编号：A0635

基于集成学习的上市公司财务数据造假识别

目录

表格和插图清单

摘要

近年来，随着我国证券市场不断扩容，不同行业、不同规模的上市公司在不断增多，金融市场监管部门的压力也不断增大。投资者可以根据其公开发布的一系列财务数据，分析其当前存在的问题和发展潜力，然而，由于投资者缺乏对上市公司财务报告造假的识别方式，上市公司可能时常会伪造财务数据，影响投资者的正确判断，干扰金融市场对企业经营质量、效益和竞争力的判断。在数字时代下，基于数据挖掘判断财报是否造假具有一定现实意义。另外的是，更是对财务数据的新动能的统计探索。

本文以第九届“泰迪杯”数据挖掘挑战赛A题提供的数据集为主，在此基础上进行基于集成学习的上市公司的造假识别探索。

首先对所给数据进行第一遍清洗，再使用集成学习的代表集成树模型XGBoost和LightGBM，并以recall（召回率）作为评价函数，分行业进行训练，得到二者公共的重要因子，完成不同行业财务造假的指标构建。

然后进行第二遍数据清洗和特征工程，分行业分别进行样本数据的采样，用Logistic回归、SVC、XGBoost、LightGBM、CatBoost以及卷积神经网络进行训练。为了提高模型泛化的能力，对模型结果进行投票，进行简单的模型融合。根据最后模型的f1-score、recall等指标得分，建立的模型具有可信度。

关键词：集成学习；财务造假；上市公司；Stacking；XGBoost

一、绪论

（一）研究背景

近年来，随着我国证券市场不断扩容，不同行业、不同规模的上市公司在不断增多，金融市场监管部门的压力也不断增大。上市公司每年会发布自己的年报、半年报、季报等财务报告信息，投资者可以根据其公开发布的一系列财务数据，分析其当前存在的问题和发展潜力，然而不时有上市公司凭着投资者对上市公司财报造造假缺乏识别手段等漏洞进行财务数据造假，影响投资者正确判断，干扰金融市场对企业经营质量、效益和竞争力的判断。

虽然相关的部门通过加强监管力度，努力设立建设正常的退市体制。但为了稳定股价，部分上市公司选择铤而走险，进行财务数据造假去吸引投资等，然后被强制退市，这给投资者带来的损失以及对金融市场健康发展的影响不容忽视。因此，对上市公司财务数据报告指标进行筛选、分析和研究，对监管部门监管、投资者利益和金融市场的健康发展来说都具有极大的意义。

伴随着信息技术的发展，数据挖掘技术在金融风险防范方面的作用越来越凸显。为确定各行业与财务数据造假相关的数据指标，比较分析不同行业上市公司相关数据指标的异同，我们在研究附件数据基础上结合了 Logistic、SVM、XGBoost、LightGBM、CatBoost、CNN 等算法对相关上市公司的财务数据进行研究，为投资选择提供思路。

（二）数据来源

本文所用原始数据集来自第九届“泰迪杯”数据挖掘挑战赛 A 题的数据，包括制造业、批发和零售业、房地产业等 19 个产业经脱敏后的有关上市公司第一年至第五年有关的财务数据如应收利息、预收款项、应付股利等指标。部分数据说明见下表 1，全部说明见所附数据包。

表 1　部分数据说明

字段名	含义
TICKER_SYMBOL	股票代码
ACT_PUBTIME	实际披露时间
PUBLISH_DATE	发布时间
END_DATE_REP	报告截止日期
END_DATE	截止日期
REPORT_TYPE	报告类型
FISCAL_PERIOD	会计区间
MERGED_FLAG	合并标志：1-合并，2-母公司
ACCOUTING_STANDARDS	会计准则
CURRENCY_CD	货币代码
CASH_C_EQUIV	货币资金
SETT_PROV	结算备付金
LOAN_TO_OTH_BANK_FI	拆出资金
TRADING_FA	交易性金融资产

（三）文献综述

用中国知网（CNKI）数据库“集成学习”“数据挖掘”“财务造假”为主题词分别进行了检索，所有相关的学术期刊或学位论文共有 3807 篇（2016 年-2021 年 4 月），主要于 2016 年开始发表，且发布文献的数量总体呈增长态势，2016 年发布 398 篇，2017 年发布 570 篇，2018 年发布 774 篇，2019 年发布 1071 篇，2020 年发布 1018 篇，2021 年 4 月为止发布 34 篇。

其中对于集成学习、数据挖掘、财务造假的内容，在深入学习与理解相关理论、并分析已有文献的基础之上，整理出主要涉及有以下几个方面：

石惠采用集成学习算法研究了基于 Stacking 的上市公司财务报告舞弊识别与预测模型。将财务报告造假当中经典的三个识别算法 BP 神经网络、支持向量机与 Logistic 回归算法相结合，成功建立集成分类器，最终得到了预测能力比较强的财务报告造假识别模型。

杜芸芸研究了财务指标以及非财务指标变量的影响，结合使用数据挖掘技术、多种对比分析，进行了基于数据挖掘的上市公司财务报告违规的研究。决策树与

Boosting 模型的结合取得了良好的效果，有效地提高了决策树的识别精确度。为提高识别的精确度、减少错误判断的可能性提供了方法和帮助。

邓启兰采用了数据挖掘方法对上市公司财务会计信息失真情况进行了识别，相关业务都是对上市公司财务会计信息是否失真的分类一般采用的数据挖掘方法。引用大量的数据作为研究样本，建立了三种不同的识别模型，快速提取所需的数据，智能地衡量财务会计信息的可靠性和真实性。

张宏斌、郭蒙的研究方法为机器学习，从我国上市公司多个业绩爆雷预警应用切入，基于文献和文本挖掘选择预测的变量，训练出了多种模型。通过完整地收集大量的数据和准确选择正确的模型，认为机器学习能够成为一种强适用性、高效率性的预测办法，可以对数据间的关系进行，研究不易解释、非线性、复杂的模型，超过传统计量经济模型的预测准确度。最终得出机器学习模型确实对上市公司业绩爆雷的预警效果较好的结论 ;预测能力与弹性网模型则集成学习模型 Bagging 和 AdaBoost 更强、更稳定。

薛巍通过对已有文献的归纳、总结了出多项指标，并基于此信息增益的特征方法来选择提取财务造假识别的相关指标，结果发现共计 14 项指标对中国上市公司财务造假识别发挥着指导作用。并将 MetaCost 这一代价敏感性学习引入到财务造假识别领域，利用随机森林的 MetaCost 算法，使财务造假公司的识别率得到了很大的提升，领先性的在财务暴雷领域取得了成功。

上述研究为基于集成学习的上市公司财务数据造假识别提供了一些思路和建议。但是从发表的文献来看，分行业分析财务数据有利于我们从大的架构中更好的体悟造假现象的产生原因和造假公司采用的手段，被挖掘处理出来的数据信息更具逻辑和严谨性，对造假数据指标的对比分析也更具条理。训练数据中模型的 f1-score、recall 等指标均未达到完全可靠的地步，但模型的结果对现实判断仍有一定参考价值为监管部门及投资者对上市财务数据监测与分析提供参考。

（四）模型假设

1.假设所获得的数据是真实可靠的。

2.假设第六年未发生重大事件和灾难或国家未推行重要政策影响上市公司企业。

（五）本文研究意义

本文研究相关上市公司财务数据,筛选各行与财务数据造假相关的数据指标并进行跟踪分析，对我们体悟投资者对上市公司财务数据稳健程度分析、监管部门对上市公司的有效监控具有很大裨益。

一方面，每个行业的敏感数据也不尽相同，分行业分析财务数据有利于我们从大的架构中更好的体悟造假现象的产生原因和造假公司采用的手段,被挖掘处理出来的数据信息更具逻辑和严谨性，对造假数据指标的对比分析也更具条理。

另一方面，研究财务造假，有利于防范类似造假手段和现象，为监管部门及投资者对上市财务数据监测与分析提供参考，对促进市场公平竞争、资本市场良性发展,削减上市公司退市风险,降低广大投资者的损失,具有很大的现实意义。

二、相关理论

（一）公司造假相关知识介绍

对于上市公司财务造假这一领域，国内外已经有很多相关研究。在对财务造假的定义方面，MBA 智库认为财务造假是：造假行为人违反国家法律、法规、制度的规定,采用各种欺诈手段在会计账务中进行弄虚作假,伪造、变造会计事项,掩盖企业真实的财务状况、经营成果与现金流量情况，从而为小团体或个人谋取私利的违法犯罪行为。美国注册公共会计师协会（AICPA）将财务报告造假定义为：公司或企业故意遗漏重大事项或在对财报中的披露进行错报，编造虚假的财务报告，或是管理当局欺诈。

上市公司财务造假手段很多，大多数造假公司会通过虚增交易、虚增资产、虚增或提前确认收入、利用过渡性科目或者隐瞒或不及时披露重大事项等手段调节利润，这些手段及行为结果将严重扰乱市场经济的正常运行，影响资本市场的健康发展，损害广大中小投资者的利益。

（二）集成学习相关算法介绍

1.Logistic Regressor

逻辑回归（Logistic Regression）又称为逻辑回归分析，均能在分类和预测两种算法中运用。它是通过对历史数据的表现分析来对发生的未来结果概率进行预测的一种基础机器学习算法。例如，我们可以将购买的概率设置为因变量，将用户的特征属性，例如性别，年龄，注册时间等设置为自变量。根据特征属性预测购买的概率。算法大致可以分为以下步骤：

①构造一个记为 h 函数的预测函数，此函数是分类函数，它能用来预测输入数据的结果。这个过程需要对数据进行分析，能知道或猜测预测函数的大概走势，比如是预测未来走势是线性函数还是非线性函数。

$$h_\theta(x)=g\left(\theta^T x\right)=\frac{1}{1+e^{-\theta^T x}} \tag{1}$$

②构造一个损失函数(loss function)并合成一个代价函数(cost function)。损失函数是说明每一个样本上，预测的输出 h 和训练数据类别（即真实值）y 间的偏差，可以说是二者之间的差（h-y），也可以说是（h-y）^{2} 或者是其他的形式。将所有训练数据的“损失”作为一个整体来考虑，其和或平均值成为代价函数，记为 J(θ)函数(此处θ参数为预测函数中的系数)。

$$J(\theta)=-\frac{1}{m}\left[\sum_{i=1}^{m} y_i \log\left(h_\theta\left(x_i\right)\right)+\left(1-y_i\right)\log\left(1-h_\theta\left(x_i\right)\right)\right] \tag{2}$$

③算出成本函数的最小值，确定使成本函数最小的参数。显然，J(θ)的值越小越好，因为这意味着预测值越接近实际值，预测函数的精度就越好，所以

J(θ)是我们的判断标准，需使其最小。求函数的最小值有不同的方法。主要的方法是梯度下降法，但也有其他优秀的算法。

$$\theta_{\mathrm{j}} = \theta_j - \alpha\sum_{i=1}^{m}\left(h_\theta\left(x_i\right) - y_i\right)x_i \tag{3}$$

2. Support Vector Machine

支持向量机，英文简称 support vector machine，一般简称为 SVM。一般来说，它是一种二元分类的机器学习模型，其基本模型的定义是全局特征空间上具有最大区间特点的线性分类器，使区间最大化是学习策略之一，它能最终转化为一个有关凸二次规划求解问题。 其大致步骤为：

①给定训练集：

$$T = \{(x1, y1), (x2, y2), \ldots, (xn, yn)\}$$

②求解二次规划问题：

$$\min_{\alpha}\ \frac{1}{2}\sum_i\sum_j \alpha_i\alpha_j y_i y_j\left(x_i \cdot x_j\right) - \sum_i \alpha_i$$

s.t. $\sum_i \alpha_i y_i = 0, \alpha_i \geq 0$

解得 $\alpha^* = \left(\alpha_1^*, \ldots, \alpha_{\mathrm{n}}^*\right)^{\mathrm{T}}$

③计算参数 w，选取一个正分量 α^* 计算 b

$$\mathrm{w}^* = \sum \alpha_{\mathrm{i}}^* \mathrm{y}_{\mathrm{i}} \mathrm{x}_{\mathrm{i}},\quad \mathrm{b}^* = \mathrm{y}_{\mathrm{j}} - \sum \alpha_{\mathrm{i}}^* \mathrm{y}_{\mathrm{i}}\left(\mathrm{x}_{\mathrm{i}} \cdot \mathrm{x}_{\mathrm{j}}\right)$$

④构造判决边界：

$g(x) = \left(w^* \cdot x\right) + b^* = 0$，由此求得决策函数：

$$\mathrm{f(x) = sgn(g(x))}$$

此外，SVM 有优点也有缺点。

优点：

①当核函数已知时，简化高维空间问题和降低问题的求解难度就较为容易。

②支持向量机可以被描述成一个凸优化问题,故可以逼近目标函数的全局最小值，通过使用已知公认的有效机器算法。

③少数的支持向量决就能定最终结果，方便捕捉关键样本，利于消除冗余的大量样本。

④因为一个虚拟变量被引入到数据的每个自有分类属性中，SVM 便可以应用于分类数据的问题中。

缺点：

①输入数据需要被进行全面地标注。

②比较困难对大规模数据训练。

③支持向量机解决场景一般是二分类问题,对于多分类的问题解决处理其效果差劲。

3．XGBoost

XGBoost 的全称为 Extreme Gradient Boosting，是一种实现 GBDT 的方式，XGBoost 中的基学习器可以是 CART(gbtree)也可以是线性分类器(gblinear)。并且，XGBoost 能够实现许多机器学习算法通过在 Gradient Boosting 框架下。XGBoost 提供了并行树升级(也称为 GBDT、GBM)，许多数据分析类科学问题都可以可以准确快速地被解决，从而为现实问题提供参考决策。

XGBoost 利用了核外计算，使数据科学从业者能够在计算机单机上处理亿万级个数据样本、数据集合。

最终，经过将这些技术结合起来创建一个端到端系统之后，该系统就可以通过小的集群系统扩展成为一个更大的数据集。

CART 决策树是 XGBoost 的子模型，XGBoost 通过 Gradient Tree Boosting 实现对连接多棵 CART 树的集成学习，最终得到成熟模型。

下面是 XGBoost 的最终模型构建。

①构造目标函数：

假设有 K 棵树，则第 i 个样本的输出为 $\hat{y}_i = \phi(\mathrm{x}_i) = \sum_{k=1}^{K} f_k(\mathrm{x}_i)$，$f_k \in \mathcal{F}$ 其中，$\mathcal{F} = \{ f(\mathbf{x}) = w_{q(\mathbf{x})} \} \left(q : \mathbb{R}^m \to T, w \in \mathbb{R}^T \right)$

因此，目标函数的构建为：

$$\mathcal{L}(\phi) = \sum_i l(\hat{y}_i, y_i) + \sum_k \Omega(f_k) \tag{4}$$

其中，$\sum_i l(\hat{y}_i, y_i)$为 loss function，$\sum_k \Omega(f_k)$为正则化项。

②叠加式的训练(Additive Training)：

样本 x_i，$\hat{y}_i^{(0)} = 0$（初始预测），$\hat{y}_i^{(1)} = \hat{y}_i^{(0)} + f_1(x_i)$，$\hat{y}_i^{(2)} = \hat{y}_i^{(0)} + f_1(x_i) + f_2(x_i) = \hat{y}_i^{(1)} + f_2(x_i)$ 等等以此类推，可以得到：$\hat{y}_i^{(K)} = \hat{y}_i^{(K-1)} + f_K(x_i)$，其中，$\hat{y}_i^{(K-1)}$ 为前 K-1 棵树的预测结果，$f_K(x_i)$ 为第 K 棵树的预测结果。

因此，目标函数可以分解为：

$$\mathcal{L}^{(K)} = \sum_{i=1}^{n} l\left(y_i, \hat{y}_i^{(K-1)} + f_K(\mathrm{x}_i) \right) + \sum_k \Omega(f_k) \tag{5}$$

因为正则化项能被解构成前 K-1 棵树的和第 K 棵树的两者复杂度的相加和，所以：$\mathcal{L}^{(K)} = \sum_{i=1}^{n} l\left(y_i, \hat{y}_i^{(K-1)} + f_K(\mathrm{x}_i) \right) + \sum_{k=1}^{K-1} \Omega(f_k) + \Omega(f_K)$，由于$\sum_{k=1}^{K-1} \Omega(f_k)$模型无法改变，当构建到第 K 棵树的时候。所以如果是常数已知的情况下，该常数在最优化的时候可以被省去，故：

$$\mathcal{L}^{(K)} = \sum_{i=1}^{n} l\left(y_i, \hat{y}_i^{(K-1)} + f_K(\mathrm{x}_i) \right) + \Omega(f_K)$$

③近似泰勒级数目标函数的使用：

$$\mathcal{L}^{(K)} \simeq \sum_{i=1}^{n} \left[l\left(y_i, \hat{y}^{(K-1)} \right) + g_i f_K(\mathrm{x}_i) + \frac{1}{2} h_i f_K^2(\mathrm{x}_i) \right] + \Omega(f_K) \tag{6}$$

其中，$g_i = \partial_{\hat{y}(t-1)} l\left(y_i, \hat{y}^{(t-1)} \right)$和$h_i = \partial^2_{\hat{y}^{(t-1)}} l\left(y_i, \hat{y}^{(t-1)} \right)$

4.LightGBM

LightGBM 由微软于 2017 年开源。参考了 XGBoost 的很多实现方法，如目标

函数的二阶泰勒展开、树叶节点值的计算、树复杂度的表示等。但在此基础上，LightGBM 采用了直方图加速方法和 Leafwise 树生长模式，因此在训练速度方面性能比 XGBoost 更优秀，训练精度保持在相同水平。在预处理阶段，利用直方图将特征值划分为桶。特征划分为桶后，只需要遍历每个桶，桶划分后计算时间复杂度为 O(#bins)。分桶操作后，可能不可能找到一个最佳的分裂阈值，但一个优点是降低了过拟合的风险。可以看出，使用特征分桶，然后搜索分割阈值，可以大大降低时间复杂度。

在对一个叶子进行分割后，需要重新计算其两个子节点的直方图(直方图的每个桶中存储着落在桶中的样本数量、样本的一阶导数和和、样本的二阶导数和)。这意味着只有子节点的特征直方图计算需要用更少的样本,然后父节点的特征直方图减去子节点得到的直方图，就可以得到直方图老一辈的子节点。这样，计算只需要在小叶子上遍历样本，就可以达到加速的目的。

5.CatBoost 介绍

Catboost 也是属于 Boosting Tree 下的梯度提升框架之一，其优点在于是对 catagory 特性的直接支持，更甚至可以支持字符串类型等类似特征。

Catboost 处理 catagory 特性的方式与 LightGBM 不同，将它们转换为数字类型的方式如下:

①重新随机将训练样本排序

②将 label 值的数据转化为整型数据

对于 regression：将 label 映射到 k 个桶中，并且将每个取值为[0, k-1]桶号来取代之前的 label，。对于 Classification：正/负样本用 1/0 表示。对于 MutiClassification：k 类编码为[0, k-1]。

③转换特征值

依次遍历每一个样本，并根据公式将 category 特征转换为数据型：

$$v = \frac{c + p}{tc + 1} \tag{7}$$

其中：v ：转换后的数值、p：平滑因子、c：遍历到当前样本时，与样本 label 相同的样本总数量、tc：遍历到当前样本的总样本数量。

6.CNN

卷积神经网络（CNN）的结构与传统神经网络(BP)公认十分的相似：两者都是由神经元组成的，神经元包括网络权值和可以学习的偏差（阈值）；每个神经元都是对输入进行相应的运算后，然后再将处理的数据输出（向前传播）；输出结果与样例输出获得错误（计算残差）；剩余传回一层一层地更新神经元的重量和偏见（反向传播）；循环重复，直到残差收敛和准确性需求得到满足。一般情况下，CNN 的结构形式是：数据层(data layer)→卷积层(convolutional layer)→池化层(pooling layer)→（池化、重复卷积层）→全连接层(fully connected layer)→结果输出层。其中最重要的就是卷积层和池化层。

卷积层简单来说就是经过卷积核进行计算的网络层。与神经网络相比，其中的卷积核便是神经网络中的的权值，卷积层等同于神经网络中的超平面。将原始图像通过卷积层的运算转换为超平面坐标系。超平面可以尽可能地集中同一类图像，多层卷积层的运算可以实现同一类图像的合并。图像特征的大小对卷积核的大小起着绝对作用，其作用路径是通过卷积核计算的网络层。超平面可以尽可能地集中同一类图像，多层卷积层的运算可以实现同一类图像的合并。

通常池化层的运算在卷积层之后再进行。经过池化层的运算，能对所得结果进行压缩，减少数据的存储空间，减少网络中的参数数量，从而减少计算资源的消耗，过拟合能被效控制。池化也称为下采样，用 $S = down(C)$ 表示、最大池、平均池等是常见的池化操作，其中最大池最为有名。顾名思义，最大池是将输入数据量（单个卷积核生成的二维结果）的每个深度切片最大化的操作。

利用反向传播算法来更新每个神经元权值的过程就是 CNN 的训练过程。本质上，CNN 是一个从输入到输出的映射。通过大量学习迭代输入输出样本，再加上使用卷积网络进行训练更新，网络便具有能拟合输入输出之间的特殊映射能力。

CNN 的训练是机器学习中的监督学习（有样本），其样本集是由输入和输出数据构成的。训练过程主要分为两个阶段:前向传播阶段和后向传播阶段(即网络权值被更新)。在前向传播阶段，数据输入到结果输出能被完成，训练完成后，在测试过程中也由网络执行这个过程。反向传播是根据输出误差和权值对公式进行修正，从而进行误差的反向传播，最后更新网络权值。网络训练过程如下：

①选择训练集，从原始样本集里随机采样 N 个样本，这 N 个样本作为下一步骤的训练样本。

②初始化四个参数，即精度控制参数、学习率、权值和阈值。

③为了计算各层网络的输出向量，从训练样本集中随机取一个样本的输入值到网络中。

④将实际的样本输出与计算的网络输出与进行比较，判断计算误差，然后使用错误来纠正更新连接的权重。

⑤中间卷积层的误差是由上层传播的错误,然后是错误的这一层是用来正确卷积的价值内核（网络重量），也就是说，它根据反向修正公式的修改重量。

⑥依次反转两个数据，即各层的网络权值和偏置（阈值）。

⑦再训练一次样本的所有数值（从输入到输出过程中计算误差，并利用误差反向来修改对应的网络权值)。

⑧经过大量更新迭代，判断实际的样本输出与计算的网络输出差值是否满足精度要求。如果没有满足，返回③并继续更新迭代。如果满足要求，执行下一步。

⑨在网络训练结束之后，保存其权值和偏置的数值。此时，整个网络训练已经彻底完成，可以认为每个权值已到达稳定值。

三、数据处理及因子筛选

（一）数据理解及预处理

首先，将附件 1 和附件 2 进行右连接，关联财务数据与其行业，其中用于训练的数据有 18060 条，需要预测，也就是第六年的数据有 4153 条。数据含有 20 个行业，但是不同行业所含数据情况有较大差异，如，制造业占据半壁江山，具体如表 X 所示，为提升模型的鲁棒性，选择将行业数量低于 XXX 归为一类，对其进行统一的数据处理和特征挖掘。

根据赛题要求，我们认为模型要尽可能减少“场外”因素的影响，因此，删除“TICKER_SYMBOL(股票代码)”，同时，也不考虑历史的造假情况。删除“REPORT_TYPE（报告类型）”,“FISCAL_PERIOD（会计区间）”,“MERGED_FLAG（合并标志）”，“ACCOUTING_STANDARDS（会计准则）”，“CURRENCY_CD（货币代码）”等五列方差为 0 的列。

（二）数据清洗

1.缺失值处理

对于缺失值处理的思路，在创新部分已有介绍，主要思想就是保持不同行业与公司之间的差异性，根据其具体情况进行缺失值处理，没有直接采用常见的描述统计指标或插值方法。在通过同一公司的中位数填充之后，统计每列的缺失值，删除缺失比例大于 90%的列，以提升模型的鲁棒性。最后，再对数据进行标准化后，将缺失值填充为-10，在集成树模型中，可以直接将此看作一类，而线性模型在处理标准化后的数据也能够拥有更快的收敛速度与更高的精确度。

2.异常值处理与标准化——RobustScale 标准化

除了少量类别数据外，大部分数据均为连续型数据，具体又可分为绝对数数据和相对数数据。为了更好的探索异常值情况，随机选择制造业中少量绝对数和相对数数据绘制直方图，以了解其数据分布，如下图 1 所示。

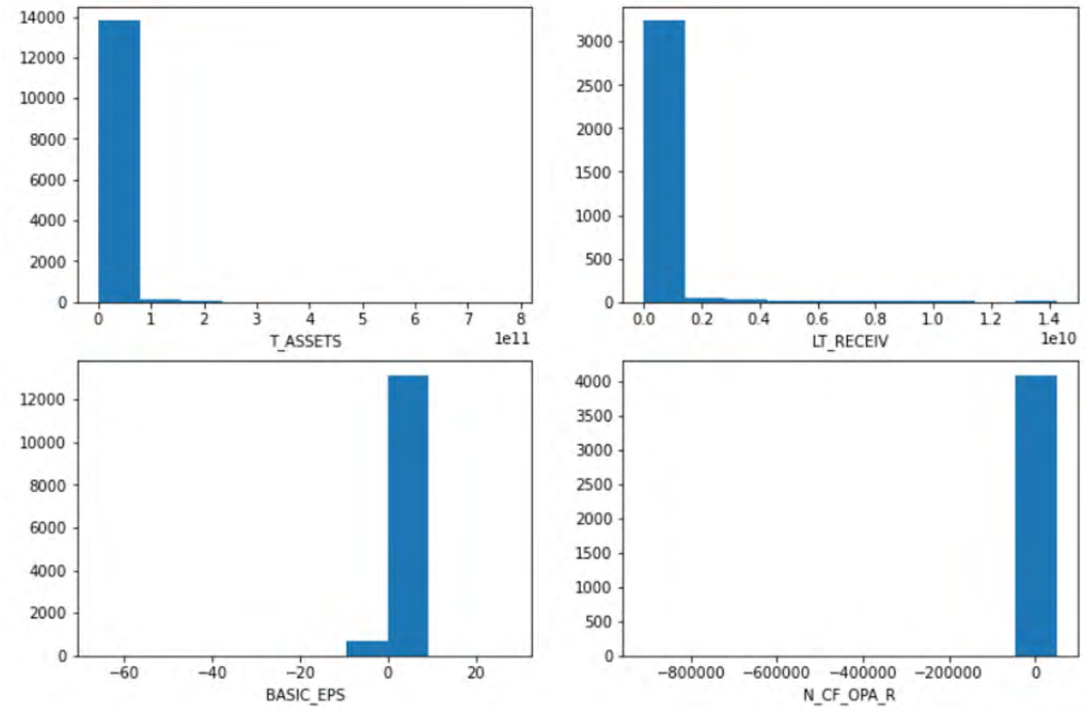

图 1　直方图

图中分别选择了 T_ASSETS(资产总计)、LT_RECEIV(长期应收款)、BASIC_EPS（每股基本收益）、N_CF_OPA_R（经营活动产生的现金流量净额/营业收入）四个指标来绘制直方图，可以看出不论是绝对数数据还是相对数数据都会呈现出偏态，图中反映出有少量异常值。值得注意的是，N_CF_OPA_R 指标存在的异常值为-909697.63，来源于 TICKER_SYMBOL 为 3852602 的第六年数据（用于预测的数据），该公司的 N_CF_OPA_R 在前面五年依次是 45.8749、9.657、61.1377、57.0686、30.9354。对于此类异常数据一方面要保持其异常的特性，另一方面也要减少其对模型精度的影响，因此，需要进行异常值处理和归一化。

一般的异常值有 3σ原则、基于箱线图的分析。Sklearn 中的预处理模块通常用于规范化，最常见的是 StandardScaler 和 MinmaxScaler。前者是根据原始

数据的均值和标准差两个指标进行标准化,后者是利用数据的基本数据最大值和最小值对原始数据进行线性变换。然而，这两种方法都不能很好地处理离群值,当数据中有许多离群值时,鲁棒 SCAL 对使用数据中心和距离具有更强的鲁棒性。

本次采用 RobustScale 标准化对数据进行处理,该方法基本处理异常值的方法来源于箱线图绘制时的 IQR（Interquartile Range），具体使用时，设置成为了 10%和 90%分位数，保持了尽可能多的信息，同时也对异常值进行了处理,随后使用了归一化操作。

（三）特征选择与构造

在数据清洗过后，首先对数据进行相关性分析，删除部分相关性高的特征。再将分行业的数据分别送入到 XGBoost 和 LightBGM 中进行训练。这两个集成模型中均含有筛选模型重要性的方法。具体来说，XGBoost 计算的是得分，通过每个单个特征与特征总和的比值得到分数，通过查阅源码得知，模型中采用 gain 作为重要性的评估指标，gain（增益）意味着相应的特征对通过与模型中的每个树采用每个特征的贡献，从而计算出来模型的相对贡献。与其他特征比较而言,此度量值的较高值体现出它对于生成预测来说更为重要；而 LightGBM 也有提供了“split”和“gain”两种方式，split 就是使用相应特征的使用次数。通过对比两个模型中的结果，得到二者共同的、排名靠前的重要特征。

筛选出重要特征之后，对争议小的特征进行交叉，作为一部分构造。同时根据数据本身的特征进行统计,比如统计每条数据的缺失情况,构造 MISSING 特征。

四、财务数据造假的指标构建

（一）基础模型的构建

对样本中的各个行业公司进行统计，如图 2 所示，其中近选择了 20 个行业中公司数量大于 100 的公司。从图中可以看出，与其他行业相比而言，制造业公司样本的数目远远高出，事实上其占比为 64%。为了模型的鲁棒性，将行业公司数量小于 100 的行业归为其他类，同时为了避免数据中的噪声，在对制造业进行模型训练时主要采用近两年的数据，而其他行业由于样本量不大，将把前五年的数据都使用在了模型的练习之中。

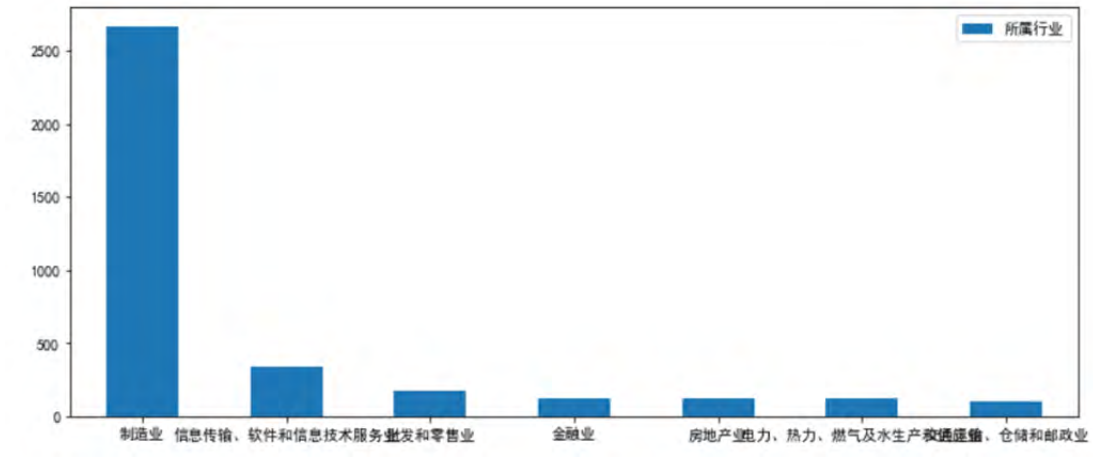

图 2　各个行业公司数量统计

（二）模型参数调优与重要特征选择

对于参数调优大致思路相当，只是根据不同模型选择不同的超参数调节，为了节省篇幅，这里就以 lgb 使用网格调参为例。

一般的调参方法有网格搜素、随机搜索和贝叶斯调参。网格搜索对所有的参数进行交叉组合，在其中找到最优的组合，该方法可以理论上可以找到最优解，但是需要对整个参数空间进行遍历，计算代价比较大。随机搜索，顾名思义，是随机的网格搜索，最大的有点便是速度快，但也可能错过最优的信息。根据本文所用的数据和模型的复杂度，最终决定选择网格搜索进行超参数调参。

具体实现上使用 sklearn 中的 grid_search 下的 GridSearchCV 方法，从名字可以看出，除了网格搜索外，该方法还含有 CV（crossvalidation）。网格搜索就是利用交叉验证的形式比较每一个参数下训练器的精度的，但是交叉验证也要求大量的计算资源，加重了网格搜索的搜索时间，好在本次数据量可以接受。同时，由于基础模型的 auc 已经接近 90%，选择了方法中自带的 f1-score 作为评价标准，通过查看源码以及测试，这了计算的是造假样本上的准确和召回率计算的 f1-score，模型初始 f1-score 为 0.09。

表 2　XGBoost 调参信息汇总

字段名	含义
TICKER_SYMBOL	股票代码
ACT_PUBTIME	实际披露时间
PUBLISH_DATE	发布时间
END_DATE_REP	报告截止日期
END_DATE	截止日期
REPORT_TYPE	报告类型
FISCAL_PERIOD	会计区间
MERGED_FLAG	合并标志：1-合并，2-母公司
ACCOUTING_STANDARDS	会计准则
CURRENCY_CD	货币代码
CASH_C_EQUIV	货币资金
SETT_PROV	结算备付金
LOAN_TO_OTH_BANK_FI	拆出资金
TRADING_FA	交易性金融资产

在对模型进行调参过后，运用模型的 plot_importanc 方法对很重要特征进行可视化，具体结果如下图 3。

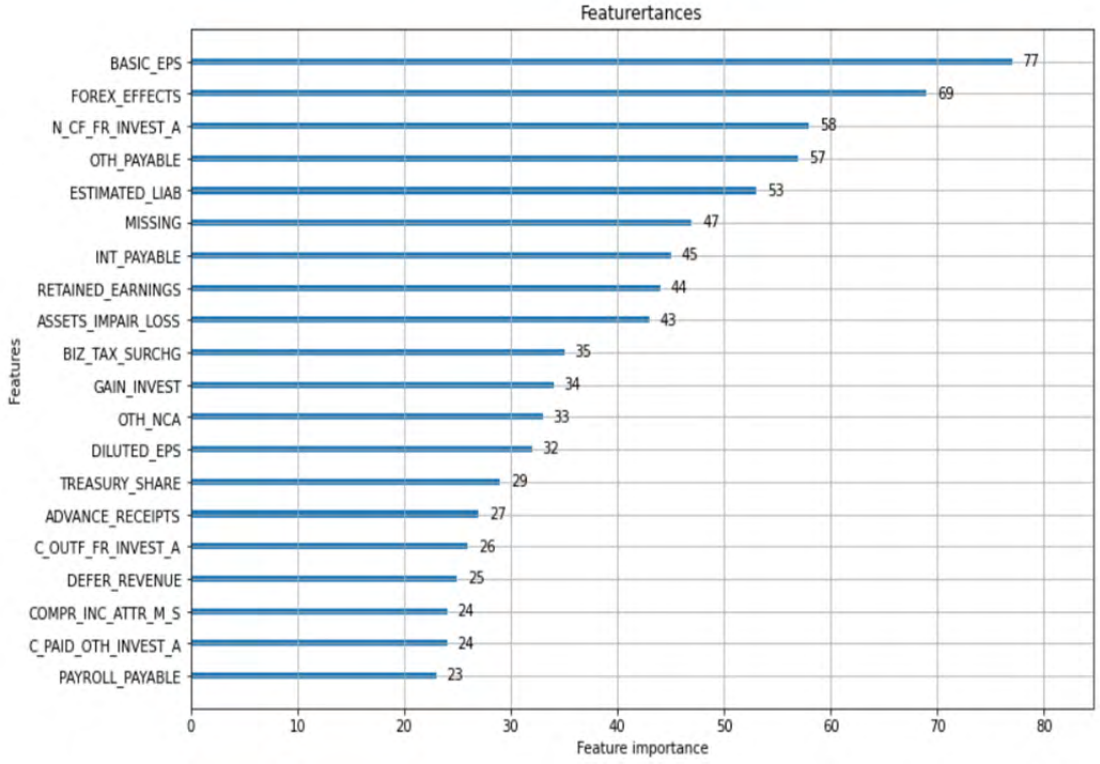

图 3　重要特征可视化

（三）不同行业上市公司相关数据指标异同

这里只给出制造业的关键指标，所有行业的指标见附录一。

表 3　制造业关键指标分类

制造业	
资产类科目	在建工程
	其他非流动资产
负债类科目	预计负债
	一年内到期的非流动负债
	其他应付款
	非流动负债合计
	预收款项
	应付利息
所有者权益类科目	减:库存股
	盈余公积
	归属于母公司所有者权益合计
损益类科目	资产减值损失
	营业外支出
	基本每股收益
	其他收益
	研发支出
	公允价值变动收益(损失以“-”号填列)
现金流量表涉及项目	吸收投资收到的现金
	收到其他与筹资活动有关的现金
	每股企业自由现金流量
	支付给职工以及为职工支付的现金

为了更好地比较不同行业的样本上市公司相关数据指标，我们根据筛选出来的每个行业异常数据所对应的财务报表项目以及直接可察觉的数值异常的财务指标，将所有的数据指标全都归总关联，为了中国《企业财务通则》中为企业规定的三种财务指标，即盈利能力指标、偿债能力指标和营运能力指标。

据分析，制造业类指示企业偿债能力的异常数据指标明显高于其他行业，这也可以从侧面反映制造业行业的敏感财务数据类别；同时，其异常数据指标中涉及损益类科目的项目也偏多，从财务角度分析，这些涉及到的指标的确有很多人

为可以操纵的灰色因素。信息传输、软件和信息技术服务业涉及损益类科目的项目在我们所分析的八类行业中数量最大，说明造假企业可能对企业盈利能力指标的操纵更多。批发和零售业向来是容易在营运能力指标上做手脚的，从我们挖掘到的异常数据指标来看，应收账款周转率、存货周转率等直接财务指标的异常很好地契合了从财务角度分析财务报表数据造假的常用分析思路。金融行业的异常数据指标显示了造假企业很大程度是从企业盈利能力指标和营运能力指标下手，欺骗报告使用者。交通运输、仓储和邮政业的异常数据指标没有涉及资产类科目的项目数据指标，营业利润与负债的比值的异常情况是让人不得不注意的地方。房地产业的异常数据指标集中显示了所有者权益类科目的项目数据和长期偿债能力指标的异常。电力、热力、燃气及水生产和供应业的异常数据中没有涉及负债类科目的项目数据指标。其他分类行业的异常数据也很明显关联到偿债能力指标、营运能力指标和盈利能力指标这三个指标，这三个指标显示财务被企业操纵、造假的可能性很大。

从异常数据指标的分析中我们能明显感受到，无论是哪个行业，应收、预付的相关款项、收入和费用的确认都有极大被企业操纵、造假的可能，造假企业最常操纵的部分便是与投资活动相关的现金流量。这些所涉及的指标背后的财务、会计程序、政策等涉及的人为因素非常多，是最容易被造假企业利用的“灰色因子”。

五、集成学习模型融合的模型构建及结果

（一）样本不均衡与采样操作

在整体数据中，正负样本的比例大约是 100：1，在进行行业划分后，比例仍不会有太大的改观，因此考虑采用采样的方法进行处理。采样一般有欠采样和过采样这两种主要思路，一般可以通过随机抽样的方法进行欠采样，弱化了过渡部分正例的影响，为了让过采样不是单纯地复制负例，可以使用

SMOTE(Synthetic Minority Oversampling Technique)、Borderline SMOTE 和 ADASYN(adaptive synthetic sampling)等方法。

SMOTE 是在随机采样的基础原理上进行改进的，通过合成少数类 K 邻近附近的样本达到采样的目的，由于 SMOTE 对所有少数类样本是同等对待，并未考查到近邻样本的类别信息，常常导致有样本混叠现象的现象出现，让最终呈现的分类效果不如意。而 Borderline SMOTE 将样本分为三类:安全类、危险类和噪声类。这种方法只对危险样本进行采样，而这些样本往往位于正样本和负样本的交界处，所以这种方法也被称为 Borderline SMOTE。同样，ADASYN 也会对一些类型的样本给予不同的权重，以产生不同数量的样本。ADASYN 的效应模型如下图所示。

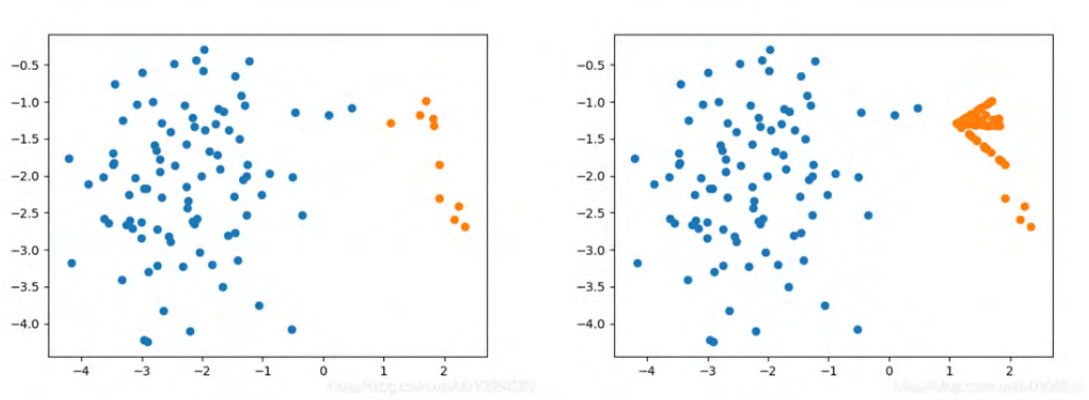

图 4　ADASYN 的效果模型

虽然模拟的例子比较简单，也有些极端(事实上正负样本不会这么泾渭分明)，因此在实际操作过后，仅有 CNN 使用到了采样的样本进行训练，其他模型均通过修改“class weights”参数或者类似参数以实现对样本不均衡问题的优化，提升模型的泛化能力。

（二）基于模型融合的预测模型构造

针对不同方式处理后的数据，进行了模型训练，在采用基础的模型融合方法——投票法进行模型融合，大致的流程如下图。

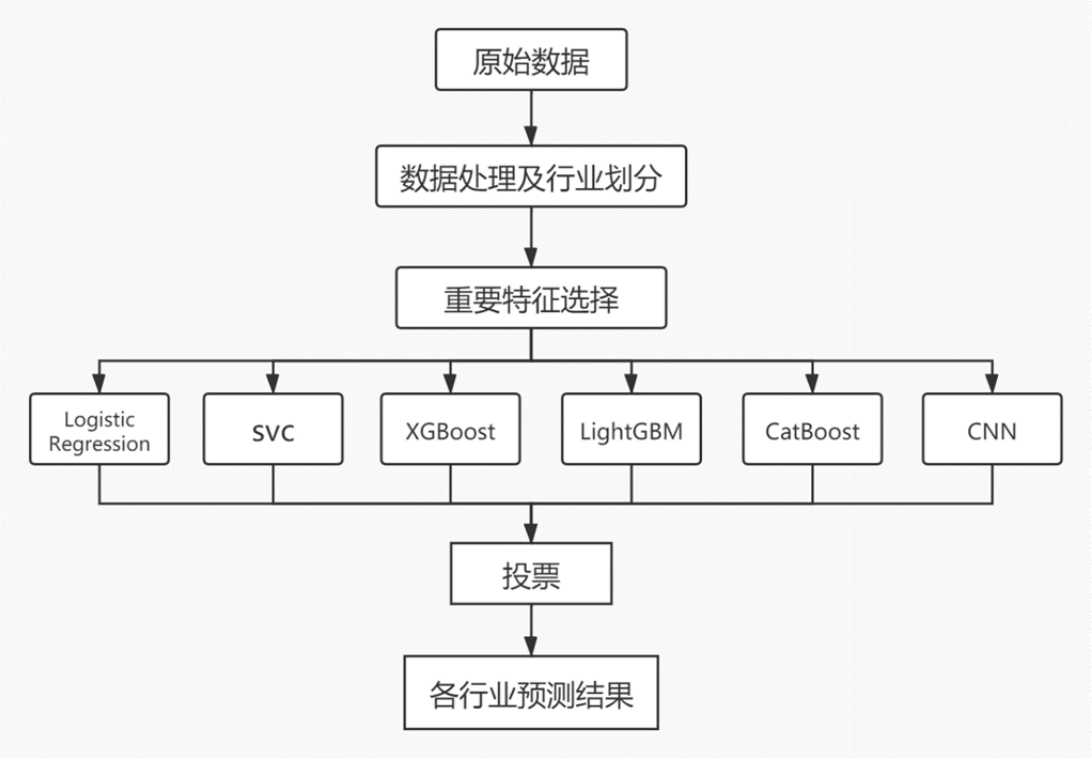

图 5　问题解决流程图

以制造业为例，融合后的模型的负类召回率为 0.21，f1-socre 为 0.16，正类 f1-socre 为 0.99。

（三）第六年各行业结果预测

在以上工作的基础上完成模型的训练后，选择第六年进行预测。首先将第六年的数据按照前面的数据清洗方式，包括缺失值填补、模型归一化，并针对不同模型进行有指向性的数据预处理，比如为神经网络进行采样操作以实现样本的均衡，最后分别送入模型，进行是否财务造假的判断。每一类根据算法默认的阈值，也即 0.5 进行类别划分，得到正负类结果。在预测结果中，制造业有 18 家，其处理后的股票代码为：376089、431877、461623、541225、630294、1385133、2149006、2444790、3285030、3418224、3424430、3646698、3816511、4157869、

4569042、4789867、4897311、1094620。其他行业中有 20 家造假，分别是 4388517、350818、1625847、3159512、4710457、4786358、79573、580788、2999962、981402、3018406、3360110、2769232、778811、85904、1516761、1892808、2402604、2844622、3262635。总数在 38 家，用于预测的公司数目为 4153。比例上大致符合训练集。但是，就过往的训练表象来看，模型的召回率一般在 20%上下。虽然对于判断为负类的公司不能全盘接受，但其存在财务造假的可能性大大增加，故可以作为重要的参考信息，再结合其他业务信息对其进行进一步判断。

六、模型评价

（一）本文主要工作与创新点

本文对上市公司财务报表信息进行了数据挖掘，着重关注了样本不均衡和 racall 作为评价指标的问题，不断在模型上进行尝试利用集成学习和深度学习的算法对是否财务造假进行了模型建立。

1.缺失值处理

由于行业和公司规模等因素的差距，原始数据中有四百万以上的缺失值，占数据总数的七成以上。但是不同行业在财务指标披露上也有所差别，比如建筑业在“应收保费”基本上均为缺失。同时，在同一行业中，由于其规模差距，部分与经营关系不太明显关系的科目或者指标往往也是缺失值。因此，对于不同行业、不同公司的缺失值不能简单地使用相应科目或者指标的总数或中位数等统计指标填充。本小组在操作过程中，将所有的数据按股票代码进行分类汇总，根据每个公司的其他年份情况进行缺失值填充。

2.不平衡数据处理

本次比赛所用数据存在样本不均衡的问题，比例达到了 100：1。在现有数据的基础上，采用了欠采样、过采样等方法进行数据增广，以提升模型的泛化能力。同时，深刻理解各个参数，比如一般集成树模型中都含有“class weights”

参数，用于修改不同类别在优化过程权重，以实现较快速度的收敛和提升模型精度。

3.评估方法改进

对于分类问题，常见的评价方法有 AUC、logloss 和 F1-score 等，这些方法兼顾模型对正负样本的分类能力，特别是在样本不均衡的二分类的问题中，其效果优于 accuracy（准确率）。但本问题的背景下，与疾病的检测类似，需要较为严格的标准，尽可能地筛选出存在问题的样本。因此，在很多模型自带的评价方法基础上，引入自定义的 recall（召回率），尽可能提高模型对负样本的检测能力。

（二）缺点和不足

本文还有很多需要改进的地方，虽然已经使用 XGBoost 和 LightGBM 得到财务造假的公共因子。但是难免出于算法选择的原因略去关键指标，可能会导致最后构建的指标和模型并不是最优。其二，忽略了对数据本身指标的深度理解与挖掘，没有集合实际的业务场景进行特征工程，最终只是不同的模型、降维方法得出重要因子，进行建模，过于依赖于算法，此外，因为有些上市公司在 5 年之内退市，会导致之前的数据无效，也可能导致了模型的不准确性。之后可以将更多的集成模型融合赋予不同的权重不断完善模型的构建。

参考文献

[1]周飞燕,金林鹏,董军.卷积神经网络研究综述[J].计算机学报,2017,40(06):1229-1251.

[2]杨曦烨.利用机器学习技术识别财务报表造假[D].南开大学,2020.

[3]陈旭,张军,陈文伟,李硕豪.卷积网络深度学习算法与实例[J].广东工业大学学报,2017,34(06):20-26.

[4]石惠.基于 Stacking 的上市公司财务报告舞弊识别与预测模型研究[D].西南财经大学,2019.

[5]杜芸芸.基于数据挖掘的上市公司财务报告违规研究[D].西南财经大学,2013.

[6]邓启兰.基于数据挖掘的上市公司会计信息失真识别模型研究[D].重庆理工大学,2019.

[7]张宏斌,郭蒙.机器学习与财务预测——来自中国上市公司业绩爆雷预警应用的经验研究[J].金融学季刊,2020,14(04):135-154.

[8]薛巍.上市公司财务欺诈识别模型研究[D].南京大学,2015.

[9]吴家香.上市公司财务舞弊的方式、特点及识别方法[J].现代审计与会计,2021(04):32-33.

[10]李朋涛.上市公司常见财务舞弊手段和防范对策研究[J].中国集体经济,2021(11):83-84.

[11]刘敏.浅谈财务会计造假与打假[J].商场现代化,2011(17):143-144.

[12]李艳艳.基于上市公司财务报表分析的企业竞争战略探析[D].山西财经大学,2012.

[13]罗雷.基于 Xgboost 方法的优惠券使用预测研究[D].南昌大学,2019.

[14]刘书丽.关于我国会计造假问题的分析[J].经济师,2005(11):53-54.

附录

制造业	
资产类科目	在建工程
	其他非流动资产
负债类科目	预计负债
	一年内到期的非流动负债
	其他应付款
	非流动负债合计
	预收款项
	应付利息
所有者权益类科目	减:库存股
	盈余公积
	归属于母公司所有者权益合计
损益类科目	资产减值损失
	营业外支出
	基本每股收益
	其他收益
	研发支出
	公允价值变动收益(损失以“-”号填列)
现金流量表涉及项目	吸收投资收到的现金
	收到其他与筹资活动有关的现金
	每股企业自由现金流量
	支付给职工以及为职工支付的现金

信息传输、软件和信息技术服务业	
资产类科目	可供出售金融资产
	应收票据
	长期股权投资
	长期应收款
负债类科目	递延所得税负债
	非流动负债合计
	其他应付款
	应付票据
所有者权益类科目	减:库存股
	资本公积
损益类科目	管理费用
	基本每股收益
	其他收益
	稀释每股收益
	销售费用
	研发支出
	营业利润(亏损以“-”号填列)
	支付的各项税费
	资产处置收益
现金流量表涉及项目	筹资活动产生的现金流量净额
	处置固定资产、无形资产 和其他长期资产收回的现金净额
	处置子公司及其他营业单位收到的现金净额
	汇率变动对现金及现金等价物的影响
	期末现金及现金等价物余额
	其中:子公司吸收少数股东投资收到的现金
	投资活动现金流入小计
	支付其他与经营活动有关的现金

批发和零售业	
资产类科目	交易性金融资产
	其他非流动资产
	应收股利
	应收票据
负债类科目	预付款项
	商誉
	应付股利
	预计负债
损益类科目	基本每股收益
	稀释每股收益
	资产处置收益
	未分配利润
	利润总额(亏损总额以“-”号填列)
	营业外支出
线径流量表涉及的项目	经营活动产生的现金流量净额
	汇率变动对现金及现金等价物的影响
	收到其他与经营活动有关的现金
	每股股东自由现金流量
	支付其他与筹资活动有关的现金
直接财务指标	归属于母公司的股东权益/带息债务
	经营活动现金流量净额/带息债务
	息税折旧摊销前利润/带息债务
	存货周转率
	应付账款周转率

金融业	
资产类科目	固定资产
负债类科目	应付利息
	短期借款
	递延所得税负债
损益类科目	支付的各项税费
	营业利润(亏损以“-”号填列)
	基本每股收益
	外币报表折算差额
	利润总额(亏损总额以“-”号填列)
	营业外支出
现金流量表涉及项目	其中:子公司支付给少数股东的股利、利润
	收到其他与投资活动有关的现金
	支付其他与投资活动有关的现金
	取得投资收益收到的现金
	收回投资收到的现金
	购建固定资产、无形资产和其他长期资产支付的现金
	其中:子公司吸收少数股东投资收到的现金
	处置子公司及其他营业单位收到的现金净额
直接财务指标	预付账款/总资产
	无形资产/总资产
	负债合计/归属于母公司的股东权益
	营业收入同比增长
	固定资产合计周转率

交通运输、仓储和邮政业	
资产类科目	应收票据
负债类科目	应付股利
	应付票据
所有者权益类科目	专项储备
现金流量表涉及项目	投资活动现金流入小计
	支付其他与筹资活动有关的现金
	取得投资收益收到的现金
直接财务指标	营业利润/流动负债
	息税折旧摊销前利润/负债合计
	营业利润/负债合计

房地产业	
资产类科目	长期股权投资
	有形净资产
负债类科目	应付账款
所有者权益类科目	盈余公积
	实收资本(或股本)
	留存收益
	每股盈余公积
	未分配利润
损益类科目	递延收益
	扣除非经常性损益后的归属于上市公司股东的净利润
	基本每股收益
	稀释每股收益
现金流量表涉及项目	处置固定资产、无形资产和 其他长期资产收回的现金净额
	其中:子公司吸收少数股东投资收到的现金
	加:期初现金及现金等价物余额
直接财务指标	营业外收入/营业总收入
	营业利润/负债合计
	经营活动现金流量净额/流动负债
	现金流量利息保障倍数
	投资活动产生的现金流量净额占比
	经营活动现金流量净额/负债合计
	货币资金/流动负债

电力、热力、燃气及水生产和供应业	
资产类科目	应收利息
	持有至到期投资
	归属于母公司所有者(或股东)的综合收益总额
所有者权益类科目	未分配利润
损益类科目	综合收益总额
现金流量表涉及项目	投资活动现金流入小计
	取得子公司及其他营业单位支付的现金净额
直接财务指标	应收账款/营业收入

其他行业 (划分按第一年行业企业的数量小于 100)	
资产类科目	可供出售金融资产
	其他流动资产
	长期借款
	在建工程
负债类科目	应交税费
	其他应付款
	其他非流动负债
	一年内到期的非流动负债
	应付债券
	应付职工薪酬
所有者权益类科目	专项储备
	少数股东权益
损益类科目	支付的各项税费
	持续经营净利润
	其中:对联营企业和合营企业的投资收益
	基本每股收益
现金流量表涉及项目	投资活动产生的现金流量净额
	支付其他与投资活动有关的现金
	处置固定资产、无形资产 和其他长期资产收回的现金净额
	分配股利、利润或偿付利息支付的现金
	投资支付的现金
	吸收投资收到的现金
	其中:子公司吸收少数股东投资收到的现金

致谢

本次统计建模论文的后期修改到最终完成的过程，我们要感谢老师的悉心指导，给予了我们细致入微的帮助和意见。我们的指导老师在忙碌之余，也一直认真地给我们提很多切实中肯的修改意见，为我们的论文撰写提供了宝贵的建议。老师在科研学习上给我们的指导、在科研态度上给我们树立的榜样，更是让我们相信了自己学习统计的能力与热情。

还要感谢我们的每一位科任老师在课堂上尽心尽责、毫无保留地传授知识，更是让我们看到了统计系老师们强大的人格魅力。另外，我们要感谢在论文数据爬取过程中给予我们帮助的同学，帮助我们一起解决数据采集过程中的困难重重。

最后，也十分感谢本组成员之间密切地分工协作，共同完成了本次统计建模论文创作！