Advances in Applied Mathematics 应用数学进展, 2021, 10(9), 2923-2932
Published Online September 2021 in Hans. http://www.hanspub.org/journal/aam
https://doi.org/10.12677/aam.2021.109306

Hans汉斯

基于CatBoost模型的上市公司财务困境预测

张朋燕，戴家佳

贵州大学数学与统计学院，贵州 贵阳
Email: 1469882658@qq.com

收稿日期：2021年7月31日；录用日期：2021年8月21日；发布日期：2021年9月2日

摘　要

本文按照财务异常与财务正常的公司数量为1:2的比例选取了2019~2020年间63家被特殊处理的上市公司和126家财务正常的上市公司。从公司发展能力、经营能力、盈利能力、股本结构、人员变动等十个方面的51个变量中利用统计方法和随机森林方法筛选出17个重要变量，在此基础上首次运用CatBoost模型进行公司财务困境问题上进行建模和预测，并与之前学者经常使用的XGBoost、随机森林和逻辑回归这三种建模方法进行对比。实验结果显示CatBoost模型在特异性、准确率和AUC值方面都要优于其他三种模型，其中准确率达到了98.2%，在准确性方面要比近年广泛运用且预测效果较好的XGBoost模型高3.5%。

关键词

变量筛选，随机森林，CatBoost，XGBoost，逻辑回归

Financial Distress Prediction of Listed Companies Based on CatBoost Model

Pengyan Zhang, Jiajia Dai

School of Mathematics and Statistics, Guizhou University, Guiyang Guizhou
Email: 1469882658@qq.com

Received: Jul. 31st, 2021; accepted: Aug. 21st, 2021; published: Sep. 2nd, 2021

Abstract

According to the ratio of 1:2 between the number of abnormal and normal financial companies, this paper selects 63 specially treated listed companies and 126 normal financial listed companies from 2019 to 2020. From the 51 variables of the company's development ability, operation ability, profitability, equity structure and personnel change, etc., 17 important variables are selected by

文章引用: 张朋燕, 戴家佳. 基于 CatBoost 模型的上市公司财务困境预测[J]. 应用数学进展, 2021, 10(9): 2923-2932.
DOI: 10.12677/aam.2021.109306

using statistical method and random forest method. On this basis, CatBoost model is used to model and predict the company's financial distress for the first time and it is compared with the three modeling methods of XGBoost, Random forest and Logistic regression that are often used by previous scholars. The experimental results show that CatBoost model is superior to the other three models in terms of specificity, accuracy and AUC value, of which the accuracy is 98.2%, which is 3.5% higher than XGBoost model, which is widely used in recent years and has good prediction effect.

Keywords

Variable Screening, Random Forest, CatBoost, XGBoost, Logistic Regression

1. 引言

公司财务困境是一个被广泛持续研究的课题，因为建立一套有效的财务困境预警系统对于企业以及其他企业利益相关者来说具有重要的现实意义。根据财务困境预警系统预测得到的财务状态评估结果，可以监控企业经营状况同时防范财务困境的出现或者财务恶化。具体来讲，企业经营者、投资者、银行等金融机构、审计人员和法律工作者等可以通过预警系统的提示，提前防范潜在的财务危机，从而调整战略降低或避免损失。

机器学习算法中逻辑回归应用最广泛、最经典。Martin [1]首次运用逻辑回归，对银行进行破产预测分析，发现多元逻辑回归模型的预测性能较好。紧随其后的 Ohlson [2]运用多元逻辑回归方法将财务危机的预测问题转化为了根据公司的部分财务特征来估计公司在一段时间内陷入财务危机的概率问题，但高度相关的财务指标可能存在严重的多重共线性问题，这是多元逻辑回归方法要解决的一个难点。邓晓岚等[3]都构建了纯财务指标逻辑回归模型和综合逻辑回归模型，结果发现后者判别效果更好。随着统计学、计算机技术及信息技术的快速发展，财务预警模型的构建在不断创新。刘可歆[4]通过因子分析计算财务指标综合因子得分，利用随机森林构造互联网上市公司财务预警模型并对模型进行优化，解决不平衡分类问题，通过多种机器学习模型比对得到最优模型，进行财务指标重要性评价。苏玉敏[5]运用 XGBoost 原理构建财务预警模型，并将预测效果与经典财务预测模型效果进行对比，得到结果为 XGBoost 模型预测准确率达 89.17%，比逻辑回归模型的预警效果要优秀很多，从而能够较好地预警公司是否将陷入危机，满足企业进行财务预警的需求。

随着机器学习的发展，2017 年俄罗斯搜索巨头 Yandex 提出了 CatBoost 这一算法，它训练速度快、泛化能力强、准确率高，因此在众多应用领域中有很优异的表现，如苗丰顺[6]在医疗数据方面使用 CatBoost 算法，该算法预测效果都优于 XGBoost 和随机森林模型，取得了显著的预测效果。Wang [7]将 CatBoost 算法应用于构建 P2P 违约预测模型，得到 CatBoost 算法的预测准确率达 96%，对实际结果的拟合效果较好。

在财务危机预测领域中，学者们对于哪个机器学习算法的性能最好还没有统一的看法。本文新构建出一个影响公司财务困境的非财务指标，即人员年变化率，其中人员年变化率 = (t − 2 年在职员工数量 − t − 3 年在职员工数量)/t − 3 年在职员工数量，这有利于拓展非财务指标的研究；另外由于 CatBoost 这种新算法特别适合样本量小、数据不平衡的情况，并且该算法目前在公司财务困境预测方面还没有应用，

因此用来研究公司困境的预测问题。

2. 模型相关理论

2.1. 随机森林

随机森林(Random Forest)是一种利用 bootstrap 自助抽样方法的集成算法，它采用决策树和 bagging 的结合模式，即在以多棵决策树并行的基础上结合成为一个强学习器。名字中的“随机”是指随机抽取样本和随机选择特征，“森林”则是指多棵不同类型的决策树像是森林一样。在构建决策树过程中需要对变量重要性进行排序，由于随机森林拥有大量决策树，将每棵决策树得到的变量重要性进行综合，可以得到最终的变量重要性排序结果，且结果比单棵决策树更加稳定、可信，因此随机森林方法可以用来进行特征选择，筛选出重要变量。

在分类问题研究中，随机森林的多棵树分类器投票机制决定最终分类结果，另外还可以给出各个变量的进行重要性评分，评估各个变量在分类中所起的作用。

2.2. CatBoost 模型

CatBoost 算法(Categorical Boosting)是在梯度提升决策树(GBDT)的框架上进行改进的算法，它能够处理好各种分类型数据，并且易于调参，是一种比 XGBoost 算法计算结果更准确更优秀的机器学习算法。

CatBoost 的初始目的是改进 GBDT 的分类特征，因为之前的处理方法是标签均值来代替对应的分类特征，计算公式为：

$$\hat{x}_k^i = \frac{\sum_{j=1}^{p-1}\left[x_{\sigma j,k} = x_{\sigma p,k}\right]Y_{\sigma j}}{\sum_{j=1}^{p-1}\left[x_{\sigma j,k} = x_{\sigma p,k}\right]}$$

这种计算方式的缺点强行用标签均值来代替相应特征值从而忽略了更多信息，可能会造成条件偏移问题。而 CatBoost 对统计量进行了改进，在原有基础上引入了先验分布项及其对应的权重，新的计算公式为：

$$\hat{x}_k^i = \frac{\sum_{j=1}^{p-1}\left[x_{\sigma j,k} = x_{\sigma p,k}\right]Y_{\sigma j} + \alpha \cdot P}{\sum_{j=1}^{p-1}\left[x_{\sigma j,k} = x_{\sigma p,k}\right] + \alpha}$$

其中 α 是先验分布项的权重($\alpha > 0$)，P 是先验项，这样做一是能够减少分类型变量中类别少的变量对数据的影响；二是能够有效的减少噪声。

CatBoost 的另一改进是对将传统梯度估计方法改进为排序提升方法(Ordered boosting)，这会得到无偏的梯度估计，降低了梯度估计误差，从而降低过拟合问题，最终达到提高模型泛化能力的目的。

2.3. XGBoost 模型

XGBoost 全称极端梯度上升(eXtreme Gradient Boosting)，也是一种对梯度提升算法的改进算法。它对损失函数做了改进，一方面是将原来的损失函数从一阶泰勒展开替换为二阶泰勒展开，将损失函数泰勒展开到二阶，另一方面是损失函数中引入了正则化项，损失函数表达式为：

$$L_t = \sum_{i=1}^{m} L\left(y_i, f_{t-1}\left(x_i\right) + h_t\left(x_i\right)\right) + \gamma J + \frac{\lambda}{2}\sum_{j=1}^{J} w_{tj}^2 > 0$$

$$L_t \approx \sum_{i=1}^{m}\left(L\left(y_i, f_{t-1}\left(x_i\right)\right) + \frac{\partial L\left(y_i, f_{t-1}\left(x_i\right)\right)}{\partial f_{t-1}\left(x_i\right)} h_t\left(x_i\right) + \frac{1}{2}\frac{\partial^2 L\left(y_i, f_{t-1}\left(x_i\right)\right)}{\partial f_{t-1}^2\left(x_i\right)} h_t^2\left(x_i\right)\right) + \gamma J + \frac{\lambda}{2}\sum_{j=1}^{J} w_{tj}^2$$

这个算法的思想是增加树，之后进行特征分裂来生长这棵树，每增加一棵树，就是学习一个新函数，进而去拟合上次预测的残差。最后根据树的结构就可以得到这个结构下的最优分数，可以通过每棵树的叶子节点去计算总分数。

2.4. 逻辑回归

逻辑回归模型(Logistic regression model)是一种广义线性模型，适用于处理结果变量为二值型变量或分类变量的问题，即结果变量 Y 可以取 0 或 1 (其中 $Y=1$ 表示事件发生，$Y=0$ 表示事件不发生)。因此当处理实际问题时结果变量为分类变量，这时线性回归模型已经不再适用，而逻辑回归就可以很好的处理这类问题，它的定义如下：

具有 p 个独立变量的向量 $x=\left(x_1,x_2,\cdots,x_p\right)$，条件概率 $P\left(Y=1\middle|x\right)=p$ 为在 x 观测值的条件下某事件发生的概率。逻辑回归模型可表示为

$$p=P\left(Y=1\middle|x\right)=\frac{1}{1+\mathrm{e}^{-g(x)}},$$

其中 $g\left(x\right)=\beta_0+\beta_1x_1+\beta_2x_2+\cdots+\beta_px_p$。事件不发生的条件概率为

$$1-p=P\left(Y=0\middle|x\right)=1-P\left(Y=1\middle|x\right)=\frac{1}{1+\mathrm{e}^{g(x)}}.$$

那么事件发生与事件不发生的概率比定义为机会比(odds)，则有

$$\text{odds}=\frac{P\left(Y=1\middle|x\right)}{P\left(Y=0\middle|x\right)}=\frac{p}{1-p}=\mathrm{e}^{g(x)}.$$

由于 $0<p<1$，可以对 odds 取对数，得到一个下面的线性函数

$$\log\left(\frac{p}{1-p}\right)=g\left(x\right)=\beta_0+\beta_1x_1+\beta_2x_2+\cdots+\beta_px_p.$$

逻辑回归模型的参数可以通过极大似然估计得到，之后将测试集数据带入机会比(odds)式子中就可以得到预测结果，通常人为的以 0.5 为界限，即 $\text{odds}\geq 0.5$ 判定为事件发生($Y=1$)，$\text{odds}<0.5$ 判定为事件不发生($Y=0$)。

3. 实证分析

3.1. 数据来源

本文所研究的数据和变量均来自 CSMAR 数据库(China Stock Market & Accounting Research Database)。在选取财务状况异常的公司样本时，先从 CSMAR 数据库中选取 2019~2020 年间被 ST (特殊处理)的不含金融类等特殊行业的所需指标完整的上市公司，并且这些公司此前从未出现过被 ST 的情况，符合条件的财务困境公司共有 63 家(2019 年 10 家，2020 年 53 家)。在确定财务困境公司与财务正常公司的配比数量时，目前并没有统一的方式，有的学者直接按照 1:1 的原则进行配比，如 Altman [8]、王克敏[9]等；有的学者按照财务困境公司占所有公司的比重进行配比的研究表明，如 Platt [10]；另外还有学者按照 1:2 的比例来确定财务困境公司和正常公司，如马铭阳[11]。为保证样本具有可比性以及研究样本能够提供充足的信息，本文在选取财务状况正常的公司样本时，按照财务异常与财务正常的公司数量为 1:2 的比例选取正常公司，并且选取的正常公司样本应与被 ST 的公司具有相近上市年限、公司规模和同一行业，选取 126 家符合条件的正常公司(2019 年 20 家，2020 年 106 家)。

以被判定为财务困境的当年为第 t 年，财务因素变量的全部数据和非财务因素变量中的国家持股比例、前十位股东持股和审计意见可直接从 CSMAR 数据库中获取，均为发生财务困境的前两年，即 t − 2 年。而非财务因素变量中的人员年变化率为自己构建的变量，需要 t − 3 和 t − 2 年的在职员工数量数据，变化率 =(t − 2 年在职员工数量 − t − 3 年在职员工数量)/t − 3 年在职员工数量。

3.2. 指标选择

财务困境预测模型的准确性和普适性依赖于构建完整的指标体系，不能遗漏掉重要的指标。另外，近些年众多学者的深入研究表明，能衡量公司发展状况的变量不只是财务因素，非财务因素也能很大程度的反映一部分信息。因此，综合近年来国内众多外学者对财务困境的研究成果[12]-[17]，本文从偿债能力、发展能力、经营能力、每股指标、现金流能力、盈利能力、股本结构、股权集中度、审计意见和人员变动这十个方面筛选出了 51 个指标，见表 1。

Table 1. Financial and non-financial index system
表 1. 财务与非财务指标体系

类型	一级指标	二级指标	时间
财务指标	偿债能力	流动比率(X_1)、速动比率(X_2)、现金比率(X_3)、经营活动产生的现金流量净额/流动负债(X_4)、资产负债率(X_5)、产权比率(X_6)、长期债务与营运资金比率(X_7)	t − 2 年
	发展能力	总资产增长率(X_8)、净资产收益率增长率(X_9)、净利润增长率(X_{10})、营业收入增长率(X_{11})、可持续增长率(X_{12})、所有者权益增长率(X_{13})、每股净资产增长率(X_{14})	t − 2 年
	经营能力	应收账款与收入比(X_{15})、应收账款周转率(X_{16})、存货与收入比(X_{17})、存货周转率(X_{18})、流动资产与收入比(X_{19})、流动资产周转率(X_{20})、固定资产与收入比(X_{21})、固定资产周转率(X_{22})、总资产周转率(X_{23})、股东权益周转率(X_{24})	t − 2 年
	每股指标	每股收益(X_{25})、每股营业收入(X_{26})、每股营业利润(X_{27})、每股净资产(X_{28})、每股负债(X_{29})、每股未分配利润(X_{30})、每股留存收益(X_{31})、每股企业自由现金流量(X_{32})、每股股东自由现金流量(X_{33})、每股现金净流量(X_{34})	t − 2 年
	现金流能力	营业收入现金含量(X_{35})、全部现金回收率(X_{36})、营运指数(X_{37})、现金适合比率(X_{38})、现金再投资比率(X_{39})、现金满足投资比率(X_{40})	t − 2 年
	盈利能力	资产报酬率(X_{41})、总资产净利润率(X_{42})、净资产收益(X_{43})、投入资本回报率(X_{44})、营业毛利率(X_{45})、营业净利率(X_{46})、销售期间费用率(X_{47})	t − 2 年
非财务指标	股本结构	国有股占比(X_{48})	t − 2 年
	股权集中	前十位股东持股(X_{49})	t − 2 年
	审计意见	审计意见类型(X_{50})	t − 2 年
	人员变动	人员年变化率(X_{51})，其中人员年变化率 =(t − 2 年在职员工数量 − t − 3 年在职员工数量)/t − 3 年在职员工数量。	t − 2 年、t − 3 年

3.3. 变量筛选

为了不遗漏重要变量，所以在上面指标选择时初步选入的 51 个变量，但变量太多不加以筛选，会使得对公司财务状况影响不大的变量进入预测模型，这将会对预测形成干扰且效率低下。因此必须筛选出有差异的重要变量。

3.3.1. 统计方法筛选

首先对个变量进行正态性检验，单样本的 K-S 检验(Kolmogorov-Smirnov 检验)可以比较频率分布与理论分布(正态分布)相符合。当 p 值 ≤ 0.05 为拒绝原假设，即不服从正态分布，这是可以利用 Mann-Whitney U 这种非参数检验方法进行变量显著性检验；当 p 值 > 0.05 时则表示变量服从正态分布，

此时可以继续利用 t 检验来对变量进行显著性检验。

利用 R 软件进行检验得到结果为：51 个变量在 K-S 检验中全部为 p 值小于 0.05，即这些变量不服从正态分布。接着对这些变量进行 Mann-Whitney U 检验，结果显示长期债务与营运资金比率(X_7)、每股负债(X_{29})、营业收入现金含量(X_{35})、前十位股东持股(X_{49})这 4 个变量未通过 Mann-Whitney U 检验，而剩余 47 个变量则通过该检验。

3.3.2. 随机森林方法筛选

随机森林方法不仅可以进行回归预测和分类预测，而且还可以计算变量重要性进行变量选择。衡量变量重要性的方法有两种，一种用置换精度降低衡量，原理为随机撤换某一个变量，如果预测精度降低就说明该变量重要；另一种用平均 Gini 指数降低衡量，即某个变量在拆分节点不纯度的总降低所体现出的重要性[18]。

上面小节的统计方法组合筛选只筛掉了 4 个变量，而剩余的 47 个变量仍旧包含许多冗余信息。因此本小节利用随机森林这种方法进行降维。最后用于模型预测的变量取两种方法的交集。利用随机森林计算指标体系中的 51 个变量的变量重要性。图 1 显示的时前 30 个重要变量。

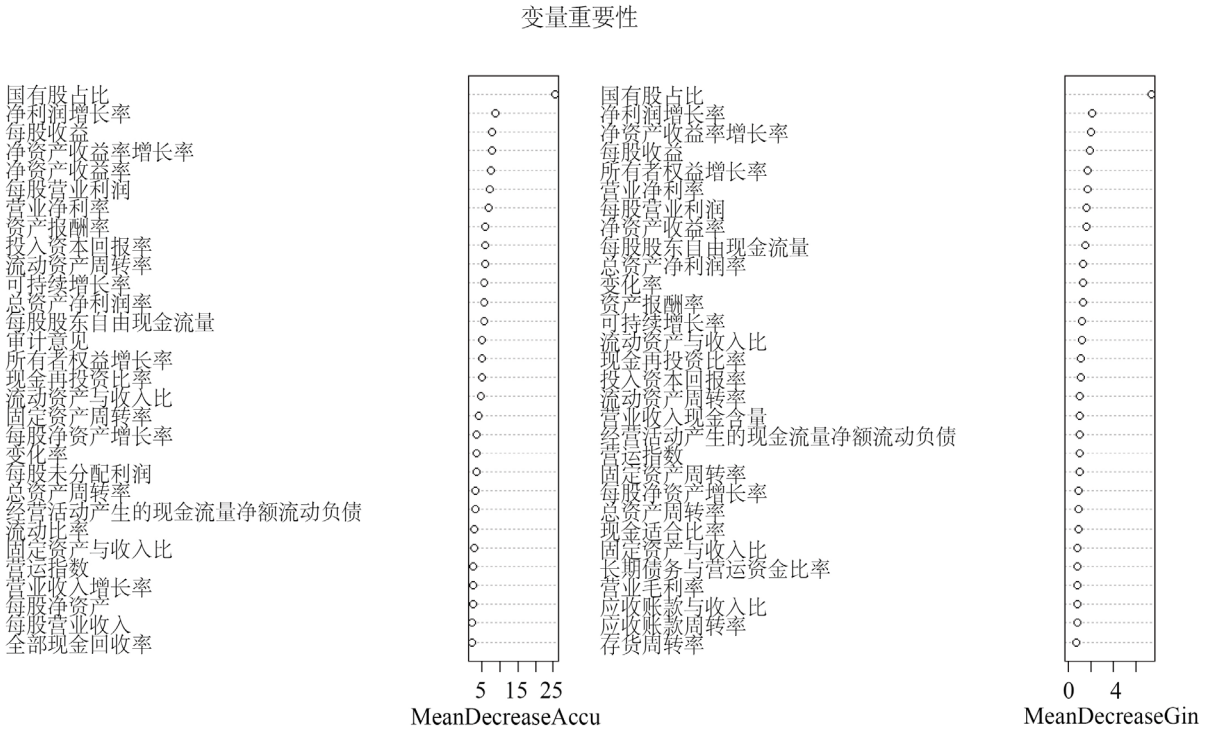

Figure 1. Importance of variables
图 1. 变量重要性

综合统计方法和随机森林两种方法的筛选，最终筛选出 17 个重要指标：净资产收益率增长率(X_9)、净利润增长率(X_{10})、可持续增长率(X_{12})、所有者权益增长率(X_{13})、流动资产与收入比(X_{19})、流动资产周转率(X_{20})、每股收益(X_{25})、每股营业利润(X_{27})、每股股东自由现金流量(X_{33})、现金再投资比率(X_{39})、资产报酬率(X_{41})、总资产净利润率(X_{42})、净资产收益率(X_{43})、投入资本回报率(X_{44})、营业净利率(X_{46})、国有股占比(X_{48})、变化率(X_{51})。

3.4. 预测模型结果对比

3.4.1. 模型建立

在完成变量筛选后，本文根据上面的筛选结果从 51 个变量中选出 17 个重要变量，其中包含 2 个非财务因素(国有股占比(X_{48})和变化率(X_{51}))和 15 个财务因素。将 189 个样本按照 7:3 的比例划分成训练集和测试集，其中训练集用于寻找规律构建模型，测试集用于做出预测，评价模型效果。实验目的是探究基于 CatBoost 这种新算法建立公司财务困境预测的优越性，并将其结果与之前学者经常使用的 XGBoost、随机森林和逻辑回归这三种建模方法的结果进行对比。

在 CatBoost 建模时，将筛选的 17 个变量作为自变量，以公司状态作为因变量，将 189 个样本以 7:3 的比例划分得到 X_train 和 X_test，相对应得到 y_train 和 y_test。由于 CatBoost 提供了一种 pool 数据结构，这样可以优化速度和内存占用。因此可以用 X_train 和 y_train 构建成 train_pool，X_test 和 y_test 构建成 test_pool。之后利用 CatBoostClassifier 函数进行分类问题的训练，使得 Logloss 达到最小得到最优参数的模型，最终使用最优模型对测试集结果进行评估。

在逻辑回归建模时，需要考虑变量间的多重相关性，因此不能直接用这些变量进行建模，本文利用逐步回归的方法进行计算，删去模型中不显著的变量，进行逻辑回归的优化，得到最终逻辑回归结果。

本文 CatBoost 算法是通过 Python 软件实现的，XGBoost、随机森林和逻辑回归这三种算法则是基于 R 软件进行分析和计算。

3.4.2. 模型评价指标

模型评价是建模完成后重要一环，用于综合评价模型的效果优劣。对于二分类问题，可基于混淆矩阵和 ROC 曲线评价。

混淆矩阵是是以矩阵形式来表现实际和预测分类效果，在此基础上定义了特异性、灵敏性、准确性，具体如表 2 所示。

Table 2. Confusion matrix
表 2. 混淆矩阵

实际	预测	
	0	1
0	TN	FP
1	FN	TP

特异性 $=\mathrm{TN}/(\mathrm{TN}+\mathrm{FP})$，又叫做真负类率；

灵敏性 $=\mathrm{TP}/(\mathrm{FN}+\mathrm{TP})$，又叫做真正类率；

准确性 $=(\mathrm{TN}+\mathrm{TP})/(\mathrm{TN}+\mathrm{FP}+\mathrm{FN}+\mathrm{TP})$。

ROC 曲线全称叫做接收者操作特征曲线，适用于分类模型中作为衡量指标，在做决策的时候能够给出客观中立的建议。ROC 曲线越靠左上角表明分类效果越理想，而如果是对角线状态则表明该模型没有分类效果，因此叫做无识别曲线。AUC 值表示的是 ROC 曲线下面的面积，取值范围为[0, 1]。当 $0.5<\mathrm{AUC}\le 1$ 时表明该模型是有预测效果的，优于随机猜想。

3.4.3. 实验结果

经过对数据进行数据处理和变量筛选，选出重要变量后进行模型预测，得到 CatBoost 模型的预测结果，同时也使用了之前学者经常使用的 XGBoost、随机森林和逻辑回归这三种建模方法，通过测试集预

测结果得到混淆矩阵和 ROC 曲线图，进而分别得到这四种模型的预测结果，如表 3 和图 2 所示。

Table 3. Analysis of prediction results of four models
表 3. 四种模型的预测结果分析

预测模型	特异性	灵敏性	准确性	AUC 值
CatBoost	1.000	0.933	98.2%	0.967
XGBoost	0.944	0.952	94.7%	0.948
随机森林	0.944	0.857	91.2%	0.901
逻辑回归	0.861	0.476	71.9%	0.669

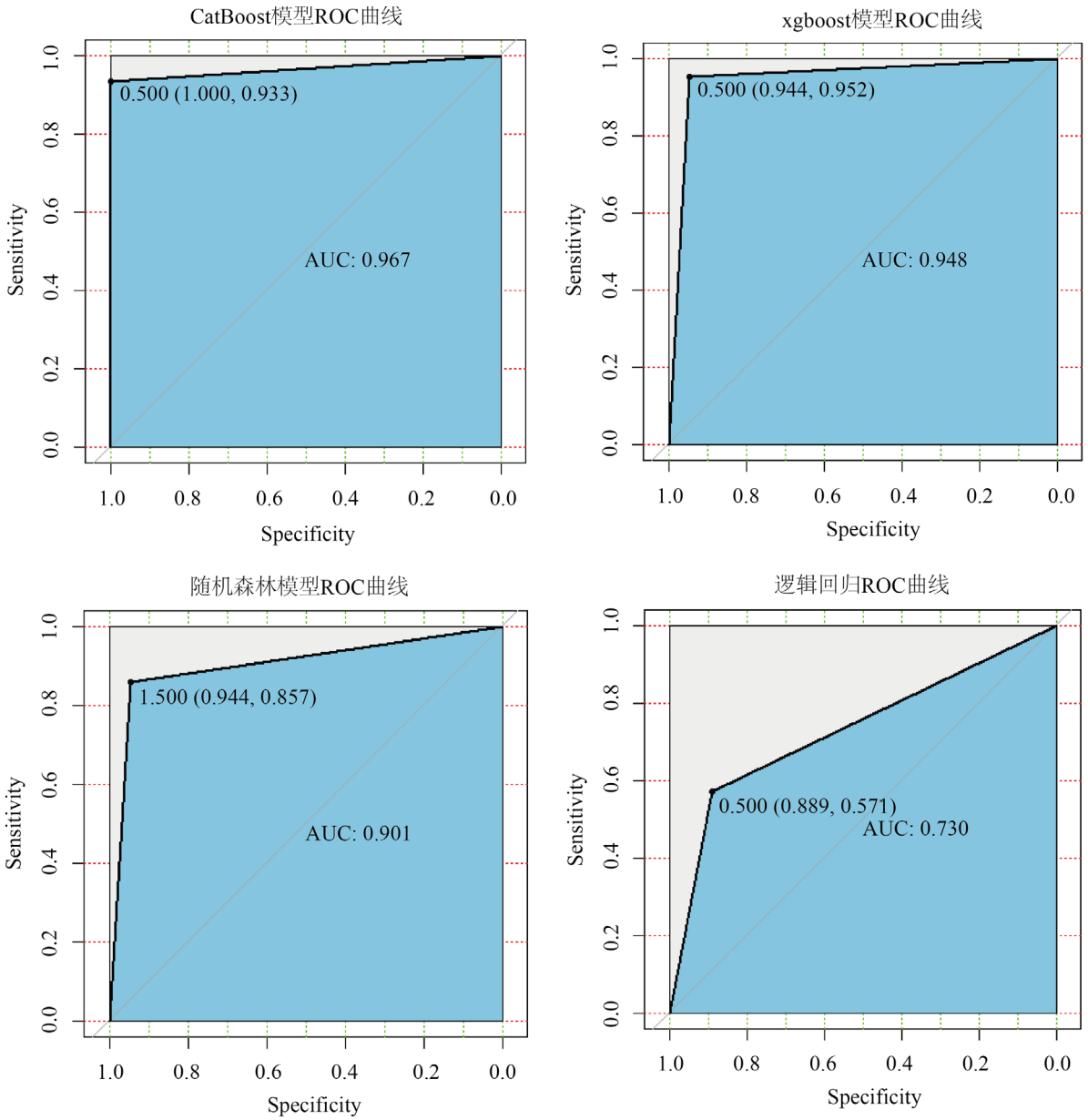

Figure 2. Comparison of ROC curves of four models
图 2. 四种模型 ROC 曲线对比图

模型预测结果可以通过特异性、灵敏性、准确性和 AUC 值度量，从图 2 和表 3 可以看出：(1) 特异性表示的是在测试集中实际为财务正常的公司里被正确的判定为财务正常的比率。在特异性方面，CatBoost 模型为 1，效果最好没有误判，XGBoost 和随机森林模型效果次之，逻辑回归效果最差，为 0.861。(2) 灵敏性表示的是在测试集中实际为财务困难的公司里被正确的判定为财务困难的比率。在灵敏性方面，XGBoost 结果略优于 CatBoost 模型，但效果也已经很好，而逻辑回归效果最差。(3) 准确性表示的是在测试集中的所有公司里被正确的判定为财务困难和财务正常的比率，是一个总体准确率，从结果中可以看出 CatBoost 模型准确率最高，达到了 98.2%，XGBoost 和随机森林模型效果稍弱。(4) AUC 值反映的是模型的分类性能的好坏，客观反映了分类的精度。从上面结果中可以看出 CatBoost 模型优于其他三种模型，AUC 值最大，从 ROC 曲线对比图中可以看出四种模型的 AUC 都大于 0.5，都具有预测价值，尤其是 CatBoost 模型，预测效果最好。

综合来讲，CatBoost 模型在特异性、准确率和 AUC 值方面都要优于其他三种模型，虽然在灵敏性方面略低于 XGBoost 模型，但灵敏度依旧很高。综合来讲，CatBoost 模型表现最好，都要优于几年来学者经常运用的 XGBoost 模型和经典的逻辑回归模型，在准确性方面要比近年广泛运用且预测效果较好的 XGBoost 模型高 3.5%。

4. 结论

本文以上市公司为研究对象，之前学者的研究证实影响公司财务状况的因素很多，不仅是财务因素，同时还有非财务因素。一般的非财务因素是从数据库中给出的几个方面进行研究，而员工离职情况也在另一方面反映着公司的运行情况，但这些数据不会对外公布，因此本文新构建出人员年变化率这一变量反映公司人员的变动情况，计算方式为人员年变化率 $=(t-2$ 年在职员工数量 $-t-3$ 年在职员工数量$)/t-3$ 年在职员工数量。为避免遗漏重要变量，本文先从发展能力、经营能力、盈利能力、股本结构、人员变动等十个方面的选中 51 个变量作为初始指标体系，再利用统计方法和随机森林方法去除冗余变量，最终从中筛选出 17 个重要变量。由于 CatBoost 这种新算法特别适合样本量小、数据不平衡的情况，具有训练速度快、泛化能力强、准确率高的优点，并且该算法目前在公司财务困境预测方面还没有应用，因此运用 CatBoost 模型进行公司财务困境建模和预测，并与之前学者经常使用的 XGBoost、随机森林和逻辑回归这三种建模方法进行对比。

实验结果显示 CatBoost 模型在特异性、准确率和 AUC 值方面都要优于其他三种模型，虽然在灵敏性方面略低于 XGBoost 模型，但灵敏度依旧很高，达到了 0.933。综合来讲，CatBoost 模型表现最好，都要优于几年来学者经常运用的 XGBoost 模型和经典的逻辑回归模型，在准确性方面要比近年广泛运用且预测效果较好的 XGBoost 模型高 3.5%，因此 CatBoost 模型在研究公司困境问题中具有很优异的表现，能够提供有效的参考性。

参考文献

[1] Daniel, M. (1977) Early Warning of Bank Failure: A Logit Regression Approach. *Journal of Banking & Finance*, **1**, 249-276. https://doi.org/10.1016/0378-4266(77)90022-X

[2] Ohlson, J.A. (1980) Financial Ratios and the Probabilistic Prediction of Bankruptcy. *Journal of Accounting Research*, **18**, 109-131. https://doi.org/10.2307/2490395

[3] 邓晓岚, 王宗军, 李红侠, 杨忠诚. 非财务视角下的财务困境预警——对中国上市公司的实证研究[J]. 管理科学, 2006(3): 71-80.

[4] 刘可歆. 基于随机森林的互联网上市公司财务预警研究[D]: [硕士学位论文]. 天津: 天津财经大学, 2019.

[5] 苏玉敏. 基于随机森林与 XGBoost 的上市公司财务预警研究[D]: [硕士学位论文]. 哈尔滨: 哈尔滨工业大学, 2019.

[6] 苗丰顺. 基于 CatBoost 算法的糖尿病预测方法[J]. 计算机系统应用, 2019, 28(9): 215-218.

[7] 王斌, 冯慧芬, 王芳, 等. 基于机器学习的 CatBoost 模型在预测重症手足口病中的应用[J]. 中国感染控制杂志, 2019, 18(1): 18-22.

[8] Altman, E.I. (1968) Financial Ratios, Discriminant Analysis and the Prediction of Corporate Bankruptcy. *The Journal of Finance*, **23**, 589-609. https://doi.org/10.1111/j.1540-6261.1968.tb00843.x

[9] 王克敏, 姬美光. 基于财务与非财务指标的亏损公司财务预警研究——以公司 ST 为例[J]. 财经研究, 2006(7): 63-72.

[10] Platt, H. and Platt, M. (2002) Predicting Corporate Financial Distress: Reflections on Choice-Based Sample Bias. *Journal of Economics and Finance*, **26**, 184-199. https://doi.org/10.1007/BF02755985

[11] 马铭阳. 基于量子鸟群-人工神经网络的企业财务危机预警模型研究[D]: [硕士学位论文]. 大连: 东北财经大学, 2019.

[12] 吴世农, 卢贤义. 我国上市公司财务困境的预测模型研究[J]. 经济研究, 2001(6): 46-55.

[13] 田宝新, 王建琼. 基于财务与非财务要素的上市公司财务困境预警实证研究[J]. 金融评论, 2017, 9(5): 103-115, 126.

[14] 孙沛. 制造业上市公司财务危机预警研究[D]: [硕士学位论文]. 太原: 山西财经大学, 2019.

[15] 宋英瑞. 我国信息技术行业上市公司财务危机预警研究[D]: [硕士学位论文]. 武汉: 中南财经政法大学, 2019.

[16] 赵泽. 大数据背景下的 M 新能源公司财务危机预警研究[D]: [硕士学位论文]. 西安: 西安石油大学, 2020.

[17] 吴静. 基于集成算法的多分类财务危机预警[D]: [硕士学位论文]. 厦门: 厦门大学, 2019.

[18] 吴喜之, 张敏. 应用回归及分类——基于 R 与 Python 的实现[M]. 第 2 版. 北京: 中国人民大学出版社, 2020.