信用债发行人信息披露与债券违约的关联性研究
——基于机器学习梯度提升模型

李　杰　孟祥军

摘　要：本文以我国信用债市场的全样本数据为基础，应用过采样技术克服违约样本不均衡所带来的限制，搭建适合我国实际情况的机器学习梯度提升模型（CatBoost、XGBoost 和 LightGBM 模型），并对模型在我国信用债市场的可用性和模型预测效果表现进行比较研究，为判断模型的识别区分能力及今后相关模型研究选择提供参考。同时，本文建立了适用于我国信用债实际情况的分级预警模型，对于信用债投资者和监管者有效预测企业未来的债务违约风险具有参考意义。

关键词：信用债　信息披露　机器学习　违约预警

一、研究背景与贡献

（一）研究背景

1. 信用债对支持实体经济发展具有重要作用

近年来，我国债券市场规模迅速扩大，截至 2023 年第一季度末，我国银行间市场及北京、上海、深圳交易所共有各类存量信用债 57 761 只、债券余额 58.97 万亿元。近十年以来，我国存量信用债的存续数量和规模分别增长逾 12 倍和 6 倍，年几何平均增速均超过 20%（图 1~2）。信用债市场对支持国家战略和实体经济起到了至关重要的作用。

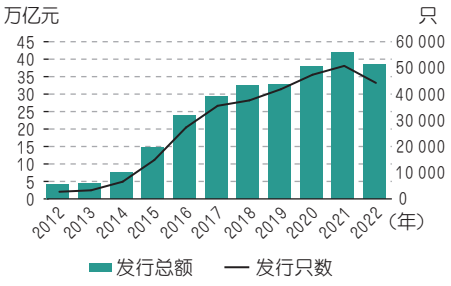
2. 近年来信用债违约呈明显上升趋势

自 2014 年“11 超日债”成为我国债券市场上首例违约的企业债券之后，我国债券市场违约债券的规模与数量有所增加。特别是 2018 年以来，违约户数、债券违约只数均呈上升趋势（图 3）。2023 年第一季度末，新增违约户数 38 户、违约债券 74 只、违约金额约 730 亿元。

3. 对债券违约预警有效信息的识别与筛选至关重要但具有挑战性

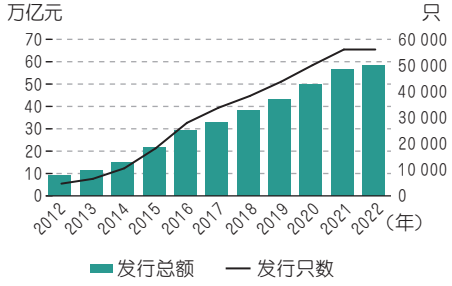
债券违约不仅令投资者损失巨大，而且频繁或突发的大额债券违约可能引发市场恐慌，不利于债券市场的正常运行和稳定发展。毫无疑问，向信用债投资者、监管者提供有用的信息，对于支持其做出投资或管理决策，以及尽早发现和化解可能出现的违约风险，都具有重要意义。但是，面对浩如烟海的各类信息，特别是复杂的财务信息，哪些信息与债券违约可能存

李杰、孟祥军，信永中和会计师事务所金融服务合伙人。本文系 2021 年度“NAFMII 研究计划”课题结项成果。课题主持人：李杰，课题组成员：孟祥军、罗玉成、徐扬、贺鹏、耿志强、王妍、高艺航。



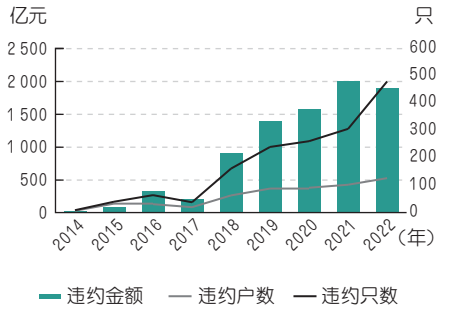
数据来源：Wind。

图 1　历年发行债券只数与规模



数据来源：Wind。

图 2　历年存量债券只数与规模



数据来源：Wind。

图 3　历年违约债券户数、只数与违约金额

在关联关系？信息具有多大程度的预警价值？预警准确性如何？除财务信息外，诉讼信息等其他信息是否对债券违约具有预警价值？这些问题始终是信用债市场投资者和监管者共同关注的问题。本文聚焦于此，对债券发行人财务信息及法律诉讼信息披露与信用债违约的关联性进行研究。

（二）研究贡献

1. 以信用债市场全样本数据为基础，识别关键预警指标，反映我国市场特征

区别于以往文献，本文以我国信用债市场的全样本数据为基础进行研究，通过对违约样本的过采样技术处理，在机器学习模型中实现模型优化，建立适用于我国信用债实际情况的 CatBoost、XGBoost、LightGBM 模型；通过训练集样本选择的标准变化，构建分级预警模型。使用样本外数据进行检验，结果表明所构建模型具有良好的违约预测能力，对于信用债投资者和监管者有效预测信用债发行人未来的违约风险具有参考意义。

2. 对机器学习模型在我国信用债违约预警识别中的可用性进行检验

本文对新兴的梯度提升模型在我国信用债违约预警识别中的可用性进行了检验，在综合性基础上对前述模型在我国信用债市场的模型预测效果表现进行了比较研究，为判断模型的识别区分能力及今后相关模型研究选择提供参考。

3. 建立信用债分级违约预警模型，提高信息披露的有效性

本文应用 CatBoost、XGBoost、LightGBM 模型，对信用债违约识别和预警中的各财务指标贡献度进行量化分析，增强了模型的可解释性，一定程度上克服了大多数机

器学习模型黑盒方法的缺陷，有助于债券市场投资者和监管者有针对性地获取和应用财务信息，更好地规范和监督债券信息披露质量，加强特质性的风险信息披露，以保护债券投资者的合法权益，促进我国信用债市场的健康发展。

二、研究方案设计与数据准备

（一）研究方案设计

本文在对违约风险预警模型的研究文献及我国信用债市场现状和违约情况进行整理的基础上，获取我国信用债市场的全样本数据进行清洗与整合；随后，构建三种典型的机器学习模型，并对各模型进行有效性检验。最后，得出信息披露与债券违约的关联性研究结论以及债券发行人信息披露建议（图 4）。

（二）指标体系设计

本文综合考虑已有的研究共识，结合指标的可靠性与可获取性，进行风险预警

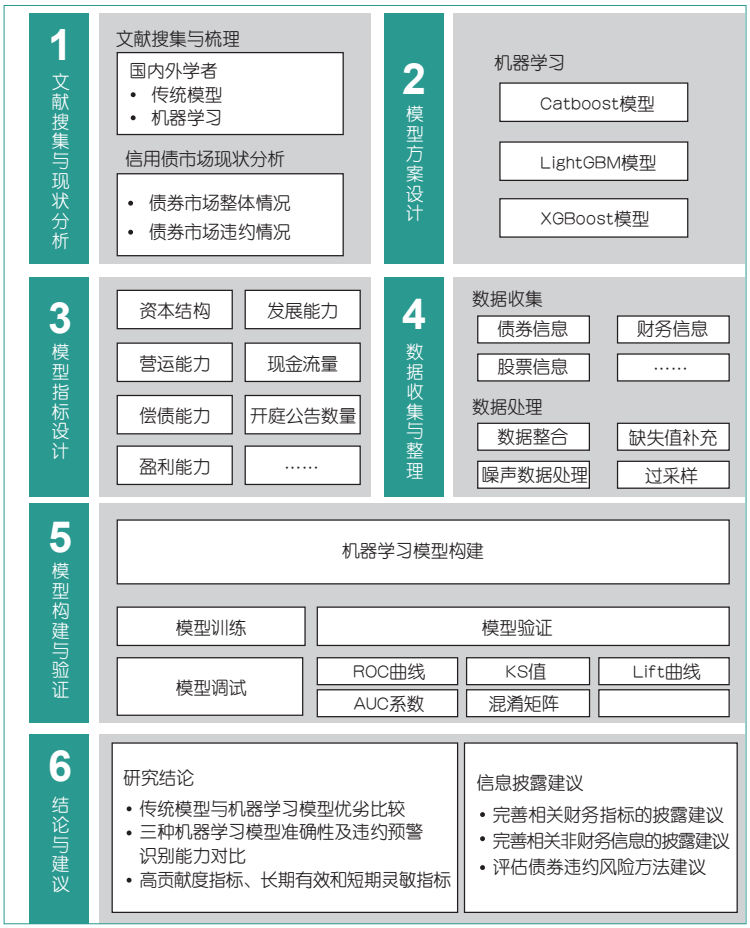


图 4　研究方案构建的总体思路

指标体系搭建。我们从资本结构、营运能力、盈利能力、现金流量、偿债能力、发展能力六个维度进行考虑，确定了 34 个指标，并根据可获得性确定了 1 个非财务指标即开庭公告数量（附表 1）[①]。

（三）违约定义与样本确定

1. 违约定义

本文所称发行人违约包括以下几种情形：本金到期未兑付；利息到期未兑付；本息到期未兑付；触发交叉违约；技术性违约；展期。

2. 模型样本确定

（1）确定规则与流程

我国信用债首次违约发生于 2014 年，截至 2020 年 12 月末，累计违约家数为 290 家。考虑到行业异质性，我们将金融债、同业存单及金融行业信用债剔除出研究样本。本文将全部信用债纳入研究范围，并将样本分类为违约样本和正常样本（表 1）。发行人财务数据来源包括同花顺、Wind、上海证券交易所、深圳证券交易所、北京证券交易所、上海清算所、中国货币网等。非财务数据即开庭公告数量来源于天眼查网站。

本文在进行信息披露与违约关联性研究时，对于违约样本分别以其首次违约时点 T 的前一年（T–1）和违约时点的前两年（T–2）的各项指标对 T 时点是否违约进行识别。如，发债主体在 2020 年发生信用债违约，以其 2019 年年末和 2018 年年末各项指标数据识别 2020 年的违约状况。对于正常样本，本文选择与违约样本相匹配的数据区间内所有未违约且财务记录完整的发行人，以其发行债券期间每一年末财务数据作为一个单独的样本，形成正常样本集。机器学习模型样本确定规则与流程如图 5 所示。

表 1　信用债各年度违约户数

起始日期	截止日期	年度	违约户数	违约只数	违约金额（亿元）
2014-01-01	2014-12-31	2014	6	7	2.80
2015-01-01	2015-12-31	2015	27	31	80.86
2016-01-01	2016-12-31	2016	34	81	264.44
2017-01-01	2017-12-31	2017	21	40	192.29
2018-01-01	2018-12-31	2018	55	143	888.79
2019-01-01	2019-12-31	2019	79	210	1 312.47
2020-01-01	2020-12-31	2020	68	197	1 701.43
2021-01-01	2021-12-03	2021	63	174	1 533.83

数据来源：同花顺。

注：违约样本数据区间为 2014 年至 2021 年 10 月末。自同花顺金融数据终端对样本数据进行获取，剔除金融行业、金融债、同业存单以及重复样本后，按照首次违约统计的违约家数为 220 家。

（2）数据处理说明

缺失值处理。违约样本数据集与正常样本数据集的财务指标均存在缺失值，由于正常发行人数量充足，将存在缺失值的样本删除后仍满足建立模型的需要，因此仅对违约样本的财务指标进行人工计算补充。

噪声数据处理。噪声数据是指数据

① 附表1~6为增强出版，中国知网—《金融市场研究》。

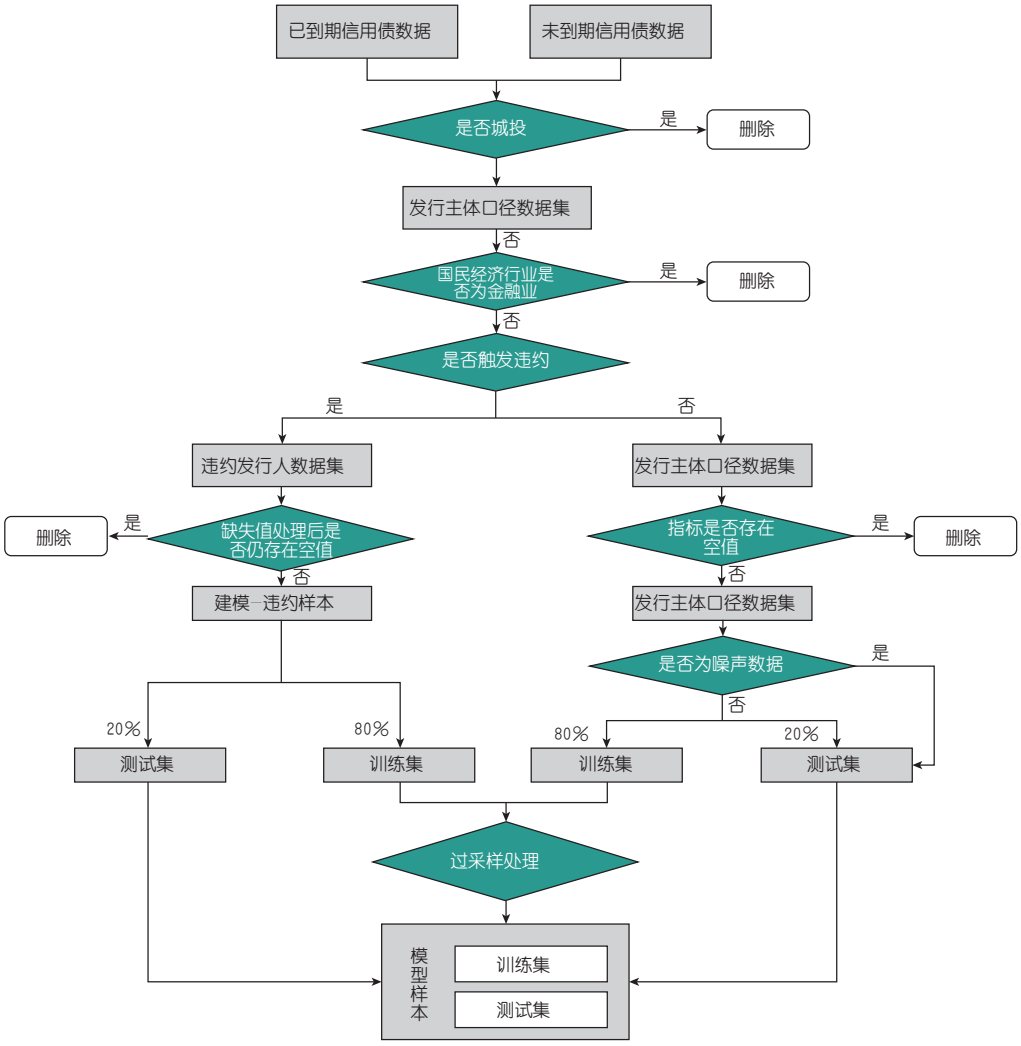


图 5　机器学习模型样本确定规则与流程

中存在着错误或异常的数据，例如，个别违约发行人中的某些指标的数据明显优于大部分正常发行人，个别正常发行人中的某些指标的数据明显劣于大部分违约发行人。这使得在样本所在的特征空间中，违约样本会存在于正常样本富集的区域，正常样本也会存在于违约样本富集的区域，从而导致严重的过拟合。为寻找出正常样本与违约样本的合适边界，研究中依次删除正常发行人中各指标低于其 5%、10%、15% 分位数的数据（反向指标分别删除高于 95%、90%、85% 的数据），得到三份正

常发行人数据集。

过采样。由于正常样本与违约样本存在着严重的不平衡，这会导致机器学习模型结果偏向于正常样本，丧失准确性。因此，对违约样本进行 SMOTE 过采样处理，过采样只适用于训练集中的数据，测试集的数据需保证都是真实的数据，不能进行过采样处理。

（3）模型样本确定

经过缺失值补充，将按照不同噪声数据处理标准得到的三份正常样本分别与 T–1 和 T–2 违约样本进行过采样处理，形成六份模型样本。经过缺失值补充、噪声数据处理和过采样处理后，样本数量可以达到模型需求。机器学习建模样本数量情况如附表 2~3 所示。

三、实证分析结果与模型应用

（一）方案设计

模型构建的环境使用 Python 语言，利用 CatBoost 包、LightGBM 包、XGBoost 包中的函数建立机器学习模型。在进行模型构建时，需要考虑模型的相关参数，模型输入的参数不同，通常会导致模型结果存在较大的差异，得到的模型可能是欠拟合或过拟合的。前述模型需不断调优的参数包括：损失函数（Loss_function），用来评估模型好坏的函数，机器学习的目标即找到使损失最小的学习器；学习速率（Learning_rate），代表模型的收敛速度，影响着模型训练的时间；树的深度（Depth），深度越深，则模型拟合数据的能力越强，但泛化能力会降低，导致模型在测试集上表现不佳；正则化系数（L2_leaf_reg），设置是为了控制模型的复杂度，过于复杂的模型会降低模型的泛化能力；目标（Eval_metric），用于验证数据的度量标准。

在对模型进行调参时，本文采用的方法为网格搜索法。该方法主要思想是人为地为需要调参的参数设置备选范围，网格搜索将在每个可能的参数组合上构建一个模型，以模型 Accuracy 值最高为标准，搜索出最优参数组合值。参数备选范围如附表 4 所示。

（二）模型结果

1. 在梯度提升模型中，CatBoost 模型对信用债违约预警识别的表现略优于 XGBoost 和 LightGBM 模型

混淆矩阵。从表 2 可见，对违约样本的识别方面，LightGBM（正确率 80%）略优于 CatBoost（正确率 76.67%）、XGBoost 模型（正确率 76.67%）。但在对正常样本的识别方面，CatBoost（正确率 80.84%）则优于 LightGBM（正确率 75.43%）、XGBoost 模型（正确率 75.25%）。总体正确率方面，CatBoost、LightGBM、XGBoost 模型分别 80.83%、75.43%、75.25%。总体而言，CatBoost 模型表现略优于 LightGBM、XGBoost 模型（图 6）。

ROC 曲线。CatBoost、LightGBM、XGBoost 模型的 AUC 系数分别为 86.57%、85.21%、85.70%，都略高于国际先进水平（85%），CatBoost 模型表现略优于 LightGBM、XGBoost 模型（图 7）。

表 2　CatBoost、LightGBM、XGBoost 模型的测试结果

违约样本				正常样本			合计		
模型名称	判断正确个数	判断错误个数	Recall	判断正确个数	判断错误个数	TNR	判断正常个数	判断错误个数	总体正确率
CatBoost	23	7	76.67%	14570	3454	80.84%	14593	3461	80.83%
LightGBM	24	6	80.00%	13595	4429	75.43%	13619	4435	75.43%
XGBoost	23	7	76.67%	13563	4461	75.25%	13586	4468	75.25%

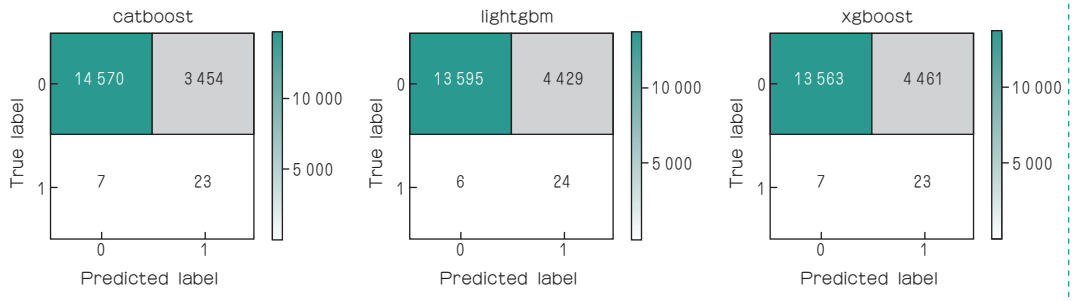


图 6　CatBoost、LightGBM、XGBoost 模型混淆矩阵

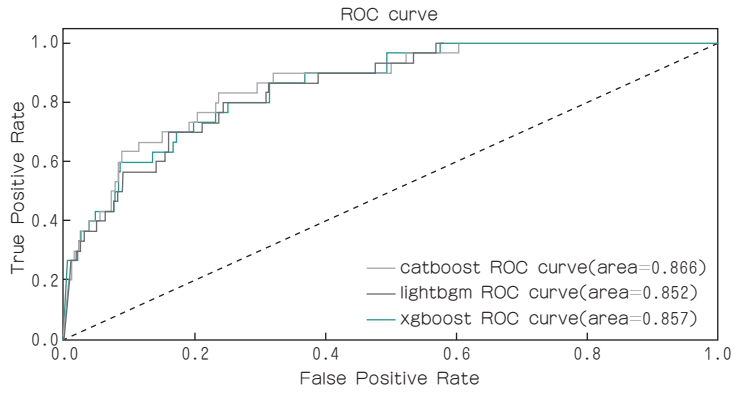


图 7　CatBoost、LightGBM、XGBoost 模型 ROC 曲线

KS 曲线。经计算，Catboost、LightGBM、XGBoost 算法的 KS 值分别为 0.57、0.50、0.53，都处于非常好的范围，表明这三个模型具有非常好的区分能力（图 8）。

Lift 曲线。Lift 曲线表明使用这三种方法的效果比不使用模型的效果提升了 5~6 倍（图 9）。

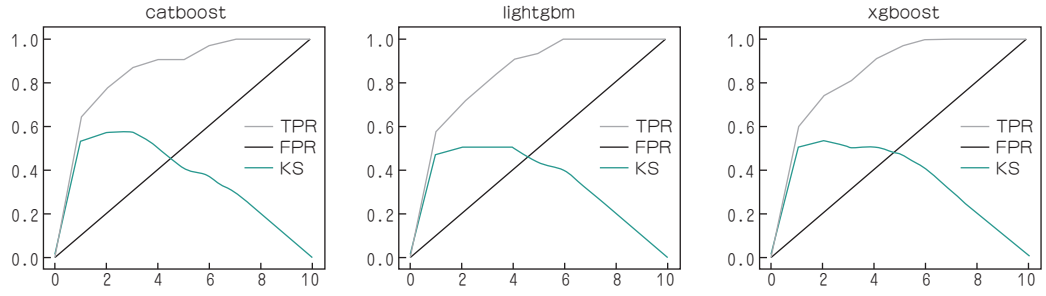


图 8　CatBoost、LightGBM、XGBoost 模型 KS 曲线

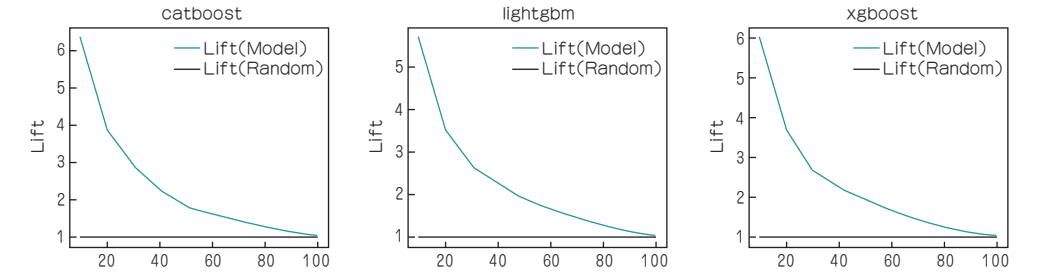


图 9　CatBoost、LightGBM、XGBoost 模型 Lift 曲线

2. 债券发行人财务信息具有违约预警价值，距离违约年份越近的样本数据，其违约预警的准确性越高

以 CatBoost 模型为例，对 T–1 时刻与 T–2 时刻模型的预警结果对比（表 3）可见，T–1 时刻的模型预警准确性均高于 T–2 时刻。即，距离违约年份越近的样本数据，其违约预警的准确性越高。

表 3　T–1 与 T–2 时刻的模型预警结果对比

模型使用的样本号	T–1		模型使用的样本号	T–2	
	Recall	TNR		Recall	TNR
样本号 1	66.67%	90.95%	样本号 4	63.33%	89.10%
样本号 2	76.67%	80.84%	样本号 5	73.33%	75.31%
样本号 3	83.33%	67.79%	样本号 6	83.33%	63.46%

3. 在同一时刻的模型中，不同财务指标对于违约预警的贡献度存在差别

以 CatBoot 为例，在 T–1 时刻模型（样本号 2）与 T–2 时刻模型（样本号 5）中，不同财务指标对于违约预警的贡献度存在差别。如附表 5 所示。

不同财务指标对于违约预警的贡献度有所差别。在 T–1 时刻和 T–2 时刻，重要性都居于前 10 位的指标共有 6 项（表 4）。即，这 6 项指标在短期和长期的违约预警方面都具有重要性，应予格外关注。

企业规模、总资产和存货周转效率、企业管理费用控制水平、融资费用与收入结构比例、现金水平增长率是始终影响债

务违约的重要性指标。

以 T–1 时刻为例，对于违约预警重要性排名前 10 位的指标合计贡献度为 56.34%，属于识别违约时应关注的重要指标（表 5）。

以 T–2 时刻为例，对于违约预警重要性排名前 10 位的指标合计贡献度为 57.21%（表 6）。具有更为长期性的预警指标，属于识别违约时应关注的重要指标。

4. 在两个时刻之间，各个指标的违约预警贡献度会发生变化，即各指标存在长期灵敏度和短期灵敏度之间存在差异

比较指标在 T–1（样本号 2）和 T–2（样本号 5）时刻的重要性变化，将短期灵敏指标定义为：在 T–1 时刻贡献度在 3% 以上，且比 T–2 时刻高 1 倍以上；将长期有效指标定义为：在 T–2 时刻贡献度在 3% 以上，且比 T–1 时刻贡献度高出 40% 以上（表 7~8）。

表 4　T–1 时刻和 T–2 时刻均重要的指标

指标类别	指标名	catboost_T–1	catboost_T–2
盈利能力	财务费用 / 营业总收入	11.14%	15.35%
现金流量	货币资金（同比增长率）	8.63%	5.22%
营运能力	存货周转率	4.67%	7.17%
发展能力	企业规模	4.52%	5.13%
盈利能力	管理费用 / 营业总收入	4.47%	5.56%
营运能力	总资产周转率	3.40%	4.86%

表 5　T–1 时刻（样本号 2）模型贡献度排名前十的指标

指标类别	指标名	catboost_T–1
盈利能力	财务费用 / 营业总收入	11.14%
现金流量	货币资金（同比增长率）	8.63%
盈利能力	净利润 / 营业总收入	5.66%
	开庭公告数量	5.65%
营运能力	存货周转率	4.67%
发展能力	企业规模	4.52%
盈利能力	管理费用 / 营业总收入	4.47%
现金流量	现金比率	4.15%
发展能力	营业收入（同比增长率）	3.94%
营运能力	流动资产周转率	3.50%

表 6　T–2 时刻（样本号 5）模型贡献度排名前十的指标

指标类别	指标名	catboost_T–2
盈利能力	财务费用 / 营业总收入	15.35%
现金流量	货币资金（同比增长率）	5.22%
营运能力	存货周转率	7.17%
发展能力	企业规模	5.13%
盈利能力	管理费用 / 营业总收入	5.56%
营运能力	总资产周转率	4.86%
现金流量	盈余现金保障倍数	3.49%
偿债能力	有形资产净值 / 总资产	3.12%
营运能力	营运资金周转率	4.28%
偿债能力	有形资产 / 带息负债	3.03%

表 7　短期灵敏指标

短期灵敏指标	T–1 时刻	T–2 时刻	指标类型
现金比率	4.15%	1.54%	现金流量
营业收入（同比增长率）	3.94%	1.10%	发展能力
流动负债权益比率	3.25%	1.36%	资本结构
净利润 / 营业总收入	5.66%	1.80%	盈利能力
开庭公告数量	5.65%	2.77%	—

表 8　长期有效指标

长期有效指标	T–1 时刻	T–2 时刻	指标类型
存货周转率	4.67%	7.17%	营运能力
总资产周转率	3.40%	4.86%	营运能力
营运资金周转率	1.81%	4.28%	营运能力

表 9　违约前的开庭公告数量信息的违约预警贡献度

指标名	catboost_T–1	catboost_T–2	T–1 贡献度 / T–2 贡献度
开庭公告数量	5.65%	2.77%	49.08%

表 10　Catboost 对于 T-1 时刻基于不同训练集构建模型的测试结果对比

样本号	测试结果								
	违约样本（全部样本）			正常样本（全部样本）			合计		
	判断正确个数	判断错误个数	Recall	判断正确个数	判断错误个数	TNR	判断正确个数	判断错误个数	总体正确率
样本号 1	130	10	92.86%	19 227	1 337	93.50%	19 357	1 347	93.49%
样本号 2	133	7	95.00%	16 826	3 738	81.82%	16 959	3 745	81.91%
样本号 3	135	5	96.43%	14 198	6 366	69.04%	14 333	6 371	69.23%

5. 债券发行人违约前的开庭公告信息具有一定的违约预警价值

从模型分析结果可见，债券发行人违约前的开庭公告数量具有一定的违约预警价值，于短期内判断违约更为灵敏（表 9）。

6. 分级预警模型

如本文“数据处理说明”所述，随着对于噪声数据处理标准的逐渐严格，对被选入训练集中的正常样本的要求越来越高，由此构建出的对应模型对于全样本测试集中违约的敏感度越高，即判断出的违约样本的数量越多（表 10）。

因此，对应样本号 1 到样本号 3 可以建立分级预警模型。即，以样本号 3 为基础的 Catboost 模型能够更早地识别具有违约风险的主体，但因其模型训练基础采用了较为完美的正常样本，也导致对正常样本的更高误判；相反，以样本号 1 为基础的 Catboost 模型对违约样本的识别准确程度更高（表 11）。

（三）样本外检验

从 Wind 获取 2021 年 1 月 1 日—2023 年 5 月符合本文违约定义的发债主体及其财务数据，使用以样本号 2 为基础的 Catboost 模型，以发债主体 T–1 时刻和 T–2 时刻的财务数据为基础判断其违约状态。将模型识别的违约判断结果与实际情况对比，以验证模型的预测有效性。

1. 使用 T–1 时刻的数据进行判断

于本文样本外，满足本文违约定义且 T–1 时刻财务数据披露完整的发债主体共有 92 户，模型将其中 87 户判别为违约，预测准确率约为 95%，模型具有较好的违约识别能力（表 12）。

2. 使用 T-2 时刻的数据进行判断

于本文样本外，满足本文违约定义且

表 11　预警级别

样本号	预警级别
样本号 1	高度预警
样本号 2	中度预警
样本号 3	初级预警

表 12　使用 T–1 时刻的数据对违约进行判断

违约类型	事件实际数量	模型判断违约数量
触发交叉违约	7	7
技术性违约	1	0
提前到期未兑付	6	6
未按时兑付本金	1	1
未按时兑付本息	14	14
未按时兑付回售款和利息	7	7
未按时兑付利息	5	5
展期	51	47
合计	92	87

表 13　使用 T–2 时刻的数据对违约进行判断

违约类型	事件实际数量	模型判断违约数量
触发交叉违约	7	7
技术性违约	1	0
提前到期未兑付	6	6
未按时兑付本金	1	1
未按时兑付本息	16	13
未按时兑付回售款和利息	5	5
未按时兑付利息	7	7
展期	52	39
合计	95	78

T–2 时刻财务数据披露完整的发债主体共有 95 户，模型将其中 78 户判别为违约，预测准确率约为 82%，模型违约识别能力良好（表 13）。

四、研究结论与建议

（一）研究结论

本文以 2014 年我国信用债市场出现首次违约至 2022 年末全部违约债券及匹配的未违约债券发行人为研究样本，并对于违约样本采用过采样技术，同时纳入债券发行人违约前的开庭公告数量信息，建立适用于我国信用债实际情况的梯度提升算法系列中的 CatBoost、XGBoost、LightGBM 模型。对于信用债违约识别和预警中的各财务指标贡献度进行量化分析，增强了模型的可解释性，一定程度上克服了违约样本不足和大多数机器学习模型黑盒方法的缺陷。本文所构建模型具有良好的违约预测能力。有助于债券市场投资者和监管者有针对性地获取和应用财务信息，加强特质性的风险信息披露，促进我国信用债市场的健康发展。本文的主要研究结论如下。

第一，债券发行人财务信息具有违约预警价值。距离违约年份越近的样本数据，其违约预警的准确性越高。

第二，在同一时刻的模型中，不同财务指标对于违约预警的贡献度存在差别。重要性排名在前 10 位的指标分属于 4 类指标：盈利能力、发展能力、营运能力、现金流量类指标。在 T–1 时刻和 T–2 时刻，重要性都居于前 10 位的指标共有 6 项（表 4），这 6 项指标在短期和长期的违约预警方面都具有重要性，应予格外关注。

第三，各指标存在长期有效和短期灵敏之间的差异。以基于 CatBoost 梯度提升算法的结果为例，具体如表 14 所示。

表 14　短期灵敏与长期有效指标

指标灵敏度	指标类别	指标名称
短期灵敏指标	现金流量	现金比率
	发展能力	营业收入（同比增长率）
	资本结构	流动负债权益比率
	盈利能力	净利润 / 营业总收入
	—	开庭公告数量
长期有效指标	营运能力	存货周转率
	营运能力	总资产周转率
	营运能力	营运资金周转率

第四，CatBoost、XGBoost、LightGB 模型均显示债券发行人违约前的诉讼数量信息具有一定的违约预警价值，且属于短期灵敏性指标。

第五，前述模型应用于我国信用债违约预警环境时的模型表现有所不同。对违约样本的识别方面，LightGBM（正确率 80%）略优于 CatBoost（正确率 76.67%）、XGBoost 模型（正确率 76.67%）。但在对正常样本的识别方面，CatBoost（正确率 80.84%）则优于 LightGBM（正确率 75.43%）、XGBoost 模型（正确率 76.25%）。总体正确率方面，CatBoost、LightGBM、XGBoost 模型分别 80.83%、75.43%、75.25%。总体而言，CatBoost 模型表现略优于 LightGBM、XGBoost 模型。

（二）建议

首先，对比我国现行的债券信息披露要求，建议补充部分对违约预警作用较好的财务指标和非财务指标。我国对于信用债的信息披露给予了高度的重视并颁布了多项监管法规，要求强化风险揭示，旨在提高信用类债券信息的披露质量。例如，2020 年 12 月 25 日，人民银行、发展改革委、证监会联合发布了《公司信用类债券信息披露管理办法》，自 2021 年 5 月 1 日施行。该管理办法对于募集说明书及定期报告中要求披露的信息进行了规范。该管理办法中，有 67 处提到“财务”、41 处提到“会计”，足见对财务会计信息的重视。此外，该管理办法还要求披露风险信息、发行条款、募集资金用途、企业基本情况、企业信用、担保等一系列非财务信息。建议对比我国现行的债券信息披露要求，补充部分对违约预警作用较好的财务指标和非财务指标。例如，强化对违约预警短期灵敏指标和长期有效指标的披露要求。对于个别难以通过公开信息渠道获取或进行验证的指标，建议在债券信息披露要求中进行适当调整和补充。

其次，债券发行人违约前的开庭公告信息具有一定的违约预警价值。建议在债券发行人信息披露中，应补充要求披露类似信息。

再次，由于距离违约时间越近的样本数据，其违约预警的准确性越高。因此，建议监管机构结合对信用债发行人的预警评级结果，对预警级别较高的信用债发行人要求更高的信息披露频率，例如，按季度或按月进行披露。

最后，通过数据处理技术，机器学习

模型能在一定程度上克服我国以往债券违约数据不足而对模型搭建产生的限制，可能在提前违约预警识别能力方面对传统模型进行有所补充。因此，建议债券投资者在对信用债违约风险分析中，积极探索和完善机器学习模型工具的使用，同时结合传统模型工具，利用发行人披露的信息，对债券违约风险进行评估，减少非理性行为。N

学术编辑：卢超群

参考文献

[1] Barboza F,Kimura H,Altman E.Machine learning Models and Bankruptcy Prediction[J].Expert Systems with Applications,2017,83:405-417.

[2] Farquad M A H,Bose I.Preprocessing Unbalanced Data Using Support Vector Machine[J].Decision Support Systems,2012,53(1):226-233.

[3] Hamori S,Kawai M,Kume T,Murakami Y,Watanabe C.Ensemble Learning or Deep Learning? Application to Default Risk Analysis[J].Journal of Risk and Financial Management,2018,11(1):12.

[4] Li Li.The Analysis of Repayment of Default Bonds: Evidence from China[J].Journal of Applied Finance & Banking,2020,10(2):101-118.

[5] Vicente García,Ana I Marqués,J Salvador Sánchez.Exploring the Synergetic Effects of Sample Types on the Performance of Ensembles for Credit Risk and Corporate Bankruptcy Prediction[J].Information Fusion,2019,47:88-101.

[6] 陈丹彤,陈志明.基于KMV模型的债券违约风险度量[J].市场周刊,2019(03):116-118.

[7] 付世豪.基于Logit回归的公司违约概率预测[J].金融经济,2019(02):108-109.

[8] 吉林省中小企业融资问题研究课题组.企业信用评级在商业保理中的应用研究——基于KMV-BP模型[J].经济视角,2017(02):60-67.

[9] 刘玎琳,赵湘莲,田月红.基于KMV模型的农业上市公司信用风险测度研究[J].数学的实践与认识,2014(12):32-39.

[10] 张奇,胡蓝艺,王珏.基于Logit与SVM的银行业信用风险预警模型研究[J].系统工程理论与实践,2015(07):1784-1790.

The Link Between Credit Bond Issuer Disclosure and Bond Defaults

LI Jie　MENG Xiangjun

(ShineWing Accounting Firm)

Abstract　This paper takes a full sample of China's credit bond market as its basis and applies the oversampling technique to overcome limitations caused by unbalanced default samples. It then builds emerging gradient boosting models (CatBoost, XGBoost and LightGBM) that are suitable for China's actual situation. A comparative study was conducted on the usability and prediction performance of the aforementioned models in China's credit bond market. The paper provides a reference for judging the identification and differentiation ability of the models and the future focus of relevant research. It also establishes a hierarchical early warning model applicable to the actual situation of credit bonds in China as a reference for credit bond investors and regulators to effectively predict enterprise bond default risk.

Keywords　Credit Bond, Information Disclosure, Machine Learning, Default Warning

JEL Classification　D53　E17　M41