doi: 10.19920/j.cnki.jmsc.2024.11.009

基于最优临界点的债券违约动态预警研究①

迟国泰[1]，杨佳琦[1,2*]，周　颖[1]
(1. 大连理工大学经济管理学院，大连 116024; 2. 北京航空航天大学经济管理学院，北京 100191)

摘要： 债券违约风险预警就是根据企业财务因素、非财务因素和外部宏观因素对未来的债券违约状态做出预测.对于不同的变量组合，违约预测的效果是不同的，势必存在一个最优的指标组合，能够最大限度地减少违约预测误差.对于不同的违约判别临界点，其违约预测效果不同，势必存在一个最优的违约判别临界点，最大限度地把违约与否的债券区分开来.本研究采用随机森林模型进行指标组合的遴选，采用 Logit 回归建立违约预测模型.本研究的创新，一是在最优指标组合的遴选上，对于不同的决策树棵数，在第二类错误最小的前提下，通过 *AUC* 最大反推得到了一个最优的随机森林; 通过对最优随机森林中节点指标的不同组合的对比，找到 *AUC* 最大的一个指标组合.二是在最优违约判别临界点的确定上，以第一类错误与第二类错误的加权之和最小为目标函数，反推逻辑回归的最优临界点.三是本模型的预测精度高于支持向量机模型、梯度提升迭代决策树、神经网络等 7 种流行的大数据预测模型.基于 2014 年—2018 年中国上市公司发行的债券数据，实证研究表明，对中国债券的中短期违约预测均有影响的关键指标为“货币资金/短期债务”、“净利润”、“发行主体发行债券数量”、“行业景气指数”和“行业企业家信心指数”等五个指标.对短期违约预测有影响的关键指标为“货币资产”、“速动比率”、“固定资产投资价格指数”、“货币供应量 M_0”.对中期违约预测有影响的关键指标为“发行主体注册资本”、“债券到期偿还量”、“债券到期指数”.

关键词： 债券违约; 违约预测; 最优指标组合; 最优违约临界点; 随机森林
中图分类号： F713.36　　**文献标识码：** A　　**文章编号：** 1007-9807(2024)11-0136-23

0　引　言

债券违约风险预警就是根据企业财务因素、非财务因素和外部宏观因素对未来的债券违约状态作出预测.债券违约是指债券的发行主体不能按照达成的债券协议履行义务的行为.债券违约风险预测作为一个热门话题在债券投资者、商业银行和公司经营等各种利益相关者的决策中扮演着重要角色.

在过去的十几年中，随着中国经济的迅猛发展，债券市场也一直处于扩张期.近年来，中国的债券市场已经稳居全球第二大债券市场.债券违约预警是一个非常值得研究的科学问题.

在债券的违约风险预测中至少涉及以下两个问题.

第一个问题是违约预警临界点的选择.典型的大数据方法，例如支持向量机模型(Support Vector Machine，SVM)的临界点是 0; 当模型预测的违约概率大于等于 0 时，债券被判为违约债券，小于 0 的被判为非违约债券.K 近邻聚类模型(K-Nearest Neighbor Algorithm，KNN)的临界点为 0.5; 当模型预测的违约概率大于 0.5 时，债券被

① 收稿日期：2020-07-27; 修订日期：2023-02-20.
基金项目：辽宁省社会科学规划基金资助项目(LZ1BGL011).
通讯作者：杨佳琦(1995—)，女，辽宁大连人，博士生.Email: yangjiaqi@buaa.edu.cn

判为违约债券，小于 0.5 时被判为非违约债券.不言而喻，当临界点不同的时候，对同一组样本进行预测，判别出来的违约债券与非违约债券个数就会不同，进而模型的预测精度就不同.那么，势必存在一个最优的临界点，能把违约债券与非违约债券最大程度地区分开来.

第二个问题是最优指标组合的遴选.在当前大数据环境下，指标的数量十分庞大，选用不同指标组合对同一笔债券进行预测，得到的违约概率就会不同，进而会影响违约的预测结果.那么，势必有一个最优的指标组合，能够使模型的预测精度达到最高.

本研究在最优临界点的确定上与 Perols 等、Tomczak 和 Zięba 最为相似，但在目标函数的选择上有所不同.Perols 等[1] 以误判损失最小为标准，Tomczak 和 Zięba[2] 以 *G-mean* 最大为标准，本研究则是以第一类错误与第二类错误加权错误率最小为标准，确定最优临界点.

本研究在最优指标组合遴选上与 Park 和 Kim[3] 最为相似，但在选择最优决策树棵数的标准不同.Park 和 Kim 的研究是以不同棵数时模型判别的总精度(*Accuracy*) 最大为标准，最终选择了以 100 棵决策树构成的随机森林模型作为指标筛选的模型.对于最终指标组合的确定，是人为地删除重要程度在后 20% 的指标[3] .本研究则是以第二类错误最小和 *AUC* 最大为标准确定最优决策树棵数，并用向前选择的方法确定最优的指标体系.

本研究的特色与创新，一是在最优指标组合的遴选上，对于不同的决策树棵数，在第二类错误最小的前提下，通过 *AUC* 最大反推得到了一个最优的随机森林；通过对最优随机森林中节点指标的不同组合的对比，找到 *AUC* 最大的一个指标组合.

二是在最优违约判别临界点的确定上，对于第一类错误与第二类错误的不同配比，以第一类错误与第二类错误的加权之和最小为目标函数，反推逻辑回归模型的最优临界点.

三是本模型的预测精度高于支持向量机模型、梯度提升迭代决策树、神经网络等 7 种流行的大数据预测模型.

实证研究发现对中国中短期违约预测均有影响的关键指标为“货币资金/短期债务”、“净利润”、“发行主体发行债券数量”、“行业景气指数”和“行业企业家信心指数”等五个指标.对短期违约预测有影响的关键指标为“货币资产”、“速动比率”、“固定资产投资价格指数”、“货币供应量 M_0”.对中期违约预测有影响的关键指标为“发行主体注册资本”、“债券到期偿还量”、“债券到期指数”.

1　文献综述

1.1　债券违约预警的研究现状

Mizen 和 Tsoukas[4] 基于有序概率选择模型预测债券公司的业务和财务风险评级，证实了过去和现在的状态对未来的风险预测有实质性的影响.Jabeur 等[5] 验证了非参数模型可以更适用于建立债券风险预警模型.我国债券市场自 2014 年发生首单实质性违约后，学界和市场对中国信用评级的质疑和担忧不减反增[6] ，市场开始重新评估债券的违约风险[7] .刘海龙[8] 分析了国内外对于债券违约风险预警模型的构建方法后，提出了从债权终止角度建立债券流动性风险度量模型并建立以隐含违约率为核心的识别债务危机的指标体系等研究方向，为后来的研究提供了参考.俞宁子等[9] 分析宏观经济和违约企业案例，总结引起违约的原因包括财务变动、宏观经济趋势、行业形势、企业内部原因、突发舆情等因素，并基于以上指标建立量化的违约风险预警模型.

现有研究在中国债券市场乃至世界债券市场的债权违约预警方面均有出色的研究，但是研究的方向更偏向于理论分析.本研究基于中国债券市场中的上市公司数据，提出较为实用的违约风险预警模型.这种适用性主要体现在两个方面，一是最优指标组合的遴选；二是违约预测模型最优违约判别临界点的确定.

1.2　指标组合筛选的研究现状

指标筛选是通过在给定的数据集上选择具有更大区分能力的更具代表性的指标来减少不相关或多余指标的过程[10] .Ma 等[11] 在 P2P 在线贷款违约预测中采用了一种全面的递归特征消除(RFE) 方法从原始变量中选择最佳变量组合，实

验结果表明该方法在识别贷款违约方面具有很高的精度.Liang 等[12]在破产预测中比较了五种著名的特征选择方法,包括:基于过滤器的逐步判别分析(SDA),逐步逻辑回归(SLR)和 T 检验以及基于包装器方法的遗传算法(GA)和递归特征消除(RFE),成功降低合并财务比率和公司治理指标特征的范围.Oreski 和 Oreski[13]将神经网络与遗传算法结合的方法(HGA-NN)用于指标筛选,并通过实证证明了该指标筛选方法在零售信用风险评估中的特征选择和判别方面具有良好的应用前景.Xia 等[14]利用基于相对特征重要性评分的特征选择系统来缓解噪声的影响,消除冗余指标.Papouskova 和 Haje[15]使用多目标演化特征选择(MOEFS)方法,该方法使用一种称为 ENORA 的 MOEFS 算法(基于进化的非支配径向槽的算法)来最小化所选择特征的数量和违约预测模型的误判代价.

现有研究在指标遴选方面取得了长足的进步,也有研究考虑了指标组合与预测精度之间的关系,但是他们没有区别对待第一类错误和第二类错误.本研究采用改进的随机森林算法,以第二类错误最小、AUC 最大为目标进行最优指标组合的遴选.

1.3 最优临界点的研究现状

模型判别临界点的改变可能会使结果大相径庭.Tsai 等[16]提出了基于总利润最大化为临界点的指标筛选模型.Verbraken 等[17]考虑了银行在向客户授信中获得的预期盈亏,并根据最大化银行预期利润计算模型的临界点对客户进行了分类.Tomczak 和 Zięba[2]以 G-$mean$ 最大为目标求模型的最优临界点.Mohammadi 和 Zangeneh[18]以 ROC 曲线下的最大面积 AUC 为基准,确定了神经网络模型的最优分类临界点.Perols 等[1]基于最小化判错成本估算了最优临界点.Carvalho 和 Brans-cum[19]基于非参数贝叶斯方法估计了最佳阈值.Kouvelis 和 Zhao[20]讨论了用于设置正/零利率的供应商信用等级阈值的存在,并为特定零售商建立了供应商的信用等级漏洞.迟国泰等[21]在对中国 14 家全部上市银行的违约概率预测时,以理论违约率与实际违约率的总体差异最小为目标,对 KMV 模型中的长期负债系数的最优值进行求解,提高了预测精度.

现有研究主要从利润最大化,成本最小化或模型判别精度来进行最优临界点的求解.在模型精度方面没有考虑第一类错误与第二类错误的差别,所以本研究以第一类错误与第二类错误加权判错率最小为目标,求解债券违约判别模型的最优临界点,其中权重配比通过判别的总精度最大来确定.这类研究,是在现有研究的基础上展开的,它会提高预测的精度.

1.4 判别模型的研究现状

Davis 和 Karim[22]对比了 Logit 模型和 EWS 信号提取方法,发现 Logit 模型作为全球化的提前预警系统有着更好的性能.Celik 和 Karatepe[23]采用人工神经网络模型来预测不良贷款相对于总贷款的比率、资本相对于资产的比率、利润相对于资产的比率以及权益相对于资产的比率.迟国泰等[24]以非违约企业的数据到正理想点的距离代数和最小为第一个目标函数,以违约企业的数据到负理想点的距离代数和最小为第二个目标函数,构建多目标规划模型,以对我国小型工业企业违约情况进行了判别.Jones 等[25]指出在信用风险和破产研究中最常使用的还是标准的 Logit 模型.

本研究在模型选择上与蓝本文献 Hilscher 和 Wilson[26]相同,也选择用传统的逻辑回归模型进行违约预测.与其不同的是同时还根据判错率最小为原则,反推最优临界点,以改进逻辑回归模型.另一方面,选择 Logit 模型作为基准,是因为它在其他信用评分设置中作为判别模型的表现相对较好,而且相对容易理解,在金融服务部门广泛使用.

2 违约判别模型的构建

2.1 债券违约预测的标准

在债券违约预测中,无论是最优指标组合的遴选,还是最优违约判别临界点的确定,亦或是不同模型之间的对比分析,论文均要用到下述标准.

设 TP 表示债券实际违约被预测为违约的数量.FN 表示债券实际违约被预测为非违约的数量.FP 表示债券实际非违约被预测为违约的数量.TN 表示债券实际非违约被预测为非违约的数量.则可以得到混淆矩阵如表 1 所示.表 1 混淆矩

阵中的第一象限为债券实际违约,被预测为违约的债券个数.其他象限类推.

表 1　违约预测模型的混淆矩阵

Table 1 Confusion matrix of default prediction model

实际情况	预测情况		合计
	违约(1)	非违约(0)	
违约(1)	true positive (*TP*)	false negative (*FN*)	实际违约的债券数量 *TP+FN*
非违约(0)	false positive (*FP*)	true negative (*TN*)	实际非违约的债券数量 *FP+TN*
合计	被预测为违约的债券数量 *TP+FP*	被预测为非违约的债券数量 *FN+TN*	总债券数 *N*

1) 总体预测精度(*Accuracy*),也称为判对率,是度量判别模型性能的关键指标,样本的总体预测精度越高,说明模型的性能越好.其定义如式(1)所示.

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{1}$$

2) 召回率(*Recall*)表示的是样本中的正例有多少被预测正确了的比例,召回率越高,说明模型的性能越好.其定义如式(2)所示.

$$Aecall = \frac{TP}{TP + FN} \tag{2}$$

3) *G-mean*(*Geometric mean*)是非违约样本判对率与违约样本判对率的几何均值,*G-mean* 值越大,表明模型的整体效果越好.其定义如式(3)所示.

$$G\text{-}mean = \sqrt{\frac{TN}{TN + FP} \times \frac{TP}{TP + FN}} \tag{3}$$

4) *F* 值(*F-measure*)是另一种判别模型的综合评价标准,*F-measure* 值越大,表示模型的整体效果越好.其定义如式(4)所示.

$$F\text{-}measure = \frac{2TP}{2TP + FN + FP} \tag{4}$$

5) *BM*(book maker informedness)是两类样本预测的准确率之和减一,*BM* 值越大,表示模型的预测效果越好.其定义如式(5)所示.

$$BM = \frac{TP}{TP + FN} + \frac{TN}{TN + FP} - 1 \tag{5}$$

6) *AUC*(Area Under the ROC Curve)是 *ROC*(Receiver Operating Characteristic Curve)曲线下面积,*AUC* 值能够更好地衡量各个模型之间的优势关系.当 *AUC* 值越接近 1,说明模型的效果越好;当 *AUC* 值小于 0.5 时,认为模型无效[28].

7) 第一类错误(*Type-I Error*)表示非违约样本的判错比例,即将非违约样本预测为违约样本的数量在实际非违约样本数量中的占比,*Type-I Error* 越小,表明模型的效果越好.其定义如式(6)所示.

$$Type\text{-}I\ Error = \frac{FP}{TN + FP} \tag{6}$$

8) 第二类错误(*Type-II Error*)表示违约样本判错的比例,即将违约样本预测为非违约样本的数量在实际违约样本数量中的占比,*Type-II Error* 越小,表明模型的效果越好.其定义如式(7)所示.

$$Type\text{-}II\ Error = \frac{FN}{TP + FN} \tag{7}$$

2.2　债券违约预警核心问题与原理

2.2.1　核心问题和解决问题的思路

1) 问题 1: 对于不同的指标组合,违约预测的效果是不同的,势必存在一个最优的指标组合、能够最大限度地减少违约预测误差.

解决问题 1 的思路: 在最优指标组合的遴选上,对于不同的决策树棵数,在第二类错误最小的前提下,通过 *AUC* 最大反推得到了一个最优的随机森林;根据最优随机森林中指标的重要程度排序,用向前选择法对所有指标进行不同组合,找到 *AUC* 最大的一个指标组合.

2) 问题 2: 对于不同的违约判别临界点,其违约预测精度不同,势必存在一个最优的违约判别临界点,最大限度地把违约与否的债券区分开来.

解决问题 2 的思路: 在最优违约判别临界点的确定上,对于第一类错误与第二类错误的不同配比,以第一类错误与第二类错误的加权之和最小为目标函数,反推逻辑回归的最优临界点.

2.2.2　债券违约预测原理

通过 $t-m$ 时刻的指标数据 x_{t-m} 和 t 时刻债券的违约状态 y_t 建立逻辑回归模型,通过 t 时刻的指标数据 x_t 来预测 $t+m$ 时刻的违约状态 y_{t+m}.这里 $t = 0,1,2,3,4,5$(季度);当 $t = 0$ 时,为违约判别;当 $t =$ 其他数值时,为违约预测.

通过最优棵数的随机森林得到不同决策树的

节点指标，对节点指标进行重要性排序，对排序后的指标运用前项组合求解最优指标组合，通过逻辑回归方程来预测债券的违约概率和违约状态.

其中，在构建随机森林模型时考虑到了最优随机森林的棵数；在构建逻辑回归模型时确定了最优违约判别的临界点.

由于中国债券数据库中违约债券远少于非违约债券，违约债券类样本和非违约债券类样本有明显的差异，样本不平衡问题严重，所以本研究采用合成少数类过采样技术(Synthetic Minority Oversampling Technique，SMOTE)来进行非平衡样本处理[28]. 相比于传统的随机抽样过采样方法，SMOTE 方法避免了少数类样本中存在大量重复样本，更有效地概括了少数群体的特征，缓解了过度拟合的问题.

2.3 基于随机森林模型的指标组合遴选

不论是采用大数据模型中的任何方法，指标组合不同，违约判别的结果也就不同，甚至大相径庭.因此，遴选与确定一个违约预测精度最高、违约判别误差最小的最优指标组合，势在必行.

建立随机森林时所采用的指标为海选的全部 M 个指标，因为此时最优指标组合尚未出现.不言而喻，经过决策树节点指标生成的规则，一棵决策树所包含的指标个数是远远小于海选指标的个数的.

事实上，本研究基于随机森林的指标筛选包含两个内容，一是随机森林中最优决策树棵数的确定，二是在随机森林中最优决策树棵数的判别模型中，基于第二类错误最小、AUC 最大的指标组合的确定.

基于随机森林模型的指标组合遴选原理如下.

模型 1　随机森林 1(1 棵决策树) —— 第二类错误 $Type\text{-}II\ Error_1$，预测精度 AUC_1；

模型 2　随机森林 2(2 棵决策树) —— 第二类错误 $Type\text{-}II\ Error_2$，预测精度 AUC_2；

…，…，…

模型 r　随机森林 r(r 棵决策树) —— 第二类错误 $Type\text{-}II\ Error_r$，预测精度 $AUCr$；

上述原理的说明如下：第一个基准(Benchmark)为第二类错误 $Type\text{-}II\ Error$，第二个基准(Benchmark)为预测精度 AUC. 当上边多个模型存在两个或两个以上的第二类错误相等且最小时，则选择对应的 AUC 最大的那个模型来进行最优指标组合的遴选.

2.3.1 随机森林中最优决策树棵数 tr 的确定

通过随机森林构造指标组合，涉及到两个方面：一是需要建立最优随机森林模型，二是需要进行最优随机森林节点指标的重要程度的计算. 前者涉及 2.3.1，后者涉及到下文的式(8) ~ 式(14)和 2.3.2.

步骤 1　构建第一棵决策树

步骤 1.1　确定第一棵树的样本

经过了样本的 SMOTE 平衡处理后，得到了新的样本 U_1，这个 U_1 共 N 个债券，其中违约与非违约债券个数相等. 在 N 个债券中，利用全部 M 个海选指标，进行有放回随机抽样，共抽取 N 次，由此形成了第一棵决策树的样本 1.

步骤 1.2　样本基尼指数的计算

设 GI 表示债券样本集的基尼指数. N 表示样本集中债券的总数量. N_1 表示样本集中违约债券的数量. 则债券样本集的基尼指数 GI 可以表示为[29]

$$GI = 1 - \left(\frac{N_1}{N}\right)^2 - \left(1 - \frac{N_1}{N}\right)^2 \tag{8}$$

式(8)第二项中的 $\frac{N_1}{N}$ 表示违约债券所占的比例. 第三项中的 $\left(1 - \frac{N_1}{N}\right)$ 表示非违约债券所占的比例.

式(8)可以化简为 $GI = -2 \times \left(\frac{N_1}{N} - 0.5\right)^2 + 0.5$. 可知

当 $\frac{N_1}{N} = 0.5$ 时，即违约债券所占的比例为 0.5 时，GI 达到最大为 0.5.

当 $\frac{N_1}{N} < 0.5$ 时，GI 随着 $\frac{N_1}{N}$ 增大而增大，当 $\frac{N_1}{N} = 0$ 时，违约债券所占的比例为 0，即当样本集中全部债券为非违约债券时，GI 达到最小为 0.

当 $\frac{N_1}{N} > 0.5$ 时，GI 随着 $\frac{N_1}{N}$ 增大而减小，当 $\frac{N_1}{N} = 1$ 时，违约债券所占的比例为 1，即当样本集中全部

债券为违约债券时，GI 达到最小为 0.

因此，样本集的基尼指数 GI 越小，说明样本集的纯度越高.

步骤 1.3　第一棵决策树第一个节点指标的确定

在所有 M 个指标中，有放回地随机选择 $m(m=\sqrt{M})$ 个指标，由此形成第一棵决策树的第一个判别节点的候选指标集合 1[29].

若 $\sqrt{M}=1$，即 $m=1$，那就是只抽取了 1 次，此时这 1 个指标就是节点的指标.

若 $\sqrt{M}\neq 1$，例如下文所有 M 个指标个数为 123，则需要在 $m=\sqrt{M}$ 的 11 个指标中进行选择，选择的标准是选择基尼指数最小的指标作为节点指标.

指标的基尼指数计算方法如下.

对于一个指标 x_i 样本 1 中共有 N 个债券，对于每个债券都有一个数值，把这 N 个数值依次作为临界点，就有 N 个临界点.

用指标 x_i 的第一个数值作为第一个临界点. 将指标 x_i 大于临界点的债券放在左边节点，将指标 x_i 小于等于临界点的债券放在右边节点.

设 GI_l 表示左边节点债券样本集的基尼指数. N_l 表示左边节点债券样本集总数. N_{l1} 表示左边节点债券样本集中违约债券的数量数. 根据式(8)可以得到左边节点债券样本集的基尼指数 GI_l.

设 GI_r 表示右边节点债券样本集的基尼指数. N_r 表示右边节点债券样本集总数. N_{r1} 表示右边节点债券样本集中违约债券的数量数. 根据式(8)可以得到右边节点债券样本集的基尼指数 GI_r.

设 $GI(i)_{(1)}$ 表示以指标 x_i 的第一个数值作为划分节点进行判别的基尼指数. N_l，N_r，GI_l，GI_r 的定义与上文相同. 则有

$$GI(i)_{(1)}=\frac{N_1}{N_i+N_r}GI_{l(1)}+\frac{N_r}{N_i+N_r}GI_{r(1)} \tag{9}$$

式(9)[29] 等号右端第一项中的 $\frac{N_l}{N_l+N_r}$ 表示左边节点债券数量占债券总数量的比例. 第二项中的 $\frac{N_r}{N_l+N_r}$ 表示右边节点债券数量占债券总数量的比例. 所以，式(9)的含义是用左边节点债券样本集和右边节点债券样本集的基尼指数的加权和，表示以指标 x_i 的第一个数值作为划分节点进行判别的基尼指数.

依次使用指标 x_i 第二个数值，第三个数值，…，第 N 个数值作为临界点，进行判别，计算出以第 i 个指标的每个临界点进行判别的 $GI(i)_{(a)}(a=1,2,\cdots,N)$. 取判别的基尼指数 $GI(i)_{(a)}(a=1,2,\cdots,N)$ 中最小的数值，作为指标 x_i 判别的基尼指数，记为 $GI(i)$.

同理可以得到所有 m 个待选指标判别的基尼指数 $GI(1)$，$GI(2)$，…，$GI(m)$，选择这 m 个指标中基尼指数最小的一个指标作为决策树的顶点，即第一个节点，将样本 1 划分为左右两个债券集合.

由此，确定了第一棵决策树的第一个节点的指标和对应的临界点.

步骤 1.4　其他节点的确定

把大于步骤 1.3 确定的顶层节点的临界点的债券放入左边的分支，小于等于临界点的债券放入右边分支. 对于左边的债券集合，重新在所有 M 个指标中，进行第二次有放回地随机选择 $m(m=\sqrt{M})$ 个指标. 仿照步骤 1.3 计算步骤，可以得到基尼指数最小所对应的节点指标及其临界点，这个节点就是顶层节点下边链接的左边节点. 同理，可求右边节点的指标及其临界点. 由此得到顶点节点下边链接的右边节点.

步骤 1.5　叶子节点的确定

叶子节点的确定标准有二，一是该节点无法继续分裂，即节点中只包含一个样本. 只包含一个样本的原因是随机森林中的决策树都是让它最大程度地生长. 二是该节点中的所有样本属于同一个类别，此时节点样本集的基尼指数为 0.

叶子结点违约状态的确定. 叶子节点的违约状态，等于它所包含的那一个或那一类债券的违约状态.

债券违约判别的标准与债券所处于叶子节点的违约状态相同.

由此建立了随机森林中的第一棵决策树.

步骤 2　构建的其他决策树

同理按照构建第一棵决策树的步骤 1.1 至步骤 1.5，可以构建出第 i 棵决策树($i=2,\cdots,500$). 现有文献中，通常以默认值 500 作为随机森林中

决策树的棵数[3,31].

步骤 3 确定随机森林中最优决策树棵数 tr

步骤 3.1 计算随机森林模型的判别精度

由于每棵决策树所选用的训练样本不同，每棵树所用的指标不同，所以每棵树的形状不同. 每棵决策树对一笔债券都有一个判别结果，最后对 i 个判别结果采用投票的方式决定随机森林模型最终的判别结果.

设 $f(x_i)$ 表示随机森林模型的输出结果. $h_i(x_i)$ 表示第 i 棵决策树的输出结果. majority vote 表示多数投票，则

$$f(x_i) = \text{majority vote}\{h_i(x_i)\}_{i=1}^{Ntree} \tag{10}$$

式(10)的含义为随机森林模型的输出结果为其中每棵决策树结果的众数. 例如随机森林中共有 100 棵树，其中有 90 棵把债券 x_i 判为违约债券，有 10 棵把债券 x_i 判为非违约债券，则随机森林模型的输出结果就是把债券 x_i 判为违约债券.

用式(10)对债券进行判别，得到了第 i 个债券的理论判别结果 $\hat{y}_i$，把 $\hat{y}_i$ 与债券的实际违约状态 y_i 进行对比，则会统计得出表 1 的混淆矩阵中的频数.

根据得到的混淆矩阵中的数量债券实际违约被预测为违约的数量(TP)，债券实际违约被预测为非违约的数量(FN)，债券实际非违约被预测为违约的数量(FP)，债券实际非违约被预测为非违约的数量(TN)，可以计算出模型的精度指标，如式(7)的第二类错误和 AUC 的数值.

对于债券违约预测而言，将违约样本预测为非违约的错误代价远比将非违约样本预测为违约样本的代价要大. 在选择最优的决策树棵数时，本研究以违约样本错判成非违约样本的比率(第二类错误，*Type-II Error*)最小和 AUC 最大为标准，确定随机森林中决策树的最优棵数，目标函数如式(11)所示.

$$\begin{cases}\min\ Type\text{-}II\ Error\\ \max\ AUC\end{cases} \tag{11}$$

式(11)的经济学含义是，在追求第二类错误最小的情况下取最大的 AUC.

根据式(11)的目标函数，能够保证第二类错误最小的时候 AUC 最大. 由于第二类错误是把违约债券判为非违约债券的比率，如果第二类错误大，会带来很大损失，因此在目标函数中，保证它最小有重大意义. AUC 最大是违约判别中的流行的标准，AUC 越大综合的判别精度越高.

步骤 3.2 最优决策树棵数 tr 的确定

本研究的最优决策树棵数确定采用两个标准.

1) *Type-II Error* 最小标准

遍历决策树棵数从 20 ~ 500 时模型对应的精度，得到不同决策树棵数的第二类错误 $Type\text{-}II\ Error_{20}$，$Type\text{-}II\ Error_{21}$，$Type\text{-}II\ Error_{22}$，…，$Type\text{-}II\ Error_{500}$. 从中选择最小的 *Type-II Error* 所对应的决策树棵数，就是第一次筛选的最优决策树棵数 Tr_i. 应当指出，此时对应的第一次筛选的决策树棵数可能不止一个，因为不止一个决策树的棵数对应于第二类错误最小. 由此得到了一个向量$\{Tr_1, Tr_2, \cdots, Tr_m\}$，$m$ 表示第二类错误最小的随机森林个数.

2) AUC 最大标准

经过以上 1) 步骤得到第二类错误最小的决策树棵数向量$\{Tr_1, Tr_2, \cdots, Tr_m\}$中，分别计算它们的 AUC 值，AUC 值最大对应的那个随机森林的决策树棵数，就是最优的决策树棵数 Tr^*.

本研究的随机森林模型与现有文献中的差别是，本研究以随机森林模型的判别精度第二类错误最小、AUC 最大为原则，确定随机森林中决策树的最优棵数.

20 棵决策树的组合就是构建一个 20 棵决策树的随机森林. 50 棵决策树组合与 20 棵决策树组合没有任何关系，它是重新构建 50 棵决策树的随机森林. 类推，本研究建立了 20 棵 ~ 500 棵决策树的组合.

第一个基准(Benchmark)为第二类错误 *Type-II Error*，第二个基准(Benchmark)为预测精度 AUC. 当上边多个模型存在两个或两个以上的第二类错误相等且最小时，则选择对应的 AUC 最大的那个模型来进行最优指标组合的遴选.

本研究通过遍历由 20 棵 ~ 500 棵决策树组成的不同的随机森林的第二类错误，选择第二类错误最小的那些随机森林，并在第二类错误最小的那些随机森林中，遴选唯一的 AUC 精度最大的那个随机森林作为最优随机森林，同时也就确定了最优的决策树棵数. 例如下文实证中当预测窗口 $m = 1$ 时，随机森林的最优决策树的棵数为

tr =80.类推,由此得到所有预测窗口 m = 0,1,2,3,4,5 时的随机森林的最优决策树棵数.

2.3.2　确定基于最优棵数的指标组合

步骤 4　计算指标 x_i 的重要性评分

一个随机森林由多棵决策树组成,而每一棵决策树又有多个节点,故一个由多棵决策树组成的随机森林就有更多的节点.对这些节点指标的每一个节点指标进行重要性评分,并把评分结果标准化,根据评分的标准化数值大小进行指标排序.

步骤 4.1　同一指标一个节点重要性评价的确定

设 $VIM_{i(c)}^{GI}$ 表示第 i 个指标 x_i 在节点 c 处的重要性评分(variable importance measures) .$GI_{(c)}$ 表示在节点 c 上方的债券样本集的基尼指数.$GI_{(c)l}$ 表示在节点 c 下方根据指标 x_i 将债券样本集分为两类,节点左边债券样本集的基尼指数.$GI_{(c)r}$ 表示在节点 c 下方根据指标 x_i 将债券样本集分为两类,节点右边债券样本集的基尼指数.则指标 x_i 在节点位置 c 的重要性评分 $VIM_{i(c)}^{GI}$ 为

$$VIM_{i(c)}^{GI} = GI_{(c)} - GI_{(c)l} - GI_{(c)r} \tag{12}$$

式(12) 的含义是用根据指标 x_i 分类前后债券样本集的基尼指数变化量来衡量指标 x_i 的重要程度.

式(12) 中等号右端第一项 $GI_{(c)}$ 是节点 c 上方债券样本集的基尼指数,在节点给定的情况下,是一个定值.第二项 $GI_{(c)l}$ 和第三项 $GI_{(c)r}$ 分别是根据指标 x_i 划分后的两个债券样本集的基尼指数,越小越好.因此式(12) 计算出的 $VIM_{i(c)}^{GI}$ 越大说明这个节点位置 c 处的指标 x_i 越重要.

在一个随机森林中: 对于不同的决策树,会存在相同的指标.对不同的决策树,会存在同一个指标的不同节点,因为,随机森林的指标是随机抽取的.式(12) 计算的是这任意一种情况时的一个指标一个节点的重要性.以负债率这个指标为例,当一个决策树的两个以上节点都是负债率指标,则每一个节点都要分别按式(12) 计算.

步骤 4.2　同一指标多个节点重要性评价的确定

将指标 x_i 在第 j 棵决策树中每次作为节点的基于基尼指数的重要性评分相加,可以得到指标 x_i 在第 j 棵决策树中的重要性评分 VIM_{ij}^{GI}.

设 VIM_i^{GI} 表示指标 x_i 在随机森林模型中的重要性评分. tr 表示随机森林模型中的最优棵数. VIM_{ij}^{GI} 表示指标 x_i 在第 j 棵决策树中的重要性评分. 则指标 x_i 在随机森林模型中的重要性评分为

$$VIM_i^{GI} = \sum_{j=1}^{tr} VIM_{ij}^{GI} \tag{13}$$

式(13) 的含义为将指标 x_i 在随机森林中每棵决策树重要性评分相加来表示指标 x_i 在随机森林模型中的重要性评分.

式(12) 与式(13) 的区别在于前者仅仅是一棵树的一个节点,而后者是由多棵树组成的一个随机森林的重要性评分.

同理,设随机森林中不同决策树的节点指标的总指标个数为 M_T,可以得到每个指标的重要性评分 VIM_i^{GI}(i = 1,2,…,M_T) .

步骤 4.3　同一指标多个节点重要性的标准化

把 M_T 个指标的重要性评分 VIM_i^{GI}(i = 1,2,…,M_T) 进行归一化处理.

设 VIM_i 表示指标 x_i 在随机森林模型中基于基尼指数的重要性评分.M_T 表示指标个数.则

$$VIM_i = \frac{VIM_i^{GI}}{\sum_{h=1}^{M_T} VIM_h^{GI}} \tag{14}$$

式(14) 是对指标 x_i 的重要性评分进行归一化处理,作为指标 x_i 的基于基尼指数的重要性评分.

步骤 5　最优指标组合的确定

步骤 5.1　指标组合的构建

对于一个特定的窗口,设随机森林不同决策树的节点指标的总指标个数为 M_T,将 M_T 个指标按照其基于基尼指数的重要性评分 VIM_i(i = 1,2,…,M_T) 由大到小排序,得到指标的重要程度排序.

根据重要性排序,采用向前选择的方法[30],构造不同的组合.依次将 VIM_i 最大的和第二大的指标构成指标组合 1; 把 VIM_i 最大的前三个指标构成指标组合 2; 把 VIM_i 最大的前四个指标构成指标组合 3; 以此类推,由此形成了 k 个指标组合,指标组合 k(k = 1,2,…,M_T - 1) .

步骤 5.2 确定最优指标组合

对于指标组合 1，应用其对应的窗口的最优随机森林模型，通过训练样本建立模型，把测试样本带入模型，在测试样本中：把模型预测的第 i 个债券的理论违约状态 $\hat{y}_i$ 与其实际违约状态 y_i 进行对比，则可以方便地得到第二类 $Type\text{-}II\ Error_1$ 和 AUC_1.

同理，可以得到第 k 个指标组合的第二类 $Type\text{-}II\ Error_k$ 和 AUC_k，$k = 1,2,\cdots,M_T - 1$.

在上述的 $M_T - 1$ 个指标组合的预测误差 $Type\text{-}II\ Error$ 中，取 $\min\ Tpye\text{-}II\ Error$.

当 $\min\ Tpye\text{-}II\ Error$ 只有一个时，则最小的第二类错误对应的那个指标组合就是最优指标组合.

当 $\min\ Tpye\text{-}II\ Error$ 不止一个时，则需要再对比这些最小第二类错误的多个指标组合的精度 AUC，取 $AUC^* = \max AUC_j$ 对应的那个指标组合就是最优指标组合.

本研究与蓝本研究文献 Park 和 Kim[3] 在指标组合遴选上比较相似，但在选择最优决策树棵数的标准不同. Park 和 Kim 的研究是以不同棵数时模型判别的总精度（$Accuracy$）最大为标准，最终选择了以 100 棵决策树构成的随机森林模型作为指标筛选的模型. 对于最终指标组合的确定，是人为地删除重要程度在后 20% 的指标[3]. 本研究则是以第二类错误最小和 AUC 最大为标准确定最优决策树棵数，并用向前选择的方法确定最优的指标体系.

本研究与 Huang 等[30] 的不同也有二：一是前项选择所采用的方法不同，本研究用的随机森林，后者应用的是逻辑回归. 二是研究的问题不同，本研究研究的是债券的违约预测，而后者研究的是产品定价.

本研究与 Xia 等[14] 的不同还有二：一是指标组合遴选的思路不同. Xia 等采用主观选取指标进行公司的信用评分，本研究采用的是通过 AUC 最大在众多的指标组合中、反推最优的一个指标组合. 二是研究的问题不同. Xia 等研究的是对公司的信用评分，本研究研究的是债券的违约预测.

本研究的特色在于通过第二类错误最小和 AUC 最大的两个标准在众多的随机森林中遴选出一个最优的随机森林，通过向前选择法在 M_T 个指标构成的（$M_T - 1$）个指标组合中，寻找最优的随机森林预测精度 AUC 最大的那一个最优的指标组合.

2.4 债券违约预警模型的构建

2.4.1 构建基于 Logit 的违约预警方程

步骤 6 构建基于 Logit 的违约预警方程

设 $P_{j,t}(y_t = 1 \mid y_{t-m} = 0)$ 表示债券 j 在第 $t - m$ 期不违约、在第 t 期违约的违约概率（这里 t 为第 t 个季度，下同），简称 $P_{j,t}$；这是一个考虑第一次违约的条件概率，其中，在第 $t - m$ 期不违约，则其违约状态 $y_{t-m} = 0$；在第 t 期违约，则其违约状态 $y_t = 1$. 显见，$P_{j,t}$ 为债券 j 在第 t 期的违约的概率，则（$1 - P_{j,t}$）是债券 j 在第 t 期不违约的概率. α、β 表示待估参数. $x_{i,j}^{t-m}$ 表示债券 j 第 $t - m$ 期的第 i 个指标值（$m = 1,2,3,4,5$），是上文步骤 5 中的最优指标组合. n 表示指标总数. 那么根据 Logit 模型[32] 可以将债券的违约概率 $P_{j,t}$ 表示为

$$P_{j,t}(y_t = 1 \mid y_{t-m} = 0) = \left(1 + \exp\left(-\left(\alpha + \beta \sum_{i=1}^{n} x_{i,j}^{t-m}\right)\right)\right)^{-1} \quad (15)$$

再设 $\hat{y}_{j,t}$ 是债券预测的违约状态，预测为违约表示为“1”，即 $\hat{y}_{j,t} = 1 = 1$；预测为非违约表示为“0”，即 $\hat{y}_{j,t} = 0$. P_t^* 表示逻辑回归模型的最优违约判别临界点. 则式（15）的示性函数为

$$\begin{cases} \hat{y}_{j,t} = 1, P_{j,t} > P_t^* \\ \hat{y}_{j,t} = 0, P_{j,t} \leqslant P_t^* \end{cases} \quad (16)$$

式（15）和式（16）则构成了本研究的基于逻辑回归的债券违约预警模型. 式（15）中的系数 α、系数 β 利用最大似然估计法计算. 其中，式（15）和式（16）的第 j 个债券第 i 个指标 x_{ij} 系由上文 2.3.2 的最优指标组合确定；其最优违约判别临界点 P_t^* 由下文 2.4.2 的以加权错误率最小为目标函数的离散函数的优化模型确定.

本研究通过 $t - m$ 期的指标数据 x_{t-m} 和第 t 期的违约状态 y_t 建立预测模型，达到用第 t 期的指标数据 x_t 预测第 $t + m$ 期债券违约状态的目的. 这也是式（15）～式（16）的经济学含义.

本研究的模型与现有研究的区别至少有三.

一是预测能力不同. 现有研究[26] 在式（15）中的 $m = 0$ 或 $m = 1$，前者仅仅是验证，没有预警作

用，后者仅仅具有未来 1 期的预测期限. 众所周知，除了极少数的短期债券外，绝大多数都是远远超过 1 期的债券，因此，对未来第 0 期或第 1 期的违约预测，其应用价值十分有限. 而本研究的 $m = 1,2,\cdots,5$，其预测期限远远大于现有研究.

二是违约判别临界点不同. 现有研究[20] 逻辑回归的违约判别临界点 $P_t^* = 0.5$，即当 $P_{j,t} > 0.5$ 时，将债券判为违约债券，$P_{j,t} \leqslant 0.5$ 时，将债券判为非违约债券. 众所周知，采用不同的违约判别临界点，被逻辑回归模型判别成违约(或非违约) 的债券个数就不一样，其判别精度就大不相同. 因此、势必存在一个最优的违约判别临界点使违约判别精度最高或判别误差最小. 本研究式(16) 中的 P_t^* 就是下文 2.4.3 通过判别误差最小导出的最优临界点. 下文的实证研究也表明，在第 $t+m$ 期($m = 1,2,\cdots,5$) 的预测模型中，最优的临界点都不等于 0.5，甚至相去甚远.

三是指标体系不同. 现有研究大多采用从更早的文献中借用的 k 个指标，把 x^i($i = 1,2,\cdots,k$) 代入逻辑回归模型. 不言而喻，指标不同，逻辑回归判别或预测的结果就大不一样，甚至大相径庭. 本研究式(15) 中的指标或变量，是在上文通过第二类错误最小和 AUC 最大的目标函数，在众多的随机森林中遴选出最优的随机森林，根据最优的随机森林得到所有指标的重要程度排序，用向前选择法对所有指标进行不同组合，找到 AUC 最大的一个指标组合，作为最优的指标组合. 以下文 3.3.2 实证中对未来 1 个季度的时间窗口($t+1$) 的违约预测为例，最优指标组合的指标个数为 29 个，原来的指标个数为 123 个，大数据的降维效果达到了 76%{ [(29 - 123) /123] = - 0.76 = - 76%}.

综上，本研究式(15) 和式(16) 建立的逻辑回归模型的特色如下.

1) 在最优指标组合的遴选上，对于不同的决策树棵数，在第二类错误最小的前提下，通过 AUC 最大反推得到了一个最优的随机森林; 根据最优随机森林中指标的重要程度排序，用向前选择法对所有指标进行不同组合，找到 AUC 最大的一个指标组合.

2) 通过违约预测误差第一类错误与第二类错误加权错误率最小且第二类错误更小，反推不同预测期限 m 的最优违约判别临界点 P_m^*，减少了把违约债券误判为非违约债券带来的投资损失，且能够最大限度地把违约债券和非违约债券区分开来. 改变了流行的逻辑回归模型以 0.5 作为临界点的做法. 式(15) ~ 式(16) 的这个特色，将在下文式(17) 的数学规划模型中详述.

3) 以 t 季度的债券违约状态 y_t 和 $t-m$ 季度的指标数据 x_{t-m} 建立违约预测模型，达到了用第 t 季度 的指标数据 x_t，预测第 $t+m$ 季度的违约状态 y_{t+m}.

2.4.2　最优违约判别临界点目标函数的构建

步骤 7　构建求解最优违约判别临界点的目标函数

设 Q 表示加权错误率. λ 表示常数. FP 表示债券实际非违约被预测为违约的数量. TN 表示债券实际非违约被预测为非违约的数量. FN 表示债券实际违约被预测为非违约的数量. TP 表示债券实际违约被预测为违约的数量.

$$\min Q = \frac{FP}{NT + FP} + \lambda \times \frac{FN}{TP + FN} \quad (17-1)$$

$$\text{s.t. } P_{j,t}(y_t = 1 \mid y_{t-m} = 0) = \left(1+\exp\left(-\left(\alpha+\beta\sum_{i=1}^{n} x_{i,j}^{t-m}\right)\right)\right)^{-1} \quad (17-2)$$

$$\begin{cases}\hat{y}_{j,t}=1, & P_{j,t}>P_m^h\\ \hat{y}_{j,t}=0, & P_{j,t}\leqslant P_m^h\end{cases} \quad (17-3)$$

$$FP=\text{card}\{x_i \mid y_i=0, \hat{y}_i=1\} \quad (17-4)$$

$$TN=\text{card}\{x_i \mid y_i=0, \hat{y}_i=0\} \quad (17-5)$$

$$FN=\text{card}\{x_i \mid y_i=1, \hat{y}_i=0\} \quad (17-6)$$

$$TP=\text{card}\{x_i \mid y_i=1, \hat{y}_i=1\} \quad (17-7)$$

式(17-1) 是以第一类错误和第二类错误的加权错误率最小为目标函数.

式(17-2) 中的债券 j 第 $t-m$ 期的第 i 个指标值 $x_{i,j}^{t-m}$，指标为上文步骤 5 中遴选的最优指标组合.

式(17-3) 中的 P_m^h 为第 h 个临界点，它是以(0,1) 的开区间为端点，步长为 10^{-4} 的 9 999 个临界点，它在每一次数学规划的求解过程中，只代入一个数值. 显而易见，上边数学规划的方程组要求解 9 999 次.

式(17-2) 和式(17-3) 的其他参数，与式(15) 和式(16) 同.

显然,对于相同的违约概率 $P_{j,t}$ 而言,式(17-3)不同的临界点 P_m^h 可能会导致债券被划分为截然不同的两种状态.

式(17-4)~式(17-7)的左端各个参数如上文表 1 的混淆矩阵所示. card{·} 表示在求出债券的违约概率 $P_{j,t}$ 后,根据违约判别临界点 P_m^h,可以将债券划分为违约债券和非违约债券.例如式(17-4)表示非违约的债券被判别为违约的个数,如表 1 混淆矩阵的第 2 行第 1 列的所示,式(17-5)~式(17-7)类推.

式(17)的数学规划,是在众多违约判别临界点 P_m^h 中,以第一类错误和第二类错误的加权错误率最小为目标,反推一个最优的判别临界点 P_t^*,将违约债券和非违约债券最大程度地去分开.

式(17)的数学规划与现有研究的区别主要有二.

一是最佳违约判别临界点的标准不同. Tomczak 和 Zięba[2] 以 *G-mean* 最大为标准,Perols 等[1] 以误判损失最小为标准.本研究则是以第一类错误和第二类错误的加权错误率之和最小为标准,反推不同预测期限 m 对应的 Logit 回归模型最佳违约判别临界点 P_t^*.

二是临界点使用的场合不同.文献[1,2]中最优临界点不论在预测期间 m 为任何期限时,都使用相同的一个最优临界点.事实上,不同时期的样本是不一样的,其最优临界点也不可能相同.本研究则通过不同时期的样本建立式(17)的数学规划,得到不同时期的最优临界点 P_m^*.下文表 5 第 3 列的实证结果也表明,使用不同预测期限 m 对应的截面样本建模,得到的最佳违约判别临界点与流行文献的 0.5 大相径庭.当 m=5 时,临界点为 P_m^*=P_5^* = 0.198 2,偏离 0.5 临界点竟高达 -60.36% [(0.198 2-0.5)/0.5 = -60.36%];当 m=1 时,临界点为 P_m^* = P_1^* = 0.472 1,偏离 0.5 临界点也达 5.58% [(0.472 1-0.5)/0.5 = 5.58%].

本研究的特色是通过违约预测误差第一类错误与第二类错误加权错误率最小且第二类错误更小,反推不同预测期限 m 的最优违约判别临界点 P_m^*,减少了把违约债券误判为非违约债券带来的投资损失,且能够最大限度地把违约债券和非违约债券区分开来.改变了流行的逻辑回归模型以 0.5 作为临界点的做法. Web Of Science 数据库 2010 年—2019 年 10 年的文献表明,现有研究 31.8%使用 0.5 做临界点,亟待改进.

对于债券违约判别而言,将非违约债券判为违约债券(第一类错误,*Type-I Error*)仅会增加投资人的防范意识,不会造成实际亏损.而将违约债券判为非违约债券(第二类错误,*Type-II Error*),可能会导致投资人不能及时抛售债券或错误入手了将亏损的债券,会造成实际亏损.本研究通过取 λ 大于等于 1 求解式(17),达到了加权错误最小且第二类错误更小的目的.因为,第二类错误比第一类错误更严重,在考察时应该给第二类错误赋予更大的权重.

2.4.3 最优违约判别临界点 P_m^* 的求解过程

步骤 8 最优违约判别临界点 P_m^* 的求解

步骤 8.1 参数 λ = 1,第一个临界点 P_m^{11} 时 Q_{11} 的确定

取第一个临界点 P_m^{11} = 0.000 1,λ = 1,代入式(17),得到 Q_{11}.

步骤 8.2 参数 λ = 1,第 h 个临界点 P_m^{h1} 时最小加权平均成本 Q_{h1}^* 的确定

在参数 λ = 1 时,令第 h 个临界点 P_m^{h1} 以(0,1)的开区间为端点,以步长 10^{-4} 形成 9 999 个临界点.

把 λ = 1,以及 9 999 个 P_m^{h1} 分别代入式(17),得到 9 999 个 Q_{h1}.取 Min($Q_{1,1}$,$Q_{2,1}$,…,$Q_{9\,999,1}$),则得到了一个 Q_{h1}^*.

步骤 8.3 参数 λ = k 时,第 h 个临界点 P_m^{hk} 时最小加权平均成本 Q_{hk}^* 的确定

取 λ = 1,2,3,4,5 五个数值,仍然取步骤 8.2 的 9 999 个数值,仿照步骤 8.1~步骤 8.2,则会得到 $Q_{h\lambda}^* = Q_{h1}^*, Q_{h2}^*, Q_{h3}^*, Q_{h4}^*, Q_{h5}^*$.

步骤 8.4 最优违约判别临界点 P_m^* 的确定

对于步骤 8.3 中的每一个 $Q_{h\lambda}^* = Q_1^*$,Q_{h2}^*,Q_{h3}^*,Q_{h4}^*,Q_h^*,都可以找到其对应的临界点 $P_m^{h\lambda} = P_m^{1-\lambda}, P_{m2}^{\lambda}, P_{m3}^{\lambda}, P_{m4}^{\lambda}, P_{m5}^{\lambda}$.把每一个 P_{mh}^{λ} 用于对逻辑回归模型的判别,都可以得到一个判别精度 ACC,取判别精度 ACC^{λ} 最大所对应的临界点,就是最优的违约判别临界点 P_m^*.

不言而喻,在上述过程中,也就找到了与 P_m^*

对应的 λ_m^*.

步骤 8.5　不同预测期限最优违约判别临界点的确定

现有研究没有对各期的最优临界点进行区分[1,2].在步骤 8.1 至步骤 8.4 的最优违约判别临界点 P_m^* 的求解中，需要用样本 $t-m$ 建模，当 m 取不同值的时候，最优违约判别临界点 P_m^* 不一样.本研究 $t-m(m=0,1,\cdots,5)$，共代入六期的数据会得到六个 $P_m^*(m=0,1,\cdots,5)$.

3　债券的违约风险预警实证

3.1　数据来源和样本选取

3.1.1　数据来源

本研究选取的中国债券数据来源于 Wind 数据库中的已到期债券(2014 年—2018 年共 20 个季度)、宏观环境数据来源于中华人民共和国国家统计局(National Bureau of Statistics)官方网站和中国经济社会发展统计 CNKI 数据库.共提取指标 124 个，其中包括财务指标 29 个，非财务指标有 29 个，宏观指标 65 个，违约状态指标 1 个.

3.1.2　样本选取

在选取样本时，由于中国债券市场到 2014 年才出现首例实质性债券违约事件，因此，本研究所考察的样本为 2014 年—2018 年上市公司发行的已到期债券.

3.2　样本数据处理

1) 原始样本预处理

针对原始数据进行剔除异常值、补空值的预处理.

2) 构建 $t-m$ 时间窗口样本

对数据经过预处理后，构造了 6 组时间窗口样本，即 $t-m(m=0,1,\cdots,5)$ 个季度的时间窗口样本(下同)，每组时间窗口样本中是第 t 个季度的违约状态和提前 m 个季度的指标数据.每组 $t-m$期时间窗口样本都包含相同的 $N_T=7\ 025$ 笔债券，其中非违约债券总数为 $N_{T0}=6\ 951$，违约债券总数为 $N_{T1}=74$.

3) 样本划分与处理

在原始样本 N_T 中随机抽取 35%的违约样本和非违约样本，就形成了训练样本 N_{tr}.然后，在剩余的样本中，再随机抽取 35%的违约样本和非违约样本，就形成了验证样本 N_{va}，最后剩下的 30%的违约样本和非违约样本，就是测试样本 N_{te}，也就是表 2 的样本Ⅲ：测试样本 N_{te}.

由于在本研究的 $N_T=7\ 025$ 笔债券中，非违约债券总数为 $N_{T0}=6\ 951$，违约债券总数为 $N_{T1}=74$；比率为 89∶1，为严重的不平衡样本.故，需要进行 SMOTE 的平衡处理.其中，在计算出违约债券 x_i 和其他所有违约债券之间的欧式距离时，要找出距离最小的 k 个，本研究依据参考文献取 $k=5$[28].整个 SMOTE 处理，可以方便地通过 MATLAB 编程得到处理结果.由此得到了表 2 中的全部样本.

表 2　样本分类

Table 2 Sample classification

样本分类	Ⅰ训练样本 N_{tr}	Ⅱ验证样本 N_{va}	Ⅲ测试样本 N_{te}
(1) 非违约	$N_{tr}^0=2\ 433$	$N_{va}^0=2\ 433$	$N_{te}^0=2\ 085$
(2) 违约	$N_{tr}^1=2\ 433$	$N_{va}^1=2\ 433$	$N_{te}^1=22$
(3) 合计	$N_{tr}=4\ 866$	$N_{va}=4\ 866$	$N_{te}=2\ 107$

4) 样本标准化

将处理后的平衡的样本Ⅰ训练样本 N_{tr}、样本Ⅱ验证样本 N_{va} 和样本Ⅲ测试样本 N_{te} 对应的数据进行标准化.

3.3　基于随机森林的指标筛选

将标准化后的数据按照上文 2.3 的步骤进行最优指标组合的遴选.下文的数据均以 $t-1$ 期为例.

3.3.1　最优决策树棵数的确定

步骤 1　第一棵决策树的构建

步骤 1.1　确定样本

对标准化后的 123 个指标 4 866 个债券的数据样本 N_{tr}，进行有放回地抽取 4 866 次，构成第一棵决策树的训练样本 $T^{(1)}$.

步骤 1.2　样本基尼指数的计算

在 123 个指标中有放回地抽取 11 次($\sqrt{123}\approx 11$)，得到由 11 个指标组成的待选指标集合 F_1.

按照式(8)和式(9)，根据 F_1 和 $T^{(1)}$，分别计算这 11 个指标的基尼指数.

步骤 1.3　确定第一个节点指标

根据步骤 1.2 的计算结果，选择基尼指数最小的指标作为第一棵决策树的根节点，求解到根

节点的临界点.

步骤 1.4 其他节点的确定

把大于根节点临界点的债券放入左边的分支,得到样本 $T_{1l}^{(1)}$.把小于等于根节点临界点的债券放入右边的分支,得到样本 $T_{1r}^{(1)}$.

重新再在 123 个指标中有放回地抽取 11 次,得到由 11 个指标组成的待选指标集合 F_2.

按照式(8)和式(9),根据 F_2 和 $T_{1l}^{(1)}$,分别计算这 11 个指标的基尼指数.其中最小的基尼指数对应的指标,就是左端节点的指标.类推,可以得到右端节点的指标.

步骤 1.5 叶子节点的确定

在各个节点上重复步骤 1.4,就得到了决策树的全部节点.整个过程是让决策树最大程度地生长直至叶子节点.被决定为叶子节点的标准是,该节点无法继续分裂,即节点中只包含一个样本,或该节点中的所有样本属于同一个类别,此时节点的基尼指数为 0.

步骤 2 构建其他的决策树

同理可以构建出第 i 棵决策树($i=2,\cdots,500$).

步骤 3 确定随机森林中最优决策树棵数 tr

步骤 3.1 计算随机森林的判别精度

仿照步骤 1.1 至步骤 1.5,可以建立 20 棵决策树.将这 20 棵决策树构成随机森林模型,模型的输出结果根据每个决策树输出结果投票决定.将标准化后的验证样本 N_{va} 代入每一棵决策树,则得到了由 20 棵决策树构成的 20 个预测模型,故每一个债券都会有 20 个预测结果.把这 20 个预测结果代入上文的式(10),则得到了一个由投票法预测的最终结果.对于样本 N_{va} 的不同债券,则得到了测试集的判别结果 $\hat{y}_{ij}^{(1)}$,与债券实际情况 $y_{ij}^{(1)}(j=1,2,\cdots,4\ 866)$ 对比,则可以得到的第二类错误和 AUC 值,记为 $Type\text{-}II\ Error_{20}=0.252\ 3$,$AUC_{20}=0.876\ 5$.

同理,可以依次建立由 21 棵决策树组成的第二组决策树模型.依此类推,一直建立到第 481 组、由 500 棵决策树组成的随机森林模型.可以得到每一组决策树的预测精度 AUC_i 和误差 $Type\text{-}II\ Error_i$.

步骤 3.2 最优决策树棵数的确定

对于 20 棵,21 棵,…,500 棵决策树组成的随机森林,其中由 80 棵决策树组合构成的随机森林指标最好:$Type\text{-}II\ Error_{\min}=Type\text{-}II\ Error_{80}=0$,$AUC_{\max}=AUC_{80}=1$.故由 80 棵决策树构成的随机森林为最优的随机森林模型,最优的决策树棵数为 80 棵.

3.3.2 最优指标组合的建立

步骤 4 计算指标的重要性评分

在由 80 棵决策树构成的随机森林中,计算各个指标的基尼指数.根据基尼指数越大越重要的原则,指标重要程度排序依次为:货币资金/短期债务,净资产,发行主体发行债券数量,…,息票品种等 123 个指标.

步骤 5 最优指标组合的确定

步骤 5.1 第 i 个指标组合的构建

在步骤 4 的指标序关系中:把排序前 2 位的指标构成指标组合 1;把排序前 3 位的指标构成指标组合 2;…,把排序前 50 位的指标构成指标组合 49.

步骤 5.2 最优指标组合的确定

应用步骤 5.1 中的第 i 个指标组合,步骤3.2 中的决策树棵数 = 80,标准化后的训练样本 N_{tr},建立随机森林模型,并运用投票法进行预测,得到了由第 i 个指标组合的预测精度和误差($i=1,2,\cdots,49$).其中,指标组合 28 的参数最优:$\min\ Type\text{-}II\ Error=Type\text{-}II\ Error_{28}=0$,$\max\ AUC=AUC_{28}=1$.故,步骤 5.1 中的指标组合 28,即排序前 29 个指标构成的指标组合为最优指标组合.

将指标名称列入表 3 的第(3)列,第(2)列为指标对应的准则层.

应该指出,上文系用标准化后的 $t-1$ 时间窗口的样本得到的最优指标组合.其他 $t-m$($m=0,2,3,4,5$)时同理,只不过是样本数据有所不同.

由于时间窗口的大小不同,各个指标在各个时间窗口的解释力不同,所以各时间窗口的指标体系会有不同.

最后得到 $t-m$($m=0,1,\cdots,5$)各期随机森林中最优决策树棵数分别是:52、80、20、37、20、318.

最优指标组合的确定:经过最优指标组合遴

选后,在 $t-m(m=0,1,\cdots,5)$ 的时间窗口样本中分别从 123 个指标遴选出的指标数量为 29、29、23、27、13 和 21.表 3 仅列出 $t-1$ 时的指标,如前所述,这里的窗口 m 为季度.

表 3　$t-1$ 期的指标筛选结果及 Logit 方程系数

Table 3 The optimal indicators system and equation coefficient of logit regression in the time $t-1$

(1) 序号	(2) 准则层		(3) 指标名称	(4) 系数
1	企业内部财务因素	偿债能力	总资产	-2.479 3
2			货币资产	-2.371 6
3			净资产	-3.052 7
4			流动比率	-3.487 3
5			速动比率	2.156 7
6			货币资金/短期债务	1.169 0
7			货币资金/总债务	-9.215 8
8		盈利能力	净利润	5.392 5
9			主营业务利润率	0.049 8
10			总资产报酬率	-5.023 4
11			净资产回报率	-3.111 3
12		运营能力	经营活动现金流	-0.657 7
13			筹资活动现金流	-1.722 5
14	非财务因素		发行主体注册资本	-1.399 2
15			发行主体发行债券数量	-7.874 5
16			大股东类型	-2.236 9
17			是否城投债	-2.788 5
18			城投行政级别	-2.143 0
19	企业外部宏观因素		行业景气指数	-2.402 2
20			分行业企业家信心指数	5.579 7
21			固定资产投资价格指数	-3.279 3
22			小额贷款公司: 机构数量	-2.579 4
23			小额贷款公司: 贷款余额	4.098 8
24			短期贷款基准利率	3.405 7
25			中长期贷款基准利率	1.051 7
26			货币供应量 M_0	0.661 5
27			货币供应量 M_0同比增长率	-1.161 0
28			地区生产总值: 第三产业	-2.785 1
29			城市商品零售价格指数	2.297 3
30	常数项			3.010 2

采用第 $t-m(m=0,1,\cdots,5)$ 个季度的指标和第 t 个季度的违约状态建模,采用第 t 个季度的指标预测第 $t+m(m=0,1,\cdots,5)$ 个季度的违约状态,研究表明

对中短期违约预测均有影响的关键指标为: "货币资金/短期债务"、"净利润"、"发行主体发行债券数量"、"行业景气指数"和"行业企业家信心指数"; 这五个指标在 6 个时间窗口样本均被选入最优指标体系.

对短期违约预测有影响的关键指标为 "货币资产"，"速动比率"，"固定资产投资价格指数"、"货币供应量 M_0"; 这 4 个指标仅在当期违约判别和未来第 1 期,第 2 期的预测中有效($m=0,1,2$).

对中期违约预测有影响的关键指标为 "发行主体注册资本"，"债券到期偿还量"，"债券到期只数"; 这 3 个指标仅预测未来第 3 期,第 4 期,第 5 期的预测中有效($m=3,4,5$).

3.4　基于最优临界点的债券违约预警模型的构建

3.4.1　构建基于 Logit 的违约预警方程

步骤 6　构建基于 Logit 的违约预警方程

以预测期限 $m=1$ 个季度的违约预警方程为例,把表 3 的 29 个指标 4 866 个债券的标准化后的训练样本 N_{tr}对应的矩阵 x_{ij}和对应的向量 Y_j代入式(15),用最大似然估计法计算估计得到 Logit 回归方程的系数 β_1和截距项 α_1,列入表 3 的第(4)列.上述过程可以通过 python 程序实现.得到的方程如式(18)所示

$$P_{j,t}(y_t=1\mid y_{t-1}=0)=\left(1+\exp\left(-\left(\begin{matrix}3.010\,2-2.479\,3x_1-2.371\,6x_2-\\ \cdots-2.785\,1x_{28}+2.297\,3x_{29}\end{matrix}\right)\right)\right)^{-1} \tag{18}$$

式(18)为预测窗口 $m=1$ 时的预测方程,当 m 为其他数值时,类推.限于篇幅、不赘述.

3.4.2　最优判别临界点的确定

步骤 7　构建求解最优违约判别临界点的目标函数

按照 2.4.2 的方法构造求解以第一类错误和第二类错误的加权错误率最小为目标函数.

步骤 8　最优违约判别临界点 P_m^* 的求解

步骤 8.1　参数 $\lambda=1$,第一个临界点 P_m^{11} 时

Q_{11}的确定

先令式 17-1 中参数 $\lambda=1$，再令判别临界点 $P_1^{1,1}=0.000\ 1$.

根据上文 2.4.2，把标准化后的 29 个指标 4 866 个债券的验证样本 N_{va} 对应的矩阵 x_{ij} 和对应的向量 Y_j 的数据代入式(18)，可得预测期限为 1 期时这 4 866 笔债券的违约概率 $\{P_{1,t}, P_{2,t},\cdots, P_{4\,866,t}\}=\{0.056\ 7, 0.087\ 6,\cdots, 0.786\ 5\}$，将 $P_{j,t}$ 依次与临界点 $P_1^{1,1}=0.000\ 1$ 做比较，当 $P_{j,t}\geqslant 0.000\ 1$ 时，$\hat{y}_{j,t}=1$，当 $P_{j,t}<0.000\ 1$ 时，$\hat{y}_{j,t}=0$，于是可以得到这 4 866 笔债券的违约状态的估计值 $\hat{\boldsymbol{y}}_{j,t}=\{\hat{y}_{1,t},\hat{y}_{2,t},\cdots,\hat{y}_{4\,866,t}\}=\{1,1,\cdots,1\}$.

把预测的违约状态 $\hat{y}_{j,t}$ 与实际的违约状态 Y_j 进行对比并统计其频数，就会得到表 1 的混淆矩阵中的各个参数 FP, TN, FN, TP，将得到的参数：$\lambda=1, P_1^{1,1}=0.000\ 1$，可得此时的加权误判率 $Q_{11}^{*}=0.981\ 7$.

步骤 8.2 参数 $\lambda=1$，第 h 个临界点 P_m^{h1} 时最小加权平均成本 Q_{h1}^{*} 的确定

在参数 $\lambda=1$ 时，令第 h 个临界点 P_m^{h1} 以(0,1)的开区间为端点，以步长 10^{-4} 形成 9 999 个临界点.

把 $\lambda=1$，以及 9 999 个 P_m^{h1} 分别代入式(17)，得到 9 999 个 Q_{h1}. 取 $\mathrm{Min}(Q_{1,1}, Q_{2,1},\cdots, Q_{9\,999,1})$，则得到了一个 Q_{h1}^{*}.

步骤 8.3 参数 $\lambda=k$ 时，第 h 个临界点 P_m^{hk} 时最小加权平均成本 Q_{hk}^{*} 的确定

取 $\lambda=1,2,\cdots,5$ 五个数值，仍然取步骤 8.2 的 9 999 个数值，仿照步骤 8.1 和步骤 8.2，则会得到 $Q_{h\lambda}^{*}=Q_{h1}^{*}, Q_{h2}^{*}, Q_{h3}^{*}, Q_{h4}^{*}, Q_{h5}^{*}$.

步骤 8.4 最优违约判别临界点 P_m^{*} 的确定

对于步骤 8.3 中的每一个 $Q_{h\lambda}^{*}=Q_{h1}^{*}, Q_{h2}^{*}, Q_{h3}^{*}, Q_{h4}^{*}, Q_{h5}^{*}$，都可以找到其对应的临界点 $P_{mh}^{\lambda}=P_m^{1\lambda}, P_{m2}^{\lambda}, P_{m3}^{\lambda}, P_{m4}^{\lambda}, P_{m5}^{\lambda}$. 把每一个 P_{mh}^{λ} 用于式(18)的判别，都可以得到一个判别精度 ACC^{λ}，计算结果如表 4 所示.

比较表 4 第(4)列第 1 行~第 5 行的值可以发现，判别精度 $ACC^{*}=\max ACC^{\lambda}=ACC^{3}=0.948\ 5$ 最大所对应的临界点，就是最优的违约判别临界点 $P_m^{*}=P_1^{*}=0.472\ 1$ 和最优的加权平均成本比率系数 $\lambda_m^{*}=3$.

表 4 $t-1$ 期 λ 取不同值时的预测精度对比汇总

Table 4 Comparison of prediction accuracy when λ taking different values in the time $t-1$

(1) 序号	(2) λ	(3) $P_m^{*\lambda}$	(4) $Accuracy^{\lambda}$	(5) *Type-II Error*	(6) *Type-I Error*
1	1	0.688 8	0.940 3	0.047 7	**0.051 8**
2	2	0.473 0	**0.948 5**	**0.015 4**	0.089 6
3	3	**0.472 1**	**0.948 5**	**0.0154**	0.089 6
4	4	0.471 8	**0.948 1**	**0.015 4**	0.090 3
5	5	0.472 8	**0.948 5**	**0.015 4**	0.089 6

步骤 8.5 不同预测期限最优违约判别临界点的确定

上边仅仅是用 $t-m$ 的样本中 $m=1$ 进行计算的. 采用 $m=0,2,3,4,5$ 进行计算，则会得到其他预测期限的 $P_m^{*}, \lambda_m^{*}, ACC^{*}$. 上述整个计算过程，可以方便地利用遗传算法 GA(Genetic Algorithm)进行.

将预测期限 $t+m(m=0,1,\cdots,5)$ 时的 6 个最优违约判别临界点 P_m^{*}，列入表 5 第(3)列第 1 行~第 6 行.

表 5 $t+m$ ($m=0,1,\ldots,5$) 期的最优违约判别临界点 Pm^{*}

Table 5 The result of optimal cutoff point in time $t+m$ ($m=0,1,\cdots,5$)

(1) 序号	(2) 预测期限 $t+m$	(3) 违约判别临界点 P_m^{*}
1	$t+0$(判别模型)	0.603 7
2	$t+1$	0.472 1
3	$t+2$	0.421 9
4	$t+3$	0.473 4
5	$t+4$	0.259 5
6	$t+5$	0.198 2

表 5 第(3)列的研究结果表明，对中国债券市场未来不同预测期限 m 的逻辑回归模型，其最优的违约预测临界点远远偏离流行文献的 0.5. 其中，最大正向偏离为 t-0 的违约判别模型，偏离程度高达 20.74% [(0.603 7-0.5)/0.5=20.74%]；最大负向偏离为未来 $t+5$ 个季度的违约判别模

型,偏离程度高达 60.36% [0.198 2-0.5) /0.5 = -60.36%];最小负向偏离为 $t+1$ 个季度的违约预测模型,偏离程度为 5.58% [0.472 1-0.5) /0.5 = -5.58%].

研究也表明,除了 $m=0$ 时的判别模型临界点大于 0.5,当 $m=1,2,\cdots,5$ 时的临界点均小于 0.5,且随着 m 的增大临界点逐渐减小.

3.5　对比分析

3.5.1　四种情形对比

定义 "一"为本研究的创新点一,即 Logit 模型最优临界点的确定 "二"为本研究的创新点二,即最优指标组合的遴选.

表 6 中各列精度的含义及其计算公式,如上文 2.1 所示.

情形 1　有"一"有"二",其预测精度如表 6 第 1 行所示.

情形 2　有"一"无"二",其预测精度如表 6 第 2 行所示.

情形 3　无"一"有"二",其预测精度如表 6 第 3 行所示.

情形 4　无"一"无"二",其预测精度如表 6 第 4 行所示.

表 6　$t-1$ 季度违约预测模型预测结果

Table 6 Results of $t-1$ quarters default prediction model

(1) 序号	(2) 情形	(3) *Recall*	(4) *Accuracy*	(5) *F-measure*	(6) *G-mean*	(7) *AUC*	(8) *BM*	(9) *Type-II Error*	(10) *Type-I Error*
1	情况 1	**0.984 6**	0.948 1	**0.951 2**	**0.946 4**	0.981 4	**0.894 3**	**0.015 4**	0.090 3
2	情况 2	0.967 8	0.941 7	0.944 6	0.940 6	0.979 3	0.881 9	0.032 2	0.085 9
3	情况 3	0.133 3	**0.990 7**	0.235 3	0.365 1	0.981 8	0.133 3	0.866 7	**0.000 0**
4	情况 4	0.133 3	**0.990 7**	0.235 3	0.365 1	0.903 8	0.133 3	0.8667	0.000 0

根据上述四种情况及其对应的预测精度,可以得到结论如下.

1) 创新点一比创新点二好. 一方面,在表 6 中,情况 2(有"一"无"二")计算的 *Recall* 均远远大于情况 4(无"一"无"二"),对比情况 4(无"一"无"二"),情况 3(无"一"有"二")的变化不大,可以说明创新点一比创新点二好; 另一方面,比较其他综合精度指标 *F-measure*,*G-mean*,*AUC*,*BM*,*Type-II Error*,情况 2 比情况 3 好,其中,本研究重点关注的 *Type-II Error* 指标,在情况 2 也优于情况 3,表示情况 2 与情况 3 相比,将"坏人"错判别成为"好人"的可能性更小,即给投资者带来损失的可能性更小,进一步说明创新点一比创新点二好.

2) 第二个创新点的"有"比"无"好. 因为在表 6 中,情况 3 比情况 4 好,情况 3 中 *AUC* 值高于情况 4.

3) 第一个创新点的"有"比"无"好. 因为在表 6 中,情况 1 和情况 2 的 *Recall* 值均高于 0.9,较未加入第一个创新点的情况 3 和情况 4 明显提高. 同时,对于重点关注的 *Type-II Error* 指标,情况 1 和情况 2*Type-II Error* 值均比情况 3 和情况 4 有明显减小,也可以看到第一个创新点对于降低"坏人"被判别成"好人"的误判率有显著作用.

4) 同时具有本研究两个创新点的违约预测精度,显著高于只有其中一个创新点的情况和一个创新点也没有的情况. 因为在表 6 中,情况 1 的综合精度指标 *Recall*、*F-measure*、*G-mean* 以及把"坏人"判别成为"好人"会给投资者带来巨大损失的 *Type-II Error*,均比其他三种情况好.

上边是针对预测窗口 $m=1$ 时的对比结论,当 m 为其他窗口时,结论相同. 不赘述.

综上,逻辑回归最优违约临界点的确定和最优指标组合的遴选,可以极大提高违约预测的精度; 其中,最优违约预测临界点的确定比最优指标组合的遴选的作用更大.

3.5.2　不同模型的精度对比

选择对比分析模型的根据一是针对性原则. 即对比分析的模型与上文引言中的蓝本文献(Blueprint-paper)相一致或相协调. 因为,论文的特色与创新是在蓝本文献的基础上进行的,这样对比有利于反映针对性和可比性,由于蓝本文献 Perols 等[1] 的方法是支持向量机,故在表 7 第3 行的对比分析模型中选用了支持向量机. 同时由于

本研究的另一个蓝本文献 Tomczak 和 Zięba [23] 采用的方法是人工神经网络，故在表 7 第 5 行的对比分析模型中选用了人工神经网络.表 7 其他有关行的 SVM 和 ANN 类推.

选择对比分析模型的根据二是在满足精度要求的情况下的可解释性原则.下文表 7 除了本研究的模型外，还选用了 7 个对比分析模型.这 7 个常用模型分别是：具有良好解释性的逻辑回归模型(LR)、线性支持向量机模型(SVM)、线性判别模型(LDA)、也选择可解释性稍差、但人们常用的神经网络模型(ANN)、朴素贝叶斯模型(NB)、梯度提升迭代决策树模型(GBDT)和 Adaboost 模型(Ada).

基于上述原则，本研究共选取了 7 个在信用评级领域常用的判别模型与本研究提出的最优临界点的 Logit 模型(LR_A)进行对比.

采用多模型进行对比，将 $t-1$ 时间窗口的测试集 N_{te}，采用本研究最优临界点的 Logit 模型(LR_A)求出模型的预测精度，填入表 7 的第 1 行.仅用本研究的创新点二，即最优指标组合的遴选建立其他模型，其精度如第 2 行~第 8 行所示.

表 7 $t-1$ 期不同违约预测模型对比结果

Table 7 Comparison results of different default prediction models in the time $t-1$

(1) 序号	(2) $t-m$	(3) 模型	(4) *Recall*	(5) *Accuracy*	(6) *F-measure*	(7) *G-mean*	(8) *AUC*	(9) *BM*	(10) *Type-II Error*	(11) *Type-I Error*
1	$t-1$	LR_A	**0.984 6**	0.948 1	**0.951 2**	**0.946 4**	0.981 4	0.132 6	**0.015 4**	0.090 3
2		LR	0.133 3	0.990 7	0.235 3	0.365 1	0.981 8	0.327 6	0.866 7	**0.000 0**
3		SVM	0.266 7	0.992 2	0.421 1	0.516 4	0.984 7	0.333 3	0.733 3	**0.000 0**
4		LDA	0.333 3	0.986 5	0.344 8	0.575 5	0.970 2	0.634 0	0.666 7	0.006 5
5		ANN	0.533 3	0.995 0	0.695 7	0.730 3	**0.990 2**	0.400 0	0.466 7	**0.000 0**
6		NB	0.866 7	0.903 1	0.160 5	0.884 9	0.953 3	0.465 2	0.133 3	0.096 5
7		GBDT	0.666 7	**0.996 4**	0.800 0	0.816 5	0.943 4	0.400 0	0.333 3	**0.000 0**
8		Ada	0.666 7	0.994 3	0.714 3	0.815 6	0.986 3	**0.772 8**	0.333 3	0.002 2

从表 7 可以看出：一是本研究提出的最优临界点的 Logit 模型(LR_A)在 $t-1$ 期预测结果的综合精度：召回率(*Recall*)为 0.982 5，远高于其他几种方法，几何平均值(*G-mean*)为 0.946 4，*F* 值(*F-measure*)为 0.951 2 度均比其他模型高；二是把“坏人”判别成为“好人”、会给投资者带来巨大损失的第二类错误(*Type-II Error*)为 0.015 4，也均比其他模型更低；另外、本模型的其他综合精度，与其他模型不相上下.本研究的 *AUC* 均大于 0.8 [24]，精度可行.

上边是针对预测窗口 $m=1$ 时的对比结论，当 m 为其他窗口时，结论相同.不赘述.

综上，本研究提出的最优临界点的 Logit 模型(LR_A)，其综合预测精度都高于常用的逻辑回归模型(LR)，线性支持向量机模型(SVM)，线性判别模型(LDA)，神经网络模型(ANN)，朴素贝叶斯模型(NB)，梯度提升迭代决策树模型(GBDT)和 Adaboost 模型(Ada) 7 种典型的违约预测模型.采用最优临界点的 Logit 模型(LR_A)进行违约债券的预警，避免把“坏人”判为“好人”，给投资者带来巨大损失，具有重要的现实意义.

3.5.3 对比分析结论

本研究进行了两方面的对比分析.一是本研究两个创新点作用的对比分析，研究发现，逻辑回归最优违约临界点的确定(创新点一)和最优指标组合的遴选(创新点二)，都可以极大提高违约预测的精度；其中，最优违约预测临界点的确定比最优指标组合的遴选的作用更大.二是本研究模型与经典违约预测模型的精度对比分析，分析结果表明，本研究提出的最优临界点的 Logit 模型(LR_A)综合预测精度优于常用的逻辑回归模型(LR)，线性支持向量机模型(SVM)，线性判别模型(LDA)，神经网络模型(ANN)，朴素贝叶斯模型(NB)，梯度提升迭代决策树模型(GBDT)和 Adaboost 模型(Ada) 7 种典型的违约预测模型.

3.6 中国债券信用特征分析

由实证数据可知，$t-3$ 时间窗口训练的模型判别效果最好.因此，本节应用 $t-3$ 时间窗口训练

得到的模型.

这里应用本研究建立的模型，指标 x_{ij} 为 2018 年第 4 个季度的数据建立模型并进行 $t+3$ 期的预测，即预测 2019 年第 3 季度的债券违约状态.

将得到的违约概率预测值转化为标准化的信用得分，分别研究中国债券所属行业和发债主体公司属性的信用分布特征.

设 p_i 表示第 i 笔债券的违约概率预测值，则第 i 笔债券信用得分（$Score_i$）的计算如式（19）所示.

$$Score_i=(1-p_i)\times 100 \tag{19}$$

式（19）的含义是将 [0,1] 区间的违约概率值转化为 [0,100] 区间的信用得分.

将式（19）得到的信用得分标准化，得到标准化的信用得分（$SScore_i$），计算过程如式（20）所示.

$$SScore_i=\frac{Score_i-Score_{\min}}{Score_{\max}-Score_{\min}}\times 100 \tag{20}$$

式（20）的含义是将信用得分在 [0,100] 区间内以最低分为 0、最高分为 100 标准化.

3.6.1　行业分析

根据中国证监会对行业的划分标准，本研究债券行业可划分为 18 类.按照式（20）计算标准化后的信用得分，并统计不同行业信用得分的均值、最大值、最小值、标准差，将结果列入表 8 的第 3 行～第 6 行.

表 8　中国债券市场信用资质行业分布

Table 8 Distribution of credit qualification industries of listed companies in China

(1) 序号	(2) 行业	(3) 平均值	(4) 最大值	(5) 最小值	(6) 标准差
1	采矿业	76.085 7	99.999 9	0.016 8	30.742 0
2	电力、热力、燃气及水生产和供应业	81.307 5	99.999 3	0.050 5	29.988 3
3	房地产业	75.754 5	99.998 9	0.058 3	32.569 5
4	建筑业	79.887 9	99.999 7	0.023 8	30.801 0
5	交通运输、仓储和邮政业	82.162 0	99.999 8	0.069 8	28.974 2
6	教育	85.385 1	99.974 0	2.076 1	29.754 0
7	金融业	63.772 5	99.995 9	0.001 3	34.297 4
8	科学研究和技术服务业	82.483 6	99.981 9	0.027 3	30.471 2
9	农、林、牧、渔业	70.598 1	99.963 7	0.025 4	35.114 1
10	批发和零售业	81.402 8	99.999 9	0.072 8	29.149 0
11	水利、环境和公共设施管理业	74.302 8	99.999 9	0.019 3	33.791 5
12	卫生和社会工作	61.654 0	99.946 8	1.047 1	35.098 7
13	文化、体育和娱乐业	70.893 5	99.997 6	0.010 2	34.724 2
14	信息传输、软件和信息技术服务业	83.286 0	99.999 5	0.119 2	29.821 0
15	制造业	85.050 1	100.000 0	0.051 6	27.385 9
16	住宿和餐饮业	83.804 3	99.993 7	0.085 8	28.921 2
17	综合	74.894 9	99.999 4	0.016 6	32.977 0
18	租赁和商务服务业	59.982 7	99.997 0	0.022 0	35.920 9

为检验不同行业之间的信用得分的平均值是否存在显著差异，本研究采用 SPSS 软件进行克鲁斯卡尔-沃利斯检验（Kruskal-Wallis test，K-W 检验）来进行分析.克鲁斯卡尔-沃利斯检验用以检验两个以上样本是否来自同一个概率分布的一种非参数方法[33].这里用来检验不同行业间的信用得分均值是否有显著性差异.由于克鲁斯卡尔-沃利斯检验适用于多分类的数据，可以检验两种以上的行业间的信用得分均值差异，而且不要求样本服从正态分布，更符合本研究的检验需求.

假设检验摘要如表 9 所示，并将两两行业成对比较的显著性检验结果列入表 10.表 10 中的常数是克鲁斯卡尔-沃利斯检验的统计量.

表 9 发债企业所属证监会行业信用资质差异检验摘要

Table 9 Difference test results of credit score among the industries

(1) 原假设	(2) 检验	(3) 显著性	(4) 决策
在“发行主体所属证监会行业名称”的类别中，“信用得分值”的分布相同.	独立样本克鲁斯卡尔-沃利斯检验	0.000	拒绝原假设

注：显著性水平为 0.050.

表 10 不同发债企业所属证监会行业之间信用得分的差异性检验

Table 10 Difference test of credit score between the ownership of firms

序号	所属行业	(1) 金融业	(2) 卫生和社会工作	…	(17) 制造业	(18) 教育
1	金融业	—	—	…	—	—
2	卫生和社会工作	0.078	—	…	—	—
…	…	…	…	…	…	…
17	制造业	33.071***	19.159***	…	—	—
18	教育	−2.132**	−1.889*	…	−0.975	—

注：***，**，* 分别表示在 99%/95%/90% 的置信水平下显著.

1) “教育”和“制造业”债券的信用资质最好.二者差别不显著.因为表 8 中，其信用得分的均值最大，在表 10 中这两类企业之间没有显著差异，他们分别与其他企业具有显著差异.

2) “住宿和餐饮业”和“信息传输、软件和信息技术服务业”的信用资质较好.二者差别不显著.因为表 8 中，其信用得分的均值位于“教育”和“制造业”之后.在表 10 中这两类企业之间没有显著差异，他们分别与其他企业具有显著差异.

3) “金融业”和“卫生和社会工作”和“租赁和商务服务业”三个行业较差，其中“租赁和商务服务业”最差.三者之间两两差别不显著，这三个行业与其他各个行业差异显著.因为表 8 中，其信用得分的均值位于最后.在表 10 中这三类企业两两之间没有显著差异，他们分别与其他企业具有显著差异.

4) “筹资活动现金流”指标来看，租赁和商务服务业与金融业债券的筹资活动现金流数值相对较高.可能是由于这类企业的筹资活动现金流主要来源于发债，企业资本及债务规模和结构发生变化的活动较为频繁，导致信用资质不佳.

3.6.2 发债主体所有制分析

本节根据发债主体公司属性，将所有债券分为 11 类，发债主体公司属性类别分别是中央国有企业、地方国有企业、公众企业、国有企业、集体企业、民营企业、个人、外商独资企业、外资企业、中外合资企业和其他企业.

将每种属性的发债公司所发行的债券按照式 (20) 计算对应的信用得分，并统计不同发债公司属性发行债券的信用得分的平均值、最大值、最小值和标准差，将结果列入表 11 的第 3 列~第 6 列.

表 11 不同发债公司属性信用特征描述表

Table 11 Credit characteristic description of enterprise firm ownership

(1) 序号	(2) 所有权属性	(3) 平均值	(4) 最大值	(5) 最小值	(6) 标准差
1	地方国有企业	75.962 8	99.999 9	0.004 5	31.801 6
2	个人	95.231 4	99.999 2	0.575 6	15.699 5
3	公众企业	64.039 5	99.999 7	0.001 3	35.972 9
4	国有企业	88.126 2	99.798 7	3.207 9	16.904 5
5	集体企业	82.855 6	99.999 2	0.141 0	29.639 5
6	民营企业	86.929 7	100.000 0	0.071 5	26.670 3
7	其他企业	79.699 7	99.994 8	0.108 7	31.104 0
8	外商独资企业	79.567 0	99.989 0	0.096 3	29.884 9
9	外资企业	80.779 5	99.994 7	0.024 6	31.891 7
10	中外合资企业	81.557 8	99.997 9	0.036 7	29.973 8
11	中央国有企业	76.232 9	99.999 4	0.013 6	32.011 1

为检验不同属性公司发行债券的信用得分的平均值之间是否存在显著差异，仍然采用 SPSS 软件进行克鲁斯卡尔-沃利斯检验 (Kruskal-Wallis test，K-W 检验) 来进行分析.检验假设摘要如表 12 所示，并将两两成对比较的显著性检验结果列入表 13.

表 12 发债企业属性信用资质差异检验摘要

Table 12 Difference test results of credit score among the ownership of firms

(1) 原假设	(2) 检验	(3) 显著性	(4) 决策
在“发行主体公司属性”的类别中，“信用得分值”的分布相同.	独立样本克鲁斯卡尔-沃利斯检验	0.000	拒绝原假设

注：显著性水平为 0.050.

表 13　不同发债企业属性之间信用得分的差异性检验

Table 13 Difference test of credit score between the ownership of firms

序号	企业属性	(1) 公众企业	(2) 地方国有企业	…	(9) 集体企业	(10) 民营企业	(11) 个人
1	公众企业	—	—	…	—	—	—
2	地方国有企业	-60.751***	—	…	—	—	—
…	…	…	…	…	…	…	…
10	民营企业	-112.199***	-81.544***	…	7.159***	—	—
11	个人	48.701***	35.047***	…	15.343***	14.825***	—

1) 个人企业发行债券的信用资质最高.因为表 11 中,其信用得分的均值最大,在表 13 中这类企业又与其他企业具有显著差异.研究发现个人企业发行的债券的“债券回售只数”相比其他所有制类别债券都要高,提前赎回债券的行为可以大大提高企业的信用资质,这可能是个人企业债券的信用资质高的原因之一.

2) 国有企业与民营企业债券的信用水平居中.因为表 11 中,国有企业与民营企业的信用得分在个人企业之后.

3) 公众企业债券的信用资质最差.理由同上的表 11 与表 13 的分析.而公众企业发行债券的“债券回售只数”最少,进一步说明了“债券回售只数”可能是影响债券信用资质的原因之一.

4　结束语

4.1　主要研究发现

研究表明对中国中短期违约预测均有影响的关键指标为:“货币资金/短期债务”、“净利润”、“发行主体发行债券数量”、“行业景气指数”和“行业企业家信心指数”等五个指标.对短期违约预测有影响的关键指标为:“货币资产”、“速动比率”、“固定资产投资价格指数”、“货币供应量 M_0”.对中期违约预测有影响的关键指标为“发行主体注册资本”、“债券到期偿还量”、“债券到期指数”.

上述研究结论的经济学含义比较明显,例如当一个行业不景气时,这个行业的多数企业都步履维艰、则债券的清偿也就成了问题; 再如当货币供应量 M_0 不充分时,大部分企业的清偿能力都会减少.其他财务指标的经济学含义更为明显,读者可自行分析,不赘述.

4.2　主要创新点

1) 在最优指标组合的遴选上,本研究与 Park 和 Kim 最为相似,但在选择最优决策树棵数的标准不同.对于不同的决策树棵数,在第二类错误最小的前提下,通过 AUC 最大反推得到了一个最优的随机森林; 通过对最优随机森林中节点指标的不同组合的对比,找到 AUC 最大的一个指标组合,这就改变了随机森林的现有研究,决策树棵数人为设定的美中不足.任何的违约判别模型和预测模型都是与变量相关的,不失一般性,本研究事实上提供了一个变量组合选择的一般框架,不但可以判别和预测公司债券违约,对企业的其它违约也有借鉴意义.

2) 在最优违约判别临界点的确定上,与 Perols 等,Tomczak 和 Zięba 最为相似,也选择用传统的逻辑回归模型进行违约预测.与他们不同的是,以第一类错误与第二类错误加权错误率最小为标准,反推最优临界点,以改进逻辑回归模型.对于第一类错误与第二类错误的不同配比,以第一类错误与第二类错误的加权之和最小为目标函数,反推逻辑回归的最优临界点.最优的违约预测临界点远远偏离流行文献的 0.5.最优临界点的发现,极大地提高了违约债券的预测精度,对于公司债券市场可以减少投资损失,对于银行可以减少坏账,对于优质公司可以促进债券流通,使公司能够健康发展.

3) 在实证的对比分析中,本模型的预测精度

高于支持向量机模型、梯度提升迭代决策树、神经网络等 7 种流行的大数据预测模型.本研究的研究成果对降低债券违约风险,保障债券投资者、商业银行、发债公司及相关者的利益具有重要意义.

4) 在本研究的实证分析中,以第 t 个季度的债券违约状态 y_t 和 $t-m$ 季度的指标数据 x_{t-m} 建立违约预测模型,达到了用 t 季度的指标数据 x_t,预测第 $t+m$ 季度的违约状态 y_{t+m}.

参 考 文 献:

[1] Perols J L, Bowen R M, Zimmermann C, et al. Finding needles in a haystack: Using data analytics to improve fraud prediction [J]. The Accounting Review, 2017, 92(2): 221-245.

[2] Tomczak J M, Zieba M. Classification restricted boltzmann machine for comprehensible credit scoring model [J]. Expert Systems with Applications, 2015, 42(4): 1789-1796.

[3] Park C H, Kim S B. Sequential random k-nearest neighbor feature selection for high-dimensional data [J]. Expert Systems with Applications, 2015, 42(5): 2336-2342.

[4] Mizen P, Tsoukas S. Forecasting US bond default ratings allowing for previous and initial state dependence in an ordered probit model [J]. International Journal of Forecasting, 2012, 28(1): 273-287.

[5] Jabeur S B, Sadaaoui A, Sghaier A, et al. Machine learning models and cost-sensitive decision trees for bond rating prediction [J]. Journal of the Operational Research Society, 2020, 71(8): 1161-1179.

[6] 迟国泰, 于善丽. 基于违约鉴别能力最大的信用等级划分方法 [J]. 管理科学学报, 2019, 22(11): 106-126.
Chi Guotai, Yu Shanli. Credit rating division method with maximum default identification capability [J]. Journal of Management Sciences in China, 2019, 22(11): 106-126. (in Chinese)

[7] 张鹏东, 潘 越, 陈思岑, 等. 打破刚兑、债券利率市场化与企业研发决策 [J]. 管理科学学报, 2022, 25(8): 63-81.
Zhang Pengdong, Pan Yue, Chen Sicen, et al. Breaking the rigid payment, bond yield liberalization and corporate R&D decision [J]. Journal of Management Sciences in China, 2022, 25(8): 63-81. (in Chinese)

[8] 刘海龙. 债券定价与债券风险预警方法综述 [J].系统管理学报, 2016, 25(1): 1-10.
Liu Hailong. Summary of bond pricing and bond risk early warning methods [J]. Journal of System Management, 2016, 25(1): 1-10. (in Chinese)

[9] 俞宁子, 刘斯峰, 欧阳炎力, 等. 债券违约风险预警模型探究 [J]. 中国市场, 2016, (39): 18-29.
Yu Ningzi, Liu Sifeng, Ouyang Yanli, et al. Research on early warning model of bond default risk [J]. Chinese market, 2016, (39): 18-29. (in Chinese)

[10] Reunanen J, Guyon I, Elisseeff A. Overfitting in making comparisons between variable selection methods [J]. Journal of Machine Learning Research, 2003, (3): 1371-1382.

[11] Ma L, Zhao X, Zhou Z L, et al. A new aspect on P2P online lending default prediction using meta-level phone usage data in China [J]. Decision Support Systems, 2018, 111: 60-71.

[12] Liang D, Lu C C, Tsai C F, et al. Financial ratios and corporate governance indicators in bankruptcy prediction: A comprehensive study [J]. European Journal of Operational Research, 2016, 252(2): 561-272.

[13] Oreski S, Oreski G. Genetic algorithm-based heuristic for feature selection in credit risk assessment [J]. Expert Systems with Applications, 2014, 41(4): 2052-2064.

[14] Xia Y F, Li C Z, Li Y Y, et al. A boosted decision tree approach using Bayesian hyper-parameter optimization for credit scoring [J]. Expert Systems with Applications, 2017, 78(7): 225-241.

[15] Papouskova M, Hajek P. Two-stage consumer credit risk modelling using heterogeneous ensemble learning [J]. Decision

Support Systems，2019，118(5)：33-45.

[16] Tsai H T，Thomas L C，Yeh H C. An economic model for credit assessment problems using screening approaches [J]. Journal of the Operational Research Society，2005，56(7)：836-843.

[17] Verbraken T，Bravo C，Weber R，et al. Development and application of consumer credit scoring models using profit-based classification measures [J]. European Journal of Operational Research，2014，238(2)：505-513.

[18] Mohammadi N，Zangeneh M. Customer credit risk assessment using artificial neural networks [J]. International Journal of Information Technology and Computer Science，2016，8(3)：58-66.

[19] Carvalho V I D，Branscum A J. Bayesian nonparametric inference for the three-class Youden index and its associated optimal cutoff points [J]. Statistical Methods in Medical Research，2018，27(3)：689-700.

[20] Kouvelis P，Zhao W H. Who should finance the supply chain? Impact of credit ratings on supply chain decisions [J]. Manufacturing & Service Operations Management，2018，20(1)：19-35.

[21] 迟国泰，曹　勇，党均章. 基于最优负债系数的上市银行违约概率测算模型与实证 [J]. 运筹与管理，2012，21(3)：176-186.

Chi Guotai，Cao Yong，Dang Junzhang. Calculation model and demonstration of default probability of listed banks based on optimal debt coefficient [J]. Operations Research and Management，2012，21(3)：176-186. (in Chinese)

[22] Davis E P，Karim D. Comparing early warning systems for banking crises [J]. Journal of Financial Stability，2008，4(2)：89-120.

[23] Celik A E，Karatepe Y. Evaluating and forecasting banking crises through neural network models：An application for Turkish banking sector [J]. Expert Systems with Application，2007，33(4)：809-815.

[24] 迟国泰，李鸿禧，潘明道. 基于违约鉴别能力组合赋权的小企业信用评级——基于小型工业企业样本数据的实证分析 [J]. 管理科学学报，2018，21(3)：105-126.

Chi Guotai，Li Hongxi，Pan Mingdao. Small enterprises credit rating based on default identification ability of combination weighting：Based on an empirical analysis of small industrial enterprises [J]. Journal of Management Sciences in China，2018，21(3)：105-126. (in Chinese)

[25] Jones S. Corporate bankruptcy prediction：A high dimensional analysis [J]. Review of Accounting Studies，2017，22(3)：1366-1422.

[26] Hilscher J，Wilson M. Credit ratings and credit risk：Is one measure enough? [J]. Management Science，2017，63(10)：3414-3437.

[27] Vo D H，Pham B N V，Ho C M，et al. Corporate financial distress of industry level listings in vietnam [J]. Journal of Risk & Financial Management，2019，12(4)：1-17.

[28] Chawla N V，Bowyer K W，Hall L O，et al. SMOTE：Synthetic minority over-sampling technique [J]. Journal of Artificial Intelligence Research，2002，16(1)：321-357.

[29] Vens C，Costa F. Random Forest Based Feature Induction [C]. 11th IEEE International Conference on Data Mining，2011：744-753.

[30] Huang G，Luo H，Xia J. Invest in information or wing it? A model of dynamic pricing with seller learning [J]. Management Science，2019，65(12)：5556-5583.

[31] Zhang H P，Wang M H. Search for the smallest random forest [J]. Statistics and ITS Interface，2009，2(3)：381.

[32] Carmona P，Climent F，Momparler A. Predicting failure in the U.S. banking sector：An extreme gradient boosting approach [J]. International Review of Economics and Finance，2019，61(5)：304-323.

[33] Amos C，Holmes G，Strutton D. Exploring the relationship between celebrity endorser effects and advertising effectiveness：A quantitative synthesis of effect size [J]. International Journal of Advertising，2008，27(2)：209-234.

Dynamic early warning of bond default based on optimal threshold

CHI Guo-tai[1], *YANG Jia-qi*[1,2*], *ZHOU Ying*[1]

1. School of Economics and Management, Dalian University of Technology, Dalian 116024, China;
2. School of Economics and Management, Beihang University, Beijing 100191, China

Abstract: Early warning of bond default risk is to predict the future bond default status based on the enterprise's financial factors, non-financial factors, and external macro factors. For different combinations of variables, the effect of default prediction is different; there must be an optimal combination of indicators, which can minimize the error of default prediction. For different thresholds, the effect of default prediction is different, and there is bound to be an optimal threshold of default judgment, which can distinguish between the bonds that default from those that do not to the greatest extent. This paper uses the random forest model to select the feature combination, and studies the default risk of bonds based on Logit model. The first contribution of this paper is in the optimal feature selection. Under the premise of the minimum *Type-II Error*, an optimal random forest is obtained by maximizing the *AUC* for different numbers of decision trees. According to the ranking of the importance of indicators, the optimal feature combination can be obtained by forward selection to maximize *AUC*. The second contribution is in the determination of the optimal default judgment threshold. Taking the minimum weighted sum of the *Type-I Error* and *Type-II Error* as the objective function, the optimal threshold of logical regression is deduced. The third is the prediction accuracy of this model is higher than popular big data prediction models. Based on data from bonds issued by Chinese bonds listed companies from 2014 to 2018, the empirical research shows that the key indicators affecting China's medium and short-term default prediction are: Monetary capital / short-term debt, net profit, the number of bonds issued by issuers, industry prosperity index, and industry entrepreneur confidence index. The key indicators affecting short-term default prediction are: Monetary assets, quick ratio, fixed asset investment price index, and money supply M_0. The key indicators that have an impact on the medium-term default forecast are registered capital of the issuer, repayment amount of bonds at maturity, and bond maturity index.

Key words: bond default; default prediction; feature selection; optimal threshold; random forests