第九届“泰迪杯”数据挖掘挑战赛

作品单位：广东第二师范学院

作品成员：陈成泽亚 蓝嘉城 李培森

指导老师：陈兴发

基于机器学习模型预测财务造假的上市公司

【摘要】“财务造假”是指上市公司伪造财务报表，虚报、瞒报公司的部分财务数据。正是如此，通过机器学习模型对公司财务数据挖掘，对数据分析、预测财务造假有了理论上的可行性。基于对样例数据集进行了常规的缺失值、异常值、标准化等数据处理，在总 19 种行业中，对 8 种行业数据不平衡处理，包含朴素随机欠采样、朴素随机过采样以及 ADASYN 过采样法等方法。运用 SVM 及 K 折交叉验证去验证并挑选各个行业最优的均衡方法，剩余 11 种则不进行任何均衡操作。随后，采取特征选择中使用了权重法、过滤法、包裹法、嵌入法等方法提取特征，并结合实际经济学意义得到最终特征。最终通过 6 种模型预测，以 F1 分数为评价指标，进行预测哪些上市公司可能会存在造假情况。

针对问题一，在对样例数据集进行上述的特征因子选择后，可以确定出各行业与财务数据造假相关的数据指标。随后，通过构造

$$a_{ij} = \frac{b_{ij}}{c_i}$$

函数并制作热力图，可以得出不同行业上市公司财务造假相关数据指标的异同。

针对问题二、问题三，通过 LR、RF、SVM、MLP、XGBoost、GBDT、ADAboost 这 7 种模型对各个行业进行财务造假预测，得出第 6 年各个行业上市公司财务造假的情况。

其中所有行业最优预测模型的总平均准确率为 0.751，总平均 AUC 为 0.76486。

【关键字】财务造假；行业分类；机器学习；特征选择；不平衡处理

目录

一、 问题背景

随着我国经济的快速发展，证券市场不断扩容，不同行业、不同规模的上市公司不断增加，目前上市公司的数量已超过 4000 家。然而，近年来不时出现上市公司财务数据造假及暴雷的情况，2020 年还出现了流动性危机及信用债违约等问题。这些问题提醒监管部门对上市公司进行有效监控。

不法公司通过虚增交易、虚增资产、提前确认收入虚增收入、利用过渡性科目调节利润等手段到达制作虚假财务报表的方式对投资人非法集资。如今在大数据的背景下，剥离财务造假的虚假外衣，通过部分特征准确区分识别财务造假是一个必然的解决方式。为了使得投资市场的风险控制的进一步发展，保障投资人的资金安全，通过数据挖掘的方式，曝光财务造假上市公司已经是迫在眉睫的实际问题。

二、 问题分析

该问题的核心内容是关于上市公司财务数据的分析，主要目的为区分上市公司其财务数据是否造假，挖掘其中的特征因子，以此构建数学模型寻找可能存在财务造假的上市公司，为投资者尽可能降低、避免暴雷风险。

以寻找特征因子作为基础目的，再而以因子构建模型以主要目标，之后检验模型并提供符合特征的企业数据为最终目标。

模型假设成立需满足以下两个条件：

1. 模型所参考的数据真实有效

2. 预测的第 6 年不出现对上市公司有重大影响的事件

三、 数据预处理

1. 数据的预处理

原始数据集的预处理和拼接是处理庞大数据量的任务,其是数据挖掘的第一步。首先需要将各个子数据文件集中处理，数据过于分散会使得后续的数据挖掘产生不必要的麻烦,因而把各个分散的数据依据数据之间的联系将各个数据拼接起来，形成一个有关联的数据集，便于后期观察数据和调用数据。处理方法即将赛题附件中，附件一中“财务指标”数据与附件三中“所属行业”相关联，再将其与附件二整合，以此获取三个分散数据的集合，将作为一个数据集的形式以便后续的处理，在后续文章中简单成为“数据”。

其中首先依据客观判断主观排除一类对数据挖掘没有实际意义的特征,如数据中七个无相关性指标——“实际披露时间”、“发布时间”、“报告截止日期”、“报告类型”、“会计区间”、“合并标志”、“会计准则”。这类因子是每个行业固有的信息，根据不同的行业存在着不同的特定数据，与其他特征没有相关性，若作为特征因子进行筛选将会对数据处理带来不必要的精力的浪费,因而不能与其他因子一并作为挖掘的对象，将无关因子删除也在一定程度上达到简化数据、简化模型的作用。

在后续进行简单的排版整理后，最终获得 22213 行 356 列的数据，将之作为原始数据，进行后续进一步的处理。

2. 缺失值、异常值的处理

缺失值、异常值的处理是数据处理不可缺少的一环。缺失值是源于数据采集

的空缺、传输间丢失等不可控情况所导致或人为故意丢失等多种情况，如何处理空值是数据处理中恒久不变的问题。而其中异常值是由于数据传输错误所导致的，通常对其采取修改或是剔除的处理方式，但具体所采取的方式也需要依据客观上分析数据所决定。

以下主要分为三个分点叙述我们所采用的处理方法及处理结果。

2.1 缺失值处理

首先观察原始数据所存在缺失值情况，对各个行业的缺失值进行统计并计算得出其所占总数据的比例，制作柱形图（图 1），以缺失值占比的大小将各特征因子的缺失值采取不同的处理方法。另在观察原始数据时，通过观察可知原始数据中‘是否在当年造假’‘所属行业’这两组数据显然与数据挖掘无关，并不需要参与数据的处理，即无需对其进行缺失值处理。在缺失值处理环节中，将‘是否在当年造假’这一因子移出原始数据（处理后重新添加）；而‘所属行业’这一因子作为数据分类的依据，以‘所属行业’的数据异同，将各个行业的数据分组处理，目的于对不同行业采取不同的特征选择方法，其特征筛选也有所异同。

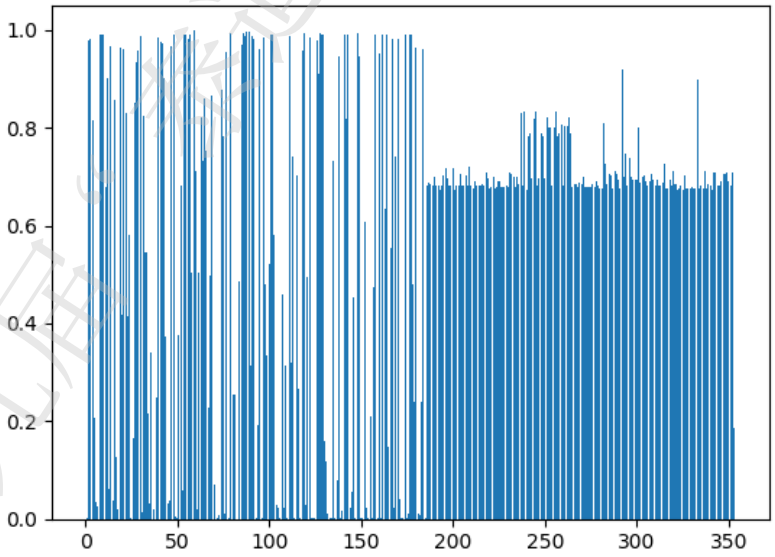

图 1 各行业数据缺失比例

依据图 1 中各行业数据缺失比例，采取了以下三种处理方式：

对于缺失比例大于 0.5 的数据，采取之间剔除的方式处理。缺失比例大于 0.5 即说明其缺失数据较多，若采用下文所使用其他方式填充，将使得整体的可信度大大降低，故需要删除。

对于缺失比例介于 0.5-0.2 之间的特征，将采取填充的方式处理。但同时考虑到每个行业的各项指标之间判断是否为缺失的标准具有差异性，不同行业之间对不同特征衡量缺失值的标准时不统一的。因此在缺失值填充这一环节，将 0 值作为空值处理，并对需要数据填充的数据进行按行业分类，并对该部分数据进行均值填充。

对于缺失比例在小于 0.2 范围的数据中，我们使用随机森林填充处理，随机森林作为一种比较新的机器学习方法，是一个包含多个决策树的分类器。其填补缺失值基本步骤，即挑选出数据中缺失值最小的一列，并用随机森林回归模型填充，再挑选出填充后缺失值最小的一列，循环使用随机森林回归模型填充。

填充结果即为缺失值处理后得到 22213 行，101 列的数据，包含填充前移出的‘是否在当年造假’因子。

2.2 异常值处理

异常值即数据中录入错误以及含有不合常理的数据，一般分为以下几个步骤：异常值检测、异常值筛选、异常值处理，一般的处理方法为：箱型图、简单统计量（比如观察极(大／小)值），3σ 原则。在统计学中，如果一个数据分布近似正态，大约 68%的数据值会在均值的一个 标准差范围内；大约 95%的数据值会在两个标准差范围内；大约 99.7%的数据值会在 3 个标准差范围内。

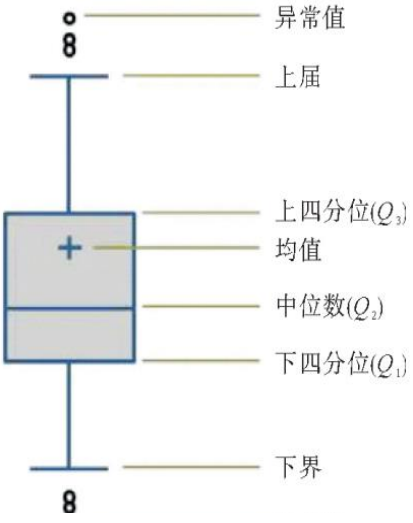

图 2 箱型图法

因此，在 3σ 原则下，定义异常点为与平均值的偏差超过 3 倍标准差的样本． 箱形图是根据数据的四分位数形成的图形化描述，这是一种简单而有效的可视化离群点的方法。 图 3 箱型图 将上下界作为数据分布的边界，高于上界或低于下界的数据点均 视为离群点或异常值．具体来说，四分位数将数据分为 3 个点和 4 个区间。

异常点定义为低于箱形图下界(或Q_1-1.5(Q_3-Q_1))或高于箱形图上界(或Q_3+1.5(Q_3-Q_1)的观测值。

具体处理方法于不同行业存在不同的结果，以下特以制造业的应收票据举例。首先使用两种方法对数据进行分析比较，然而发现原始数据显然是不服从正态分布的，若使用 3σ 法处理将使得问题解决复杂化，因此最终选择使用箱型图法处理。对制造业的数据作箱型图如下图 3，可以发现存在异常值，需要对异常值进行处理。但基于各个行业的数据不同，异常值大小的必然存在着或大或小的差异。因此需要依照不同行业作箱型图，再对异常值做出适当处理，用对应行业数据的异常值替换为该行业箱型图的上下界。

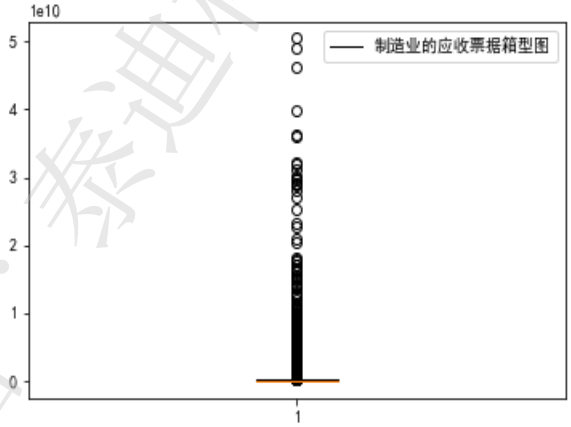

图 3 制造业‘应收票据’箱型图

2.3 标准化

在机器学习之前，由于经过异常值和缺失值处理后的各特征数据其存在或多或少的差异性。各个特征的数据方差、均值等基本统计量存在着较大差异，使得

部分特征数据会对整体模型产生较大的影响，导致其能一定程度上主导最终模型。为了各特征数据所存在的差异性，首先需要将各项数据的方差标准化。

数据的标准化的方法有以下几种主要方法：min-max 标准化、log 函数转换、atan 函数转换、z-score 标准化。其中我们通过客观数据的特点，一一比对后进行选择使用。本次的特征数据存在着特征因子存在最大值和最小值未知的情况，或有超出取值范围的离群数据的情况，这样的特征较符合 z-score 标准化方法的使用情景，故本次标准化处理我们采用的是 z-score 标准化方法。

其主要目的是为了将数据的均值化为 0，标准差化为 1，以便对特征因子筛选。标准化方差如下：

$$\bar{x} = \frac{x - \bar{x}}{std(x)}$$

其中$\bar{x}$为均值；$std(x)$为标准差

四、 机器学习算法介绍

以下逐一介绍本次机器学习算法，包括逻辑回归、支持向量机、随机森林、MLP、XGBoost、GBDT、ADAboost：

a. 逻辑回归

LR（logistic regression）是一个强大的统计学方法，它可以用一个或多个解释变量来表示一个二项式结果[9]。它通过使用逻辑函数来估计概率，从而衡量类别依赖变量和一个或多个独立变量之间的关系，后者服从累计逻辑分布[9]。

b. 支持向量机

SVM 支持向量机主要针对二分类线性分类器，方法是确定一个超平面，使得数据点到超平面的距离最大，这是一个二次规划问题。由于对任何数据集找到

它合适的映射是困难的。因此通常会从常用核函数中选择。常用的核函数有多项式核函数、高斯函数、线性函数等。[10]

c. 随机森林

随机森林是基于 Bagging 思想的集成学习模型，通过构建多个学习器共同完成学习任务，最初由 Ho 提出[1]。该模型 bagging 算法原理类似投票，每次使用一个训练集训练一个弱学习器，有放回地随机抽取 n 次后，根据不同的训练集训练出 n 个弱学习器。对于分类问题，根据所有的弱学习器的投票，进行"少数服从多数"的原则进行最终预测结果。对于回归问题，采取所有学习器的平均值作为最终结果。

d. MLP

MLP 也称多层感知器，是一种前馈人工神经网络模型，其将输入的多个数据集映射到单一的输出数据集上。MLP 多层感知器是一种前向结构的人工神经网络 ANN，映射一组输入向量到一组输出向量。MLP 可以被看做是一个有向图，有多个节点层组成，每一层全连接到下一层。除了输入节点，每个节点都是一个带有非线性激活函数的神经元。使用 BP 反向传播算法的监督学习方法来训练 MLP。MLP 是感知器的推广，克服了感知器不能对线性不可分数据进行识别的弱点。

相对于单层感知器，MLP 多层感知器输出端从一个变到了多个；输入端和输出端之间也不光只有一层，现在有两层：输出层和隐藏层。基于反向传播学习的是典型的前馈网络，其信息处理方向从输入层到各隐层再到输出层，逐层进行。隐层实现对输入空间的非线性映射，输出层实现线性分类，非线性映射方式和线性判别函数可以同时学习。

e. XGBoost

XGboost 由 GBDT、RGF 等算法改进而来是一种基于 Boosting 的集成学习算法。XGboost 的基学习器可以是决策树，也可以是线性模型。对于有少量缺失值的数据集，该模型具有较好的容错能力，可以通过稀疏感知算法自动学习出决策树的分裂方向[14]。在决策树的分裂过程中，使用贪心算法的近似算法寻找最有可能的分裂点。由于采用分布式计算，需要遍历所有 的数据集，XGboost 会更加地消耗计算机内存。[14]

f. GBDT

梯度提升树(gradient boosting decision tree，GBDT)是 2001 年 Friedman 提出的一种 boosting 算法。它是一种迭代的决策树算法，该算法由多棵决策树组成，所有树的结论加起来作为最终答案[15]。具体思想是每次建立模型是在之前建立的模型损失函数的梯度下降方向，而传统的 boosting 思想是对正确和错误的样本进行加权(每一步的结束，增加分错的点的权重，减少分对点的权重)。[15]

g. ADAboost

Adaboost 是一种迭代算法,通过对弱学习算法的加强而得到强学习算法,即通过一个包含关键特征的弱分类器集合,构建出具有理想分类能力的强分类器[16]。Adaboost 算法的优点在于它使用加权后选取的训练数据代替随机选取的训练样本, 将弱分类器联合起来,使用加权的投票机制代替平均投票机制。[16]

算法概述：

1、先通过对 N 个训练样本的学习得到一个弱学习器；

2、将分错的样本和其他的新数据一起构成一个新的 N 个的训练样本，通过

对这个样本的学习得到第二个弱分类器；

3、将 1 和 2 都分错了的样本加上其他的新样本构成另一个新的 N 个的训练样本，通过对这个样本的学习得到第三个弱分类器

4、最终经过提升的强分类器。即某个数据被分为哪一类要由各分类器权值决定。

五、 模型评价

由于数据集样本各行业财务造假公司数量极不平衡，因此需要综合考虑模型的查准率和召回率，本文采用 F1 分数作为模型训练的评价标准，同时 AUC 也可以作为辅助评价的指标。

1. F1 分数

F1 分数，是统计学中用来衡量二分类（或多任务二分类）模型精确度的一种指标。它同时兼顾了分类模型的准确率和召回率。F1 分数可以看作是模型准确率和召回率的一种加权平均，它的最大值是 1，最小值是 0，值越大意味着模型越好。

	真实性 1	真实性 2
预测 1	True Positive(TP)真阳性	False Positive(FP)假阳性
预测 2	False Negatve(FN)假阴性	True Negative(TN)真阴性

查准率，指的是预测值为 1 且真实值也为 1 的样本在预测值为 1 的所有样本中所占的比例。

$$p = \frac{TP}{TP + FP}$$

召回率，也叫查全率，指的是预测值为 1 且真实值也为 1 的样本在真实值为 1 的所有样本中所占的比例。

$$r = \frac{TP}{TP + FN}$$

F1 分数，又称为平衡 F 分数，它被定义为精确率和召回率的调和平均数。

$$F = 2 \times \frac{precision \cdot recall}{precision + recall}$$

2. AUC 值

AUC 被定义为 ROC 曲线下与坐标轴围成的面积，显然这个面积的数值不会大于 1。又由于 ROC 曲线一般都处于 y=x 这条直线的上方，所以 AUC 的取值范围在 0.5 和 1 之间。AUC 越接近 1.0，检测方法真实性越高；等于 0.5 时，则真实性最低，无应用价值。

六、 数据不平衡处理

数据集包含第 1~5 年的样本特征数据和标签，以及第 6 年的样本特征集，将第 1~4 年的样本数据作为训练集，第 5 年的样本数据作为测试集，用机器学习模型进行训练，然后再利用第 6 年的样本特征集预测出它的标签，即可得到公司财务造假情况。在此之前，需先对训练集数据进行不平衡处理。

在传统的有监督机器学习算法中，当数据集样本类别的比例不均衡时，多数类的样本会被过多地关注，这样，少数类样本的分类性能就会受到影响，因此需要对数据不平衡的问题进行处理。处理方法一般有不处理、朴素随机欠采样，朴素随机过采样以及 ADASYN 过采样法。

a. 不处理

不处理即对该数据不做任何不平衡处理。由于各个行业造假公司个数不尽相同，其中就包括 11 个行业的训练集造假个数不超过 2 个，以致于无法对该行业进行均衡处理，因而对这 11 个行业不作任何均衡处理。

其中不做平衡处理的行业包括：

住宿和餐饮业，卫生和社会工作，居民服务、修理和其他服务业，建筑业，教育，文化、体育和娱乐业，水利、环境和公共设施管理业，科学研究和技术服务业，租赁和商务服务业，综合，采矿业。共计 11 个。

b. 朴素随机欠采样

朴素随机欠采样指的是，针对不平衡数据，可以采用从多数类样本中随机选择少量样本，然后再与少数类样本合并，形成新的训练数据集，随机欠采样分为有放回和无放回两种类型

但是欠采样会导致其他问题的产生，运用这种方法可能会使最终的训练数据集丢失了部分数据，模型只学习了一部分数据。

在采取朴素随机欠采样处理之后，通过与其他方法的 AUC 值进行比较可以得出，没有一个行业适合此方法。

c. 朴素随机过采样

此方法与欠采样是相反的，朴素随机过采样是采用生成少数类样本的方法，即从少数类的样本中随机采样以增加新的样本。

采用过采样的方法同样会产生一个问题，过采样后的数据集中会出现重复样本，重复样本在训练时表现肯定较好，因此训练模型会有一定的拟合性。

在采取朴素随机过采样处理之后，同样通过比较可以得出，均衡数据时，此方法最优的行业有交通运输、仓储和邮政业，农、林、牧、渔业，电力、热力、燃气及水生产和供应业共 3 种行业。

d. ADASYN 过采样

ADASYN 自适应综合过抽样，与 Borderline SMOTE 相似，指的是对不同的少数类样本赋予不同的权重，从而生成不同数量的样本。其思想基于根据少数类数据样本的分布自适应地生成少数类数据样本的思想，与那些更容易学习的少数类样本相比，更难学习的少数类样本会生成更多的合成数据。

ADASYN 方法不仅可以减少原始不平衡数据分布带来的学习偏差，还可以自适应地将决策边界转移到难以学习的样本上。在抽取的行业指标中，适合的有信息传输、软件和信息技术服务业，制造业，房地产业，批发和零售业，金融业共 5 种。

各处理方法效果对比

通过前面 4 种处理方法可以获取各个行业不平衡处理后的样本数据，将数据

各个行业在四种不平衡处理方法的AUC值

	直接预测	朴素随机欠采样	朴素随机过采样	ADASYN过采样
交通运输、仓储和邮政业	0.5	0.75	1	0.981012658
住宿和餐饮业	0	0	0	0
信息传输、软件和信息技术服务业	0.479310345	0.714285714	0.973170732	0.983796296
农、林、牧、渔业	0.96875	0.5	0.985714286	0.954545455
制造业	0.533333333	0.529655172	0.715037114	0.788269951
卫生和社会工作	0	0	0	0
居民服务、修理和其他服务业	0	0	0	0
建筑业	0	0	0	0
房地产业	0.583425414	0.55	0.948275862	0.960784314
批发和零售业	0.573252688	0.5	0.972	0.995934959
教育	0	0	0	0
文化、体育和娱乐业	0	0	0	0
水利、环境和公共设施管理业	0	0	0	0
电力、热力、燃气及水生产和供应业	0.497175141	0.5	1	0.995238095
科学研究和技术服务业	0	0	0	0
租赁和商务服务业	0	0	0	0
综合	0	0	0	0
采矿业	0	0	0	0
金融业	0.496875	0.5	0.965909091	0.969201807

图 4 各个行业在四种不平衡处理所训练得出的 AUC 值

用 SVM 模型训练并预测检验，各个行业样本数据通过 4 种方法所得出的检验结果如图所示

其中标蓝的为各个行业在四种不平衡处理所训练得出最高的 AUC 值

七、 特征选择结果分析

在一个学习任务中，所提供的特征集中，有些特征对学习可能是有效的，有些则没有什么作用，特征选择的目的就是从特征集中挑选一组最具有统计意义的特征子集，从而达到降维的效果。

1. 特征选择方法

一般的特征选择方法有以下 4 种：权重法、过滤法、包裹法和嵌入法，其中

过滤法、包裹法和嵌入法数据以制造业为例。

i. 权重法

权重法，用随机森林进行特征重要性评估。主要是看每个特征在随机森林中的每棵树上做了多大贡献，然后取平均值，最后比较不同特征之间的贡献大小。在训练集的数据中，包含 7 种行业存在造假次数，但造假次数小于等于 2。故对此 7 种行业的特征选择进行权重法分析。

ii. 过滤法

过滤法，先采用特征选择对初始特征进行筛选，然后用筛选后的特征来训练模型。

本文选择单变量特征选择的互相关度量以及卡方检验方法来筛选特征，其原理是单独计算每个变量的统计指标，根据该指标来给特征评分，选择评分排名前 20 的特征，计算结果如表 1 所示

表 1　过滤法选择的特征

特征选择方法	选择结果
互相关度量	2 4 5 7 9 11 17 18 20 21 23 27 28 50 55 68 69 78 82 87
卡方检验	7 13 14 15 37 53 57 58 60 61 65 67 79 80 81 83 85 90 91 96

iii. 包裹法

包裹法考虑所使用的学习器，将学习器的性能作为特征子集选择的评价依据，以此选择对学习器有利的特征子集。

本文采用逻辑回归递归特征消除法和随机森林回归递归特征消除法进行特征选择，计算结果如表 2 所示

表 2　包裹法选择的特征

特征选择方法	选择结果
逻辑回归递归特征消除法	6 12 13 16 31 33 36 37 43 44 48 63 67 74 79 80 82 88 90 96
随机森林回归递归特征消除法	9 11 19 21 23 27 36 44 47 53 54 57 63 67 69 78 82 87 93 94

iv. 嵌入法

对过滤法和包裹法，特征选择过程与模型训练过程是独立进行的，而嵌入法则是综合考虑这两个过程，在学习的同时进行特征选择。

常用的嵌入法有两种：

一是基于惩罚项的特征选择法，即用正则 L1 范数作为惩罚项。L1 范数不但可以降低过拟合风险，还可以使求得的 w 有较多分量为 0。所以当希望减少特征的维度以用于其他分类器时，可以选择不为 0 的系数所对应的特征。本文采用基于 SVM 模型、Lasso 模型、逻辑回归模型的惩罚项特征选择法来选择特征，计算结果如表 3 所示。

二是树模型的特征选择法，这种方法能够用来计算特征的重要程度，因此可以用来去除不相关的特征。本文采用随机森林模型、决策树模型的树模型特征选择法来选择特征。计算结果如表 4 所示

表 3　基于惩罚项的嵌入法选取的特征

特征选择方法	选择结果
SVM 模型	7 13 14 17 32 38 44 45 61 62 66 79 80 81 82 86 87 91 92 97
Lasso 模型	7 17 32 34 38 44 45 50 59 61 62 66 77 80 82 85 86 87 91 97
逻辑回归模型	7 13 16 17 32 34 38 44 45 62 66 72 79 80 81 82 87 91 92 97

表 4　基于树模型的嵌入法选取的特征

特征选取方法	选择结果
随机森林模型	1 3 10 12 15 18 22 24 25 29 43 48 59 61 67 76 79 83 88 91
决策树模型	7 15 16 17 34 37 38 44 45 55 58 61 62 66 79 80 81 91 93 97

2. 特征选择结果分析

统计所有方法获取的特征（以制造业为例），将出现次数大于等于 4 次的特

征挑选为最终的特征，共选择了 19 个特征。结果如表 5 所示，其中包含长期股权投资、其他非流动资产、短期借款、递延所得税负债、负债合计、归属于母公司所有者权益合计、所有者权益(或股东权益)合计、负债和所有者权益(或股东权益)总计、支付其他与筹资活动有关的现金等特征，它们均具有经济学意义，故为最终特征。

表 5　最终选取的特征

选取特征结果		
长期股权投资	其他非流动资产	短期借款
递延所得税负债	负债合计	归属于母公司所有者权益合计
所有者权益(或股东权益)合计	负债和所有者权益(或股东权益)总计	其他综合收益
支付其他与筹资活动有关的现金	投资支付的现金	收到的税费返还
归属于少数股东的综合收益总额	货币资金	其他应收款
存货	在建工程	无形资产
递延所得税资产		

同理，以下表 6 为各行业与财务数据造假相关的数据指标。

表 6　各行业与财务数据造假相关的数据指标

行业	选取特征结果
交通运输、仓储和邮政业	应收票据、预付款项、长期股权投资、其他应收款、短期借款、其他流动资产、递延收益、流动资产合计、递延所得税负债、在建工程、归属于母公司所有者权益合计、递延所得税资产、少数股东权益、非流动资产合计、所有者权益(或股东权益)合计、预收款项、收到的税费返还、应付职工薪酬、资产减值损失、应交税费、持续经营净利润、购买商品、接受劳务支付的现金、货币资金
信息传输、软件和信息技术服务业	应收票据、汇率变动对现金及现金等价物的影响、长期待摊费用、其他综合收益、短期借款、持续经营净利润、递延收益、货币资金、负债合计、预付款项、归属于母公司所有者权益合计、流动资产合计、所有者权益(或股东权益)合计、在建工程、收回投资收到的现金、预收款项
农、林、牧、渔业	长期股权投资、持续经营净利润、应付票据、归属于少数股东的综合收益总额、应付利息、应收账款、递延所得税负债、其他流动资产、少数股东权益、流动资产合计、负债和所有者权益(或股东权益)总、在建工程、吸收投资收到的现金、递延所得税资产、收到的税费返还 分配股利、利润或偿付利息支付的现金
制造业	长期股权投资、投资支付的现金、其他非流动资产、收到的税费返还、短期借款、归属于少数股东的综合收益总额、递延所得税负债、货币资金、负债合计、其他应收款、归属于母公司所有者权益合计、存货、所有者权益(或股东权益)合计、在建工程、负债和所有者权益(或股东权益)总计、无形资产、其他综合收益、递延所得税资产、支付其他与筹资活动有关的现金
建筑业	可供出售金融资产、无形资产、长期待摊费用、投资活动现金流入小计、少数股东权益、支付其他与经营活动有关的现金、其他综合收益、现金及现金等价物净增加额、投资支付的现金、收到其他与经营活动有关的现金、收到的税费返还、筹资活动产生的现金流量净额、其他综合收益、期末现金及现金等价物余额、持续经营净利润、营业外支出、预付款项、营业外收入、在建工程
房地产业	应收票据、预付款项、可供出售金融资、其他应收款、资产总计、流动资产合计、投资支付的现金、无形资产、取得投资收益收到的现金、递延所得税资产、汇率变动对现金及现金等价物的影响、应付账款、持续经营净利润、盈余公积、货币资金
批发和零售业	应收票据、货币资金、递延所得税负债、存货、归属于母公司所有者权益合计、无形资产、少数股东权益、应付职工薪酬、其他综合收益、应交税费、其他综合收益、其他应付款、持续经营净利润、流动负债合计、归属于少数股东的综合收益总额
教育	长期待摊费用、取得借款收到的现金、递延收益、支付的各项税费、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、应收账款、投资活动现金流出小计、其他应收款、加:期初现金及现金等价物余额、应付账款、期末现金及现金等价物余额、预收款项、筹资活动现金流出小计、应付职工薪酬、财务费用、应交税费、销售费用、盈余公积
文化、体育和娱乐业	长期股权投资、处置固定资产、无形资产和其他长期资产收回的现金净额、递延收益、筹资活动产生的现金流量净额、收回投资收到的现金、支付给职工以及为职工支付的现金、收到其他与投资活动有关的现金、稀释每股收益、资产减值损失、利润总额(亏损总额以"－"号填列)、存货、基本每股收益、预收款项、营业利润(亏损以"－"号填列)、未分配利润、归属于母公司所有者(或股东)的综合收益总额、投资活动产生的现金流量净额、综合收益总额、分配股利、利润或偿付利息支付的现金、归属于母公司所有者(或股东)的净利润
水利、环境和公共设施管理业	收到的税费返还、基本每股收益、资产减值损失、营业利润(亏损以"－"号填列)、应收账款、归属于母公司所有者(或股东)的综合收益总额、应付账款、综合收益总额、未分配利润、归属于母公司所有者(或股东)的净利润、稀释每股收益、净利润(净亏损以"－"号填列)、营业收入、所得税费用、利润总额(亏损总额以"－"号填列)、少数股东损益
电力、热力、燃气及水生产和供应业	资产总计、资本公积、吸收投资收到的现金、未分配利润、收到其他与投资活动有关的现金、投资活动产生的现金流量净额、其他流动资产、取得借款收到的现金、流动资产合计、购买商品、接受劳务支付的现金、应付账款、销售商品、提供劳务收到的现金
科学研究和技术服务业	所有者权益(或股东权益)合计、投资活动现金流出小计、支付其他与筹资活动有关的现金、支付给职工以及为职工支付的现金、应交税费、稀释每股收益、其他应付款、营业收入、实收资本(或股本)、营业总收入、盈余公积、基本每股收益、支付其他与经营活动有关的现金、综合收益总额、销售商品、提供劳务收到的现金、归属于母公司所有者(或股东)的净利润、经营活动现金流入小计、营业成本、经营活动现金流出小计、营业总成本
综合	应收票据、盈余公积、可供出售金融资产、投资活动产生的现金流量净额、长期待摊费用、经营活动产生的现金流量净额、递延所得税负债、现金及现金等价物净增加额、汇率变动对现金及现金等价物的影响、投资收益(损失以"－"号填列)其中:对联营企业和合营企业的投资收益、所得税费用、持续经营净利润、营业外收入
采矿业	应收票据、投资活动现金流入小计、可供出售金融资产、支付其他与经营活动有关的现金、长期股权投资、收到其他与经营活动有关的现金、长期待摊费用、稀释每股收益、其他非流动资产、利润总额(亏损总额以"－"号填列)、应付票据、基本每股收益、支付其他与筹资活动有关的现金、营业利润(亏损以"－"号填列)、吸收投资收到的现金、销售费用、收到的税费返还、净利润(净亏损以"－"号填列)、持续经营净利润
金融业	应付票据 应付账款、所有者权益(或股东权益)合计、应交税费、投资支付的现金、其他应付款、收回投资收到的现金、流动负债合计、收到的税费返还、实收资本(或股本)、资产减值损失、投资活动产生的现金流量净额、货币资金、取得借款收到的现金、在建工程、投资活动现金流出小计、非流动资产合计

将特征选择方法中所介绍的各个方法选择出来的特征进行匹配，为了更好反应特征之间的相关关系，构造以下公式并作热图（如图 5）

$$a_{ij} = \frac{b_{ij}}{c_i}$$

其中，a_{ij}为第 i 种行业相对第 j 种行业的特征相似度；b_{ij}为第 i 种行业与第 j 种行业相同特征个数；c_i为第 i 种行业的特选个数

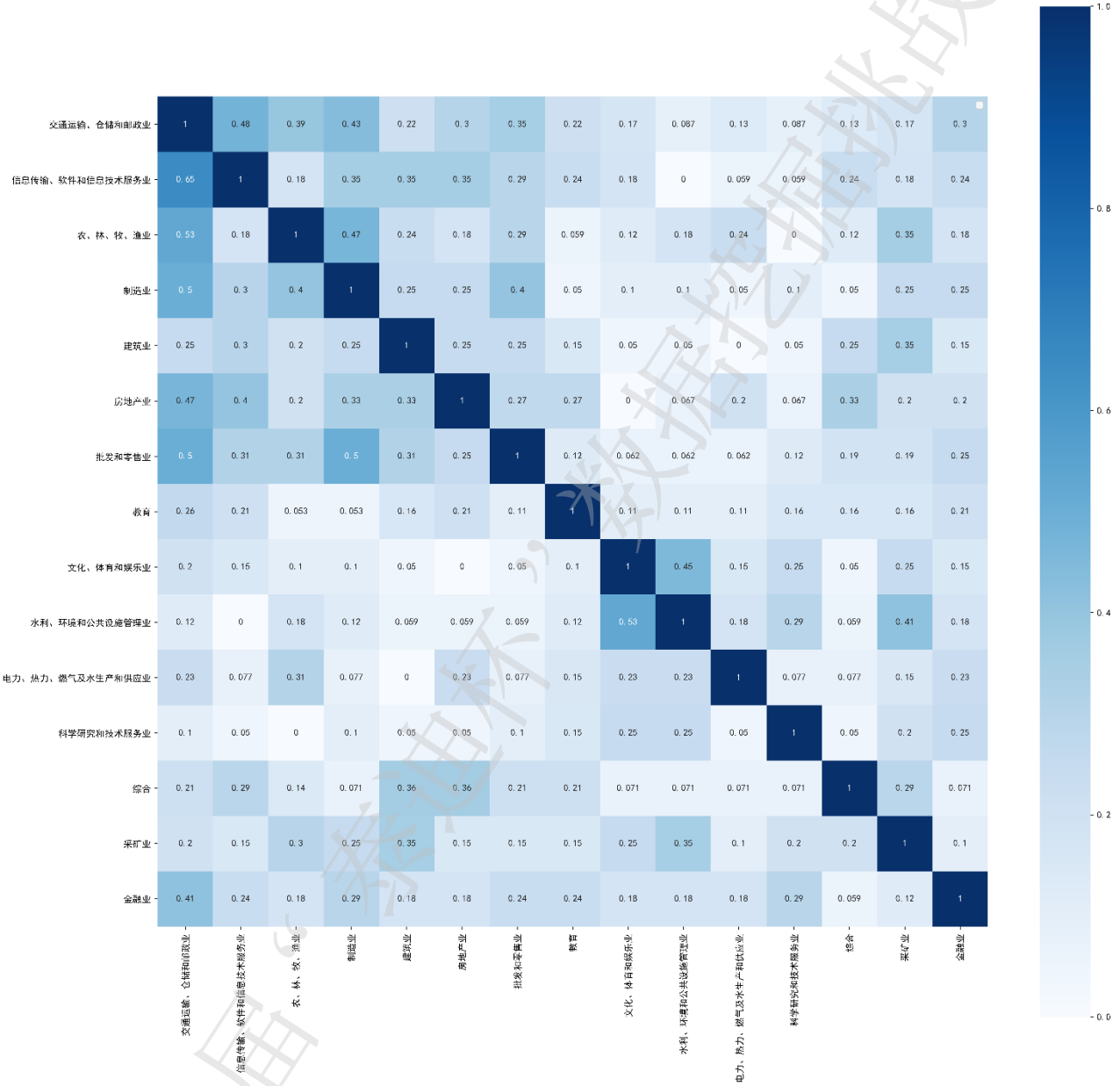

图 5　特征因子相关性热图

从图 5 中我们可以发现，各种方法选择的特征都有一定的重合度，可以很好的反映出以下两点：

1、各行业与财务造假相关的数据指标

2、不同行业上市公司相关数据指标的异同

从热图的颜色深浅可以明显得到，信息传输、软件和信息技术服务业与交通运输、仓储和邮政业两个行业的相关性是最高的，为 0.65。具体财务造假指标异同如图 6。

而像房地产业与文化、体育和娱乐业两个行业的相关性为 0，即毫无相同的造假指标。

信息传输、软件和信息技术服务业与交通运输、仓储和邮政业指标异同		
同	应收票据、预付款项、短期借款、递延收益、流动资产合计、在建工程、归属于母公司所有者权益合计、所有者权益(或股东权益)合计、持续经营净利润、货币资金、预收款项	
异	信息传输、软件和信息技术服务业	长期股权投资、其他应收款、其他流动资产、递延所得税负债、递延所得税资产、少数股东权益、非流动资产合计、收到的税费返还、应付职工薪酬、资产减值损失、应交税费、、购买商品、接受劳务支付的现金、
	交通运输、仓储和邮政业	汇率变动对现金及现金等价物的影响、长期待摊费用、其他综合收益、短期借款、递延收益、负债合计、收回投资收到的现金

图 6

八、 超参数调整

1. 超参数调整方法

本文主要选择网络搜索，也称为穷举搜索，是一种重要的调参手段，采用的方法是将所有可能的参数值组合放入学习器进行训练，选取其中表现最好的参数值组合作为最终结果，网络搜索可以保证在指定的参数范围内找到最优的参数值。

通过数据不平衡处理得到训练集的样本之后，用模型对其进行训练，通过调参后的 AUC 值作为评价指标。

2. 各模型参数调优（以交通运输、仓储和邮政业为例）

建模时先固定每个参数的初始值，再设定其调参范围，进行网络搜索和交叉验证寻找最优化结果。其中设置的初始值、范围和调参结果见各算法框架参数结果详情表

MLP 调参

参数名称	调参范围	调参结果	调参后 AUC
最大迭代次数（'max_iter）	【100,5000】	1770	
随机状态（random_state）	【1,30】	18	0.665094
隐藏层大小（hidden_layer_sizes）	【1,200】	87	

ADAboost 调参

参数名称	调参范围	调参结果	调参后 AUC
决策树数目（n_estimators）	【10,1000】	70	0.721698
学习率（learning_rate）	【0.01,1】	0.218421	

随机森林调参

参数名称	调参范围	调参结果	调参后 AUC
决策树数目（n_estimators）	【10,1000】	20	0.778302
决策树最大深度（max_depth）	【1,17】	11	

逻辑回归调参

参数名称	调参范围	调参结果	调参后 AUC
最大迭代次数（'max_iter）	【100,1000】	640	0.613208
惩罚系数（C）	【0.001,1】	0.143714	

支持向量机调参

参数名称	调参范围	调参结果	调参后 AUC
惩罚系数（C）	【0.001,1】	0.053579	0.814725
分布系数（gamma）	【0.1,1】	0.816327	

Xgboost 调参

参数名称	调参范围	调参结果	调参后 AUC
学习率（learning_rate）	【0.1,1】	0.9	0.127359
决策树数目（n_estimators）	【10,1000】	830	
决策树最大深度（max_depth）	【1,20】	6	

GBDT 调参

参数名称	调参范围	调参结果	调参后 AUC
学习率（learning_rate）	【0.1,1】	0.3	0.857436
决策树数目（n_estimators）	【10,1000】	220	
决策树最大深度（max_depth）	【1,20】	15	
随机状态（random_state）	【5,30】	16	

3. 各行业最优模型的 AUC 值

下表为各种行业最优的调参模型以及 AUC 值，其中包含 8 种行业赋值为-1，后续运用权重法进行预测。

表 7　各行业最优调参模型

行业	模型	AUC 值
交通运输、仓储和邮政业	GBDT	0.857436
信息传输、软件和信息技术服务业	ADAboost	0.668750
农、林、牧、渔业	GBDT	0.737500
制造业	GBDT	0.650000
建筑业	NULL	-1.000000
房地产业	ADAboost	0.995763
批发和零售业	RF	0.680723
教育	NULL	-1.000000
文化、体育和娱乐业	NULL	-1.000000
水利、环境和公共设施管理业	NULL	-1.000000
电力、热力、燃气及水生产和供应业	SVM	0.948718
科学研究和技术服务业	NULL	-1.000000
综合	NULL	-1.000000
采矿业	NULL	-1.000000
金融业	RF	0.5800000

九、 模型预测

本文使用 7 种模型，即 XGboost、GBDT、LR、SVM、MLP、随机森林、和 ADAboost 分别对各行业第六年的样本特征集做预测，其中包含 7 种行业使用 SVM 权重法预测，而非上述 7 种模型。

针对问题 2，发现 GBDT 模型更适合制造业去预测第六年该行业财务造假的上市公司，下图分别为制造业财务造假分布以及其他行业财务造假分布：

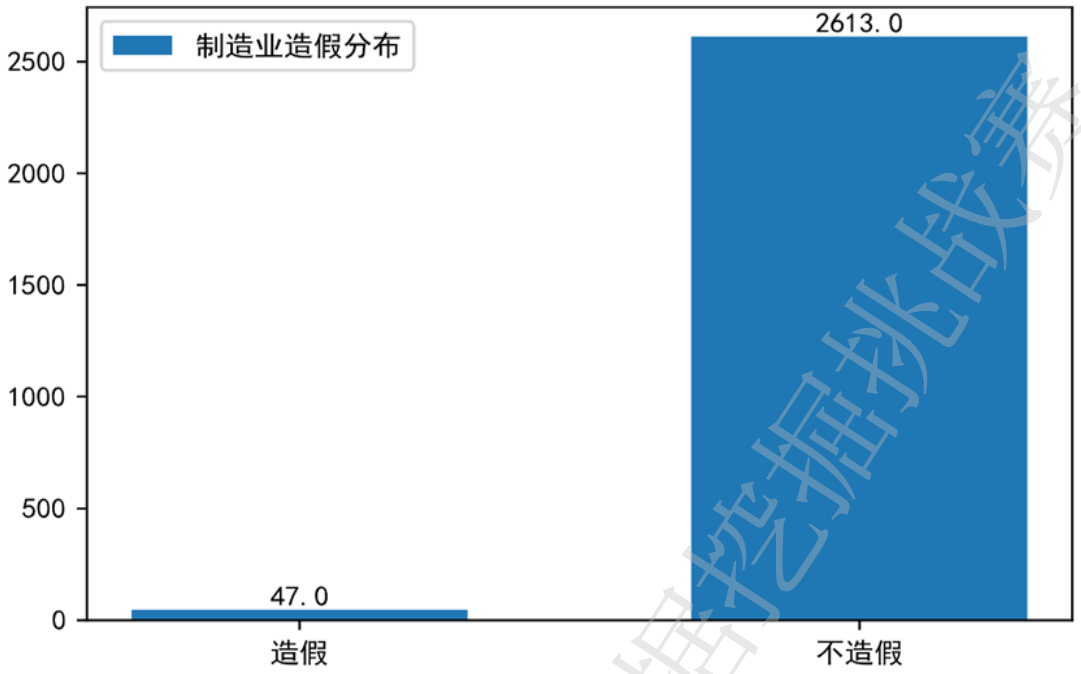

图 7 制造业财务造假分布

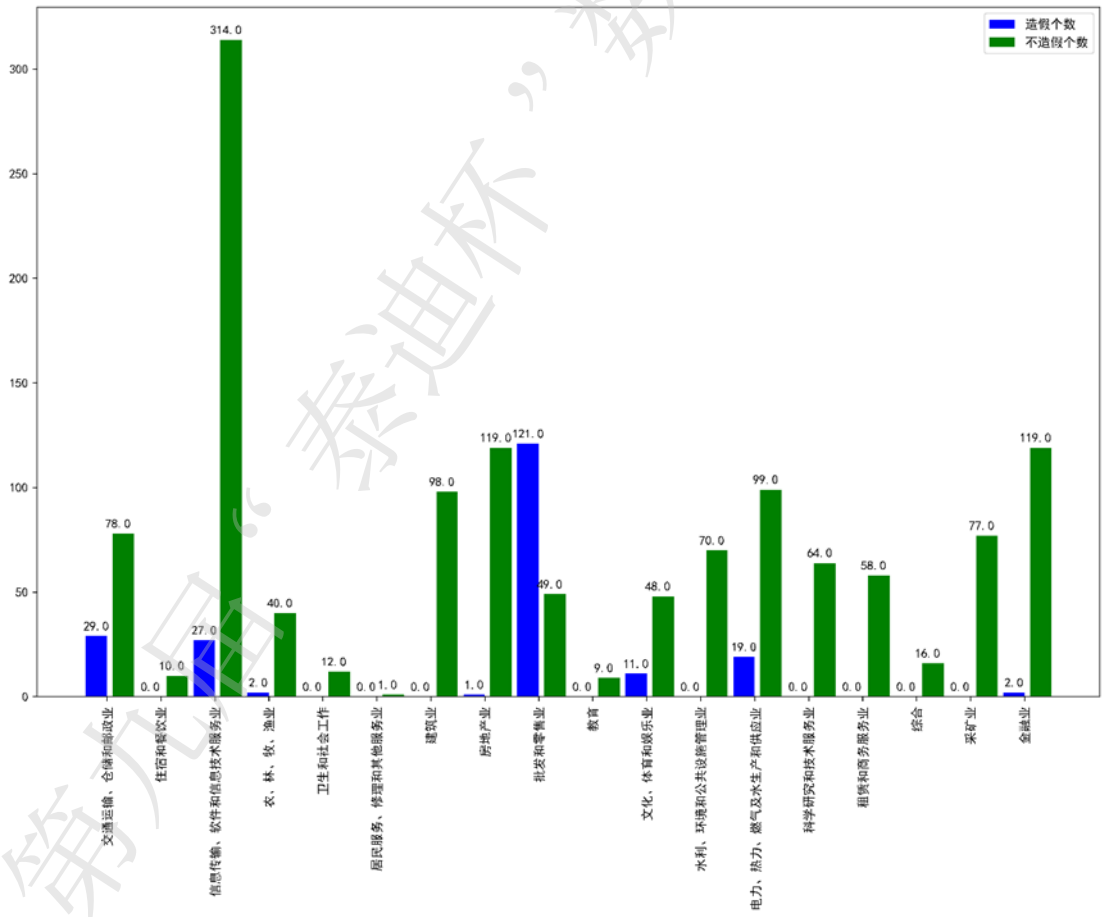

图 8 其他行业财务造假分布

制造业造假上市公司股票代码：16916, 75645, 325632, 393847, 610185, 694894, 978338, 986020, 1145258, 1148273, 1299232, 1343104, 1430872, 1468966, 1508932, 1794373, 1876786, 2077309, 2106304, 2239503, 2341416, 2503243, 2615523, 2643320, 2656200, 2673748, 2704109, 2848498, 2850127, 2873110, 2941088, 3027803, 3342777, 3344750, 3451668, 3646698, 3682953, 3716093, 3766507, 3852602, 3869439, 3914715, 4008179, 4087185, 4457516, 4864298, 4930214

其他行业同理可求。

十、 结语

本文采用机器学习模型预测各行业上市公司是否存在财务造假情况，主要目的是在投资之前获得可能存在财务造假的公司，从而及时避免损失。最终得出以下结论：

本文分别将附件 1 和附件 3 数据指标合并到附件 2，合并后的数据进行预处理操作，数据不平衡处理以及特征选择之后，使用 7 个机器学习模型进行训练预测，通过比较各模型的评价结果来筛选最优模型预测，具有一定的合理性。

本文在数据不平衡处理采用不处理、朴素随机欠采样、朴素随机过采样、ADASYN 过采样四种方式进行比较筛选，通过比较运用处理各行业表现较好的方式。

本文使用特征选择的方法提取特征，并进行综合统计后得出特征，发现选择的特征从经济学的角度具有意义，且挑选的特征进行模型训练之后，得到的结果也是较好的，因此各行业挑选出来的特征因子是有效的。根据热图也可以明显分析各行业财务造假指标的异同。

在预测模型的选择上，本文通过运用 LR、RFC、SVM、MLP、ADAboost、XGBoost 和 GBDT 这 7 种模型对各个行业的上市公司进行造假预测，选择评价结

果最好，最优的模型训练出来的数据作为结果。

预测第6年制造业上市公司财务造假情况结果：其中财务造假个数有47个。不造假个数有2613个，造假比例约为1.77%，比较合理。其他行业同理。

综上所述，本文采用的策略与方法有一定的准确性和现实意义，可以作为分析各行业上市公司是否存在财务造假的有效模型。

参考文献

[1] 李航. 统计学习方法[M]. 北京:清华大学出版社,2012.

[2] 朱福喜. 人工智能[M].3 版. 北京:清华大学出版社,2012.

[3] 王小涵,王育红. 上市公司财务造假与违规手段研究[J]. 财会通讯,2021

[4] 袁先智,周云鹏,严诚幸,刘海洋,曾途. 公司财务欺诈预警与风险特征筛选的新方法:基于人工智能算法[C]. 第十五届中国管理学年会论文集,2020

[5] 何清,李宁,罗文娟,史忠植. 大数据下的机器学习算法综述[J]. 模式识别与人工智能,2014

[6] 李邵琴,杜建强,聂 斌,熊旺平,黄灿奕,李 欢. 特征选择方法综述[J]. 计算机工程与应用,2019

[7] 刘星毅,农国才. 几种不同缺失值填充方法的比较[J]. 南宁师范高等专科学校学报,2007

[8] 晔沙. 数据缺失及其处理方法综述[J]. 网络与信息工程,2017

[9] 吕晨,程建华. 基于 Logistic 模型的上市公司财务造假识别研究[J]. 中原工学院学报,2020

[10] 胡越,罗东阳,花奎,路海明,张学工. 关于深度学习的综述与讨论[J]. 信息与控制,2018

[11] 张康林,叶春明. 基于改进随机森林的企业破产预测研究[J]. 科技促进发展,2021

[12] 何文琴,杨仕晓. 基于随机森林模型的 P2P 网贷平台违约风险量化评价研究[J]. 现代营销(下旬刊),2021

[13] 李永丽,王 浩,金喜子. 基于随机森林优化的自组织神经网络算法[J]. 吉林大学学报,2021

[14] 范诗语,耿子悦,田芮绮,杜永强. 基于集成学习的上市企业违约风险评价[J]. 统计与管理,2021

[15] 蔡文学,罗永豪,张冠湘,钟慧玲. 基于 GBDT 与 Logistic 回归融合的个人信贷风险评估模

型及实证分析[J]. 管理现代化,2017

[16] 李 翔,朱全银. Adaboost 算法改进 BP 神经网络预测研究[J]. 计算机工程与科学,2013