上市公司财务数据造假分析预测

摘要

财务数据造假是指上市公司伪造财务数据，使财务信息歪曲反映经济活动，以此达到特定利益的集团或个人的不正当违法犯罪行为。专业投资者，对上市公司多年的财务数据报告，筛选数据指标进行跟踪分析和研究，识别真伪，则能避免投资踩雷。

针对问题一，对数据进行预处理，采用人工处理异常值和切比雪夫定理及随机森林模型相结合的方法对数据进行异常值的检测和处理。由方差过滤法、卡方检验过滤法、Pearson 相关系数过滤法、递归特征消除法、基于 L1 正则项的特征选择法、基于随机森林模型的特征选择法、基于 GBDT 模型的特征选择法，在这七种特征选择模型中，通过评价指标 AUC 对比，选择效果较好的 Pearson 相关系数过滤法、递归特征消除法和基于随机森林模型的特征选择法、基于 GBDT 模型的特征选择法四个模型，选择算法特征重要性排名前 20 个重要特征，挑出四个模型共同确定的特征因子，从而得出问题一中对上市公司财务数据造假有较大影响的因子。

针对问题二，在处理数据类别不均衡上，采用周志华实验室提出的欠采样的算法 EasyEnsemble 处理数据集。在预测模型的选择上，本文选择用了基于 Stacking 集成学习模型，第一层基学习器用了 LR、随机森林、SVM、GBDT、XGBoost、Lightgbm，第二层元学习器选择了 LR，从而确定了最优的 Stacking 集成学习预测模型。Stacking 集成学习预测模型在测试集上的 AUC 得分为 86.31% ，高于所有基础分类器，可见建立的模型较为稳定，不存在严重的过拟合且效果较好，并由此模型求出第六年的预测结果。

针对问题三，在问题二的基础上，考虑其余 18 个行业的数据量过少的情况，依据行业数据量的多少将 18 个行业划分成大、中、小三个不同的类别，添加行业所属类别，利用 OneHotEncoder、LabelEncoder 扩增特征因子，基于 AUC 评价指标，在 KNN、SVM、LR、RFC、GBDT 单模型中选择训练效果较好的进行调参和预测，求出除制造业外各行业上市公司第 6 年财务数据造假的上市公司。

本文利用机器学习算法，充分使用上市公司的财务数据，融合多种算法，建立起较为稳定的 Stacking 集成学习预测模型，具有较大的参考价值和现实意义。

关键词： EasyEnsemble 算法　机器学习　AUC 指标　Stacking 集成学习预测模型

Abstract

Financial data fraud refers to listed companies forge financial data, so that financial information distorted to reflect economic activities, in order to achieve the specific interests of the group or individual's illegal and criminal behavior. Professional investors can conduct tracking analysis and research on the financial data reports of listed companies for many years, screen the data indicators, and identify the truth and falsification, so as to avoid investment thunder.

In view of problem 1, the data is preprocessed, and the method combining artificial outliers processing with Chebyshev theorem and random forest model is used to detect and process the outliers. By variance filtering method, chi-square test filtering method, Pearson correlation coefficient filtering method, recursive feature elimination method, feature selection method based on L1 regular term, feature selection method based on random forest model, feature selection method based on GBDT model, in these seven feature selection models, through the evaluation index AUC comparison, Four models, namely, Pearson correlation coefficient filtering method, recursive feature elimination method, feature selection method based on random forest model and feature selection method based on GBDT model, were selected. The top 20 important features ranked by the algorithm were selected, and the feature factors jointly determined by the four models were picked out. So we get the factors that have great influence on the financial data fraud of listed companies in the first question.

For problem two, in the process of data category imbalance, using EasyEnsemble algorithm to process the data set, prediction model selection, this paper chose to use the Stacking based integrated learning model, the first layer based learning device using LR, random forest, SVM, GBDT, XGBoost, Lightgbm, The second layer of elementary learner selects LR to determine the optimal STACKING ensemble learning prediction model. The AUC score of Stacking ensemble learning prediction model on the test set is 86.31%, which is higher than that of all the basic classifiers. It can be seen that the established model is relatively stable, there is no serious overfitting and the effect is good, and the prediction results of the sixth year are obtained from this model.

For question 3, based on question 2, considering the fact that the data amount of the remaining 18 industries is too small, the 18 industries are divided into three different categories according to the data amount of the other 18 industries. The category of the industries is added, and the feature factors are amplified by OneHotenCoder and LabelEncoder. Based on the AUC evaluation index, In the single model of KNN, SVM, LR, RFC and GBDT, the training effect is better, and the parameters are adjusted and predicted, and the listed companies in all industries except the manufacturing industry are found out that the financial data of the listed companies in the sixth year is false.

In this paper, the use of machine learning algorithm, make full use of the financial data of listed companies, fusion of a variety of algorithms, to establish a more stable Stacking integrated learning prediction model, has a greater reference value and practical significance.

Key words: EasyEnsemble algorithm Machine learning AUC index Stacking ensemble learning prediction model

目录

第 1 章 绪论

1.1 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的同时，上市公司财务数据造假及暴雷的情况越来越频繁，甚至还出现了流动性危机及信用债违约等问题，急需监管部门对上市公司进行有效监控。

建立健全的常态化退市机制是中国资本市场的必经之路。加大监管力度，对出现严重财务数据造假、丧失持续经营能力的上市公司，使其强制退市。但是上市公司的退市必定会给投资者带来损失，因此投资者在选择投资品种时，有必要对上市公司的财务数据进行深入的分析研究。

通过对上市公司多年的财务数据报告，筛选数据指标进行跟踪分析和研究，可以判断上市公司的财务是否稳定，识别财务数据真伪，避免投资踩雷，而这些需要研究者去挖掘。

1.2 问题重述

（1）根据上市公司的行业分类和上市公司的财务数据，确定出各行业与财务数据造假相关的数据指标，并分析比较不同行业上市公司相关数据指标的异同。

（2）根据制造业各上市公司的财务数据，确定出第 6 年财务数据造假的上市公司。

（3）根据上市公司财务数据中其他（除制造业外）各行业上市公司的财务数据，确定出第 6 年财务数据造假的上市公司。

1.3 本文主要工作与创新点

（1）对数据的预处理

在对数据进行预处理时，采用人工处理异常值和切比雪夫定理相结合的方法对数据进行异常值的检测，此外还进行了随机森林模型对异常样本进行了检测和处理。

（2）通过数据分析筛选对上市公司财务造假有较大影响的数据指标

本文将特征工程筛选后的特征因子数据，利用机器学习算法 Pearson 相关系数过滤法、递归特征消除（Recursive Feature Elimination）和随机森林模型的特征选择、基于 GBDT 模型的特征选择法，四种机器学习算法中特征重要性的数值得出因子重要性为前 20 的因子，再由这 4 个模型共同确定有较大影响的特征因子。

（3）机器学习算法分类预测下一年哪一些公司将会发生财务造假

对样本不平衡而言，由于提供的样本数据存在严重的样本不平衡情况，为避免出现数据

的过拟合给原来的样本数据带入噪音，采用周志华实验室提出的欠采样的算法EasyEnsemble[13]，即利用集成学习机制，将反例划分为若干个集合供不同学习器使用，这样对每个学习器来看都进行了欠采样，但是全局却不会丢失重要的信息。

对模型构建而言，本文对问题一特征工程后确定的特征因子，使用五种不同类型的机器学习算法去预测下一年哪些上市公司可能会发生财务数据造假，基于AUC指标与网格调参给机器学习算法参数调优，再用模型融合，建立Stacking集成学习预测模型。

（4）行业分类解决样本数量过少问题

在问题二的基础上，考虑其余18个行业的数据量过少的情况，先将18个行业划分成三个大的类别，添加行业所属类别，利用OneHotEncoder、LabelEncoder扩增特征因子，然后以AUC为评价指标，在KNN、SVM、LR、RFC、GBDT等单模型中选择效果较好的进行训练和预测，求出除制造业外各行业上市公司第 6 年财务数据造假的上市公司。

1.4 模型假设

（1）假设所获得的数据是真实可靠的。

（2）假设第 6 年未发生重大事件和灾难或国家未推行重要政策影响证券市场。

1.5 本文研究意义

从金融市场的角度来看，我国资本市场目前发展已经步入稳定，这就使得金融投资成为了资本投资的主要方向，不过由于经济体制以及市场化程度的发展深度，我国资本市场金融投资仍旧存在许多风险要素，必须要通过一些切实的模型预测对有可能出现的风险进行防控，维护金融市场的稳定。

从投资者的角度来看，筛选相应的数据指标，建立合理有效的模型对上市公司的财务数据是否造假进行监控，避免由于上市公司财务数据造假使投资者投资踩雷，而造成损失，维护广大投资者的利益具有一定的意义。

第 2 章 相关理论

2.1 财务数据造假相关知识介绍

财务数据造假是指上市公司事前经过安排，故意以欺诈、舞弊等手段，伪造、变造虚假财务数据，使财务信息歪曲反映经济活动，以此达到特定利益的集团或个人的不正当违法犯罪行为。

2.2 机器学习算法介绍

2.2.1 LogisticRegressor

逻辑回归是机器学习中一个非常经典的分类模型，它是一种分类方法，主要用于两分类问题（即输出只有两种，分别代表两个类别），Logistic 回归的本质是：假设数据服从这个分布，然后使用极大似然估计做参数的估计。

它将数据拟合到 sigmiod 函数，其函数形式为 $g(z)=\frac{1}{1+e^{z}}$，寻找预测函数:$h_\theta(x)=g(\theta^T x)=\frac{1}{1+e^{\theta^T x}}$，函数的值有特殊的含义，它表示结果取 1 的概率，因此对于输入 x 分类结果为类别 1 和类别 0 的概率分别为：

$$P(y=1|x,\theta)=h_\theta(x), P(y=0|x,\theta)=1-h_\theta(x)$$

构造损失函数：

$$J(\theta)=\frac{1}{m}\sum_{i=1}^{n} Cost(h_\theta(x^{(i)}),y^{(i)})$$

其中 $Cost(h_\theta(x),y)=\begin{cases}-\log\left(h_\theta(x)\right) if\ y=1\\ -\log\left(1-h_\theta(x)\right) if\ y=0\end{cases}$

求解使得$J(\theta)$函数最小并求得回归参数。

2.2.2 RandomForest

随机森林是模型集成中Bagging方法的典型代表，通过对样本或者变量的n次随机采样，就可以得到n个样本集。对于每一个样本集，可以独立训练决策树模型，对于n个决策树模型的结果，通过集合策略来得到最终的输出。需要注意的是，这n个决策树模型之间是相对独立的，并不是完全独立的，训练集之间是有交集的。 可以通过Bootstrap Sample（有放回采样）方法实现对样本的随机采样，基于公式

$$\lim_{n\to\infty} 1-\left(1-\frac{1}{n}\right)^n=1-\frac{1}{e}\approx 63.2\%$$

每次采样大约会有63.2%的样本被选中，该方法同样适用于对变量进行随机抽取随机森林的优点。

在特征和数据的随机采样方面:它可以处理很高维度(特征很多)的数据，并且不用降维，无需做特征选择。对于不平衡的数据集来说，它可以平衡误差。如果有很大一部分特征遗失，仍然可以维持准确度，可以判断出不同特征之间的互相影响(类似于控制变量法)，不容易过拟合。

在树模型的特性方面:它可以判断特征的重要程度。

在算法结构方面：它的训练速度比较快，容易做成并行方法，实现起来比较简单。下面是随机森林算法的构造过程：

1）假如有N个样本，则有放回随机选择N个样本（每次随机选择一个样本，然后返回继续选择）。这就选择好了的N个样本来训练一个决策树，作为决策树根节点处的样本；

2）当每个样本有M个属性时，在决策树的每个节点需要分裂时，随机从这M个属性中选取出m个属性，满足条件m<<M。然后从这m个属性采用某种策略（ID3，C4.5，基尼系数）来选择一个属性作为该节点的分裂属性；

3）决策树形成过程中每个节点都要按照步骤2来分裂（即如果下一次该节点选出来的那一个属性是刚刚其父节点分裂时用过的属性，则该节点已经达到叶子节点，无须继续分裂了）。一直到不能够分裂为止。注意整个决策树形成过程中没有进行剪枝。

4）按照步骤1）～3）建立大量的决策树，这样就构造出随机森林了。

2.2.3 SVM

SVM（支持向量机）是一种二类分类模型，它的基本模型是在特征空间中寻找间隔最大化的分离超平面的线性分类器，支持向量机学习方法包含构建由简至繁的模型：线性可分支持向量机（linear support vector machine in linearly separable case）、线性支持向量机（linear support vector machine）及非线性支持向量机（non-linear support vector machine）。

1.当训练样本线性可分时，通过硬间隔最大化，学习一个线性分类器，即线性可分支持向量机。

2.当训练数据近似线性可分时，引入松弛变量，通过软间隔最大化，学习一个线性分类器即线性支持向量机。

3.当训练数据近似线性不可分时，通过使用核技巧及软间隔最大化。

支持向量机的具体求解过程如下：

(1) 假设给定一个特征空间上的训练数据集：

$$T = \{(X_1, Y_1), (X_n, Y_n)\} \in (X \times Y)^n$$

其中，$x_j \in X = R^n, y_i \in Y = \{-1, +1\}(i = 1,2,\ldots,n)$，$x_i$为特征向量。

(2) 选择适当核函数$K(x_i, x_j)$以及参数C，解决优化问题：

$$min\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i y_i \alpha_j y_j K(x_i, y_j) - \sum_{j=1}^{n}\alpha_j$$

$$s.t.\sum_{i=1}^{n}\alpha_i y_i = 0, 0 \le \alpha_i \le C, i = 1,\ldots,n$$

得最优解：$\alpha^* = (\alpha_1^*, \ldots, \alpha_n^*)^T$。

(3) 选取α^*的正分量，计算样本分类阈值：

$$b^* = y_i - \sum_{i=1}^{I} y_i \alpha_i^* K(x_i, x_j)$$

(4) 构造最优判别函数：

$$f(x) = sgn[\sum_{i}^{n} y_i a_i^* K(x_i, x_j) + b^*]$$

支持向量机内积核核函数 K 的主要种类有：

① 线性内核函数 $K(x_i, y_j) = (x_i, y_j)$

② 多项式核函数 $k(x_i, x_j) = [(x_i, x_j) + 1]^q$

③ 高斯径向基核函数(RBF) $K(x_i, x_j) = \exp\left\{-\frac{||x_i - x_j||^2}{\sigma^2}\right\}$

④ 双曲正切核函数(Sigmoid 核函数) $K(x_i, x_j) = \tanh(v(x_i . x_j) + c)$

一般地，用 SVM 做分类预测时必须调整相关参数(特别是惩罚参数 c 和核函数参数 g)，这样才可以获得比较满意的预测分类精度，采用 Cross Validation 的思想可以获取最优的参数，并且有效防止过学习和欠学习状态的产生，从而能够对于测试集合的预测得到较佳的精度。

2.2.4 XGBoost

XGBoost 模型是典型 boosting 算法，是对 GBDT 模型的算法和工程改进。区别 Bagging 模型，基学习器可以并行，boosting 模型的基学习器间存在先后依赖。GBDT 是一种提升树模型，第 m 轮用一棵 arct 回归树拟合前 m－1 轮损失的负梯度，降低模型的 bias。

XGBoost 是在 GBDT 等提升算法基础上进行优化的算法，引入二阶导数信息，并加入正则项控制模型的复杂度；此外，虽然基模型的训练存在先后顺序，但每个基学习器内部的树节点分裂可以并行，XGBoost 对此进行了并行优化，实现优化目标函数以达到误差和复杂度综合最优。其原理如下：

目标函数 L(x)由误差函数 F(x)和复杂度函数Ω(x)组成：

$$L(x) = F(x) + \Omega(x) \quad m-1$$

$$L(x) = \sum_{i} I(y_i, \hat{y}_i) + \gamma T + \frac{1}{2}\lambda \sum_{J=1}^{T} ||w_j||^2$$

其中 I 是用来衡量与 y 的相近程度的可导且凸的损失函数，通过每一步增加一个基分类器，贪婪地去优化目标函数，使得每次增加都使得损失变小。然后让后一次迭代的基分类器去学习前一次遗留下来的误差。这样可以得到用于评价当前分类器性能的评价函数,如下：

$$L_m(X) = \sum_{i} I\left[y_i, \hat{y}_i^{m-1} + f_m(x_j)\right] + \Omega(f_m)$$

这个算法又可以成为前向分步优化。为了更好更快的优化此函数，可以在 附近二阶泰勒展开，泰勒展开的形式为公式。

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

令$g_i = \frac{\partial I(y_i, \hat{y}(t-1))}{\partial}$，$h_i = \frac{\partial^2 I(y_i, \hat{y}(t-1))}{\partial(\hat{y}(t-1))^2}$最后可得到目标函数，在剔除常数项后可以得到最终的表达式，如公式所示：

$$L_m(x) = \sum_{i=1}^{n} [g_i + f_m(x_i) + \frac{1}{2}h_i f_m^2(x_i)] + \Omega(f_m)$$

2.2.5 LightGBM

LightGBM (Light Gradient Boosting Machine)是一个由微软亚洲研究院分布式机器学习工具包（DMTK）团队开源的基于决策树算法的分布式梯度提升（Gradient Boosting Decision Tree，GBDT）框架，LightGBM 相较于 XGBoost，提出 Histogrm 算法，对特征进行分桶，减少查询分裂节点的事件复杂度；此外，提出 Goss 算法减少小梯度数据；同时，提出 EFB 算法捆绑互斥特征，降低特征维度，减少模型复杂度。

Lightgbm 使用了如下两种解决办法：一是 GOSS，不是使用所用的样本点来计算梯度，而是对样本进行采样来计算梯度；二是 EFB，不是使用所有的特征来进行扫描获得最佳的切分点，而是将某些特征进行捆绑在一起来降低特征的维度,是寻找最佳切分点的消耗减少。

这样大大的降低的处理样本的时间复杂度，但在精度上，通过大量的实验证明，在某些数据集上使用 Lightgbm 并不损失精度，甚至有时还会提升精度。下面就主要介绍这两种方法。

1、GOSS 算法描述

输入：训练数据，迭代步数 d，大梯度数据的采样率 a，小梯度数据的采样率 b，损失函数和若学习器的类型（一般为决策树）；

输出：训练好的强学习器；

（1）根据样本点的梯度的绝对值对它们进行降序排序；

（2）对排序后的结果选取前 a*100%的样本生成一个大梯度样本点的子集；

（3）对剩下的样本集合（1-a）*100%的样本，随机的选取 b*（1-a）*100%个样本点，生成一个小梯度样本点的集合；

（4）将大梯度样本和采样的小梯度样本合并；

（5）将小梯度样本乘上一个权重系数；

（6）使用上述的采样的样本，学习一个新的弱学习器；

（7）不断地重复（1）～（6）步骤直到达到规定的迭代次数或者收敛为止。

通过上面的算法可以在不改变数据分布的前提下不损失学习器精度的同时大大的减少模型学习的速率。

从上面的描述可知，当 a=0 时，GOSS 算法退化为随机采样算法；当 a=1 时，GOSS 算法变为采取整个样本的算法。在许多情况下，GOSS 算法训练出的模型精确度要高于随机采样算法。另一方面，采样也将会增加若学习器的多样性，从而潜在的提升了训练出的模型泛化能力。

2、EFB 算法描述

输入：特征 F，最大冲突数 K，图 G；

输出：特征捆绑集合 bundles；

（1）构造一个边带有权重的图，其权值对应于特征之间的总冲突；

（2）通过特征在图中的度来降序排序特征；

（3）检查有序列表中的每个特征，并将其分配给具有小冲突的现有 bundling（由控制），或创建新 bundling。

上述算法的时间复杂度为并且在模型训练之前仅仅被处理一次即可。在特征维度不是很大时，这样的复杂度是可以接受的。但是当样本维度较高时，这种方法就会特别的低效。所以对于此，作者又提出的另外一种更加高效的算法：按非零值计数排序，这类似于按度数排

序，因为更多的非零值通常会导致更高的冲突概率。 这仅仅改变了上述算法的排序策略，所以只是针对上述算法将按度数排序改为按非 0 值数量排序，其他不变。

第 3 章 数据预处理

3.1 数据的选取

本文通过 Jupyter Notebook 的 Python 接口读取本题提供的数据，对提供的三类数据在分析过程都将采用。

读取附件 2 数据获取了 4153 家上市公司 6 年的 363 个特征因子，其中‘FLAG’是判定上市公司当年是否造假的数据,其‘0’表示没有造假，‘1’表示造假，是典型的二分类问题，用机器学习算法来解决。再根据特征类型删除特征类型为‘object’的三个特征，’REPORT_TYPE’,’CURRENCY_CD’和’ACCOUTING_STANDARDS’。由于其所有样本的取值相同，对模型的特征构建和模型训练无益，故删除。

根据问题二和问题三的要求，分行业找出财务数据造假的上市公司，则对提供的上市公司财务数据按行业进行划分，划分后的财务数据样本数量如下图所示：

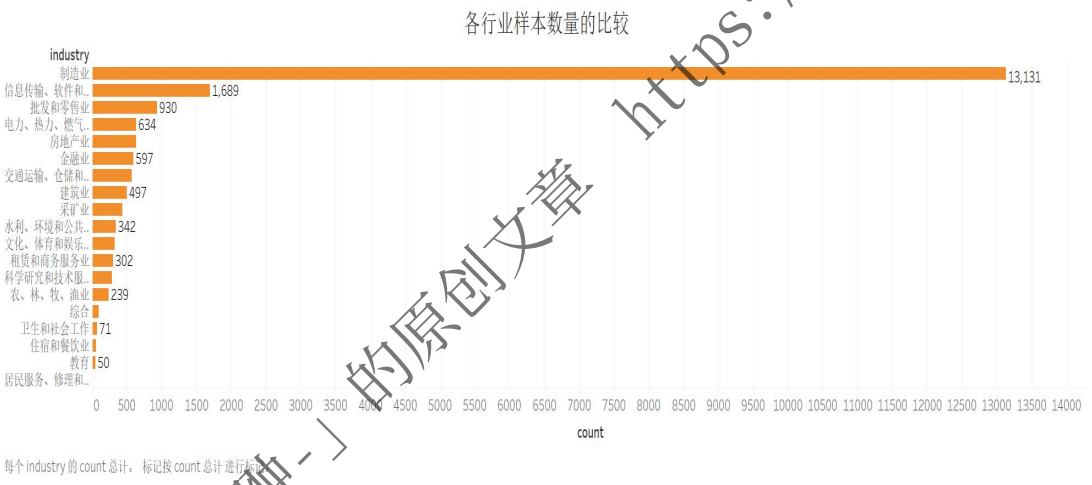

图 1 各行业样本数量的比较

3.2 数据探索

3.2.1 数据缺失情况

对提供的样本数据进行分析，通过数据的分析与探索可以发现数据缺失值严重，总体的缺失值达到了 57.15%，特征缺失比例大于 50%的有 258,占总特征数 362(除标签外)的 71.27%，特征缺失比例大于 60%的有 248,占总特征数 362(除标签外)的 68.50%，特征缺失比例大于 70%的有 115,占总特征数 362(除标签外)的 31.76%，特征缺失比例大于 80%的有 92,占总特征数 362(除标签外)的 25.41%。

3.2.2 数据类别分布情况

对提供的样本数据进行分析，发现样本严重不平衡 ，整体样本标签分布情况如下图。

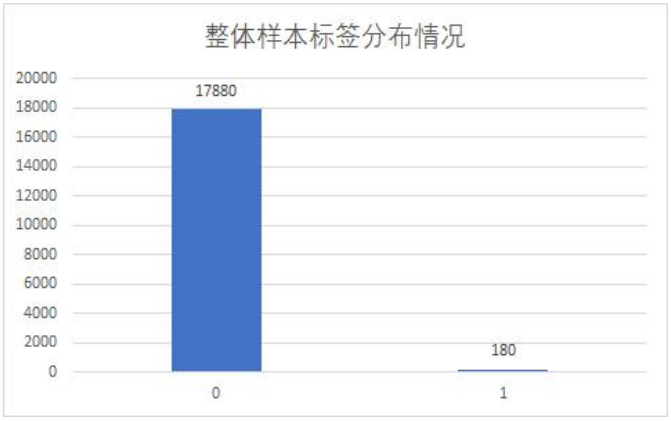

图 2 整体样本标签分布情况

常见的样本不平衡处理方法有 SMOTE、欠采样和过采样，但是考虑到可能出现过拟合的情况，对样本的不平衡现象采用采用周志华实验室提出的欠采样的算法 EasyEnsemble，即利用集成学习机制，将反例划分为若干个集合供不同学习器使用，这样对每个学习器来看都进行了欠采样，但是全局却不会丢失重要的信息。

3.2.3 公司年数据的数量情况

图 3 公司年数据的数量情况

对样本数据进行分析，发现上市公司有 5 年的数据的有 3094 家，有 4 年的数据有 367 家，有 3 年的数据有 102 家，有 2 年数据的有 227 家，仅有 1 年数据的有 362 家。

3.3 特殊数据的处理

3.3.1 异常值的检测和处理

数据来源于 4153 家上市公司,数据比较分散，无法使用箱型图做异常值的处理。以特征因子'NOTES_RECEIV','NOTES_RECEIV','AR','PREPAYMENT'为例子画出箱型图，如图所示，无法判断出异常值，因此不能采用箱型图查看异常值。

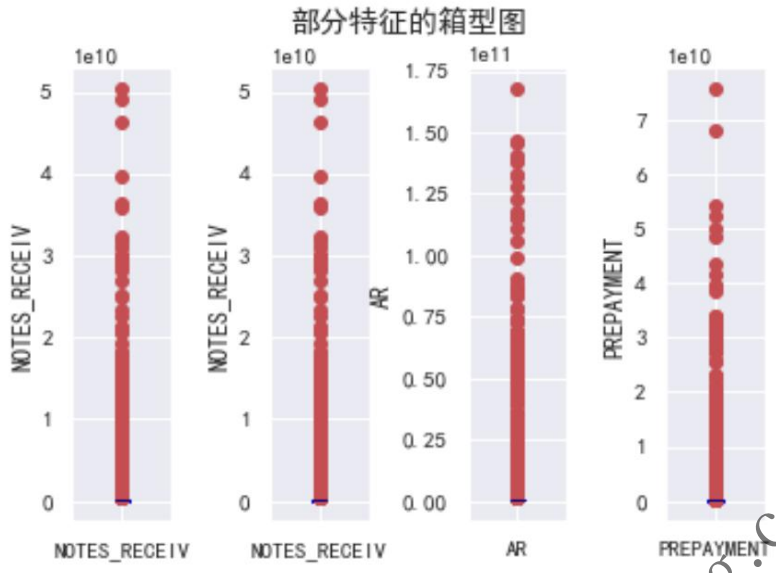

图 4 部分特征的箱型图

最终本文主要采取以下两步骤，对异常数据进行检测和处理：

步骤 1：人工处理异常值，根据附件三的单位可以知道主要有：元、%、元/股、天/次、次等，根据这些单位人工处理异常值（列如：以次为单位的特征，不可能为负数），将这些值设为空，留到后面当做缺失值处理

步骤 2：利用切比雪夫定理（5sigma 法）对异常值进行判断，将异常值设为空，留到后面当做缺失值处理。

3.3.2 缺失值的检测和处理

将附件 2 中上市公司财务数据按行业划分成 19 份数据集，各行业的样本数量图 3 所示。

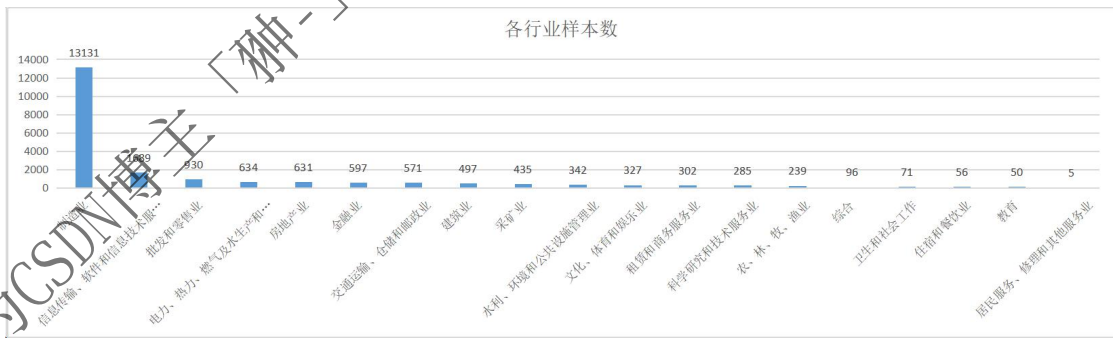

图 5 各行业数量条形图

每份行业数据集作为一份数据集分别来进行自生行业的缺失值处理，每份行业的数据集包括训练集特征因子和测试集特征因子，其训练集，测试集，总数据集的缺失值分布大致一样，如下图所示

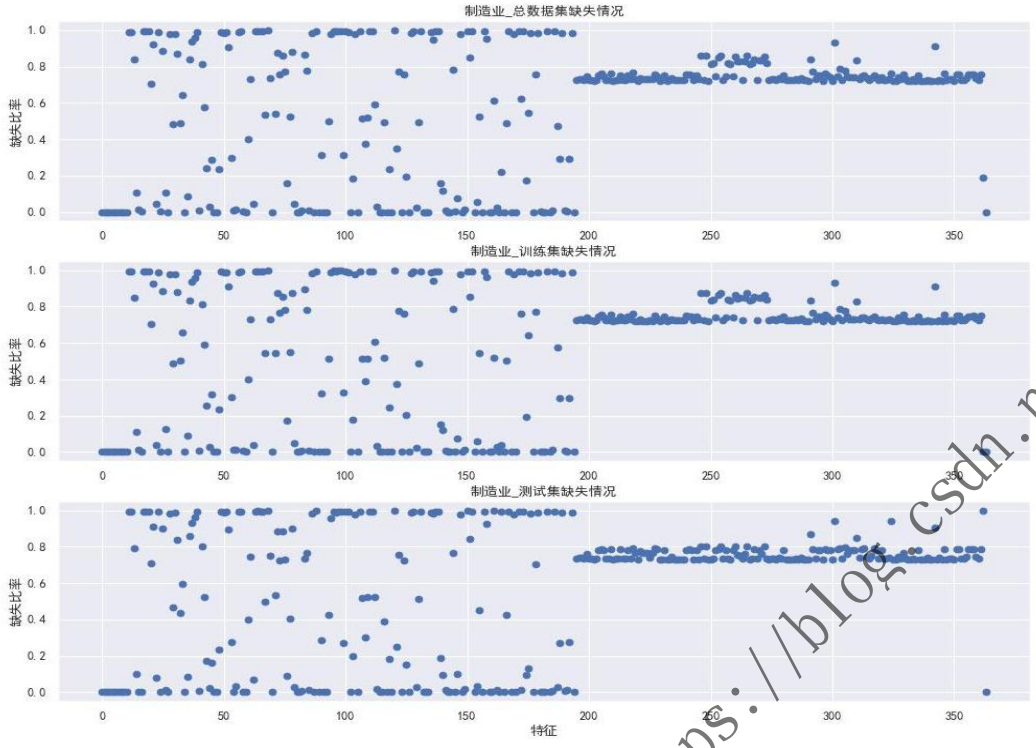

图 6 制造业总数据集、训练集、测试集缺失情况

某个特征因子的缺失值数目占总样本比例大于 50%时，我们将直接舍弃这个特征因子。因为如果把这个特征因子作为特征出现，反而可能会因此带入噪音，影响最后得出的结果。

当出现某个特征因子缺失的样本数量占总样本比例的 20%到 50%时，且为数值型特征属性时，考虑到不改变数据的分布情况，若数据偏正态分布，使用均值进行填充；若数据偏长尾分布，使用中位数进行充。

当某个特征缺失因子缺失的样本数量占总样本比例小于 20%的情况下，采用随机森林对缺失值进行填补。随机森林（Random Forest）包含多个决策树的分类器，其输出类别由个别树输出类别的众数而定，可以在不显著增加运算量的前提下提高预测精度。并且运算结果对缺失值和非平衡的数据能达到较稳健的水平。随机森林填补缺失值的步骤如下：

1）将数据中所有有缺失值的列提取建立模型，并按列缺失值个数从小到大排序。

2）将缺失值个数最小的一列提取，其他列的缺失值用 0 填补。

3）使用随机森林回归模型填补这一列的缺失值。

4）重复 2、3 步骤，直到所有缺失值被填补，得到完整数据。

对样本数据进行填补，填补后的行业样本数与特征数的关系如下图所示：

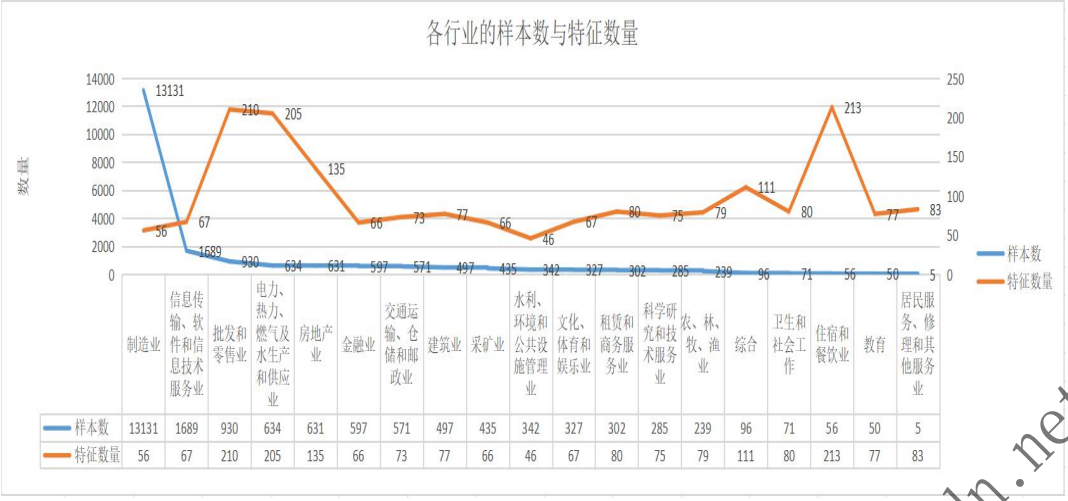

图 7 各行业的样本数与特征数量

3.3.3 异常样本的检测和处理

为了避免所给予的样本数据出现收集错误或则错误标记的情况，我们使用随机森林模型对异常样本进行检测和处理。下面是我们在异常样本进行检测中使用的流程图。

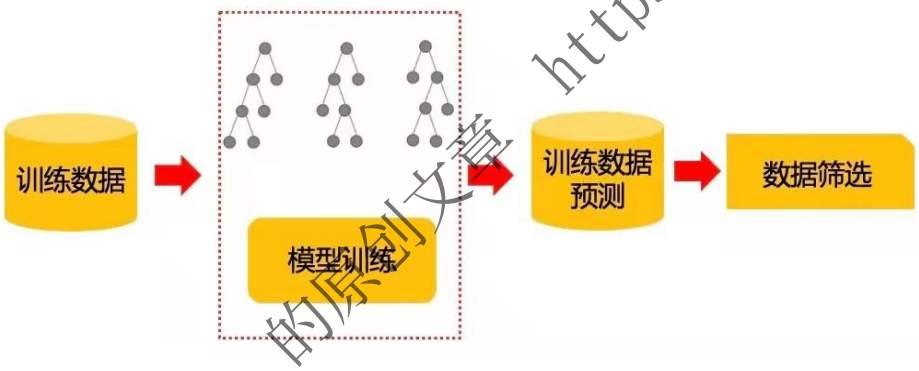

图 8 样本异常检测流程图

对应的算法步骤为：

1. 我们对训练数据进行随机森林模型的训练，得到我们的模型；
2. 使用在训练集上训练得到的模型对训练集进行预测；
3. 我们对预测结果进行特定的筛选。

为了避免异常样本对模型的训练造成影响，我们对检测处理来的异常样本进行删除处理。

3.4 数据合并

为了保证模型中训练集和测试集的特征维度一致，将训练集和测试集进行合并，进行一致的处理。

第 4 章 基于问题一的研究

4.1 相关数据指标的筛选

从所有特征中，选择出有意义、对模型有帮助的特征，以避免必须将所有特征都导入模型去训练的情况。为了让特征的选择更有差异性，选用 4 大类别的特征选择方法对特征进行筛选。

4.2 相关数据指标的筛选方法

4.2.1 Pearson 相关系数过滤法

皮尔森相关系数是一种最简单的，能帮助理解特征和响应变量之间关系的方法，该方法可以有效衡量的是变量之间的线性相关性。其结果的取值区间为[− 1 ， 1]，-1 表示完全的负相关，+1 表示完全的正相关，0 表示没有线性相关。其计算公式如下：

$$\boldsymbol{\rho}(\mathbf{X},\mathbf{Y}) = \frac{\mathbf{E}[(\mathbf{X}-\boldsymbol{\mu}\mathbf{X})(\mathbf{Y}-\boldsymbol{\mu}\mathbf{Y})]}{\boldsymbol{\sigma}\mathbf{X}\boldsymbol{\sigma}\mathbf{Y}}$$

因为正负均表示相关关系的程度，我们将 Pearson 相关系数的公式修改如下：

$$\boldsymbol{\rho}(\mathbf{X},\mathbf{Y}) = \left|\frac{\mathbf{E}[(\mathbf{X}-\boldsymbol{\mu}\mathbf{X})(\mathbf{Y}-\boldsymbol{\mu}\mathbf{Y})]}{\boldsymbol{\sigma}\mathbf{X}\boldsymbol{\sigma}\mathbf{Y}}\right|$$

使用 Pearson 相关系数过滤法，对模型特征进行训练，筛选出判断制造行业上市公司财务造假的特征因子，然后根据特征重要性排序，选择前 20 个的重要特征因子如下：END_DATE,END_DATE_REP,ACT_PUBTIME,PUBLISH_DATE,C_FR_OTH_INVEST_A,T_EQUITY_ATTR_P,C_PAID_TO_FOR_EMPL,T_SH_EQUITY,ADMIN_EXP,N_CF_OPERATE_A,C_PAID_OTH_FINAN_A,REFUND_OF_TAX,OPERATE_PROFIT,T_ASSETS,T_LIAB_EQUITY,T_CA,LT_AMOR_EXP,C_PAID_INVEST,PROC_SELL_INVEST,C_FR_OTH_OPERATE_A。

4.2.2 递归特征消除

递归特征消除，通过递归考虑越来越少的特征集来选择特征。首先，对估计器进行初始特征集训练，并且可以通过任何特定属性或可调用项来获得每个特征的重要性。然后，从当前的一组功能中删除最不重要的功能。在修剪的集上递归地重复该过程，直到最终达到所需的要选择的特征数量。

使用递归特征消除，对模型特征进行训练，筛选出判断制造行业上市公司财务造假的特征因子，然后根据特征重要性，选择前 20 个的重要特征因子如下：OTH_COMPR_INCOME、AVAIL_FOR_SALE_FA、GOING_CONCERN_NI、PROC_SELL_INVEST、NOTES_PAYABLE、INT_PAYABLE、ASSETS_IMPAIR_LOSS、ST_BORR、MINORITY_INT、REFUND_OF_TAX、LT_EQUITY_INVEST、

C_FR_OTH_INVEST_A、N_INCOME_ATTR_P、T_LIAB、OPERATE_PROFIT、T_EQUITY_ATTR_P、T_PROFIT、T_ASSETS、T_SH_EQUITY、T_COMPR_INCOME。

4.2.3 基于随机森林模型的特征选择

随机森林模型的特征选择可以通过袋外数据计算特征的重要性。计算特征的重要性的步骤如下：

1)对每一颗决策树，选择相应的袋外数据（out of bag，OOB），计算袋外数据误差，记为 errOOB1。

2)随机对袋外数据 OOB 所有样本的特征 X 加入噪声干扰，再次计算袋外数据误差，记为 errOOB2。

3)假设森林中有 N 棵树，则特征 X 的重要性$=\sum$（errOOB2 − errOOB1）/N。这个数值之所以能够说明特征的重要性是因为，如果加入随机噪声后，袋外数据准确率大幅度下降（即 errOOB2 上升），说明这个特征对于样本的预测结果有很大影响，进而说明重要程度比较高。

使用随机森林模型，对模型特征进行训练，筛选出判断制造行业上市公司财务造假的特征因子，然后根据特征重要性，选择前 20 个的重要特征因子如下：END_DATE、END_DATE_REP、ACT_PUBTIME、PUBLISH_DATE、ASSETS_IMPAIR_LOSS、T_SH_EQUITY、GOING_CONCERN_NI、T_LIAB_EQUITY、FINAN_EXP、T_EQUITY_ATTR_P、OPERATE_PROFIT、N_INCOME_ATTR_P、T_ASSETS、C_PAID_TO_FOR_EMPL、T_CA、INT_PAYABLE、C_FR_OTH_OPERATE_A、N_CE_END_BAL、N_INCOME、T_COMPR_INCOME。

4.2.4 基于 GDBT 的特征选择

主要是通过计算特征 i 在单棵树中重要度的平均值，计算公式如下：

$$\hat{J}_j^2 = \frac{1}{M}\sum_{m=1}^{M}\hat{J}_j^2(T_m)$$

其中，M 是树的数量。特征 i 在单棵树的重要度主要是通过计算按这个特征 i 分裂之后损失的减少值，其中，L 是叶子节点的数量，L-1 就是非叶子结点的数量。

使用 GDBT 的特征选择模型，对模型特征进行训练，筛选出判断制造行业上市公司财务造假的特征因子，然后根据特征重要性，选择前 20 个的重要特征因子如下：END_DATE, ASSETS_IMPAIR_LOSS、OPERATE_PROFIT、LT_AMOR_EXP、C_PAID_FOR_TAXES、C_PAID_TO_FOR_EMPL、NOTES_PAYABLE、T_SH_EQUITY、MINORITY_INT、INT_PAYABLE、FINAN_EXP、C_PAID_OTH_FINAN_A、REFUND_OF_TAX、N_INCOME、GOING_CONCERN_NI、C_INF_FR_OPERATE_A、T_EQUITY_ATTR_P、T_CA、N_CHANGE_IN_CASH、MINORITY_GAIN。

4.3 相关数据指标的选择

GBDT、由方差过滤、卡方检验过滤、Pearson 相关系数过滤法、递归特征消除 、基于 L1 正则项的特征选择、基于随机森林模型，本文在 7 个模型确定最优参数之后，在测试集上进行预测，模型训练结束后，选择效果较好的 Pearson 相关系数过滤法、递归特征消除（Recursive Feature Elimination）和基于随机森林模型的特征、GBDT 选择四个模型所得特征重要性，由于特征变量较多，我们选择这四个模型排名前 20 个重要特征。在此基础上，挑选出四个模型共同确定的特征因子，并进行分类，得到结果如下：

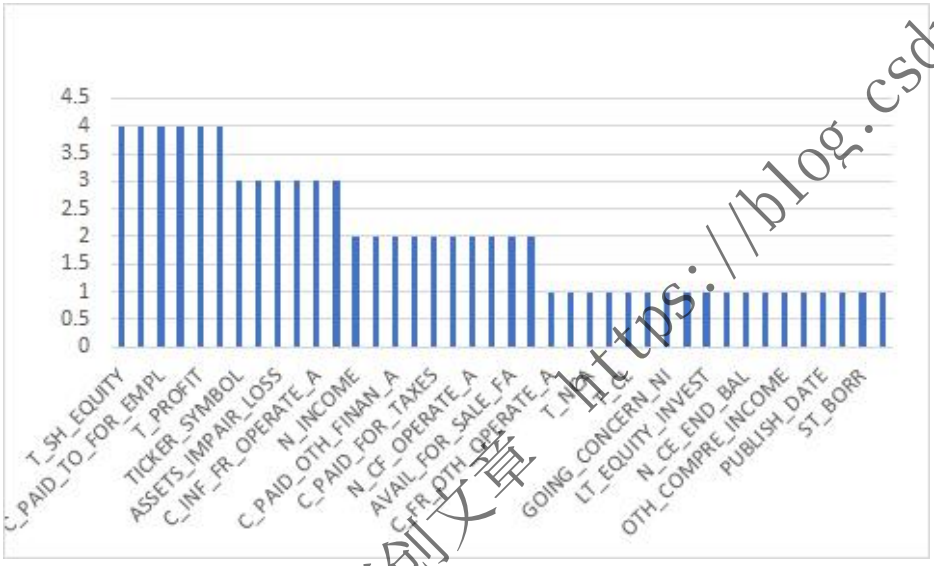

图 9 模型共同确定特征因子

对上述由 4 个模型确定出来的结果，我们计算每个特征因子在四个模型中的重复次数，当该因子重复次数在 3 次及以上，我们认为该特征有效。以制造业为例，获得的有效特征因子如下表：

表 1 制造业有效特征因子及特征因子重复次数

序号	特征因子	特征因子重复次数
0	T_SH_EQUITY	4
1	OPERATE_PROFIT	4
2	C_PAID_TO_FOR_EMPL	4
3	INT_PAYABLE	4
4	T_PROFIT	4

5	MINORITY_INT	4
6	TICKER_SYMBOL	3
7	T_EQUITY_ATTR_P	3
8	ASSETS_IMPAIR_LOSS	3
9	T_COMPR_INCOME	3
10	C_INF_FR_OPERATE_A	3
11	N_INCOME_ATTR_P	3

4.4 不同行业上市公司相关数据指标的异同

对不同行业，利用 Pearson 相关系数过滤法、递归特征消除（Recursive Feature Elimination)、基于随机森林模型的特征选择、基于 GDBT 的特征选择四种模型对模型特征进行训练，筛选出判断制造行业上市公司财务造假的特征因子，然后根据四种模型筛选出特征因子数量进行判断，当他们的在四个模型筛选出的特征因子里面出现次数达到 3 次及以上，则选取为影响判断此行业是否进行财务数据造假的特征因子，以制造业为例，下面是选取出的影响上市公司是否发生财务数据造假的特征因子。（其他行业选取出的特征因子，详情见附录表 2）

表 2 制造业有效特征因子

所属行业	指标名称
制造业	T_SH_EQUITY
制造业	OPERATE_PROFIT
制造业	C_PAID_TO_FOR_EMPL
制造业	INT_PAYABLE
制造业	T_PROFIT
制造业	MINORITY_INT
制造业	TICKER_SYMBOL
制造业	T_EQUITY_ATTR_P
制造业	ASSETS_IMPAIR_LOSS
制造业	T_COMPR_INCOME

制造业	C_INF_FR_OPERATE_A
制造业	N_INCOME_ATTR_P

通过上述操作，将 19 个行业影响本行业是否发生财务数据造假的特征因子筛选出来，得到影响不同行业的上市公司是否发生财务数据造假的特征因子的个数如下图所示。

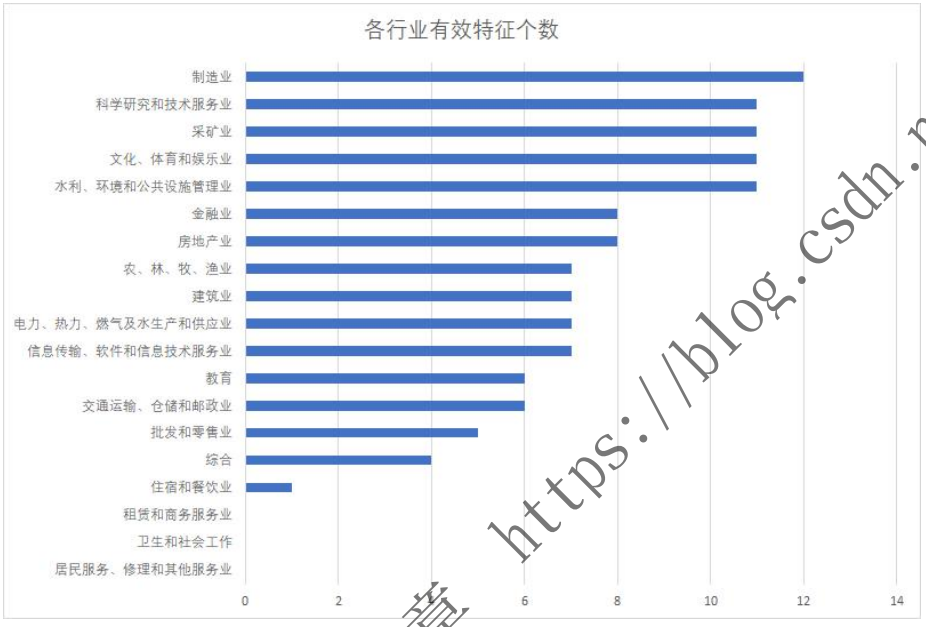

图 10 各行业有效特征个数

由上图我们可以发现，影响不同的行业是否发生财务数据造假有关的特征因子数量有明显的差异，对于制造业，科学研究和技术服务行业，采矿业，文化、体育和娱乐业，水利、环境和公共设施管理业筛选处理的相关数据特征达 11 个以上，可以通过筛选的特征及模型融合预测其所在行业的公司是否发生数据造假；而对于租赁和商务服务业，卫生和社会工作，居民服务、修理和其他服务业的相关特征仅有 0 个，难于通过筛选出的特征进行判断，需要通过模型训练进行判别其是否发生财务数据造假。

通过不同模型对上市公司财务数据造假特征因子的选取，分析所选取的不同行业间特征因子的异同。

根据对模型选取出的不同行业与发生财务数据密切相关的特征因子分析可以得出，在不同的行业之间存在某些共同的特征因子如：T_PROFIT（利润总额）、T_COMPR_INCOME（综合收益总额）、N_INCOME_ATTR_P（归属于母公司所有者或股东的净利润）、OPERATE_PROFIT（营业利润）等，这些特征因子与大部分行业上市公司是否发生财务数据造假有密切的关系。通过对这些特征因子的分析可以看出上市公司是否发生财务数据造假与公司的盈利情况密

切相关。

分析模型中选取出的其他特征因子，可以发现不同的行业间，尤其是区别较大的行业间影响是否发生财务数据造假的特征因子，他们判断是否发生财务数据造假的特征因子与本行业内容贴近。

总体分析不同行业上市公司判断是否发生财务数据造假的特征因子可以发现，无论是哪一个行业，判断上市公司是否发生财务数据造假的特征因子都与公司的盈利密切相关。判断一个上市公司是否发生财务数据造假需要通过全面的数据进行判别，不能仅通过某些方面判断需要结合行业特质、商业模式、企业所有制等进一步判断。

第 5 章 基于机器学习模型的问题二研究

5.1 模型的构建

5.1.1 测试集、训练集的划分

每种行业数据集用 Python Sklearn 中 model_selection 的 train_test_split 模型，人为划分训练集与测试集，将数据划分成训练集与测试集的比为 8:2，以制造业为例，划分结果如下：

表 3 制造业测试集、训练集划分

数据	制造业
X_train	8504 行 54 列
X_test	2127 行 54 列
y_train	{ 0.0 : 8438 : 1.0 : 66 }
y_test	{ 0.0 : 2109, 1.0 : 18 }

上表中，X_train 是训练集特征因子，X_test 是测试集特征因子，y_train 是训练集标签，y_test 是测试集标签。对训练集特征因子和训练集标签进行欠采样得到一份模型训练的数据集，数据集的标签类别比是 1:1{0:66,1:66}，这样得到模型 1，用模型 1 来预测 X_test，得到 y_pred 模型 1 预测结果，对训练集特征因子和训练集标签随机按照上面采样 1000 次，得到 1000 模型预测的 1000 个 y_pred，然后 1000 个模型结果进行集成策略的硬投票得到最终的一个模型预测结果，然后将其结果与 y_test 进行 AUC 模型评估，通过多次随机欠采样的方式构造多个模型输出多种结果，对其结果用集成策略来解决数据类型样本不均衡的问题。

5.1.2 数据标准化

在机器学习算法中，再将有着不同规格的数据转换到同一规格，或者不同分布的数据转换到某个特定分布的需求，即将数据“无量纲化”。在逻辑回归，支持向量机等中，无量纲化可以加快求解速度。常用的数据标准化有

本文使用数据标准化将数据“无量纲化”，即将数据按均值中心化后，再按标准缩放，数据就会服从均值为 0，方差为 1 的正态分布（即标准正态分布）。其计算为：

$$x^* = \frac{x - \mu}{\sigma}$$

5.1.3 模型评价指标

AUC（Area Under Curve）被定义为ROC曲线下与坐标轴围成的面积，一般我们以TPR为y 轴，以FPR为x轴，就可以得到RoC曲线。AUC的数值都不会大于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，一般就无太多应用价值了。其中关于：FPR（False Positive Rate）以及 TPR（True Positive Rate）的数学计算公式为：

$$\mathrm{FPR}=\frac{\mathrm{FR}}{(\mathrm{TN}+\mathrm{FP})}\mathrm{TPR}=\frac{\mathrm{TP}}{(\mathrm{TP}+\mathrm{FN})}$$

5.2 模型参数调优与模型重要特征

5.2.1 参数调优概念及方法

利用机器学习算法构建出来的模型，因参数的不同选择会有较大的影响，当参数选取不恰当，就容易发生欠拟合或者是过拟合的情况。为了尽可能提高模型的精度，同时题升模型的泛化能力，需要对模型进行调参。对模型进行参数调优，选择超参数时，可以根据经验微调，另外可以选择不同大小的参数，代入模型，挑选出表现效果最好的参数。

本文选择的调参方法为网格搜索，网格搜索法是指定参数值的一种穷举搜索方法，通过将估计函数的参数通过交叉验证的方法进行优化来得到最优的学习算法。即将各个参数可能的取值进行排列组合，列出所有可能的组合结果生成“网格”。然后将各组合用于SVM训练，并使用交叉验证对表现进行评估。在拟合函数尝试了所有的参数组合后，返回一个合适的分类器，自动调整至最佳参数组合，可以通过clf.best_params_获得参数值。 网格搜索可以保证在指定的参数范围内找到精度最高的参数。

5.2.2 各个模型参数调优

建模时先固定每个参数的初始值，再设定其调参范围，进行网格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，本文模型优化评价指标设为和曲线下面积（AUC）。

表4 LR的调参过程

参数名称	初始值	调参范围	调参结果	调参前AUC	调参后AUC
penalty	l1	l1、l2	l1	80.19%	80.22%
c	1	range(1,5)	3		

逻辑回归模型需要调整的参数有两个，分别为 penalty 和 C，penalty 表示正则化的方式，C 表示正则化强度的倒数，其默认值为 1，即默认正则项与损失函数的比值是 1：1。C 越小，损失函数会越小，模型对损失函数的惩罚越重，正则化的效果越强。

从 AUC 的结果 80.22%不难看出模型拟合的效果一般，难以将其作为最优算法预测，继续探索其他模型。

表 5 SVM 的调参过程

参数名称	初始值	调参范围	调参结果	调参前 AUC	调参后 AUC
kernel	linear	'rbf'、'ploy'	linear	81.13%	81.84%
C	0	range(0,1,0.1)	0.5		

SVM 需要调整的参数也有两个，分别为 kernel 和 C，kernel 表示算法中采用的核函数类型，可选参数有：‘linear’:线性核函数，‘poly’：多项式核函数，‘rbf’： 径像核函数/高斯‘sigmod’：核函数。C 表示错误项的惩罚系数。C 越大，即对 分错样本的惩罚程度越大，因此在训练样本中准确率越高，但是泛化能力降低， 也就是对测试数据的分类准确率降低。相反，减小 C 的话，容许训练样本中有一 些误分类错误样本，泛化能力强。SVM 调参后 AUC 有 81.84%较上一个模型效果较好，但拟合程度依旧不高，因而下面选用复杂程度更高的集成算法建模调参。

表 6 RFC 的调参过程

参数名称	初始值	调参范围	调参结果	调参前 AUC	调参后 AUC
n_estinator	500	range(300,800)	650	81.29%	82.37%
min_depth	8	range(3,15)	5		
min_samples_split	100	range(50,150)	70		
min_samples_leaf	20	range(10,100)	20		

随机森林调整的参数有四个，分别为 n_estinators：随机森林中树模型的数量、max_depth：树的最大深度、min_samples_split：中间节点要分枝所需要的最小样本数和 min_samples_split：叶节点要存在所需要的最小样本数。

该模型的 AUC 结果有 82.37%较 SVM 算法有略微提高，但是效果仍然一般，继续探索下列其他模型。

表 7 XGBoostd 的调参过程

参数名称	初始值	调参范围	调参结果	调参前 AUC	调参后 AUC
n_estinators	500	range(300,800)	623	81.73%	83.12%
Max_depth	5	range(3,7)	5		
gamam	0.6	range(0.5,1,0.1)	0.7		
subsample	0.8	range(0.5,1,0.1)	0.8		
Reg_alpha	1	range(0,2,0.5)	0		
Reg_lambda	0	range(0,2,0.5)	1		
Learning_rate	0.1	[0.01,0.1]	0.01		

XGBoost 算法调整第一个参数是 n_estimators，这个参数非常强大，该参数越大，模型的学习能力就会越强；下面只介绍该模型中几个相对重要参数：参数 subsample 表示随机抽样的时候抽取的样本比例，范围是(0,1]；参数 Learning_rate 表示集成中的学习率，又称为步长以控制迭代速率，常用于防止过拟合。默认是 0.1，取值范围[0,1]。XGBoost 算法调参后 AUC 有 83.12%，AUC 很接近 1.0，检测方法真实性高，表明模型有较好的拟合效果。

表 8 LightGBM 的调参过程

参数名称	初始值	调参范围	调参结果	调参前 AUC	调参后 AUC
max_depth	5	range(2,10)	5	82.79%	83.65%
num_leaves	30	range(20,50)	30		
max_bin	200	range(100,300)	173		
lambda_l1	0	range(0,1,0.1)	0		
lambda_l2	0	range(0,1,1)	0		
n_estimators	300	range(200,400)	275		
Learning_ra	0.1	[0.1,0.01]	0.01		

te					

LightGBM 的基本调参过程如下：首先选择较高的学习率，大概 0.1 附近，这 样是为了加快收敛的速度。这对于调参是很有必要的。其次是对决策树基本参数调参，最后是正则化参数调参。因此，第一步先确定学习率和迭代次数，第二步，确定 max_depth 和 num_leaves，这是提高精确度的最重要的参数。第三步，确定 min_data_in_leaf 和 max_bin。第四步，确定 feature_fraction、bagging_fraction、bagging_freq。第五步，确定 lambda_l1 和 lambda_l2。第六步，确定 min_split_gain。第七步，降低学习率，增加迭代次数，验证模型。Lightgbm 算法调参后 AUC 有 83.65%，AUC 很接近 1.0，检测方法真实性高，表明模型有较好的拟合效果。

5.3 基于模型融合的预测模型构造

5.3.1 模型选择

本文主要选择在测试集上得出的 AUC 作为评价各分类算法的指标，因此下面将五个算法调参前后各模型的测试集上得出的 AUC 对比，如表 9 所示。

表 9 各模型 AUC 值

算法	调参前	调参后
LR	80.19%	80.22%
SVM	81.13%	81.84%
RFC	81.29%	82.37%
Xgboost	81.73%	83.12%
Lightgbm	82.79%	83.65%

将调参后 AUC 的值从小到大排序，依次是 LR、SVM、RFC、Xgboost、Lightgbm ，可见发现 RFC、Xgboost、Lightgbm 这三种对 GBDT 的优化实现的算法效果远好于其他算法。

为了观察各个模型之间的效果，对每个模型分别绘制对应 ROC 曲线，来观察各模型之间的优劣。ROC 曲线是根据一系列不同的二分类方式（分界值或决定阈），以真阳性率（灵敏度）为纵坐标，假阳性率（1-特异度）为横坐标绘制的曲线。ROC 曲线将这个图划分成了两部分，为增加对比效果，再使用 ROC 曲线观察各模型之间的优劣，ROC 曲线图是反映敏感性与特异性之间关系的曲线。横坐标 X 轴为 1-特异性，也称为假阳性率 （误报率），X 轴

越接近零准确率越高；纵坐标 Y 轴称为敏感度，也称为真阳性率（敏感度），Y 轴越大代表准确率越好。根据曲线位置，把整个图划分成了两部分，曲线下方部分的面积被称为 AUC（Area Under Curve），用来表示预测准确性，AUC 值越高，也就是曲线下方面积越大，说明预测准确率越高。曲线越接左上角（X 越小，Y 越大），准确率越高。 ROC 曲线如图 11 所示。

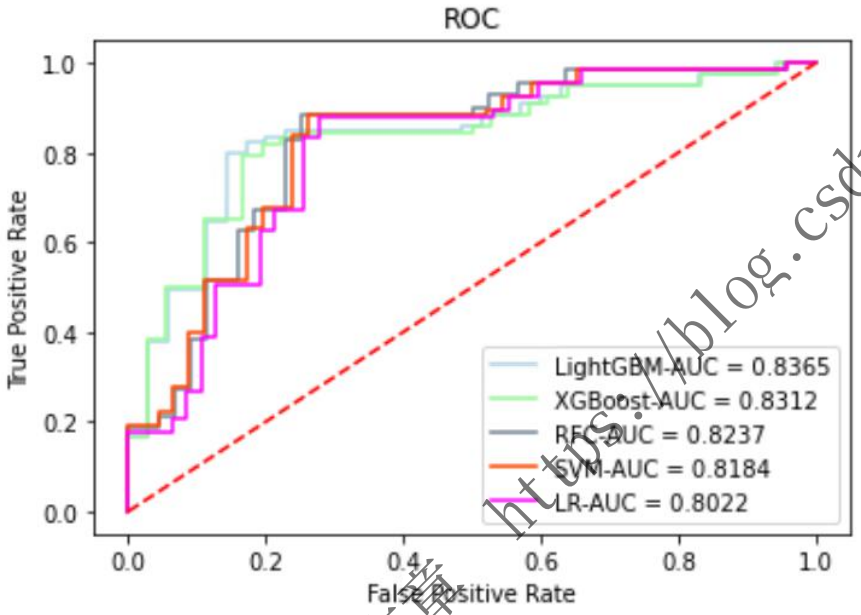

图 11 ROC 曲线

不难发现 RFC、 Xgboost、Lightgbm 这三种对 GBDT 的优化实现的算法效果远好于其他算法，其中 Lightgbm 比 RFC、XGBoost 表现的更为优秀。本文为了更好的提升模型的预测准确率和泛化能力，使用 Stacking 模型融合，将这几个学习器测结果作为新的训练集，去学习一个新的学习器。

5.3.2 模型融合的介绍

模型融合有 Stacking 和 Blending，本文使用的是 Stacking 方法进行模型融合，Stacking 是集成学习的一种，在介绍 Stacking 模型融合前首先了解集成学习的这个概念。集成学习(ensemble learning)是通过构建并结合多个学习器来完成学习任务，有时也被成为多分类系统（multi-classifer system）、基于委员会的学习等。集成学习主要有 Bagging（并行）、Boosting（串行）、Stacking（树形）、Blending，如下图 12 所示：

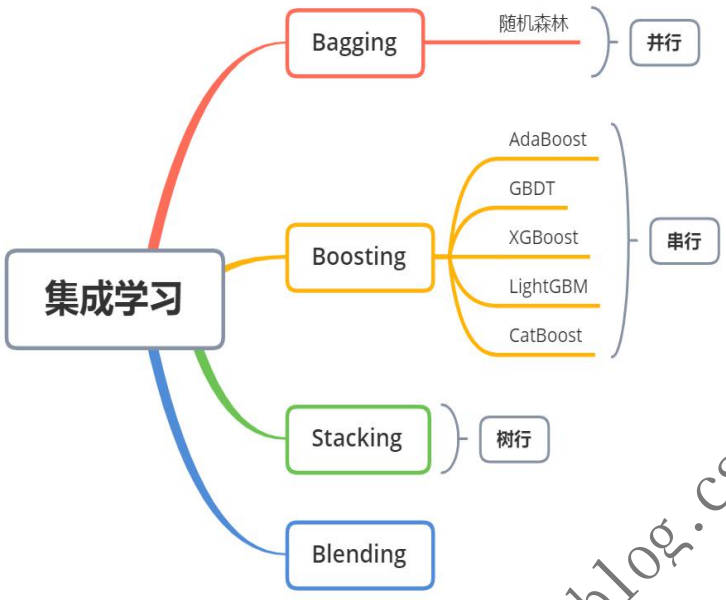

图 12 集成学习示意图

5.3.3 模型融合的过程

Stacking 基本思想：用初始训练数据学习出若干个基学习器后，将这几个学习器的预测结果作为新的训练集(第一层)，来学习一个新的学习器(第二层)。

Stacking 集成算法的好坏由两个方面决定的：一是需要基学习器尽保持独立，即基学习器之间有一定的差异性，这样集成学习才能“博采众长”，二是预测效果相近且基分类器预测效果越高，集成学习的模型效果就会越好。

本文基于经典的 Stacking 集成学习模型方法进行了两点优化改进：

①：针对本文类型极度不均衡的情况，采用周志华实验室提出的欠采样的算法 EasyEnsemble，利用集成学习机制，将反例划分为若干个集合供不同学习器使用，这样对每个学习器来看都进行了欠采样，但是全局却不会丢失重要的信息。

```
def easyensemble(df, desired_apriori, n_subsets=1000):
    train_resample = []
    for i in range(n_subsets):
        sel_train = undersampling(df, desired_apriori)
        train_resample.append(sel_train)
    return train_resample
```

图 13 核心思想代码图

②：将其每一步验证中使用的单个相同的模型改为 5 个不同的机器学习模型进行预测。

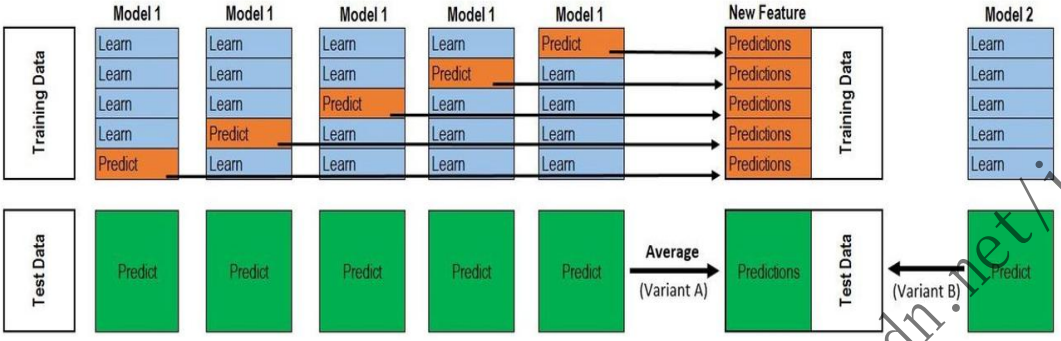

图 14 模型融合过程示意图

本文第一层使用五种算法 LR、RFC、SVM、XGBoost、LightGBM 产生第二层使用 LR 训练的数据，。过程如下：第一层将输入的训练集作 5 折交叉验证，使用 4 折作为训练集，训练一个模型并预测另外 1 折和测试集，将模型 5 折进行预测的结果合并得到如图 12 的 New Feature，将模型 5 折预测测试集的结果使用 Averaging 的方法求平均值，最终得到如图 12 Predictions 模型的预测结果。第二层将第一层 5 折交叉训练得到的 5 个 New-Feature 合并得到新的训练集，将 5 个新 Predictions 合并得到新的测试集，用新训练集和测试集构造第二层的预测器，即 LR 模型。为了降低过拟合的问题，第二层分类器采用了比较简单的的广义线性逻辑回归分类器，在特征提取的过程中，我们已经使用了复杂的非线性变换，因此在第二层输出层不需要复杂的分类器。

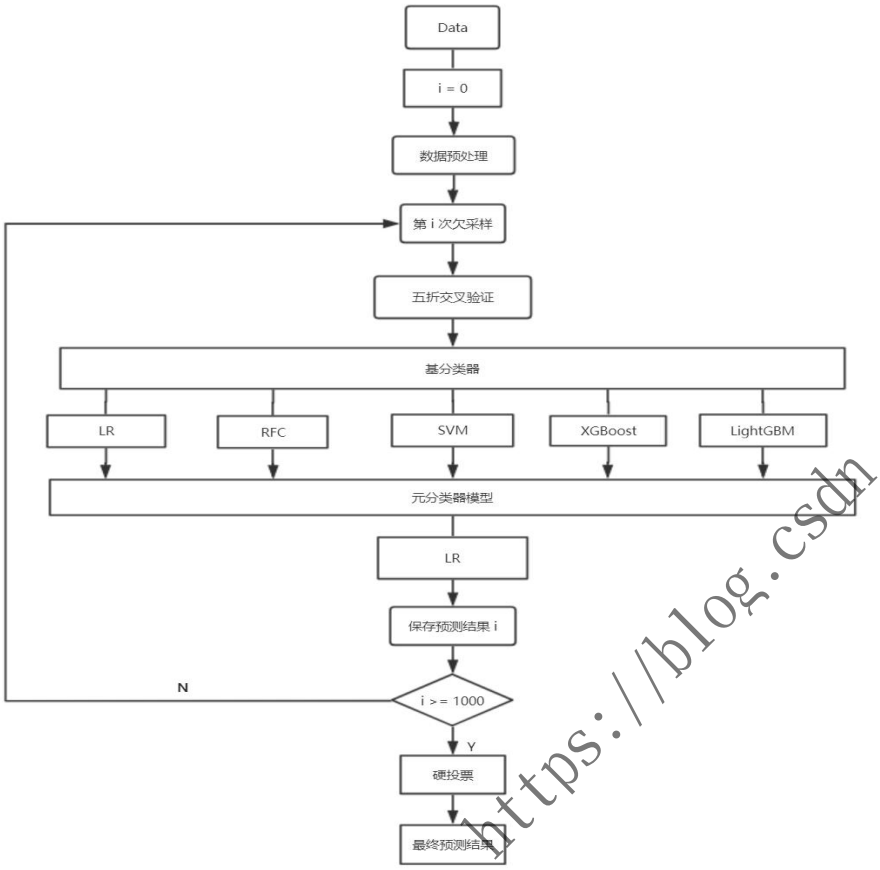

图 15 模型构建流程图

采用周志华实验室提出的欠采样的算法 EasyEnsemble，如图在第一层的基学习器对训练集进行 1000 次随机欠采样，对每次采样的得到数据集放进第一层基学习器学习，然后进行第二层的 Stacking 模型融合，得到 1000 个 Stacking 模型融合的结果，最后对 1000 个结果进行集成策略中的硬投票得到最终的测试集结果。且每份欠采样得到的数据集在模型融合的第一层使用 5 折交叉验证划分数据的方式防止过拟合的发生。

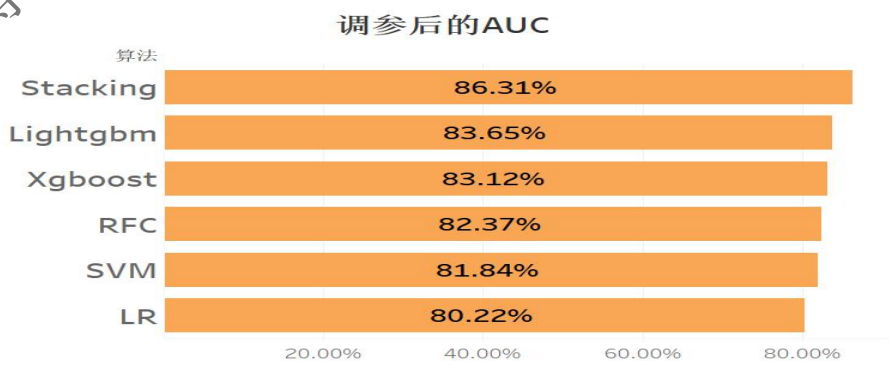

图 16 Stacking 模型融合与其他模型 AUC 对比

显而易见，基于 Stacking 的模型融合算法效果优于其他算法，证明 Stacking 集成学习模型充分整合了第一层基学习器优异的表现，是模型具有更好的预测能力和更强的泛化能力。

5.4 基于融合模型的预测第六年的决策结果

通过对上市公司是否发生财务数据造假行为 Stacking 集成学习融合模型的建立，使用第六年上市公司的财务数据，经过数据探索、数据预处理、特征工程、模型训练、模型评估、模型融合、模型预测等一系列的操作后，去判断 2500 个制造业上市公司里面是否进行财务数据造假。

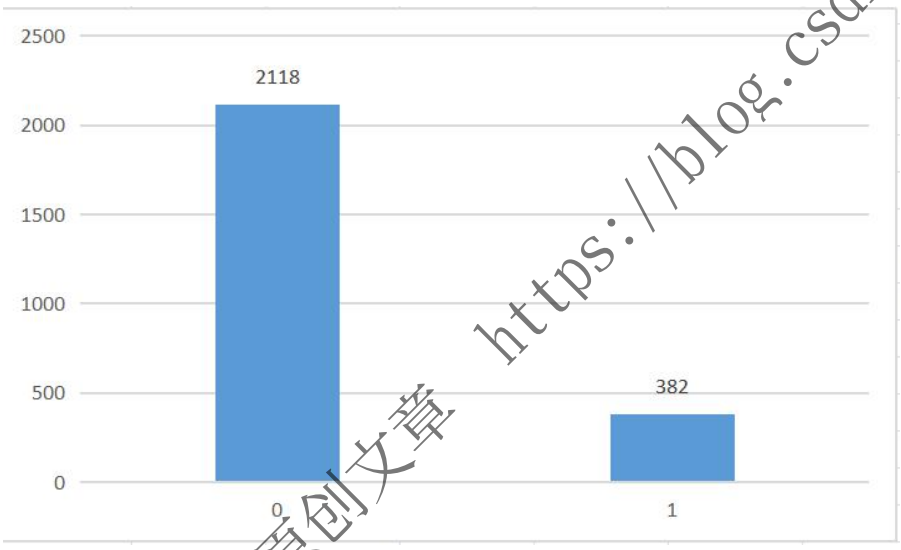

图 17 第六年上市是否发生财务数据造假直方图

得到结果：存在 1.528%的上市公司发生财务数据造假，即有 382 个制造业的上市公司发生财务数据造假，2118 个制造业的上市公司不会发生财务数据造假。为了方便观察，绘制第六年的制造业上市公司是否发生财务数据造假的 0-1 决策直方图，0 表示不发生财务数据造假，1 表示发生财务数据造假，如图 15 所示。

第 6 章 基于多种算法问题三的研究

6.1 基于模型的选择及构造

6.1.1 模型选择

考虑到其余各行业的数据量比较小，甚至在某些行业中，没有数据造假的数据样本，然而依旧使用问题 2 的方法的话，无疑加大了整体模型的复杂度，并且预测出来效果相比单个模型的预测效果更差。

对此，我们根据各各行业的数据量，将 19 个行业根据样本数据量划分为三类：将信息传输、软件和信息技术服务业，批发和零售业的数据集合并为第一类；对电力、热力、燃气及水生产和供应业，房地产业，金融业，交通运输、仓储和邮政业，建筑业，采矿业，水利、环境和公共设施管理业，文化、体育和娱乐业，租赁和商务服务业，科学研究和技术服务业，农、林、牧、渔业化的数据集合并划分为第二类；对综合，卫生和社会工作，住宿和餐饮业，教育，居民服务、修理和其他服务业化的数据集合并划分为第三类。然后对划分出来的三类样本，添加所属行业类别的特征因子，最后使用单模型进行训练和预测。

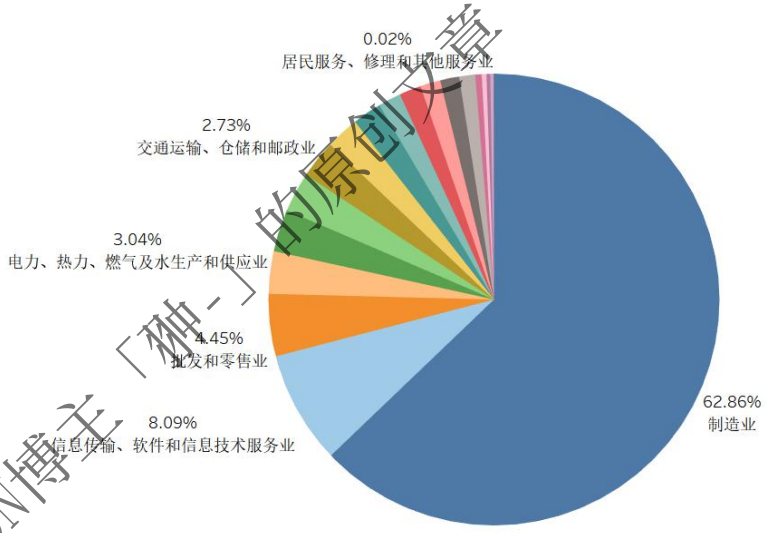

图 18 各行业数量占比扇形图

6.1.2 模型选择及评价

将划分出来的第一类行业：信息传输、软件和信息技术服务业，批发和零售业，进行模型的融合及调参，以模型 AUC 作为评价各类算法的指标，下面是五种算法调参后各模型测试集的出 AUC 对比，如表 10 所示。

表 10 第一类行业模型选择及模型评价 AUC

模型选择	模型 AUC
KNN	69.93%
SVM	74.27%
LR	78.69%
RFC	76.34%
GBDT	79.63%

将调参后的 AUC 的值从小到大排序依次是：KNN 69.93%、SVM 74.27%、RFC 76.34%、LR 78.69%、GBDT 79.63%。最终选用 AUC 指标最好的 GBDT 模型进行调参，调参前后结果如表 11 所示：

表 11 调参后第一类行业模型选择及模型评价 AUC

调参前 AUC	调参后 AUC
79.63%	82.73%

将划分出来的第二类行业：电力、热力、燃气及水生产和供应业，房地产业，金融业，交通运输、仓储和邮政业，建筑业，采矿业，水利、环境和公共设施管理业，文化、体育和娱乐业，租赁和商务服务业，科学研究和技术服务业，农、林、牧、渔业化，进行模型的融合及调参，以模型 AUC 作为评价各类算法的指标，下面是五种算法调参后各模型测试集的出 AUC 对比，如表 12 所示。

表 12 第二类行业模型选择及模型评价 AUC

模型选择	模型 AUC
KNN	71.21%
SVM	76.53%
LR	80.06%
RFC	80.37%
GBDT	81.68%

将调参后的 AUC 的值从小到大排序依次是：KNN 71.21%、SVM 76.53%、LR 80.06%、RFC

80.37%、GBDT 81.68%。最终选用 AUC 指标最好的 GBDT 模型进行调参，调参前后结果如表 13 所示。

表 13 调参后第二类行业模型选择及模型评价 AUC

调参前 AUC	调参后 AUC
81.68%	83.43%

将划分出来的第三类行业：综合，卫生和社会工作，住宿和餐饮业，教育，居民服务、修理和其他服务业化，进行模型的融合及调参，以模型 AUC 作为评价各类算法的指标，下面是五种算法调参后各模型测试集的出 AUC 对比，如表 14 所示。

表 14 第三类行业模型选择及模型评价 AUC

模型选择	模型 AUC
KNN	69.21%
SVM	73.15%
LR	79.78%
RFC	76.34%
GBDT	78.36%

将调参后的 AUC 的值从小到大排序依次是：KNN 69.21%、SVM 73.15%、RFC 76.34%、GBDT 78.36%、LR 79.78%。最终选用 AUC 指标最好的 LR 模型进行调参，调参前后结果如下：

表 15 调参后第三类行业模型选择及模型评价 AUC

调参前 AUC	调参后 AUC
79.78%	80.53%

6.2 基于融合模型的预测第六年的决策结果

通过将 18 个行业划分成三个大的类别，添加行业所属类别,利用 OneHotEncoder、LabelEncoder 扩增特征因子，然后以 AUC 为评价指标，在 KNN、SVM、LR、RFC、GBDT 等单模型中选择效果较好的进行训练和预测，挑选出效果最好的模型，求出除制造业外各行业上市公司第 6 年财务数据造假的上市公司。

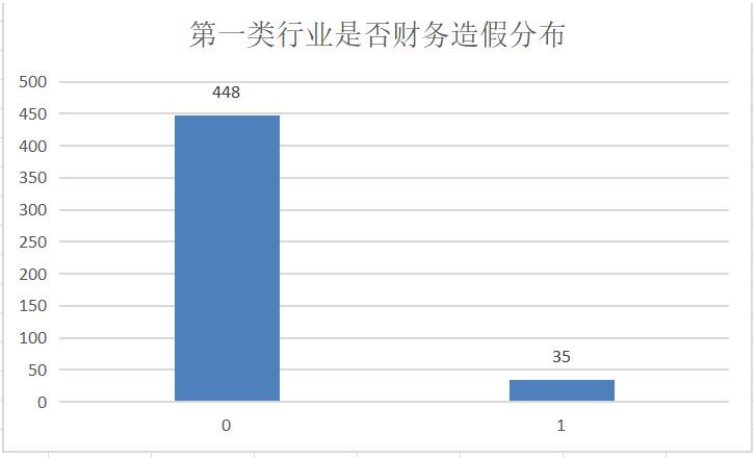

图 19 第一类行业是否财务数据造假分布

第一类行业 483 家上市公司中，有 448 家上市公司财务数据良好，有 35 家公司进行财务数据造假，财务数据造假公司的比例达到 7.25%。

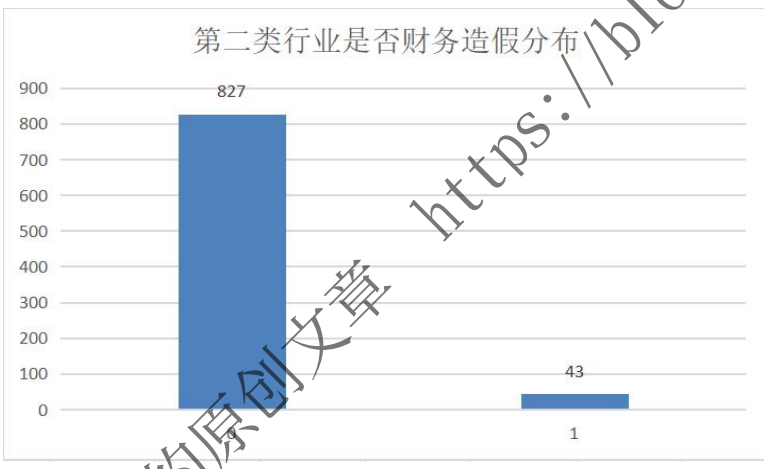

图 20 第二类行业是否财务数据造假分布

第二类行业 870 家上市公司中，有 827 家上市公司财务数据良好，有 43 家公司进行财务数据造假，财务数据造假公司的比例达到 4.94%。

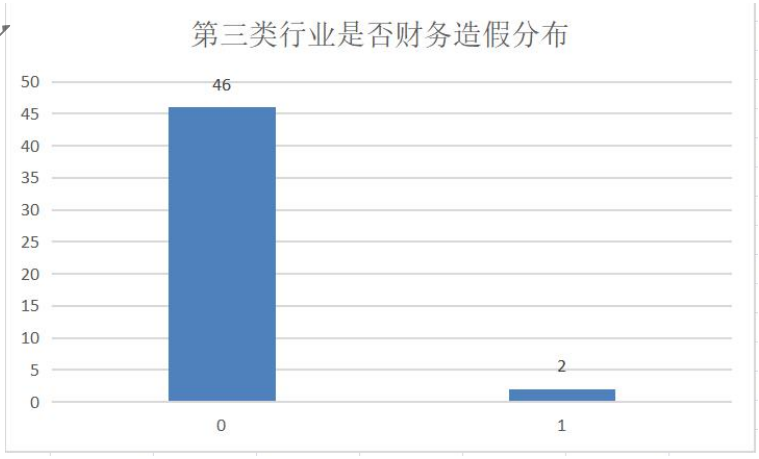

图 21 第三类行业是否财务数据造假分布

第三类行业 48 家上市公司中，有 46 家上市公司财务数据良好，有 2 家公司进行财务数据造假，财务数据造假公司的比例达到 4.17%。

通过对上市公司是否发生财务数据造假行为模型的建立，使用第六年上市公司的财务数据，经过数据探索、数据预处理、特征工程、模型训练、模型评估、模型融合、模型预测等一系列的操作后，去判断除制造业外上市公司里面是否进行财务数据造假，经过模型的检验与预测具有良好的效果。故，本文采用的策略与方法有一定的准确性和现实意义，可以作为预测上市公司下一年是否发生财务数据造假方案的有效模型。

第 7 章 总结

本文通过结合影响上市公司是否财务数据造假的实际情况与进行机器学习算法进行建模预测，主要目的是构建最优的分类预测模型预测上市公司是否财务造假。

一、在对数据进行预处理时，采用人工处理异常值和切比雪夫定理相结合的方法对数据进行异常值的检测，此外还进行了随机森林模型对异常样本进行了检测和处理。由于样本数据存在严重的样本不平衡情况，为避免出现数据的过拟合给原来的样本数据带入噪音，采用周志华实验室提出的欠采样的算法 EasyEnsemble，具有良好的处理效果。

二、使用方差过滤法、卡方检验过滤法、Pearson 相关系数过滤法、递归特征消除法、基于 L1 正则项的特征选择法、基于随机森林模型的特征选择法、基于 GBDT 模型的特征选择法，在这七种机器学习模型中，通过评价指标 AUC 对比，选择效果较好的 Pearson 相关系数过滤法、递归特征消除法和基于随机森林模型的特征选择法、基于 GBDT 模型的特征选择法四个模型，选择算法特征重要性排名前 20 个重要特征，挑出四个模型共同确定的特征因子，作为问题一的结果。

三、由 Pearson 相关系数过滤法、递归特征消除法和基于随机森林模型的特征选择法、基于 GBDT 模型的特征选择法四个模型共同确定的特征因子为：T_SH_EQUITY、OPERATE_PROFIT、C_PAID_TO_FOR_EMPL、INT_PAYABLE、T_PROFIT、MINORITY_INT、TICKER_SYMBOL、ASSETS_IMPAIR_LOSS、T_COMPR_INCOME、C_INF_FR_OPERATE_A、N_INCOME_ATTR_P。

四、在预测模型的选择上，本文选择用了基于 Stacking 集成学习模型，第一层基学习器用了 LR、随机森林、SVM、GBDT、XGBoost、Lightgbm，第二层元学习器选择了 LR，从而确定了最优的 Stacking 集成学习预测模型。Stacking 集成学习预测模型在测试集上的 AUC 得分为 86.31%，高于所有基础分类器，可见建立的模型较为稳定，不存在严重的过拟合且效果较好。

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，可以作为预测上市公司下一年是否发生财务数据造假方案的有效模型。

参考文献

[1] 肖爽. 中国农业上市公司财务造假特征研究[D].中南财经政法大学,2019.

[2] 刘子扬. 基于机器学习的信贷风控研究[D].南京邮电大学,2020.

[3] 李荟.检测财务数据造假的新工具——奔福德定律的妙用[J].商业会计,2012(20):46-48.

[4] 詹红雁.上市公司财务造假常用的手段、识别及防范建议[J].商业会计,2018(08):99-101.

[5] 滕小芹. 关于中国上市公司财务数据的研究—运用 M-score 模型[D].上海交通大学,2013.

[6] 曹正凤.随机森林算法优化研究[D].北京:首都经济贸易大学,2014.

[7] 徐慧丽.基于随机森林的多阶段集成学习方法[J].高师理科学刊,2018(2).

[8] The use of accounting flexibility to reduce labor renegotiation costs and manage earnings[J] . Julia D'Souza,John Jacob,K. Ramesh. Journal of Accounting and Economics . 2001 (2)

[9] Incentives and Penalties Related to Earnings Overstatements That Violate GAAP[J] . Messod D. Beneish. The Accounting Review . 1999 (4)

[10] Reversal of fortune dividend signaling and the disappearance of sustained earnings growth[J] . Harry DeAngelo,Linda DeAngelo,Douglas J. Skinner. Journal of Financial Economics . 1996 (3)

[11] Stakeholders' implicit claims and accounting method choice[J] . Robert M Bowen,Larry DuCharme,D Shores. Journal of Accounting and Economics . 1995 (3)

[12] Labor Union Contract Negotiations and Accounting Choices[J] . Susan E. Liberty,Jerold L. Zimmerman. The Accounting Review . 1986 (4)

[13] Liu Xu-Ying,Wu Jianxin,Zhou Zhi-Hua. Exploratory undersampling for class-imbalance learning.[J]. IEEE transactions on systems, man, and cybernetics. Part B, Cybernetics : a publication of the IEEE Systems, Man, and Cybernetics Society,2009,39(2).

附录

表 1

relevant_p	relevant_RFE	relevant_tree	relevant_GBDT
T_SH_EQUITY	T_SH_EQUITY	T_SH_EQUITY	T_SH_EQUITY
OPERATE_PROFIT	OPERATE_PROFIT	OPERATE_PROFIT	OPERATE_PROFIT
T_CA	GOING_CONCERN_NI	T_CA	T_CA
END_DATE	T_ASSETS	GOING_CONCERN_NI	GOING_CONCERN_NI
C_PAID_OTH_FINAN_A	REFUND_OF_TAX	END_DATE	END_DATE
T_ASSETS	T_EQUITY_ATTR_P	C_PAID_OTH_FINAN_A	C_PAID_OTH_FINAN_A
REFUND_OF_TAX	ASSETS_IMPAIR_LOSS	T_ASSETS	REFUND_OF_TAX
T_EQUITY_ATTR_P	OTH_COMPR_INCOME	T_EQUITY_ATTR_P	ASSETS_IMPAIR_LOSS
END_DATE_REP	AVAIL_FOR_SALE_FA	ASSETS_IMPAIR_LOSS	C_PAID_TO_FOR_EMPL
ACT_PUBTIME	PROC_SELL_INVEST	END_DATE_REP	LT_AMOR_EXP
PUBLISH_DATE	NOTES_PAYABLE	PUBLISH_DATE	C_PAID_FOR_TAXES
C_FR_OTH_INVEST_A	INT_PAYABLE	ACT_PUBTIME	MINORITY_INT
C_PAID_TO_FOR_EMPL	ST_BORR	FINAN_EXP	INT_PAYABLE
ADMIN_EXP	MINORITY_INT	T_LIAB_EQUITY	NOTES_PAYABLE
N_CF_OPERATE_A	LT_EQUITY_INVEST	N_INCOME_ATTR_P	FINAN_EXP
T_LIAB_EQUITY	C_FR_OTH_INVEST_A	ADMIN_EXP	N_CHANGE_IN_CASH
LT_AMOR_EXP	N_INCOME_ATTR_P	T_NCA	N_INCOME
C_PAID_INVEST	T_LIAB	T_PROFIT	COMPR_INC_ATTR_M_S
PROC_SELL_INVEST	T_PROFIT	N_CE_END_BAL	OTH_NCA
C_FR_OTH_OPERATE_A	T_COMPR_INCOME	C_FR_OTH_OPERATE_A	MINORITY_GAIN

（注：蓝色标记为同时被三种或者三种算法选上的特征）

表 2

所属行业	指标名称
制造业	T_SH_EQUITY

制造业	OPERATE_PROFIT
制造业	C_PAID_TO_FOR_EMPL
制造业	INT_PAYABLE
制造业	T_PROFIT
制造业	MINORITY_INT
制造业	TICKER_SYMBOL
制造业	T_EQUITY_ATTR_P
制造业	ASSETS_IMPAIR_LOSS
制造业	T_COMPR_INCOME
制造业	C_INF_FR_OPERATE_A
制造业	N_INCOME_ATTR_P
信息传输、软件和信息技术服务业	C_FR_OTH_OPERATE_A
信息传输、软件和信息技术服务业	C_PAID_G_S
信息传输、软件和信息技术服务业	ASSETS_IMPAIR_LOSS
信息传输、软件和信息技术服务业	C_PAID_FOR_OTH_OP_A
信息传输、软件和信息技术服务业	N_INCOME_ATTR_P
信息传输、软件和信息技术服务业	REFUND_OF_TAX
信息传输、软件和信息技术服务业	AVAIL_FOR_SALE_FA
批发和零售业	COMPR_INC_ATTR_P
批发和零售业	DIV_PAYABLE
批发和零售业	N_INCOME_ATTR_P
批发和零售业	OPERATE_PROFIT
批发和零售业	T_PROFIT
电力、热力、燃气及水生产和供应业	T_COMPR_INCOME
电力、热力、燃气及水生产和供应业	T_CL
电力、热力、燃气及水生产和供应业	C_PAID_OTH_FINAN_A
电力、热力、燃气及水生产和供应业	AVAIL_FOR_SALE_FA
电力、热力、燃气及水生产和供应业	RETAINED_EARNINGS
电力、热力、燃气及水生产和供应业	C_FR_OTH_INVEST_A

电力、热力、燃气及水生产和供应业	C_PAID_INVEST
房地产业	OP_CL
房地产业	OP_TL
房地产业	T_EQUITY_ATTR_P
房地产业	BIZ_TAX_SURCHG
房地产业	INV_INC_TR
房地产业	N_CE_END_BAL
房地产业	RETAINED_EARNINGS
房地产业	N_TAN_A_TL
金融业	T_COMPR_INCOME
金融业	N_CE_BEG_BAL
金融业	T_PROFIT
金融业	C_FR_OTH_INVEST_A
金融业	NOPERATE_EXP
金融业	T_REVENUE
金融业	N_CE_END_BAL
金融业	N_INCOME
交通运输、仓储和邮政业	GOING_CONCERN_NI
交通运输、仓储和邮政业	RETAINED_EARNINGS
交通运输、仓储和邮政业	FINAN_EXP
交通运输、仓储和邮政业	NCL_WITHIN_1Y
交通运输、仓储和邮政业	AVAIL_FOR_SALE_FA
交通运输、仓储和邮政业	DIV_PAYABLE
建筑业	GOING_CONCERN_NI
建筑业	MINORITY_INT
建筑业	FINAN_EXP
建筑业	GOODWILL
建筑业	LT_BORR
建筑业	C_FR_OTH_OPERATE_A

建筑业	C_PAID_TO_FOR_EMPL
采矿业	OTH_NCA
采矿业	N_INCOME_ATTR_P
采矿业	COMPR_INC_ATTR_P
采矿业	N_INCOME
采矿业	REFUND_OF_TAX
采矿业	T_PROFIT
采矿业	T_COMPR_INCOME
采矿业	C_PAID_OTH_FINAN_A
采矿业	OPERATE_PROFIT
采矿业	TICKER_SYMBOL
采矿业	NOTES_PAYABLE
水利、环境和公共设施管理业	T_SH_EQUITY
水利、环境和公共设施管理业	LT_EQUITY_INVEST
水利、环境和公共设施管理业	GAIN_INVEST
水利、环境和公共设施管理业	REFUND_OF_TAX
水利、环境和公共设施管理业	C_FR_OTH_FINAN_A
水利、环境和公共设施管理业	PAID_IN_CAPITAL
水利、环境和公共设施管理业	PREPAYMENT
水利、环境和公共设施管理业	RETAINED_EARNINGS
水利、环境和公共设施管理业	LT_BORR
水利、环境和公共设施管理业	CASH_C_EQUIV
水利、环境和公共设施管理业	T_LIAB
文化、体育和娱乐业	T_PROFIT
文化、体育和娱乐业	T_COMPR_INCOME
文化、体育和娱乐业	COMPR_INC_ATTR_P
文化、体育和娱乐业	N_INCOME
文化、体育和娱乐业	N_INCOME_ATTR_P
文化、体育和娱乐业	OPERATE_PROFIT

文化、体育和娱乐业	PROC_SELL_INVEST
文化、体育和娱乐业	AVAIL_FOR_SALE_FA
文化、体育和娱乐业	C_FR_OTH_INVEST_A
文化、体育和娱乐业	C_PAID_OTH_INVEST_A
文化、体育和娱乐业	C_PAID_FOR_DEBTS
科学研究和技术服务业	C_PAID_INVEST
科学研究和技术服务业	PAID_IN_CAPITAL
科学研究和技术服务业	N_INCOME
科学研究和技术服务业	COGS
科学研究和技术服务业	T_REVENUE
科学研究和技术服务业	RETAINED_EARNINGS
科学研究和技术服务业	CIP
科学研究和技术服务业	OPERATE_PROFIT
科学研究和技术服务业	T_COMPR_INCOME
科学研究和技术服务业	C_PAID_G_S
科学研究和技术服务业	T_PROFIT
农、林、牧、渔业	AVAIL_FOR_SALE_FA
农、林、牧、渔业	SURPLUS_RESER
农、林、牧、渔业	ADMIN_EXP
农、林、牧、渔业	N_CHANGE_IN_CASH
农、林、牧、渔业	PUR_FIX_ASSETS_OTH
农、林、牧、渔业	REFUND_OF_TAX
农、林、牧、渔业	N_CF_FR_INVEST_A
住宿和餐饮业	CA_TA
教育	AP
教育	N_CE_BEG_BAL
教育	T_NCA
教育	AR
教育	N_CE_END_BAL

教育	C_PAID_FOR_DEBTS
综合	N_CHANGE_IN_CASH
综合	NCL_WITHIN_1Y
综合	N_CF_OPA_CL
综合	N_CF_OPA_ND