书馆角落里看着我的论文结尾，身边窸窸窣窣的翻书声、噼里啪啦的键盘声，都变得好不真实。我突然很羡慕还能继续在这个明亮宽敞的图书馆里学习的同学们，每个投入的神情里，我感受到的都是无穷的力量，而我不知道今后的我是否还能有这样的坚毅和梦想。

回望我这一路走来，被从迷茫和困境里拉起的幸运时刻太多，爱和温暖包裹着我并推着我勇敢前进。

古之学者必有师。感谢我的导师赵雪梅教授和文玉锋老师，我的论文也是在二位的指导下完成的。不单纯因为和我父母年龄相仿而让我感到如爸妈一般的爱，更是因为他们待我就如同孩子一般，那种发自内心的关怀、孜孜不倦的教导，照亮了每个困顿的日子里的我。初见赵老师，觉得浑身透露的都是学术气息和严谨认真，我甚至都不敢有过多眼神交流，有什么事都找外向幽默的文老师。直到参加比赛与赵老师相处后，赵老师对我尽心尽力的指导和帮助，赛后的一次次鼓励和肯定，让我猛然发现赵老师其实是细腻又明媚、耐心又平和的大家长。两位老师潜心科研、德高望重，热爱学习并终身学习是老师们做的的亲身表率。他们也是我的精神导师，让我知道不要闷头钻进学习里就失去了自身的活力，忽视了周遭世界的迷人和美丽。何其有幸能够成为赵文师门的一员，两位导师对我的影响将使我受益终身。感谢学院所有老师的栽培和教导，每位老师在课堂上的风采都历历在目，篇幅有限，恕不能一一致谢。

十月胎恩重，三生报答轻。感谢我的父母，二十多年来我都生活在无比幸福的家庭里，他们的善良和积极感染着我，让我从小就开朗又乐观。不论我做什么选择，他们从来没有过一次反对，即使可能观念有别，也只会坐下来和我谈谈他们的看法，尊重我所有的思考和判断，任我在广阔的天空中尽情翱翔。想给我最好的一切，却觉得自己能力有限给不了我太多，而在我看来，我已经拥有世界上最好最棒的父母，如果能够互相挑选，我和他们一定是毫不犹豫的双向选择。宽容、理解、守护……父母的恩情，岂是这三言两语就能道得明的。

高山流水遇知音，彩云追月得知己。感谢身边陪伴我的所有朋友，你们是我自己选择的家人，不论身在何地，都让我感觉有依靠和牵挂。感谢计算机学院的

杨杰同学，谢谢你的包容以及在生活和学习中对我的陪伴和帮助；感谢我的闺中密友僮僮（赵思僮），世界上的另一个我，谢谢你倾听我所有的喜悦和悲伤，未来还有说不完的话、吃不完的美食要一起分享；感谢单纯可爱的梁雨晴，你的直率和真诚让我感到温暖；感谢机车后座永远留有我位置的卢（卢俐均），和你一起在五号楼看过的无数次月亮挂在我心中最柔软的地方……会计专硕班的所有同学都是可爱又亲切的朋友，遗憾没能与大家多些相处，真心希望大家前程似锦，有美好的未来。感谢我多年的好朋友姚翔铧、胡老师（胡映辰）、丁丁（梁丁毅），你们像我的心理医生，不论什么烦心事都能让我跳出苦闷，豁然开朗；小美（沈权美）、piao（卜英芝）、玉玲（李玉玲）、大大（周筱），你们让我感觉不论在哪漂泊，总有张开的怀抱在等我回家……

无数日夜像放映机开始播放一样在我脑海里浮现，图书馆临近关门的钟声响起，大家陆续起身，相继离开，我才猛然从缥缈的思绪中回过神来。感恩我拥有和失去的一切，是所有的美好和遗憾铸造了当下独一无二的我。