E-Commerce Letters 电子商务评论, 2024, 13(2), 1991-2006
Published Online May 2024 in Hans. https://www.hanspub.org/journal/ecl
https://doi.org/10.12677/ecl.2024.132241

Hans汉斯

基于模型融合的上市公司财务造假的预测

仪　梦[1*]，吴丽丽[2#]

[1]甘肃农业大学理学院，甘肃　兰州
[2]甘肃农业大学信息科学技术学院，甘肃　兰州

收稿日期：2024年3月3日；录用日期：2024年3月29日；发布日期：2024年5月22日

摘　要

我国上市公司财务报告造假的问题一直伴随着市场的发展。针对此问题，构造了基于分类模型的上市公司财务造假的预测研究。通过数据的预处理和机器学习算法模型，以及统计学的一些方法，建立了一套完整的分析预测模型。首先按照行业分为大类，将年数据中缺失率达到50%以上的指标剔除，剩下缺失的数据选用0来填充。对于日数据提取每股指标，并按年进行均值化，将均值化后的日数据整合为年数据并且提取出特征因子，最终通过降维的思想筛选出对上市财务造假有较大影响的因子。通过确定的因子，将特征因子初步处理，并且进行标准化，通过使用三大类特征选择的方法，使系统的特定指标进一步优化，接着用主成分降维，正则化特征提取，最终用决策树分类模型、线性判别模型、梯度提升分类模型、支持向量机模型四种分类模型进行分类预测。

关键词

财务造假，随机森林，组合决策树，逻辑回归模型，曲线下面积

Prediction of Financial Fraud of Listed Companies Based on Model Fusion

Meng Yi[1*], Lili Wu[2#]

[1]College of Sciences, Gansu Agricultural University, Lanzhou Gansu
[2]College of Information Science and Technology, Gansu Agricultural University, Lanzhou Gansu

Received: Mar. 3rd, 2024; accepted: Mar. 29th, 2024; published: May 22nd, 2024

Abstract

The problem of financial report fraud of listed companies in our country has been accompanied by

*第一作者。
#通讯作者。

文章引用: 仪梦, 吴丽丽. 基于模型融合的上市公司财务造假的预测[J]. 电子商务评论, 2024, 13(2): 1991-2006.
DOI: 10.12677/ecl.2024.132241

the development of the market. In order to solve this problem, this paper constructs a forecasting research on financial fraud of listed companies based on classification model. Through the data preprocessing and machine learning algorithm model, as well as some statistical methods, we established a complete set of analysis and prediction model. First of all, according to the industry it is divided into large categories. The annual data missing rate of more than 50% of the indicators removed. The remaining missing data selected to fill 0. For the daily data, we extract the per-share index and average it every year, then integrate the average daily data into the number of years. We integrate the averaged daily data into annual data and extract the characteristic factors. Finally, through the thought of dimensionality reduction, the factors that have a greater impact on listed financial fraud are screened out. By using the method of feature selection of three categories, the specific index of the system is further optimized, and then the principal component is used to reduce the dimension. Finally, decision tree classification model, linear discriminant model, gradient promotion classification model and support vector machine model were used to predict the classification.

Keywords

Financial Fraud, Random Forest Model, XGBoost, Logistic Regression Model, AUC

1. 引言

我国上市公司财务报告造假行为严重扰乱了中国的社会经济秩序，发生的上市公司财务报告造假案件引发了社会各界对财务造假问题的深刻反思[1]。上市公司财务报告舞弊问题[2]在当今全球市场上非常普遍。

愈演愈烈的上市公司财务报告舞弊的危害性主要表现在：(1) 严重影响了社会市场经济秩序[3]，侵害了国家会计法规和会计制度，破坏了我国法制化社会的进程。(2) 使尚在发展中的证券市场饱受虚假财务信息[4]的冲击，使投资者蒙受损失，加剧了市场的投机行为，使投资者对市场的规范化运营的信心严重缺失，影响了社会的安定。(3) 政府机构和上市公司的高管因受到虚假财务信息[5]的误导而导致无法做出正确的决策，在影响上市公司正常运营的同时也会影响政府机构对社会经济资源的合理配置。(4) 上市公司通过虚减收入和虚增费用等手段偷逃国家税款，使国家税款严重流失。(5) 上市公司股东、债权人和职工的合法权益受到侵犯，对这些群体造成的经济损失不可估量。(6) 滋生了大批的贪污腐败人员，严重影响了社会财富的分配，影响了社会的稳定和谐。

2. 模型原理

2.1. 随机森林(Random Forest Model)

在机器学习中，随机森林[6]由许多的决策树组成，因为这些决策树的形成采用了随机的方法，因此也叫做随机决策树。随机森林中的树之间是没有关联的。当测试数据进入随机森林时，其实就是让每一颗决策树进行分类，最后取所有决策树中分类结果最多的那类为最终的结果。因此随机森林是一个包含多个决策树的分类器，并且其输出的类别是由个别树输出的类别的众数而定。

2.2. 组合决策树(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 [7]是一种提升树模型，第 m 轮用一棵回归树拟合前 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。

2.3. 逻辑回归模型

逻辑回归[8]是机器学习中一个应用非常广泛的分类模型，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)。对于逻辑回归的损失函数构成的模型可能过拟合的问题，正则化是结构风险最小化策略的实现，是在经验风险上加一个正则化项或惩罚项。正则项可以取不同的形式，在回归问题中取平方损失，就是参数的 L2 范数，也可以取 L1 范数。取平方损失时，模型的损失函数变为

$$J(\theta)=\frac{1}{2m}\sum_{i=1}^{m}\left(h_\theta(x_i)-y_i\right)^2-\lambda\sum_{i=1}^{m}\theta_j^2$$

其中 λ 为正则项系数，正则化后的梯度下降算法的更新变为：$\theta_j=\theta_j-\frac{\alpha}{m}\sum_{i=1}^{m}\left(h_\theta(x_i)-y_i\right)x_i^j-\frac{\lambda}{m}\alpha\theta_j$，从而完成对事件发生概率的预测。

2.4. LightGBM 模型

Light GBM [9]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。

2.5. 模型评价指标

机器学习算法评价指标有很多种，本文模型优化评价指标设为曲线下面积(AUC)。AUC (Area Under Curve)被定义为 ROC 曲线下的面积。其中，ROC 曲线全称为受试者工作特征曲线(Receiver Operating Characteristic)，它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(敏感性)为纵坐标，假阳性率(1-特异性)为横坐标绘制的曲线。AUC 就是衡量学习器优劣的一种性能指标。AUC [10]可通过对 ROC 曲线下各部分的面积求和而得。AUC 值越大的分类器，正确率越高。

ROC 曲线[11]能很容易的查出任意阈值对学习器的泛化性能影响。此曲线有助于选择最佳的阈值。ROC 曲线越靠近左上角，模型的准确性就越高。最靠近左上角的 ROC 曲线上的点是分类错误最少的最好阈值，其假正例和假反例总数最少。还可以对不同的学习器比较性能。将各个学习器的 ROC 曲线绘制到同一坐标中，直观地鉴别优劣，靠近左上角的 ROC 曲所代表的学习器准确性最高。

该方法简单、直观、通过图示可观察分析学习器的准确性，并可用肉眼做出判断。ROC 曲线将真正例率和假正例率以图示方法结合在一起，可准确反映某种学习器真正例率和假正例率的关系，是检测准确性的综合代表。ROC 曲线不固定阈值，允许中间状态的存在，利于使用者结合专业知识，权衡漏诊与误诊的影响，选择一个更加的阈值作为诊断参考值。

3. 模型构建

3.1. 数据处理

对于年数据，读取数据获得了 5 年的 217 个特征因子，将特征因子“是否造假”改名为“本年是否

造假”，用机器学习算法解决二分类问题。

本文中的数据来自于对上市公司多年的财务数据报告，将部分异常值 0 更改为缺失值，之后再填充缺失值。如果某个特征因子缺失的样本占总数极大，直接舍弃；如果某个特征因子缺失的样本适中，且为数值型特征属性，用 0、平均数或众数进行缺失值填补；如果某个特征因子缺失的样本较少，使用随机森林填补缺失值。

由于选取的数据包含不同类别的因子，并且不同因子之间在取值范围与量化度量单位等方面存在较大差异。因此需要将因子数据进行标准化处理，数据标准化：将特征数据的分布调整成标准正态分布，也叫高斯分布[12]，也就是使得数据的均值为 0，方差为 1，其转换函数为：

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

其中 μ 是原始数据的均值，而 σ 为原始数据的标准差。标准化的过程为两步：去均值的中心化(均值变为 0)；方差的规模化(方差变为 1)，注意，标准化去均值，方差规模化是针对一个特征维度来做的，而不是针对样本。标准化的原因在于如果有些特征的方差过大，则会主导目标函数从而使参数估计器无法正确地去学习其他特征。本文通过创建一组特征数据，每一行表示一个样本，每一列表示一个特征，来进行标准化。

3.2. 特征因子处理

通过分类方法将所有的行业分为 19 大类。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。

过滤方法通常用作预处理步骤，特征选择完全独立于任何机器学习算法。是根据各种统计检验中的分数以及相关性的各项指标来选择特征。方差过滤[13]通过特征本身的方差来筛选特征的类。本文先消除方差为 0 的特征，将最初的 216 个指标降到 153 个指标，剩下空缺的数据选用 0 来填充。

互信息法[14]是用来捕捉每个特征与标签之间的任意关系(包括线性和非线性关系)的过滤方法。互信息法可以找出任意关系。互信息法不返回 p 值或 F 值类似的统计量，它返回“每个特征与目标之间的互信息量的估计”，这个估计量在[0, 1]之间取值，为 0 则表示两个变量独立，为 1 则表示两个变量完全相关。经典的互信息也是评价定性自变量对定性因变量的相关性的，为了处理定量数据，最大信息系数法被提出，互信息计算公式如下：

$$I(X;Y) = \sum_{x\in X}\sum_{y\in Y} p(x,y)\log\frac{p(x,y)}{p(x)p(y)}$$

3.3. 参数调优

机器学习模型参数众多，参数选择不恰当，就会出现欠拟合或者过拟合的问题。本文选择不同大小的参数，带入模型中，挑选表现最好的参数。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)。

3.3.1. 随机森林调参

随机森林调整的参数有三个，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。该模型的 AUC 结果有 84.2%，有显著提升，见表 1：

Table 1. RFC parameter procedure
表 1. RFC 调参过程

参数名称	初始值	调参范围	结果	调参后的 AUC
n_estinators	100	[100, 50, 10, 1]	50	0.751
max_depth	8	[8, 5, 3, 2]	2	0.781
min_samples_split	5	[5, 3, 2, 2]	2	0.842

我们做出 RFC 调参过程的 ROC 曲线，又称接受者操作特征曲线。AUC 指标表示 ROC 曲线下的面积，见图 1：

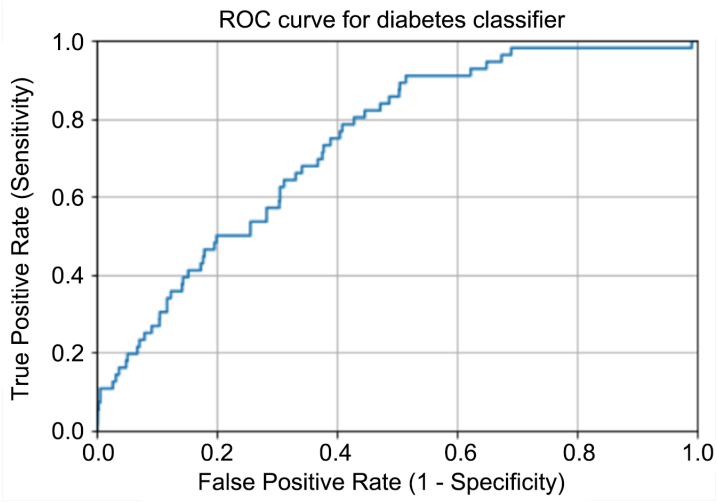

Figure 1. ROC curve
图 1. ROC 曲线图

由 RFC 调参过程，我们做出对 153 个指标重要性的柱状图，可以很清晰的看到其中凸显的指标极为对上市公司造假有较大影响的因子。我们导出 excel 数据得出前 10 的指标为 TEAP_TA、ADMIN_EXP、PAID_IN_CAPITAL、CASH_C_EQUIV、OPA_PROFIT、N_CF_OPA_PROPT、AR、BIZ_TAX_SURCHG、C_PAID_OTH_INVEST_A、RESER_PS，如图 2：

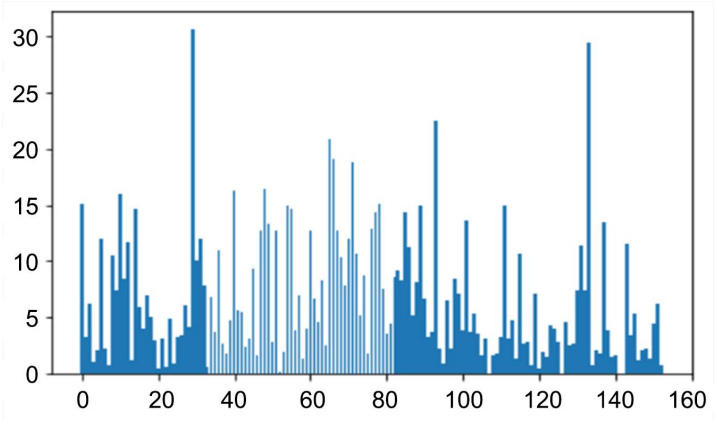

Figure 2. RFC indicator importance histogram
图 2. RFC 指标重要性柱状图

3.3.2. 逻辑回归调参过程

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。见表 2：

Table 2. Logistic regression of parametric process
表 2. 逻辑回归调参过程

参数名称	初始值	调参范围	结果	调参后的 AUC
penalty	0.3	[0.3, 0.2, 0.1]	0.2	0.763
C_penalty	1	[1]	1	0.7859

从 AUC 的结果 78.6%，我们继续探索其他模型。

3.3.3. XGBoost 调参

XGBoost 算法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]。如表 3：

Table 3. XGBoost invokes the procedure
表 3. XGBoost 调参过程

参数名称	初始值	调参范围	结果	调参后的 AUC
colsample_bytree	1	[1]	1	0.601
Min_child_weight	1	[1]	1	0.613
Max_depth	6	[6, 5, 4]	4	0.620
gamma	0.3	[0.3, 0.2, 0.1]	0.1	0.643
subsample	1	[1]	1	0.723
learning_rate	0.3	[0.3, 0.2, 0.1]	0.1	0.745

做出 XGBoost 调参过程的 ROC 曲线，如图 3：

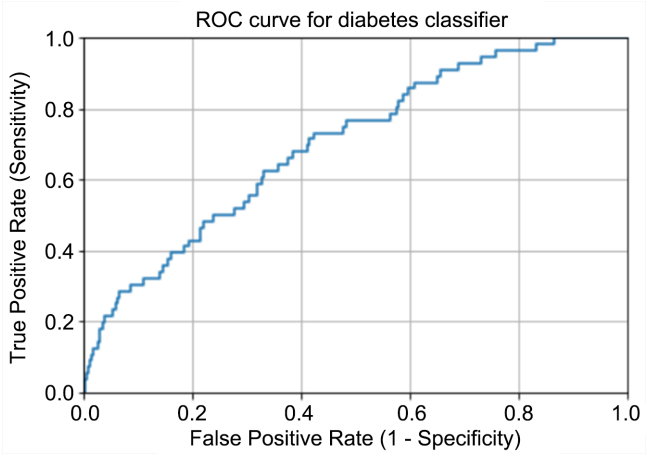

Figure 3. ROC curve
图 3. ROC 曲线图

XGBoost 算法调参后 AUC 有 75%，AUC 接近 1.0，检测方法真实性高，表明模型有较好的拟合效果。为得出对上市公司造假有较大影响的因子，我们将特征因子作图，对算法的重要性从大到小排列，其特征因子和其重要性数值如图 4 所示。

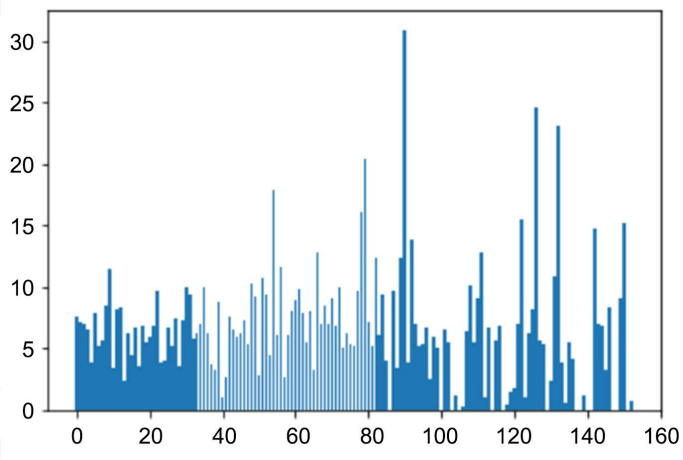

Figure 4. XGBoost metric importance histogram
图 4. XGBoost 指标重要性柱状图

我们求出在 XGBoost 算法中特征重要性为前 30 的特征因子，其中前 10 的指标为 T_REVENUE、N_WORK_CAPITAL、REVENUE、OTH_COMPRE_INCOME、CAP_FIX_RATIO、OP_PS、T_CL、COGS、ST_BORR、C_PAID_OTH_INVEST_A。

3.3.4. LightGBM 调参过程

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 78.6%，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。如表 4 所示：

Table 4. LightGBM invokes the procedure
表 4. LightGBM 调参过程

参数名称	初始值	调参范围	结果	调参后的 AUC
colsample_bytree	1	[1]	1	0.653
Min_child_weight	1	[1]	1	0.701
Max_depth	6	[6, 5, 4]	5	0.709
gamma	0.3	[0.3, 0.2, 0.1]	0.2	0.721
subsample	1	[1]	1	0.772
learning_rate	0.3	[0.3, 0.2, 0.1]	0.1	0.786

做出 lightGBM 调参过程的 ROC 曲线，如图 5：

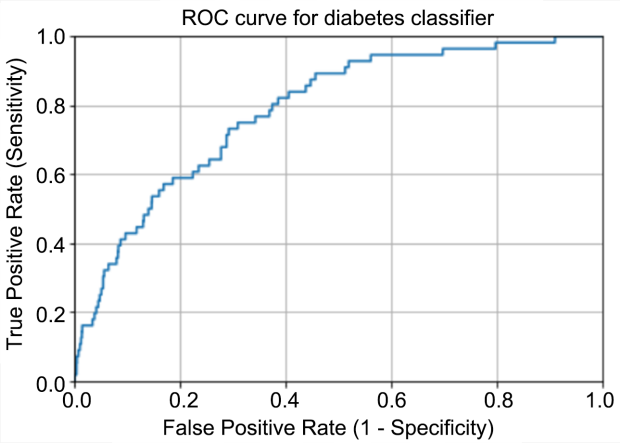

Figure 5. ROC curve
图 5. ROC 曲线图

LightGBM 算法调参后 AUC 有 75%，AUC 接近 1.0，检测方法真实性高，表明模型有较好的拟合效果。为得出对上市公司造假有较大影响的因子，我们将特征因子作图，对算法的重要性从大到小排列，其特征因子和其重要性数值如下图 6 所示。

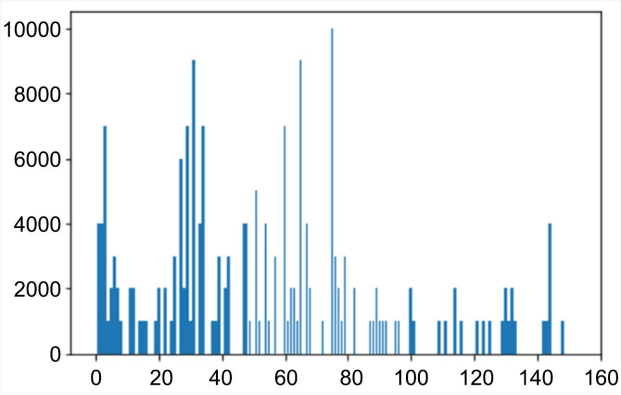

Figure 6. LightGBM indicator importance histogram
图 6. LightGBM 指标重要性柱状图

我们求出在 XGBoost 算法中特征重要性为前 30 的特征因子[10]中，其中前 10 的指标为 REVENUE、COGS 、 PAID_IN_CAPITAL 、 RETAINED_EARNINGS 、 T_PROFIT 、 AR 、 DEFER_REVENUE 、MINORITY_INT、PROC_SELL_INVEST、C_OUTF_OPERATE_A。

3.3.5. CATBoost 调参过程

CatBoost 算法调整参数 Learning_rate、Depth 和 Bagging_temperature。Learning_rate 不再阐述，参数 Depth 表示树的深度，bagging_temperature 表示贝叶斯套袋控制强度，区间[0, 1]，默认为 1。如表 5 所示：

Table 5. CATBoost invokes the procedure
表 5. CATBoost 调参过程

参数名称	初始值	调参范围	结果	调参后的 AUC
Learning_rate	0.3	[0.3, 0.2, 0.1]	0.2	0.671
Depth	4	[4, 2, 1]	1	0.7859

CATBoost 算法调参后 AUC 有 79%，AUC 接近 1.0，检测方法真实性高，表明模型有较好的拟合效果。为得出对上市公司造假有较大影响的因子，我们将特征因子作图，对算法的重要性从大到小排列，其特征因子和其重要性数值如下图 7 所示：

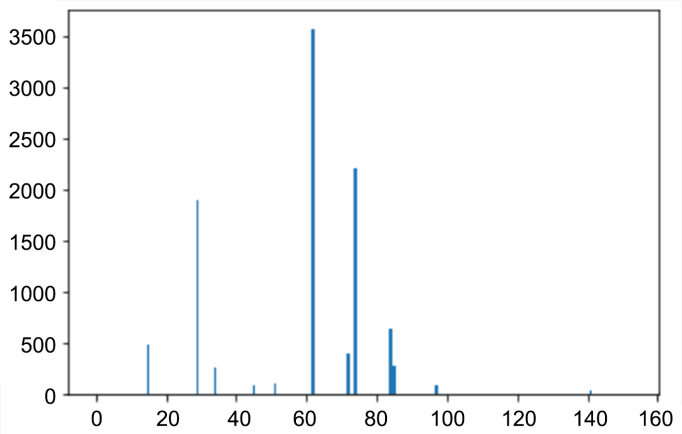

Figure 7. CATBoost indicator importance histogram
图 7. CATBoost 指标重要性柱状图

3.4. 确定对决策影响较大的因子

本文在上述 4 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，利用四个模型所得特征重要性，由于特征变量较多，我们已经选择了这四个模型排名前 25 个重要特征。在此基础上，挑选出三个模型共同确定的重要因子，并进行分类，得到结果如下表 6：

Table 6. Results
表 6. 结果

指标排序	造假指标	指标(中文)
1	**REVENUE**	收入
2	**COGS**	主营业务成本
3	**T_PROFIT**	净利润
4	**AR**	应收账款
5	**MINORITY_INT**	非控股股东权益
6	**R_TR**	内部收益率
7	**T_REVENUE**	总收入
8	**T_REVENUE_YOY**	总收入同比变化
9	**LT_BORR**	财务分析报告
10	**PAID_IN_CAPITAL**	实收资本
11	**CAP_FIX_RATIO**	固定资本比率
12	**DEFER_REVENUE**	递延收入
13	**T_CL**	总现金流入
14	**C_OUTF_OPERATE_A**	
15	**OTH_RECEIV**	其他收入
16	**SELL_EXP**	卖进费用

续表

17	**C_INF_FR_FINAN_A**	
18	**PREPAYMENT**	预付账款
19	**RETAINED_EARNINGS**	留存收益
20	**BIZ_TAX_SURCHG**	商业税公司
21	**DIV_PAYABLE**	财务比率
22	**N_CF_OPERATE_A**	现金流量净现值
23	**ST_BORR**	ST 股财报
24	**C_FR_BORR**	现金转账财报
25	**C_PAID_FOR_TAXES**	支付的各项税费

4. 基于模型融合的预测模型构造

4.1. 融合模型对比

本文用了 Stacking 模型融合方法，即为使用树行计算方法的集成学习方法。由于人解决问题的思维是树形的，将模型树行化符合问题本身的逻辑，精确率和召回率呈稳态正相关。因此采用树行计算方法的 Stacking 方法可以整合不同模型的最好表现，使模型融合更加科学化，用以提升模型的预测准确率和泛化能力。Stacking 融合模型一般分为 2 层内容。第 1 层模型主要用于产生第 2 层模型的训练集数据(TrainingData)。产生过程如下：首先，训练集数据内容是用一个基础模型进行 k 折交叉验证的结果。k 折交叉验证，就是先拿出 k-1 折作为训练数据，另外一折作为测试数据(TestingData)。每一个交叉验证产生的预测结果组合起来，作为第 2 层模型的训练集数据。另外，还要对数据集原来的整个训练数据进行预测，这个过程会生成 k 个预测数据集，对于这部分数据，本文将数据集各部分相加取平均作为下一层模型的测试集数据。第 2 层学习模型采用非线性模型，通过将第 1 层模型输出的结果作为训练数据训练模型，得到新的预测结果。通过将新的预测结果和第 2 层模型的测试数据集进行对比，观察预测准确度。如图 8 所示：

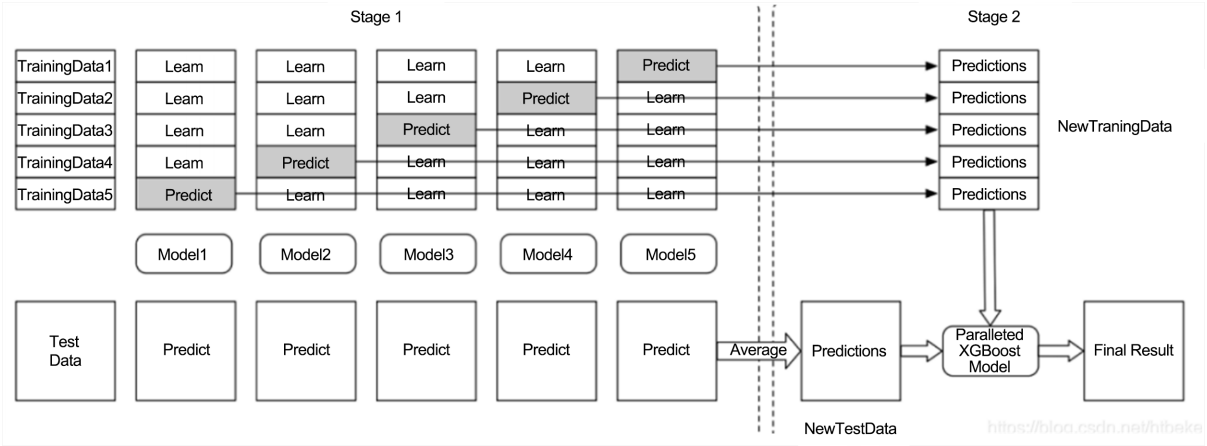

Figure 8. Stacking flowchart
图 8. Stacking 流程图

Stacking 集成学习算法的效果好坏取决于两个方面：一个是基分类器的预测效果，通常基分类器的预测效果越好，集成学习模型的预测效果越好；另一个是基分类器之间需要有一定的差异性，因为每个

模型的主要关注点不同，这样集成才能使每个基学习器充分发挥其优点。本文基于经典的 Stacking 模型融合方法进行改进：1) 将其每一步验证中使用的单个相同模型，改为 5 个不同的机器学习模型进行预测；2) 使用第 1 层所有基分类器所产生的类别概率值作为第 2 层模型的输入。

Stacking 相比 LinearBlending 来说，更加强大，然而也更容易过拟合。本文采用 5 折交叉验证法求解，用到的层次为两个层次，这个方法叫做 stacking 堆叠。

第一层用到了 XGBoost、Light GBM、随机森林、逻辑回归及 CatBoost 五种模型，5 种算法不仅有充分的理论支撑，而且在科学研究中正扮演着重要的角色。第 2 层元学习器同样选择学习能力较强的 Light GBM 算法，用于对第 1 层基学习器的集成，并且使用 7 折交叉验证划分数据的方式防止过拟合的发生，综上所述，本文基于 Stacking 集成学习的分类模型第 1 层基学习器选择 LGBM、RFC、XGBoost、Light GBM、Catboost，第 2 层元学习器选择 light GBM，模型结构如图 9 所示，通过此方法可以得到第六年的预测结果，整理得到如下饼图：

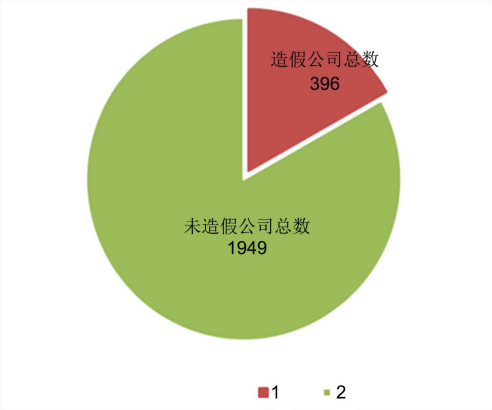

Figure 9. Pie chart of the total number of manufacturing enterprise whether to fake
图 9. 制造企业是否造假企业总数饼图

可以得到，第六年造假公司有 396 家，未造假公司有 1949 家，如下表 7 所示：

Table 7. Number of forgery company
表 7. 造假公司编号

	1	2	3	4	5	6
1	4019	461623	1047928	1452732	1898974	2363841
2	4213	467739	1050124	1463106	1908690	2364212
3	31035	511361	1060623	1475529	1915230	2374556
4	50357	535187	1072740	1499870	1917077	2383853
5	73338	548784	1087133	1508169	1917131	2391741
6	75645	569860	1102339	1508508	1968470	2396884
7	79499	583371	1103553	1512805	1975301	2435254
8	81575	607310	1114489	1533276	1993579	2435904
9	97965	612944	1126031	1534091	2011661	2436184
10	131509	630294	1133137	1536458	2024615	2442440
11	135231	635943	1136701	1545745	2034931	2444790

续表

12	155465	637281	1153502	1578921	2040251	2445197
13	166534	641851	1162478	1634387	2043333	2450989
14	166659	657368	1165331	1644007	2046826	2456492
15	189771	668840	1230774	1663579	2111801	2500113
16	198262	684783	1231430	1689193	2112284	2527742
17	284358	694894	1233831	1722591	2113648	2544142
18	287574	721454	1241747	1741317	2135452	2564968
19	302307	725938	1251220	1787928	2136960	2572928
20	305202	770959	1254216	1790666	2165071	2625107
21	312467	772122	1257862	1794373	2178858	2626870
22	313420	797107	1263748	1802096	2180481	2634733
23	325632	811328	1273541	1802653	2214474	2657564
24	342901	820979	1286992	1811362	2219887	2662475
25	349646	823133	1304725	1818220	2233158	2665725
26	349920	824674	1331316	1826967	2239503	2670976
27	376089	859910	1346405	1850533	2242597	2673748
28	399309	877559	1347287	1854599	2246642	2674558
29	407079	894508	1370722	1856192	2267190	2675963
30	413331	917791	1373516	1868870	2268406	2680698
31	413459	923620	1381078	1874358	2282829	2706709
32	422647	947351	1394196	1876786	2294861	2727652
33	431877	965153	1409661	1878137	2318006	2729166
34	449715	993505	1430872	1885796	2323008	2730037
35	449726	1009709	1432553	1891959	2347419	2731130
36	455179	1044654	1444042	1896420	2357255	2745140
37	2761144	4772233	4514761	4074874	4354943	3090508
38	2777010	4773317	4520607	4095867	4362849	3110158
39	2790290	4773872	4563832	4110715	4376158	3113925
40	2792275	4775806	4569042	4136067	4390993	3122352
41	2795783	4785950	4589100	4157869	4415670	3164095
42	2798174	4789867	4593225	4161524	4425487	3195475
43	2811617	4795842	4594694	4174154	4429333	3205897
44	2826195	4835102	4601114	4178256	4445741	3214052
45	2838489	4838928	4621970	4186839	4451155	3217869
46	2847211	4857304	4638024	4191281	4455960	3220876
47	2847618	4859692	4640362	4200913	4456759	3248102
48	2881859	4882136	4652939	4211486	4471087	3254864
49	2990894	4890706	4662982	4239770	4471107	3267677
50	3017801	4897311	4663150	4259116	4474326	3277491
51	3020358	4898392	4687189	4265385	4475502	3287799

续表

52	3027989	4941663	4694403	4272413	4479025	3315757
53	3028649	4947245	4694503	4285172	4490413	3328655
54	3033695	4955770	4715729	4286287	4507053	3336683
55	3040312	4961319	4720778	4290778	3882975	3363404
56	3052129	4968617	4724087	4314899	3906022	3379549
57	3056128	4986535	4731575	4337932	3910370	3412883
58	3085552	4992858	4735999	3695289	3923399	3418224
59	3087276	4998808	4736790	3695809	3925532	3423888
60	3587014	3646332	4758209	3728995	3946194	3424430
61	3596878	3646698	3538572	3783229	3983344	3425757
62	3601904	3679000	3539128	3796888	4004646	3435688
63	3612127	3682953	3563206	3810059	4041432	3440528
64	3644236	3686245	3565648	3814950	4050050	3444357
65	3835191	3691837	3570756	3816511	4065876	3456641
66	3517308	3857935	3536395	3819610	3850081	3458266

Stacking 集成学习的分类模型即为：第一阶段，将训练数据均匀地分成 5 份，使用“留一法”训练 5 个逻辑回归模型，用这 5 个模型分别去预测剩下的一份训练数据和测试数据，将 5 份预测的训练数据合并，可以得到一份新的训练数据 NewTrainingData，将 5 份预测的测试数据采用均值法合并，得到一份新的测试数据 NewTestData。用同样的方法再分别训练随机森林、gbdt、XGBoost，light GBM，新的训练和测试数据上，就可以得到 5 个模型的分数。第二阶段，将上一阶段的 NewTraningData 作为训练数据，NewTestData 作为测试数据，重新训练一个 XGBoost 模型，得到最终的预测分数。这种方法可以避免过拟合，学习出特征之间组合的信息，提高预测的准确率。

本文第 1 层基学习器选择 XGBoost、Light GBM、随机森林、逻辑回归及 CatBoost 第 2 层元学习器。选择了 light GBM，从而确定了最终最优的预测模型。Stacking 集成模型在测试集上的得分高于所有基础分类器，Stacking 集成模型在测试集上的得分高于所有基础分类器，其 AUC 得分为 78.5%，可得出其 ROC 曲线图如下图 10：

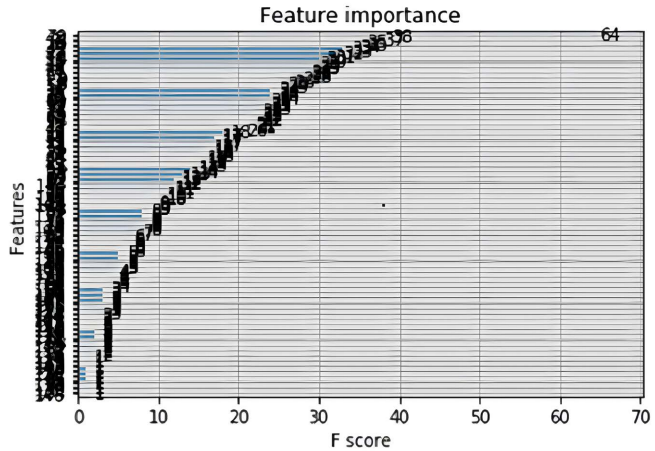

Figure 10. ROC curve
图 10. ROC 曲线图

4.2. 结果讨论

通过对下一年我们选取出的非制造是否高送转的 Stacking 集成学习融合模型的建立，使用第 5 年数据和基础数据数据预处理并特征选择后的合并数据中的特征因子去预测 1323 家非制造企业第 6 年是否造假。如图 11 所示：

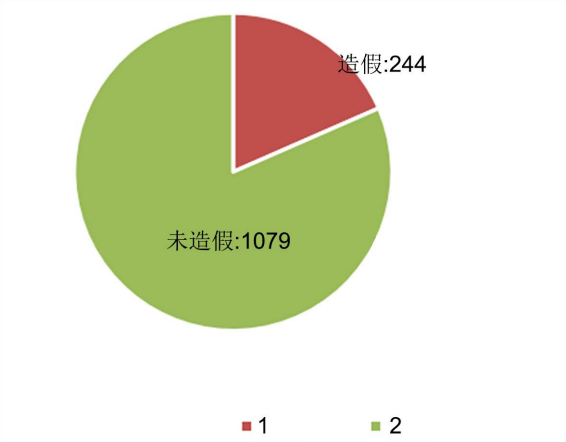

Figure 11. Pie chart of the total number of non-manufacturing enterprise whether to fake
图 11. 非制造企业是否造假企业总数饼图

得到预测结果：存在 18.4%非制造企业会在第六年出现造假现象，即有 244 个上市企业造假，1079 个企业不会造假。为方便观察，绘制第 6 年企业是否会造假，0 表示不会造假，1 表示执行高送转，如上图 11 所示。

利用 Stacking 集成学习融合模型，我们使用第 5 年数据和基础数据数据处理特征选择后的合并数据中的特征因子去预测 1323 家非制造企业可能造假的具体企业，得出以下结果，如表 8：

Table 8. Specific counterfeiting enterprises of non-manufacturing enterprises
表 8. 非制造企业具体造假企业

	1	2	3	4	5	6
1	10083	1682405	3326606	933562	2574856	4330281
2	38184	1682715	3352332	962308	2584388	4357417
3	62331	1725654	3363628	962466	2603195	4367525
4	73662	1726484	3387626	974393	2611171	4382140
5	79573	1731244	3389413	985873	2640939	4388517
6	84271	1818571	3429869	991606	2659712	4431385
7	85904	1820481	3445220	996153	2665107	4432826
8	98122	1824625	3461548	996441	2682553	4433656
9	162606	1856393	3463726	1083660	2705618	4442582
10	169426	1874536	3468727	1115066	2745216	4445647
11	195183	1892808	3473078	1182264	2747835	4455639
12	214798	1907820	3488598	1201147	2769232	4456365
13	232852	1922317	3499832	1203334	2790878	4480980

续表

14	249234	1936514	3511485	1209362	2796741	4564086
15	266963	1970382	3530822	1227673	2812981	4570965
16	286880	1976303	3578030	1269286	2816927	4647549
17	297926	2017677	3583157	1313450	2829042	4661966
18	301092	2040254	3618063	1345499	2829669	4669213
19	686427	2399070	4187334	1656791	3262635	4887547
20	693604	2402604	4192287	4938803	825566	2497957
21	315965	2042348	3640933	1359300	2843139	4678393
22	316520	2049023	3648636	1362836	2844622	4680149
23	325727	2070248	3683463	1373528	2849337	4690399
24	334688	2114004	3719746	1379943	2895524	4710457
25	350818	2134524	3804323	1471977	2918873	4727328
25	364305	2151056	3864884	1490049	2923338	4734851
27	389141	2172483	3872511	1513417	2944368	4741756
28	390627	2176533	3889707	1516761	3056391	4749099
29	394067	2183203	3942923	1522220	3058236	4755501
30	431111	2185630	3973471	1524574	3065911	4763536
31	491001	2221349	3975089	1525496	3118775	4786358
32	545184	2250948	3980150	1568226	3137668	4800652
33	597339	2303318	3984945	1596218	3174326	4825493
34	620668	2312291	4031377	1625847	3176303	4839163
35	628392	2313379	4089079	1633079	3177888	4840667
36	655040	2332077	4102514	1644358	3182805	4841735
37	657586	2339115	4117272	1645542	3189773	4858120
38	672970	2379742	4147734	1656299	3208224	4868990
39	759602	2417596	4198575	4942421	874870	2558544
40	786249	2435283	4214318	4946014	4237084	812418
41	2488211	4229984	4968271	4245500		

由此可以看出：此融合模型不仅预测了上市公司财务造假数量，也预测了具体造假公司，为打击财务造假问题提供了具体可行的参考结果。

基金项目

甘肃省科技计划项目(定西地区农村电子商务营销综合能力提升) (20CX9NA095)。

参考文献

[1] 高利芳, 李艺玮. 职务舞弊的内部审计困境与准则完善[J]. 财经问题研究, 2019(8): 104-112.

[2] 戴丹苗, 刘锡良. 中概股公司财务舞弊的文献综述[J]. 金融发展研究, 2017(1): 11-19.

[3] 邢小艳. 基于模式识别的“高送转”投资策略研究[D]: [硕士学位论文]. 广州: 华南理工大学. 2016: 810.

[4] 桂萍, 王婷. 高管变更、内部控制质量与公司财务造假[J]. 财会月刊, 2018(10): 85-87.

[5] 黄世忠. 上市公司财务造假的八因八策[J]. 财务与会计, 2019(16): 4-11.

[6] 朱卫东, 苏剑, 武子豪. 高管特征与真实盈余管理——基于随机森林的实证[J]. 会计之友, 2022(12): 100-107.

[7] 段亚萍. 基于 XGBoost 算法优化 BP 神经网络的新能源汽车专利价值评估[D]: [硕士学位论文]. 重庆: 重庆理工大学, 2023. https://doi.org/10.27753/d.cnki.gcqgx.2023.000216

[8] 张涵夏. 适用于线性回归和逻辑回归的场景分析[J]. 自动化与仪器仪表, 2022(10): 1-4+8. https://doi.org/10.14016/j.cnki.1001-9227.2022.10.001

[9] 薛慧. 基于 LightGBM 模型的电力上市公司财务风险预警研究[D]: [硕士学位论文]. 成都: 西南财经大学, 2021. https://doi.org/10.27412/d.cnki.gxncu.2021.002834

[10] 陈业辉, 郑克立, 陈立中, 等. 口服他克莫司血药浓度-时间曲线下面积[J]. 中华泌尿外科杂志, 2004(11): 29-31.

[11] 李子言. 大数据背景下 ROC 曲线介绍与应用[J]. 科教导刊, 2021(14): 81-84. https://doi.org/10.16400/j.cnki.kjdk.2021.14.026

[12] 崔智泉. 浅谈高斯分布的原理和应用[J]. 中国校外教育, 2018(16): 63-64.

[13] 贺怀清, 贾洁, 刘浩翰. 基于方差过滤的改进多通路 Metropolis 光线传输算法[J]. 计算机辅助设计与图形学学报, 2018, 30(6): 1082-1088.

[14] 龚晓彦. 基于互信息的医学图像配准算法研究[D]: [硕士学位论文]. 秦皇岛: 燕山大学, 2010.