陈湘州,刘佳.基于 LR-RF-XGBoost 的债券违约风险预警[J]. 湖南科技大学学报(自然科学版), 2024, 39(1):115-124. doi:10.13582/j.cnki.1672-9102.2024.01.014

CHEN X Z, LIU J. Early Warning Research on Bond Default Risk Based on LR-RF-XGBoost [J]. Journal of Hunan University of Science and Technology (Natural Science Edition), 2024, 39(1):115-124. doi:10.13582/j.cnki.1672-9102.2024.01.014

基于 LR-RF-XGBoost 的债券违约风险预警

陈湘州[1,2*],刘佳[1]

(1.湖南科技大学 商学院,湖南 湘潭 411201; 2.湖南省新型工业化研究基地,湖南 湘潭 411201)

摘 要:随着市场经济的迅猛发展,各国的债券市场也相继成长,并趋向于多元化发展.然而,在这一发展过程中,中国的债券违约事件屡见不鲜且愈演愈烈,极大地阻碍了市场活力.以发行企业债券、公司债券、短期融资债券以及中期债券的公司为研究主体,提出 LR-RF-XGBoost 债券违约预警模型,该模型基于软投票法将逻辑回归(Logistic Regression)、随机森林(Random Forest)、极端梯度提升算法(Extreme Gradient Boosting)相融合,对样本的财务指标及非财务指标数据进行研究.研究结果发现:LR-RF-XGBoost 融合模型相比于其他单一预警模型泛化能力更强,准确率高达 95.3%.该方法有利于为投资者以及债券市场监督部门提供可靠的预测信息,帮助企业及早识别风险,为债券市场的健康发展提供保障.

关键词:债券违约;逻辑回归 ;随机森林;极端梯度提升

中图分类号:F832.51;F275 **文献标志码**:A **文章编号**:1672-9102(2024)01-0115-10

Early Warning Research on Bond Default Risk Based on LR-RF-XGBoost

CHEN Xiangzhou[1,2], LIU Jia[1]

(1. School of Business, Hunan University of Science and Technology, Xiangtan 411021, China;
2. Hunan Provincial New Industrialization Research Base, Xiangtan 411201, China)

Abstract: With the rapid development of the market economy, the Chinese bond market has also grown and gradually developed into an integral part of the market economy. However, in the course of this development, bond defaults have been common and increasing, greatly hampering market dynamics. This paper takes companies issuing corporate bonds, corporate bonds, short-term financing bonds and medium-term bonds as the research subjects, and proposes the LR-RF-XGBoost bond default warning model, which is based on a soft voting method combining Logistic Regression, Random Forest and Extreme Gradient Boosting algorithm, study of financial indicators as well as non-financial indicators for the sample. Results show that the LR-RF-XGBoost fusion model has a higher generalisation capability than other single warning models, with an accuracy of 95.3%. The method is useful in providing investors and bond market supervisory authorities with reliable predictive information, and can better help companies themselves to identify risks early, providing protection for the healthy development of the bond market.

Keywords: bond defaults; logistic regression; random forest; extreme gradient boosting

收稿日期:2023-02-22

基金项目:国家社会科学基金资助项目(13BJY057)

* **通信作者**,E-mail: xzhou605@ sohu.com

根据 Wind 数据库统计,自 2014 年 3 月起,截至 2022 年 11 月,违约的债券高达 1 010 只,违约金额达到 8 215.79 亿元,其中主体评级在 A-等级及以上的有 294 只,占比高达 29.11%;从 2019 年开始,甚至出现了 AAA 等级债券发行主体的违约现象.债券违约事件的频频爆发,不仅会给债权人带来投资风险,遭受损失,还会降低企业自身的信誉,不利于今后的融资,投资者甚至会对发行主体的运营能力以及债券市场的规范化治理产生怀疑.因此,对发行主体进行债券违约风险预警研究显得尤为重要,这不仅有助于企业及时发现问题,规避风险,保护利益相关方的权益,更为中国债券市场的健康稳定发展奠定基础.

随着信息技术的不断发展,目前已有许多学者使用机器学习对债券违约风险预警模型进行研究,大致可以分为 2 类:

一类是通过机器学习单一模型进行债券违约风险预警.例如,HARRIS[1] 基于支持向量机建立了信用评分模型;YANG 等[2] 基于 Logit 和 Probit 方法,以财务报表为导向,对社交平台的评论进行文本分析,预测出信用风险的概率;SOHN 等[3] 使用模糊逻辑回归技术进行信用评分,预测出企业贷款违约的可能性;HUANG 等[4] 为探索我国中小企业的信用风险,运用几种常用的神经网络模型进行训练,并调整取得最佳参数对测试集进行验证,比较分析发现,概率神经网络(PNN)的正确率和 AUC 值最高,且第二类误差最小;王未卿等[5] 使用随机森林模型对影响城投债的信用利差因素进行研究分析;韩璐等[6] 将支持向量机非线性模型应用到信用违约预测中,可以获得违约概率等信息;陈学彬等[7] 将擅长于处理时间序列数据的长短期记忆网络模型(LSTM)应用于我国个体信用债违约风险的预测中;霍艳[8] 使用改进后的 Logit 算法评测发行主体的信用评级,并预测债券违约的风险.

第二类是基于机器学习融合模型进行债券违约风险预警.ZHANG 等[9] 将上采样(SMOTE)和极端梯度提升算法相融合用于债券违约预测,发现在处理不平衡样本时具有显著效果;YU 等[10] 将深度置信网络(DBN)与支持向量机融合对信用风险进行评估,首先使用装袋法生成可变的训练集用于平衡子集大小,再使用支持向量机进行计算,最后将深度置信网络模型作为集成算法融合,实验结果表明该融合模型在处理不平衡样本有着显著效果;陈学彬等[11] 将 LSTM 模型与 MCM 经典分析方法相融合用于预测债券违约;肖毅等[12] 基于集成思想,使用卷积神经网络(CNN)提取出具有高影响力的特征,再结合长短期记忆网络模型进行财务风险预测.

根据已有研究发现,许多学者运用单一模型进行债券违约预测,但其预测精准度不高.肖艳丽等[13] 认为,单一预警模型不能充分挖掘有用的数据信息,可能影响预测准确性,导致预测结果存在偏差;WU 等[14] 认为融合模型能够弱化单一模型的缺点,且预测效果更优,融合模型的运用能推动债券违约预测技术的发展.为弱化单一模型的缺点以及提高债券违约预测的准确度,本文构建了 LR-RF-XGBoost 融合模型进行债券违约预测,研究工作如下:(1)将逻辑回归、随机森林、极端梯度提升通过软投票法相融合,更加全面综合地对债券违约进行预测研究;(2)与常用的单一机器学习分类模型做对比实验,验证该融合模型的优越性;(3) 排列出各个特征对模型预测结果影响程度的大小,建议企业在日常运营活动中着重关注重要指标.

1 研究设计

1.1 债券违约风险预警模型的建立

集成算法经过模型融合,得到的效果通常胜过当前先进的单一集成算法,因此,本文的债券违约风险预警模型由各个分类评估器融合而成,该模型的流程图如图 1 所示.在债券样本数据集中,每个样本包含 m 个特征(x_n,$n=1,2,\cdots,m$)和标签(y_i,$y_i=0,1$)这 2 个部分,债券违约样本的标签是 1,正常样本的标签是 0.为了保证模型预测的准确性,将所有数据按 80%∶20%的比例划分为训练集与测试集,2 个数据集相互独立.训练集的数据主要用于模型的训练,寻找最佳参数,而测试集则是对训练好的模型进行测试,检验其预测的精准度.

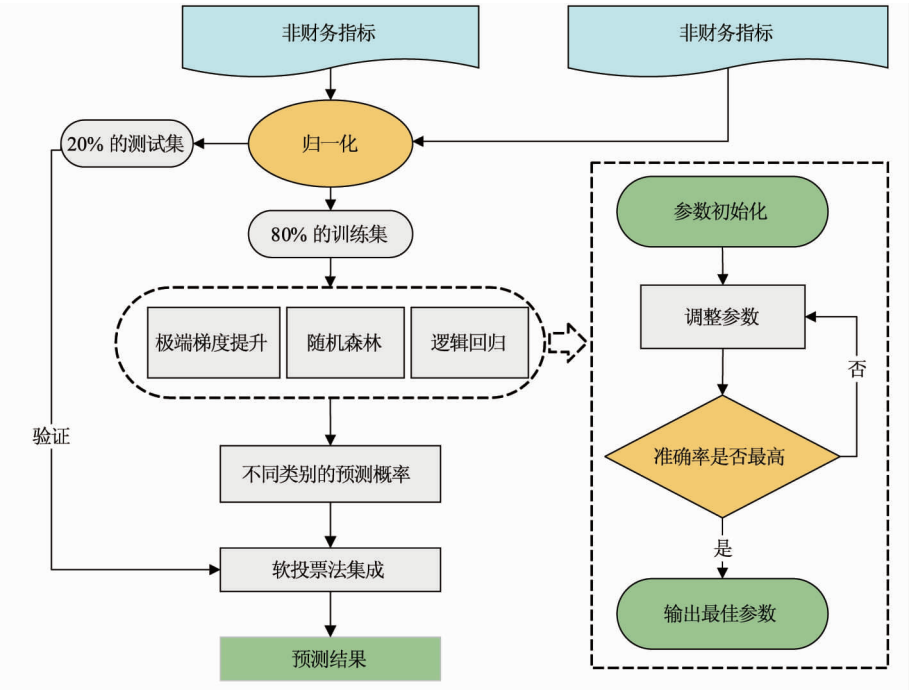

图 1　基于 LR-RF-XGBoost 债券违约预警融合模型

1.1.1　数据预处理

在进行模型训练之前，需要处理数据量纲不同的问题，而数据的预处理能够帮助模型提升预测的精准度，采用归一化方法，通过式(1)将数据映射到区间 [0,1] .

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)}. \tag{1}$$

式中：x 为数据特征；x' 为经过归一化方法处理后的数据特征.

1.1.2　LR-RF-XGBoost 模型构建

LR-RF-XGBoost 模型是运用逻辑回归，随机森林和极端梯度提升 3 种算法分别预测债券违约与正常的概率，再通过运算得到每类概率的加权平均值，概率最高的类别便是模型预测的结果.该模型的具体步骤如下：

第一步：在模型融合之前，将几种分类算法单独运行一次，选出预测结果最高的算法作为融合模型的基础，通过比较，本文选择随机森林作为基准.随机森林属于集成算法中的装袋法(Bagging)，而装袋法的核心思想是建立多个相互独立的评估器，根据平均或者多数原则来决定集成评估器的预测结果，因此随机森林是将多棵决策树集成的一种算法.其中衡量决策树预测结果的标准是不纯度，不纯度越小，分类效果越好，本文使用信息熵(Entropy)作为不纯度的衡量指标，其计算公式为

$$\text{Entropy}(t) = -\sum_{i=1}^{c-1} p(i|t)\log 2^{p(i|t)}. \tag{2}$$

式中：t 为树的节点；i 为标签的分类；$p(i|t)$ 为标签分类在节点 t 上的比例.

森林中的每一棵决策树都是相互独立的，分别对数据集进行判断，然后将得到的结果根据多数原则作为随机森林的分类结果[15].每棵决策树都将对样本进行分类，但由于数据集的特征维度较高，分类过程可视化后过于繁杂，因此本文只展示部分图片，以便于解释决策树的生长过程，如图 2 所示，每个节点都会有相应的判断条件对数据进行分割，最终叶节点得出分类结果.

第二步：构建多组分类器进行融合，本文运用软投票法(Soft Voting)将各个基评估器预测不同类别的概率进行加和，概率最高的类别就是投票的结果，最终发现随机森林、逻辑回归和极端梯度提升算法融合后的预测结果是所有组合中最高的，且相比于单一的随机森林，精准度有了一定的提升.

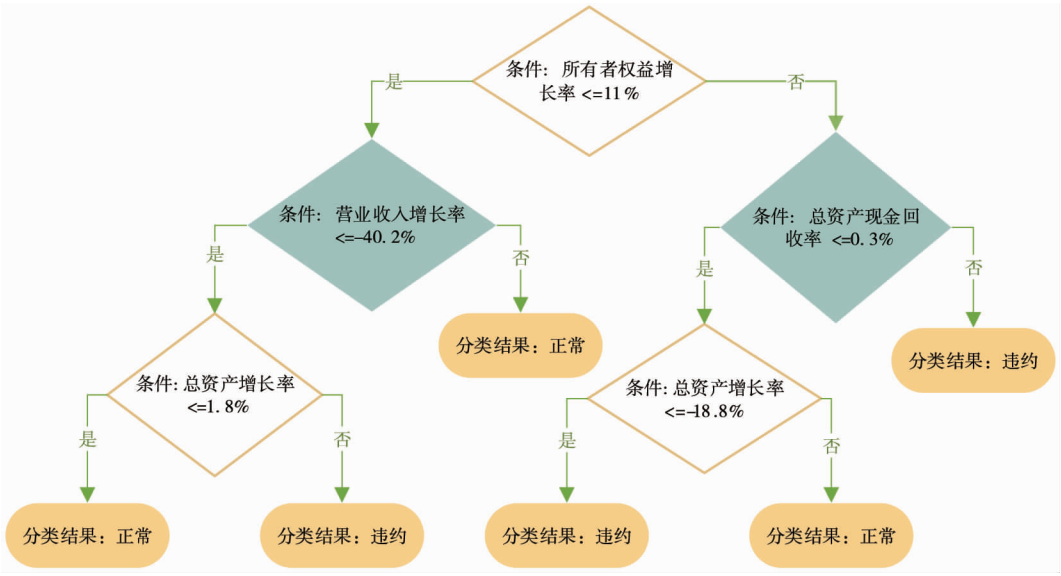

图 2 单棵决策树分类结果

Logistic 回归是由线性回归变换而来的,实质上是一种线性的分类器,通常用于二分类,通过 Sigmoid 函数将连续性变量转化为离散型变量,返回(0,1)区间的概率值[16],其对应的逻辑回归公式为

$$f(z)=\frac{1}{1+\mathrm{e}^{-z}}; \tag{3}$$

$$z=\sum_{n=1}^{m}x_n^{(i)}\omega_n+b=\boldsymbol{x}\boldsymbol{\omega}^{\mathrm{T}}+b. \tag{4}$$

式中:$\boldsymbol{x}$ 为特征向量;$\boldsymbol{\omega}$ 为回归系数;b 为偏置.

Logistic 回归通过训练数据集以追求损失函数的最小化,二元交叉熵损失(Binary Cross Entropy Loss)是二分类问题中常见的损失函数,其公式为

$$\mathrm{loss}=-\sum_{i=1}^{n}[y_i\log\hat{y}_i+(1-y_i)(1-\hat{y}_i)]. \tag{5}$$

式中:y_i 为样本 i 上真实的标签;$\hat{y}_i$ 为 Logistic 回归模型的预测值.

极端梯度提升属于集成算法中的提升法(Boosting),在众多研究领域中被广泛使用,其运行速度和模型性能比其他方法更优异[17-18],极端梯度提升中的每一个基评估器是相关的,对于在上一次模型判断错误的样本,会增加它在下一次模型中的采样概率,即增加被判错样本的抽样权重,因此每一个模型都是按照顺序建立的.极端梯度提升算法与随机森林一样,也是基于决策树进行改善的,不同的是,该算法中每一棵树都是相关的,每增加一棵树便会学习一个新的函数,来拟合上个模型预测的残差[17],其公式为

$$\hat{y}_i^{(0)}=0; \tag{6}$$

$$\hat{y}_i^{(1)}=\hat{y}_i^{(0)}+f_1(x_i); \tag{7}$$

$$\hat{y}_i^{(2)}=\hat{y}_i^{(1)}+f_2(x_i); \tag{8}$$

$$\cdots$$

$$\hat{y}_i^{(t)}=\hat{y}_i^{(t-1)}+f_t(x_i). \tag{9}$$

式中:$f_t(x_i)$ 为每次增加的决策树;$\hat{y}_i^{(t)}$ 为第 t 轮模型的预测值,而每次选取树的标准是为了降低损失函数.

XGBoost 的目标函数(损失函数)为

$$\mathrm{Obj}^{(t)}=\sum_{i=1}^{n}l(y_i,\hat{y}_i^{(t-1)}+f_t(x_i))+\Omega(f_t)+C. \tag{10}$$

式中:$l(\cdot)$ 为损失函数.经过泰勒展开上述公式并加入惩罚函数得

$$\mathrm{Obj}^{(t)}=\sum_{i=1}^{n}\left[\gamma T+G_j\omega_j+\frac{1}{2}(H_j+\lambda)\omega_j^2\right]. \tag{11}$$

式中：γ，λ 为对叶子节点的惩罚力度；T 为叶子节点的个数；G_j 为对损失函数一阶偏导的值；ω_j 为最终叶子节点的值；H_j 为对损失函数二阶偏导的值.

1.2　债券违约风险预警模型的评估

为了全面验证模型的优越性，本文不仅使用了 5 折交叉验证来衡量模型泛化能力的强弱，即模型预测未知数据集的能力，还运用测试集验证模型预测的准确率，最后运用 ROC 曲线以及 AUC 值对整体分类的效果进行评价.

1.2.1　K 折交叉验证

K 折交叉验证（K-fold Cross Validation）是将训练集分为 K 个样本，第一次训练时，将第一个样本作为测试集，剩下的 $K-1$ 个样本作为训练集；第二次则将第二个样本作为测试集，剩下的 $K-1$ 个样本作为训练集，以此类推，最终将第 K 个样本作为测试集[18-19]，如图 3 所示.

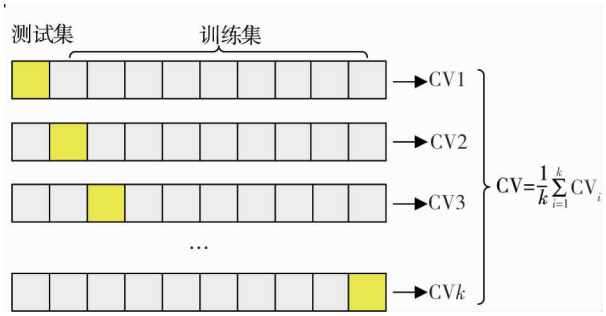

图 3　K 折交叉验证

1.2.2　混淆矩阵及 ROC 曲线

债券违约风险预测是一个二分类的问题，预警模型可能将正常样本预测为正常（TP），将正常样本预测为违约（FN），将违约样本预测为违约（TN），将违约样本预测为正常（FP），这 4 种情况便构成了混淆矩阵，如表 1 所示.

表 1　混淆矩阵

混淆矩阵	预测违约	预测正常
实际违约	TN	FP
实际正常	FN	TP

在混淆矩阵的基础上衍生出假正类率（FPR）和召回率（TPR）2 个重要的分类评估指标，FPR 又叫假阳性率，具体计算公式见式（12），TPR 又称真阳性率，具体计算公式见式（13），TPR 值越高说明实际违约样本被预测出来的概率越高.TPR 和 FPR 分别是基于实际表现 1 和 0 出发的，即分别在实际的正样本和负样本中来观察相关概率问题，因此无论样本是否平衡，都不会受到影响，所以选择这 2 个指标是构成 ROC 曲线（Receiver Operating Characteristic Curve）的基础.

$$\mathrm{FPR}=\frac{\mathrm{FP}}{\mathrm{TN}+\mathrm{FP}};\tag{12}$$

$$\mathrm{TPR}=\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}.\tag{13}$$

ROC 曲线又称感受性曲线，早期用于雷达检测，以区分信号和噪音，后有学者将其运用于模型预测准确性检验，以 FPR 为横坐标，TPR 为纵坐标，是以假正类率和召回率的点连成的线，图像越靠近左上角，曲线越陡峭，AUC 值（ROC 曲线以下的面积）越接近 1，说明模型的分类效果越好[20-23].

2　债券违约预警模型分析

2.1　样本采集及数据来源

本文的研究对象为在中国证券市场上发行企业债券、公司债券、短期融资债券以及中期债券的公司，

包括上市公司和非上市公司,选取在2014年3月—2022年3月发生信用债券违约的公司作为债券违约主体,违约当年记作t年,采用$t-3$年~$t-1$年的数据作为违约主体的样本数据集,并对当年的债券违约情况进行预测(蒋敏等[21]).将缺失值较多的样本删除,个别缺失数据采用均值法补齐,得到了4 042个样本,其中违约样本为554个.数据主要来源于Wind数据库和锐思数据库.

2.2 指标特征的选取

在已有研究的基础上[22-25],本文从盈利能力、偿债能力、经营能力、现金流动能力、发展能力、资本结构这6个方面选取了18个财务指标,虽然许多论文在进行风险预警研究时只考虑了财务指标,但是非财务指标在风险预测中对结果也具有显著影响力(林宇等[23]),如企业性质在债券违约预测中具有很大的贡献度(生柳荣等[24]),外部审计意见在财务风险预测方面具有重要的参考价值(MU1OZ-LZQUIERDO等[25]),考虑到特征的重要性与多样性,因此,增加了3个非财务指标,综上所选的指标如表2所示.

表2 债务违约风险预警特征指标

分类	名称	变量	计算公式
		非财务指标	
企业信息	企业性质	$X1$	—
	审计意见	$X2$	—
	企业规模	$X3$	—
		财务指标	
偿债能力	现金负债总额比率	$X4$	经营现金净流量/负债总和
	现金流动负债比率	$X5$	经营现金净流量/流动负债
	流动比率	$X6$	流动资产/流动负债
	速动比率	$X7$	速动资产/流动负债
	长期负债比率	$X8$	长期负债/资产总额
盈亏能力	总资产净利润率	$X9$	净利润/企业资产平均总额
	净资产收益率	$X10$	净利润/平均股东权益
经营能力	应收账款周转率	$X11$	营业收入/应收账款平均余额
	流动资产周转率	$X12$	营业收入/平均流动资产总额
	总资产周转率	$X13$	营业收入/平均资产总额
	营运资金	$X14$	流动资产-流动负债
现金流动能力	销售现金比率	$X15$	经营现金净流量/主营业务收入
	总资产现金回收率	$X16$	经营现金净流量/平均资产总额
成长能力	营业收入增长率	$X17$	本年增加额/上年营业收入总额
	总资产增长率	$X18$	总资产的增长额/年初资产总额
	所有者权益增长率	$X19$	本期增加额/上期所有者权益合计
资本结构	流动负债率	$X20$	流动负债总额/总负债
	有形资产负债率	$X21$	负债总额/(总资产-无形资产)

注:1.发行主体是国企时企业性质取值0,发行主体是非国企时取值1;

2.标准无保留审计意见时审计意见取值0,非标准无保留意见时取值1;

3.公司规模是总资产取自然对数

2.3 调整参数及预测分析

在机器学习中,模型预测的精准度很大程度上依赖于超参数的调整,因此,为了得到模型的最优结构,通过学习曲线来选择最佳的超参数值,LR-RF-XGBoost融合模型中调整的超参数包括:

1)正则化,其目的是为了防止模型过拟合,常用的有L1正则化和L2正则化这2种,通过实验对比该模型的正则项选择为L2范式,公式为

$$J(\theta)_{L2} = CJ(\theta) + \sqrt{\sum_{i=1}^{n} \theta_j^2}, j \geqslant 1 \tag{14}$$

式中：$J(\theta)$ 为没有加正则项的损失函数；$J(\theta)_{L2}$ 为增加了 L2 正则化的损失函数；C 是控制正则化的程度，经调整后，最优值取 0.1；n 为特征总数；j 为每个参数.

2）决策树的棵数（n_estimators），模型最佳的决策树的棵数为 100 棵；

3）树的最大深度（max depth），决策树过度分枝可能会导致实验结果过拟合，通过设置一个阈值来控制树生长的深度，经过实验比较，该模型的最佳深度为 17；

4）学习率（learning rate），学习率越大，迭代的速度就越快，算法很快达到极限，可能难以收敛到最佳值；学习率越小，越有可能找到最佳精确值，但是迭代速度缓慢、耗时较长，而最佳的学习率，能够协调预测精准度高和训练耗时短的矛盾，完整的迭代决策树公式为

$$\hat{y}_i^{(k+1)} = \hat{y}_i^{(k)} + \eta f_{k+1}(x_i). \tag{15}$$

式中：η 为迭代决策树的步长，又称学习率，模型通过训练获得最佳学习率为 0.9.

运用学习曲线调整各个超参数的过程如图 4 所示.

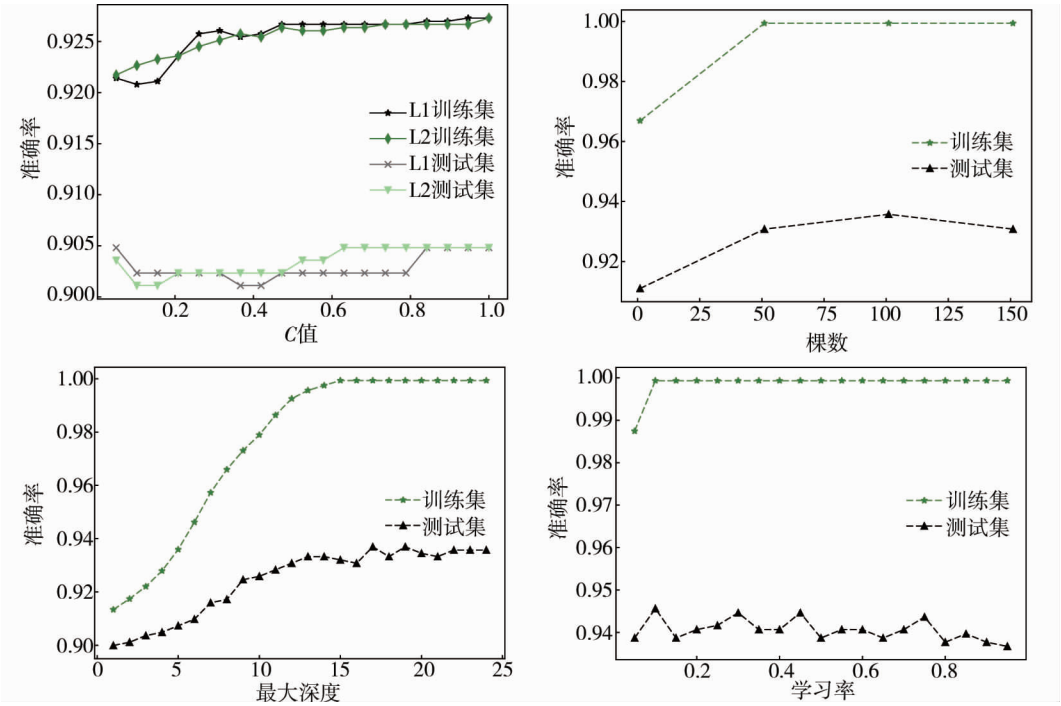

图 4　参数调整过程

通过对模型参数调整，债券违约分类得到很好的预测结果，五折交叉验证的准确率为 94.12%，测试集的准确率高达 95.30%，最后计算得出 AUC 值为 0.92，该融合模型预测的精准度较高，且 AUC 值接近于 1. 因此，根据实验结果，有理由相信该融合模型能够有效地判断一个债券样本是否会违约.

为了更直观地了解实验预测的效果，将 LR-RF-XGBoost 融合模型在训练集以及测试集的预测结果通过散点图可视化（见图 5 和图 6），图 5 和图 6 中圈代表模型的预测值，星代表样本的真实值，纵坐标 0 表示正常样本的标签，1 代表违约样本的标签. 当圈与星重合时，说明模型预测的结果与样本的真实标签一致；当圈与星分离时，说明模型对该样本预测错误. 图 5 为训练集的预测效果对比图，共有 3 234 个样本，由于数据集较大，可视化后的点与点之间过于密集，于是随机放大第 2 000～2 010 样本的预测值与真实值（见图 5 中的小图），大部分预测值与真实值重合，只有极少数的真实值与预测值不符，说明模型在训练集上的预测效果较好. 图 6 为测试集的预测效果对比图，共有 808 个样本，准确率高达 95.3%，从图 6 中可以看出：绝大部分的预测值与真实值重合，随机放大第 400～420 样本的预测值与真实值，其中只有 1 个样本预测错误，可见该模型在测试集上达到了很好的预测效果.

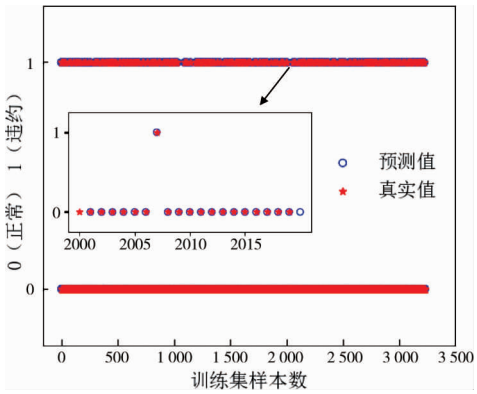

图 5 融合模型训练集预测结果

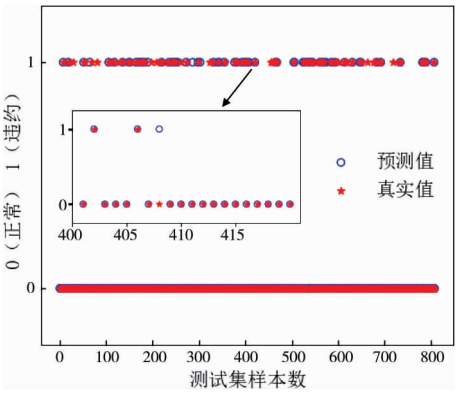

图 6 融合模型训练集预测结果

2.4 模型预测结果对比

为验证本文所构建的 LR-RF-XGBoost 模型的优越性,与目前已有的债券违约预警单一模型以及常用于分类的评估器相比较,如逻辑回归,随机森林,支持向量机,朴素贝叶斯单一模型,评比指标主要包括:5 折交叉验证结果,测试集准确率,AUC 值以及 ROC 曲线,对比结果如表 3 所示.

表 3 不同预警模型预测结果比较 %

模型	LR-RF-XGBoost	随机森林	逻辑回归	支持向量机	朴素贝叶斯
5 折交叉验证	94.12	93.56	92.23	91.80	89.48
测试集准确率	95.30	93.72	90.48	91.59	89.00

从表 3 得出:LR-RF-XGBoost 融合模型的预测效果最佳,在 5 折交叉验证的预测结果以及测试集上的准确率都领先于其他的单一模型,说明了该债券违约风险预警融合模型的泛化能力较强,预测正确的可能性较大.虽然随机森林预测模型在测试集上的准确率为 93.72%,仅次于 LR-RF-XGBoost 融合模型的准确率 95.30%,但是随机森林预测模型在测试集上的错误率为 6.28%,而 LR-RF-XGBoost 融合模型的错误率为 4.7%,该融合模型有效地将错误率降低了 25.16%.

ROC 曲线是评估一个分类器好坏的有效指标,曲线越靠近左上角,说明预测的准确率越高,从图 7 可以看出:相比于其他模型,LR-RF-XGBoost 模型的 ROC 曲线靠近左上角,预测的错误率最低.从图 8 可以看出:LR-RF-XGBoost 融合模型的 AUC 值是 0.921,而随机森林的 AUC 值是 0.916,虽然两者相差不大,但是该数据仍然能证明融合模型的预测效果最佳.逻辑回归、支持向量机和朴素贝叶斯的 AUC 值分别为0.873,0.821,0.791,预测效果不好,精确度并不突出.

对比结果表明:LR-RF-XGBoost 模型在 5 折交叉验证、测试集准确率、AUC 值以及 ROC 曲线上的实验效果最好,通过软投票法将逻辑回归、随机森林和极端梯度提升算法融合,显著提升了预测精准度.综合各个评估指标得出:基于 LR-RF-XGBoost 的债券违约预警融合模型,相比于其他单一预警模型具有更优的预测效果.

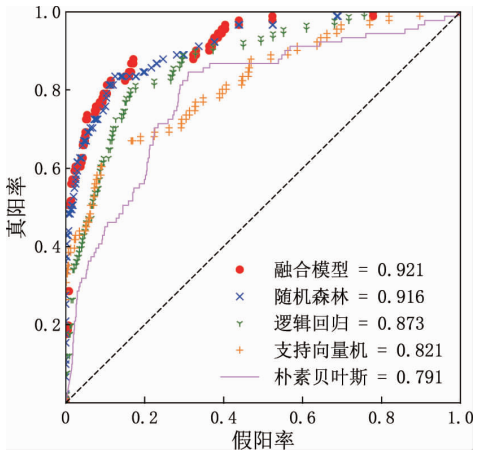

图 7 不同预警模型预测 ROC 曲线比较

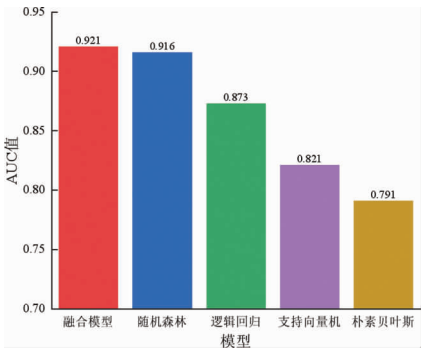

图 8 不同预警模型预测 AUC 值比较

2.5　特征指标重要性分析

融合模型的精准预测是以对各个指标的正确分析为基础的,本文的 21 个特征指标对模型预测结果的影响程度各不相同,而企业的日常运营管理很难同时兼顾全部指标的变化动态,因此,找出影响因子较大的特征,不仅有助于监管部门和企业在日常运营活动中进行事先监管控制,还能为企业节约时间精力,投身于创造价值最大的活动中.该融合模型中的 feature_importances 方法解决了这一难题,获取了各个特征重要性,如表 4 所示.

表 4　特征重要性及排序

排序	特征名称	特征重要性参数	排序	特征名称	特征重要性参数
1	企业属性	0.107 2	12	总资产增长率	0.040 3
2	审计意见	0.077 8	13	营业收入增长率	0.039 9
3	所有者权益增长率	0.059 6	14	公司规模	0.038 3
4	长期负债比率	0.056 7	15	总资产现金回收率	0.038 0
5	应收账款周转率	0.055 7	16	速动比率	0.036 6
6	总资产净利润率	0.053 0	17	流动比率	0.036 2
7	净资产收益率	0.050 9	18	营运资金	0.035 9
8	有形资产负债率	0.047 3	19	销售现金比率	0.035 9
9	流动资产周转率	0.043 7	20	现金负债总额比率	0.032 9
10	总资产周转率	0.041 9	21	现金流动负债比率	0.030 7
11	流动负债率	0.041 4			

注:特征重要性参数越高,说明该指标特征对模型预测结果的影响程度越大,但所有的特征重要性数值相加并不等于 1

在 21 个特征指标当中,企业属性的特征重要性参数是 0.107 2,对预测结果影响最大;其次是审计意见,特征重要性参数为 0.077 8;所有者权益增长率、长期负债比率、应收账款周转率以及总资产净利润率的特征重要性参数相差不大,仅次于审计意见的重要性;其他指标对结果的影响程度相对较小,其中现金流动负债比率对预测结果的影响最小.通过对各个指标的特征重要性进行排序,观察到企业属性、审计意见和所有者权益增长率对债券违约预测结果的影响较大,着重对以上几个指标进行分析,并在日常经营活动中严格把控,降低债券违约风险.

1)企业属性对预测结果的影响程度最大,因此,在判断一个发行主体是否会发生债券违约时,首先应当考虑其属性分类,具有国有企业性质的发行主体发生债券违约的概率较低,在全面监管的同时,需要有相关监管部门加大对非国有企业发行主体的监管力度.

2)审计意见的重要性排第二,影响程度较高,应当作为判断债券违约的重要条件,独立的会计师事务所对发行主体出具无保留意见审计报告,则发生债券违约的风险较低,当出具非标准意见的审计报告时,应当加强对该发行主体的监控管理.

3)所有者权益增长率、长期负债比率、应收账款周转率以及总资产净利润率对预测结果也具有重大的影响力,当这些指标出现异常波动、大幅度下跌或上涨情况时,应当注意发行主体可能存在债券违约的风险.

3　结论

1)与其他单一模型进行比较,LR-RF-XGBoost 融合模型的准确率更高、泛化能力更强,不仅弱化了单一模型的缺点,还提高了预测精准度,对于创新债券违约预警模型有一定的参考价值,为债券违约预警研究提供了一种新的方法.

2)LR-RF-XGBoost 模型获得了 21 个指标对预测结果的重要性,指标重要性排序有助于企业在日常运营管理活动中重点关注重要指标的异常波动情况,做到事先监控,事中把控,降低债券违约风险.

3)然而该模型只能有效预测出债券发行主体是否违约,不能精准推断出违约具体时间,这对企业如何制定有效且具体的风险规避措施提出了挑战,因此,如何识别出债券违约具体时间是企业亟待解决的问题,也是该学术领域中的重要研究方向.

参考文献:

[1] HARRIS T. Quantitative credit risk assessment using support vector machines: broad versus narrow default definitions[J]. Expert Systems with Applications, 2013, 40(11): 4404-4413.

[2] YANG Y, GU J, ZHOU Z F. Credit risk evaluation based on social media[J]. Environmental Research, 2016, 148: 582-585.

[3] SOHN S Y, KIM D H, YOON J H. Technology credit scoring model with fuzzy logistic regression[J]. Applied Soft Computing, 2016, 43(C): 150-158.

[4] HUANG X B, LIU X L, REN Y Q. Enterprise credit risk evaluation based on neural network algorithm[J]. Cognitive Systems Research, 2018, 52: 317-324.

[5] 王未卿, 肖勇贵, 李霞. 基于随机森林回归模型的城投债信用利差影响因素研究[J]. 数学的实践与认识, 2020, 50(12): 311-320.

[6] 韩璐, 韩立岩. 正交支持向量机及其在信用评分中的应用[J]. 管理工程学报, 2017, 31(2): 128-136.

[7] 陈学彬, 武靖, 徐明东. 我国信用债个体违约风险测度与防范: 基于 LSTM 深度学习模型[J]. 复旦学报(社会科学版), 2021, 63(3): 159-173.

[8] 霍艳. 信用评级与信用债违约实证研究: 基于机器学习算法[J]. 金融监管研究, 2022(3): 15-35.

[9] ZHANG Y, CHEN L. A study on forecasting the default risk of bond based on XGboost algorithm and over-sampling method [J]. Theoretical Economics Letters, 2021, 11(2): 258-267.

[10] YU L, ZHOU R, TANG L, et al. A DBN - Based Resampling SVM Ensemble Learning Paradigm for Credit Classification with Imbalanced Data[J].Applied Soft Computing,2018, 69: 192-202.

[11] 陈学彬, 武靖, 徐明东. 基于 LSTM 和 MCM 的债券违约风险分析[J]. 新金融, 2021(6): 54-59.

[12] 肖毅, 熊凯伦, 张希. 基于 TEI@ I 方法论的企业财务风险预警模型研究[J]. 管理评论, 2020, 32(7): 226-235.

[13] 肖艳丽, 向有涛. 企业债券违约风险预警: 基于 GWO-XGBoost 方法[J]. 上海金融, 2021(10): 44-54.

[14] WU C Y, WANG J Z, CHEN X J, et al. A novel hybrid system based on multi-objective optimization for wind speed forecasting[J]. Renewable Energy, 2020, 146: 149-165.

[15] PAL M. Random forest classifier for remote sensing classification[J]. International Journal of Remote Sensing, 2005, 26(1): 217-222.

[16] PENG C Y J, LEE K L, INGERSOLL G M. An introduction to logistic regression analysis and reporting[J]. The Journal of Educational Research, 2002, 96(1): 3-14.

[17] CHEN T, GUESTRIN C. Xgboost: A scalable tree boosting system[C]//Proceedings of the 22nd Acm Sigkdd International Conference on Knowledge Discovery and Data Mining, 2016: 785-794.

[18] 文冰梅,赵联,文黄磊. AIC 准则与留一法交叉验证渐近等价的证明[J].统计与决策.2022,38(6): 40-43..

[19] FUSHIKI T. Estimation of prediction error by using K-fold cross-validation[J]. Statistics and Computing, 2011, 21(2): 137-146.

[20] 王强, 陈景武, 陈广. ROC 曲线在累积比数 logit 判别模型效果评价中的应用[J]. 数理统计与管理, 2008, 27(3): 453-457.

[21] 蒋敏, 周炜, 史济川, 等. 基于 fsQCA 的上市企业债券违约影响因素研究[J]. 管理学报, 2021, 18(7): 1076-1085.

[22] 肖艳丽,向有涛.企业债券违约风险预警:基于 GWO-XGBoost 方法[J].系统管理学报.2021(10): 44-54.

[23] 林宇, 吴庆贺, 李昊, 等. 基于 Twin-SVR 的公司违约风险预测研究[J]. 管理评论, 2019, 31(11): 33-43.

[24] 生柳荣, 陈海华, 胡施聪, 等. 企业债券信用风险预警模型及其运用[J]. 投资研究, 2019, 38(6): 25-35.

[25] MUÑOZ-IZQUIERDO N, CAMACHO-MIÑANO M D M, SEGOVIA-VARGAS M J, et al. Is the external audit report useful for bankruptcy prediction? evidence using artificial intelligence[J]. International Journal of Financial Studies, 2019, 7(2): 20.