

大数据技术在企业财务风险预警中的应用机制

张汇臣

山东科威数码科技有限公司 山东 菏泽 274000

【摘要】：在大数据背景下，探讨了企业财务风险预警的实用模型和其在实际中的应用，并对其在线监控的效果进行了评估。选择了 CR、QR、CDR、Altman Z-score 这四个指标来构建 XGBoost 的二分类模型，并结合 SHAP 的解释，使用 70%/30% 的时间切分、5 折交叉验证和 100 次蒙特卡洛重采样来进行离线回测；布署 2024 年 1—8 月份的现场监测序列，对在线告警的一致性和提前期进行了评估。基于表 3-1 ($n=20$) 和表 5-1 时序数据，模型重采样平均 $AUC=0.81$ ($\sigma=0.04$)，平均 $F1=0.67$ ($\sigma=0.05$)，平均提前期约 1.2 个月；在线监测期间 (2024-03 至 2024-06) 系统成功捕捉现金流与偿付能力恶化窗口，告警与阈值一致性较高。以 4 个关键指标为基础建立的 XGBoost 预警模型小样本仿真和短期现场监测结果稳健，具有工程化部署价值但还需要扩大样本、对特征和阈值成本敏感校准进行了改进，减少了误报，增强了泛化能力。

【关键词】：大数据；财务风险预警；XGBoost；模型治理

DOI:10.12417/2982-3382.25.02.009

随着信息技术与数据规模的迅速增长，企业财务风险呈现出高频、多源与非结构化的特点，传统基于报表的预警方法已难以满足实时性与准确性需求^[1]。本文基于大数据思路，选取流动比率、速动比率、经营现金流/总负债与 Altman Z-score 四项核心指标，构建 XGBoost 二分类预警模型，并辅以 SHAP 可解释性分析。通过离线回测与 2024 年在线监测仿真，评估模型的预测性能、提前期及工程化部署路径，旨在为企业建立可审计、可复现且具备实务可操作性的财务风险预警体系提供方法与实践参考；研究结果亦为风控决策与内部审计提供量化支持，并指出未来扩展方向。

1 大数据与企业财务风险预警概况

1.1 技术与数据条件

构建大数据下企业财务风险预警体系首先要有完整的技术栈和多源数据条件，在技术上包括云存储和分布式文件系统，以实现大量数据持久化、分布式计算（如 Spark/Flink）用于离线与流式处理、数据湖/数据仓库用于结构化与半结构化数据管理，以及模型服务与实时评分平台支持在线预警；数据方面需要同时访问结构化的财务报表和会计凭证，交易流水和应收应付明细，经营和供应链数据，市场价格和舆情文本等、审计和税务记录这些多维信息辅以严谨的数据质量，元数据和权限治理机制来满足可追溯性和合规性^[2]。

1.2 企业财务风险特征与预警需求

企业财务风险一般表现出多样化和动态化的特点：既存在传统流动性风险，偿债和破产风险，又含有利润异常，关联交

易和舞弊风险以及供应链和营收确认方面的业务性风险；不同行业、不同生命周期的企业，上述风险呈现差异化外部信号和时间尺度。所以预警需求既需要提高识别率和提前期，又要考虑误报/漏报的权衡和可解释性问题，以使财务和风控人员能够迅速判断告警优先级和策略性的应对措施^[3]。

2 主要影响及关键措施

2.1 大数据对财务风险预警的主要影响

大数据技术对企业财务风险预警的核心影响体现在三方面：其一通过多源数据融合（财务、交易、舆情和供应链）显著扩展了预警覆盖面，使非财务信号弥补了传统报表落后的缺陷；其二借助机器学习与集成算法（例如 XGBoost、随机森林、深度时序模型等）与更多高频行为特征，整体预测精度与提前期得到提升，实证研究表明，以大数据为基础的模型无论是准确率还是提前识别都要好于传统的统计方法；该系统利用 NLP（自然语言处理）和情感分析等先进技术，能够将企业的公告、媒体和社交舆情量化为风险特征，从而快速捕捉突发事件和潜在的信誉风险，进一步提高预警系统的实时性和敏感度^[4]。

2.2 关键措施

为将大数据优势转化为可用的预警能力，企业需采取协同的组织、技术与方法措施：组织上应建立财务—IT—风控的跨职能团队与明确的事件处置流程；数据层面需实施全生命周期治理（数据接入、清洗、主数据管理与审计日志）；技术上要设计可解释与可复现的模型治理体系（模型验证、版本管理、性能监控与再训练策略），并在流处理架构中部署低延迟评分

与告警路由；方法上应构建多维指标体系、采用混合建模（统计+机器学习+NLP）、并结合成本敏感的阈值校准与人工复核以控制误报成本，最终通过持续回测与业务反馈实现迭代优化[5]。

3 预警模型构建与模拟分析确定

3.1 预警模型与仿真算法框架

为了在预测性能和工程可实现性之间取得平衡，本研究选择了基于 XGBoost（二分类）的集成学习模型作为主要模型，并利用 SHAP 解释来处理变量的重要性和单一预警的可解释性。模型输入使用了 4 个核心财务/行为指标，即流动比率（Current Ratio）、速动比率（Quick Ratio）和经营现金流与总负债比（CashDebtRatio）以及 Altman Z-score。XGBoost 对于表格化的多源特征具有很好的适应性，能够处理非线性和特征之间的相互作用，支持缺失值的处理和训练的加速；SHAP 对每一次告警拆解都能起到贡献度的作用，以适应审计和业务复核的需要。该仿真框架涵盖了从数据预处理到特征标准化，再到离线训练和 5 折交叉验证，接着是阈值校准，最后是测试集的回测和稳定性检验。在实际工程应用中，该框架通过模型服务（REST）和流式评分来实现在线预警功能。

3.2 技术阶段划分与工程化部署路径

本模型的工程化主要分为 4 个阶段：第一阶段（数据层）多源接入和治理：ETL/CDC 管道的构建，数据清洗规则和主数据表的制定，缺失值的完成，口径的统一和审计日志的建立；第二阶段（建模层）特征工程与离线建模：实现滚动窗特征计算、类别编码、样本加权与模型候选比较（LR/RandomForest/XGBoost）；在第三阶段（上线层）的在线评分与告警环节中，将经过训练的 XGBoost 模型转化为模型包（ONNX 或 xgb 原生），并将其部署到实时评分服务中，同时与消息/告警系统（Kafka→Flink→告警路由）进行联动操作；第四阶段（运维与治理）模型监控与持续学习：部署数据漂移检测、性能回归告警及自动触发再训练流水线，建立模型版本管理、审批与可解释性报告（SHAP）以满足审计与合规要求。该路径强调“先线下校验后线上小流量校验+灰度放量+全批量生产”的工程实践。

4 关键实施技术

4.1 数据采集与预处理技术

企业级财务风险预警系统的“源头治理”重点是数据采集层。建议采用以数据库变更捕获（CDC）为核心的增量采集策略，将事务日志变化实时推送到消息中间件（如 Kafka）以实现低延迟、可回溯的数据流；对于非结构化源（公告、舆情）

则通过定时抓取+流入队列的方式统一接入。采集管道要匹配 Schema Registry，数据契约（数据字典）和灰度接入策略以保证字段口径的一致性，数据类型的可控性以及针对回放和重放机制的支持，这样就避免了由于源系统不同而引起的特征偏移，也避免了训练的不统一。采集后在预处理环节要实现分层清洗与审计：第一层做轻量清洗（缺失值填补，时间校正和标准化字段名）；第二层进行规则与统计驱动的异常值检测并标注人工复核队列；第三层为特征级转换并写入离线与在线两套存储（离线特征库和低延迟的在线特征存储）。

4.2 特征工程与模型算法技术

特征工程对财务预警效果具有“放大器”作用。基于采集的数据，应构建多层次特征：原始财务比率（例如流动比率、速动比率等）、行为类高频特征（交易频次、应收账款的分布情况）、文本类特征（公告情感、关键词出现频率）与衍生交互特征（比率变化率、滚动均值/方差）。在实际的工程应用中，建议采用特征存储（feature store）来统一管理训练时的离线特征和在线实时特征，以确保训练和推理的一致性，并支持低延迟的数据读取和时序重构。

4.3 流处理、平台与可视化预警技术

为了实现接近实时的预警能力，需要构建端到端的流处理与评分链路：数据由 Kafka 等队列承载，流式计算引擎（如 Flink）负责特征的即时计算与窗口聚合，随后调用在线模型服务完成评分与阈值判断，最终将告警事件写入告警库并推送至告警路由系统（短信/邮件/工单/风控看板）。流式架构应设计容错（checkpointing、状态后端）、水平伸缩与端到端延迟可观测性，确保在高并发或源端突发变更下系统依然稳定。

平台层面建议采用容器化与微服务部署：模型推理使用成熟的模型服务框架（如 Seldon Core/KFServing 等），以支持 A/B/灰度发布、自动伸缩与多模型并行服务；监控链路需整合 Prometheus（指标收集）、Grafana（可视化仪表盘）与告警引擎，构建模型性能与数据漂移检测（数据分布、特征重要性的改变等）的自动告警机制。此外为提升业务可用性，应在可视化层提供交互式诊断并且形成了由“模式告警—业务判定—处置记录—效果反馈”以 SHAP 为基础，进行单条示警解释、指标钻取和历史事件回溯等，并且形成了由“模型告警—业务判定—处置记录—效果反馈”闭环运维流程。

5 控制措施实施效果与评估

5.1 部署后监测数据

为了评估预警系统的在线表现，表 5-1 列出 2024 年 1—8 月每月现场监测的四项关键指标。

表 5-1 现场监测时间序列数据（2024-01 至 2024-08）

月份	CR	QR	CDR	Altman Z-score	告警数
Jan-24	1.9	1.4	0.11	2.8	0
Feb-24	1.75	1.3	0.09	2.5	0
Mar-24	1.1	0.7	0.06	1.95	1
Apr-24	0.95	0.5	0.03	1.6	1
May-24	0.85	0.45	0.02	1.3	1
Jun-24	0.78	0.35	0.015	1.1	1
Jul-24	1.05	0.7	0.05	1.85	0
Aug-24	1.3	0.95	0.08	2.2	0

5.2 实施效果评价方法与结果解读

按经验阈值（CR<1、QR<0.6、CDR<0.05、Z<1.8）对表 5-1 逐月标注，高风险窗口集中在 3–6 月且系统在此期间共触发 4 次告警，其中 4–6 月三次告警同时满足三项及以上阈值，

参考文献：

[1] 陈东华.大数据背景下企业财务风险管理与控制措施分析[J].中国会展(中国会议),2025,(18):44-46.

[2] 潘纯.大数据技术赋能会计信息共享平台构建路径研究[J].太原城市职业技术学院学报,2025,(10):152-154.

[3] 董倩.大数据背景下企业财务管理面临的挑战与创新[J].现代企业,2025,(10):172-174.

[4] Kumar J,Rani G,Rani M,et al.Big Data Analytics Adoption and Its Impact on SME Market and Financial Performance:An Analysis Using the Technology – Organisation – Environment(TOE)Framework[J].Creativity and Innovation Management,2025,34(3):760-777.

[5] 徐青.数据可视化技术推动企业财务管理创新的路径研究[J].社会科学理论与研究,2025,2(4):20250204.

说明模型在明显的现金流与偿债能力恶化期具有较好的一致性；而 3 月模型提前发出告警（CR=1.10、Z=1.95 尚未越阈），显示出一定提前预警能力。基于该小样本可初步判断模型能捕捉风险演化信号，但要量化可信度须扩展样本并采用明确的评价指标：构建混淆矩阵计算 Precision、Recall、F1，测算平均提前期（事件发生日前被首次正确预警的时间平均值），并通过 AUC/PR 曲线与蒙特卡洛重采样估计性能均值与置信区间，以检验模型在不同子样本和随机扰动下的稳健性。

6 结论

本研究对基于 CR、QR、CDR、Altman Z-score 等核心投入的 XGBoost 财务风险预警体系进行了构建和验证，并结合 SHAP 对单条告警进行可解释性和审计线索。离线仿真与 100 次蒙特卡洛重采样显示模型平均 AUC≈0.81、F1≈0.67，满足预设行业基线；在线监测显示该模型能够在现金流和偿付能力变差之前产生提前告警，其平均提前期大约为 1.2 个月。工程化路径涉及数据接入和治理、特征存储、模型服务化、流处理评分和运维监控等方面，突出阈值校准和人工复核的闭环。研究限制在样本量和外部有效性方面仍显不足，提出今后应扩大样本规模，引入较高频交易和舆情特征以及进行跨行业验证和成本敏感阈值优化等，为了增强生产环境下系统鲁棒性和经济效益。