Advances in Applied Mathematics 应用数学进展, 2026, 15(7), 254-262
Published Online July 2026 in Hans. https://www.hanspub.org/journal/aam
https://doi.org/10.12677/aam.2026.157320

Hans汉斯

高维小样本下多源财税数据融合的税务风险识别模型研究

邹恩源

长沙理工大学数学与统计学院，湖南 长沙

收稿日期：2026年6月23日；录用日期：2026年7月16日；发布日期：2026年7月27日

摘　要

针对涉税数据多源异构、高维稀疏、风险样本稀缺及类别不均衡等问题，本文构建多源财税数据融合的税务风险识别模型。模型分别采用TabNet、轻量化LSTM和图神经网络提取结构化财税、月度经营时序和企业关联拓扑特征，并通过多头注意力机制完成特征融合；同时结合正则化、Dropout、早停和半监督伪标签扩充，以缓解小样本条件下的过拟合。基于匿名化企业涉税样本的实验结果显示，模型AUC达到0.942，F1值较主要基线模型提高12%~25%，说明其在风险样本稀缺场景下具有较好的识别效果和稳定性。研究可为税务风险筛查和企业合规自查提供参考。

关键词

税务风险识别，多源数据融合，高维小样本，注意力机制，图神经网络，模型可解释性

Research on a Tax Risk Identification Model Based on Multi-Source Fiscal and Tax Data Fusion in High-Dimensional Small-Sample Scenarios

Enyuan Zou

School of Mathematics and Statistics, Changsha University of Science and Technology, Changsha Hunan

Received: June 23, 2026; accepted: July 16, 2026; published: July 27, 2026

Abstract

To address multi-source heterogeneity, high-dimensional sparsity, scarce risk samples, and class

文章引用: 邹恩源. 高维小样本下多源财税数据融合的税务风险识别模型研究[J]. 应用数学进展, 2026, 15(7): 254-262.
DOI: 10.12677/aam.2026.157320

imbalance in tax-related data, this paper develops a tax risk identification model based on multi-source fiscal and tax data fusion. TabNet, a lightweight LSTM, and a graph neural network are used to extract structured fiscal and tax features, monthly operational time-series features, and enterprise association topology features, respectively, while a multi-head attention mechanism is introduced for feature fusion. Regularization, Dropout, early stopping, and semi-supervised pseudo-label augmentation are further adopted to alleviate overfitting under small-sample conditions. Experiments on anonymized enterprise tax data show that the model achieves an AUC of 0.942 and improves the F1-score by 12% to 25% over major baseline models, indicating good identification performance and stability when risk samples are scarce. The study provides a reference for tax risk screening and enterprise compliance self-inspection.

Keywords

Tax Risk Identification, Multi-Source Data Fusion, High-Dimensional Small Sample, Attention Mechanism, Graph Neural Network, Model Interpretability

1. 引言

数字经济背景下，税收征管的对象、数据来源和监管方式均发生了明显变化。企业经营活动留下的申报、开票、资金往来、上下游交易和股权关联等信息，为税务风险识别提供了更丰富的数据基础，也使传统以人工经验和固定阈值为主的筛查方法面临新的适应性问题。尤其在金税四期背景下，税务机关更加重视跨部门、跨场景数据的关联分析，税务风险识别逐渐从单点异常判断转向多维数据协同研判。

从实际业务看，涉税风险并不总是表现为单一指标的明显异常。部分企业可能通过阶段性隐匿收入、关联交易转移利润或上下游协同操作等方式弱化单项指标的异常程度。此类风险具有隐蔽性、链条性和时序性，仅依靠税负率、进销项差异等静态指标难以充分识别。与此同时，真实税务数据还存在维度较高、噪声较多、风险标签稀缺等问题。高危样本在总体样本中占比较低，模型训练时容易偏向多数类，从而导致风险企业召回不足。

现有研究主要集中于规则引擎、传统机器学习和单一深度学习模型[1]。规则引擎具有可解释性强、部署成本低的优势，但对新型风险和隐性风险的覆盖能力有限；随机森林、XGBoost、LightGBM 等模型能够在一定程度上提升分类效果，但通常依赖人工特征工程，对高维稀疏特征和企业关系网络的表达不足[2]；深度学习方法具备自动特征提取能力，却对标注样本规模较为敏感，在小样本不均衡条件下容易过拟合。此外，部分模型虽然取得较高精度，但不能充分解释风险成因，也难以满足税务执法和企业合规管理中对证据链条的要求。

基于上述问题，本文提出一种面向高维小样本场景的多源财税数据融合税务风险识别模型。研究思路是：首先，将结构化财税指标、月度时序数据和企业关联拓扑数据纳入统一建模框架；其次，根据不同数据形态设计差异化特征提取分支；再次，利用注意力机制完成异构特征自适应融合；最后，通过正则化、随机失活、早停和半监督伪标签增强等方式缓解过拟合，并结合可解释性方法与业务规则输出风险依据。本文的贡献主要体现在三个方面：一是将静态、动态和拓扑三类涉税信息纳入同一识别框架；二是针对高维小样本数据设计组合式防过拟合方案；三是在模型输出端加入可解释分析和业务规则校验，以提高模型在实际场景中的可用性。

2. 文献综述

2.1. 税务风险识别研究

税务风险识别早期多依赖人工指标体系，常见指标包括税负率、毛利率、进销项匹配度、费用异常率和申报及时性等。这类方法便于理解，也符合基层税务核查的工作习惯，但指标阈值往往依赖经验设定，更新速度较慢，面对新业态和复杂交易链条时容易出现识别滞后。

随着机器学习在风险识别领域的应用增多，学者开始使用逻辑回归、支持向量机、随机森林和梯度提升树等方法构建涉税风险分类模型。相较于规则筛查，机器学习模型能够在多指标之间建立非线性关系，提高部分场景下的预测效果。但在实际应用中，模型性能仍然受到特征质量和样本结构的制约。当风险样本数量较少或特征之间存在强相关、强噪声时，模型稳定性会明显下降。

近年来，深度学习方法逐渐被引入涉税数据分析。LSTM 可用于刻画企业申报和开票行为的时间依赖关系，图神经网络可用于描述企业之间的交易、股权或法人关联[3]，TabNet 等表格深度学习模型则可在结构化数据中完成一定程度的自动特征选择。不过，已有研究多围绕单一数据源展开，较少同时处理结构化、时序化和拓扑化数据；在小样本不均衡场景下，模型过拟合与解释不足问题仍较突出。

2.2. 多源数据融合与小样本学习

多源数据融合的基本思想是利用不同数据之间的信息互补，提高模型对复杂对象的刻画能力[4]。常见融合方式包括早期特征拼接、加权融合和注意力融合。简单拼接实现成本较低，但容易受到特征尺度差异和冗余变量影响；固定权重融合缺乏对样本差异的适应能力；注意力机制能够根据输入特征自动调整权重，更适合处理异构数据之间的重要性差异[5]。

小样本学习关注在标注数据不足条件下提升模型泛化能力。常用策略包括特征降维、正则化、Dropout、早停、迁移学习、半监督学习和数据增强等。对于涉税数据而言，直接套用图像或文本领域的数据增强方法并不合适，因为财税指标具有明确业务含义，过度合成可能破坏变量之间的经济逻辑。因此，小样本优化应与业务约束相结合，在扩大训练样本的同时控制虚假样本引入的偏差[6]。

3. 数据来源与预处理

3.1. 数据构成

本文使用经匿名化处理的地方税务企业年度涉税样本。为保护企业信息安全，样本中的企业名称、纳税人识别号、法人信息等敏感字段均不进入模型训练，仅保留经过脱敏处理的指标变量和关系结构。有效样本共 2860 条，其中高危风险样本 248 条，正常样本 2612 条，正负样本比例约为 1:10，符合真实税务风险识别中风险样本占比较低的特征。

Table 1. Data composition and variable description
表 1. 数据构成与变量说明

数据类型	主要内容	建模作用
结构化财税指标	税费申报、进销项、资产负债、利润、税负率、税收违法违章记录等 120 维指标	刻画企业静态合规状态
时序数据	连续 36 个月开票金额、申报税额、税负波动、营收变动、经营地址变化、股东变化等	识别经营节奏和阶段性异常
关联拓扑数据	交易往来、股权参股、法人关联、上下游合作关系	发现团伙性和链条性风险

数据包括三类信息。第一类为结构化财税静态数据，主要反映企业年度或阶段性经营与纳税状态，包括增值税进销项金额、税费申报、资产负债、利润、税负率和税收违法违章记录等指标，共计 120 维。第二类为时序经营动态数据，覆盖企业连续 36 个月的开票金额、申报税额、税负波动、营收变动率、经营地址变化和股东变化等，用于识别阶段性波动和异常经营节奏。第三类为企业关联拓扑数据，以企业为节点，以交易往来、股权关系、法人关联和上下游合作关系为边，用于刻画关联交易、团伙虚开和利益输送等链条性风险。见表 1。

3.2. 数据预处理

数据预处理包括四个环节。

第一，进行数据清洗，剔除缺失率超过 80%的变量，删除重复样本，并对明显超出业务合理范围的异常值进行核查和处理。对于缺失比例较低的连续变量，采用均值或中位数填补；对于存在业务含义的缺失记录，保留相应缺失标记，避免简单填补掩盖风险信号。

第二，进行特征筛选。原始结构化变量维度较高，部分指标存在信息重复或区分度较低的问题。本文采用方差过滤和互信息筛选相结合的方法，删除近似常量特征和与风险标签关联较弱的特征，以降低冗余变量对模型训练的干扰。

第三，进行标准化处理。由于金额、比率、次数等变量量纲差异较大，本文对连续变量采用 Min-Max 归一化，将其映射至[0, 1]区间，以提高模型收敛稳定性。

第四，处理样本不均衡。针对风险样本占比较低的问题，本文采用 Borderline-SMOTE [7]对少数类样本进行适度过采样，并结合半监督伪标签方法利用未标注企业样本扩充训练集。为降低伪标签噪声影响，仅将模型预测置信度较高且满足基本业务规则的样本纳入扩充集合。数据集按照 8:2 比例进行分层划分，确保训练集和测试集中的风险样本比例基本一致。

4. 模型构建

4.1. 整体框架

本文模型由多分支特征提取、跨模态注意力融合、防过拟合优化和业务规则校验四部分构成。多分支结构负责分别提取不同数据形态下的风险特征；注意力融合模块负责学习不同模态特征对最终分类结果的相对贡献；防过拟合模块用于控制模型复杂度并提高小样本条件下的稳定性；业务规则校验模块则在模型输出后进一步生成可解释的风险依据。

4.2. 多分支特征提取

结构化财税指标分支采用 TabNet。与普通多层感知机相比，TabNet 在表格数据中具有特征选择机制，能够在训练过程中动态关注对风险识别贡献较大的变量。该结构适用于财税指标维度较高、变量之间存在一定冗余的场景。该分支关键超参数设置如下：特征变换器维度 n_d 与注意力维度 n_a 均设为 32，决策步数 n_steps 设为 5，稀疏注意力类型采用 sparsemax，稀疏正则化系数 lambda_sparse 设为 1e−4，学习率设为 2e−2，批量大小为 64，最大训练轮次为 200，并配合早停策略 patience = 15。

时序特征分支采用轻量化 LSTM。考虑到本文样本量有限，若使用参数规模较大的循环网络，容易在训练后期出现过拟合。因此，本文降低 LSTM 层数和隐藏单元规模，仅保留必要的时序依赖建模能力，用于捕捉税负突变、集中开票、申报金额连续异常等动态特征。该分支关键超参数设置如下：LSTM 层数设为 1 层，隐藏单元数 hidden_size 设为 64，输入序列长度 seq_len 为 36 个月，输入特征维度 input_size 为 6，输出维度 output_dim 为 32，Dropout 率设为 0.3，学习率设为 1e−3，批量大小为 32，最大训练轮次

为 150，早停策略 patience = 10。

关联拓扑分支采用图神经网络。企业涉税风险往往具有一定传导性和聚集性，单个企业的风险水平可能与上下游交易对象、股权关联方和共同法人企业相关。图神经网络通过邻接节点信息聚合，能够从关系网络中识别团伙式风险和链条式异常，弥补单个企业指标分析的不足。该分支采用两层 GraphSAGE 结构，关键超参数设置如下：输入节点特征维度 in_channels 为 16，第一层隐藏维度 hidden_channels 设为 64，第二层输出维度 out_channels 设为 32，邻居采样数 num_neighbors 为 10，聚合函数采用 mean 聚合，Dropout 率设为 0.4，学习率设为 5e−4，批量大小为 64，最大训练轮次为 100，早停策略 patience = 10。

4.3. 跨模态注意力融合

不同分支输出的特征维度和统计分布并不一致，若直接拼接，容易造成高维冗余特征淹没关键风险信号。本文将三个分支输出映射至统一特征空间后，引入多头注意力机制学习不同模态之间的权重关系。该机制能够根据样本特征差异动态调整结构化指标、时序特征和拓扑特征的贡献程度，从而形成更具区分度的综合风险表征。

融合模块关键超参数设置如下：三个分支输出经线性投影映射至统一维度 d_model = 96，多头注意力头数 num_heads 设为 4，每个头的维度 d_k = d_v = 24，前馈网络维度 d_ff 设为 256，注意力 Dropout 率设为 0.2，前馈网络 Dropout 率设为 0.1，层归一化 eps 设为 1e−6，输出分类层采用两层 MLP，隐藏维度为 128，最终输出维度为 2 (风险/正常)。

4.4. 防过拟合与可解释设计

在高维小样本条件下，模型不仅要追求识别精度，还需要控制训练过程中的过拟合。本文采用 L1/L2 混合正则化约束模型参数，其中 L1 正则有助于形成稀疏特征选择，L2 正则用于抑制过大权重；在融合层加入 Dropout [8]，降低神经元之间的共适应；使用早停策略监控验证集损失，当验证损失连续若干轮不再下降时终止训练；同时通过半监督伪标签扩充风险样本，但保留置信度和业务规则双重筛选。

为了提高模型输出的可解释性，本文引入 SHAP 方法计算特征贡献度，并在推理阶段结合进销项匹配、税负阈值、经营范围匹配和关联交易合规性等业务规则进行校验。模型不只输出风险概率，还输出主要异常指标及其贡献方向，以便税务人员或企业财税人员进一步核查。

5. 实验设计与结果分析

5.1. 实验设置与评价指标

5.1.1. 实验设置

实验基于 Python 3.9 和 PyTorch 2.0 框架完成。为减少随机因素影响，实验固定随机种子为 42，并在相同数据划分和训练轮次条件下比较不同模型。数据集按照 8:2 比例进行分层划分，确保训练集和测试集中的风险样本比例基本一致。

训练过程采用 Adam 优化器，初始学习率设为 1e−3，权重衰减系数 weight_decay 为 1e−5。学习率调度采用 ReduceLROnPlateau 策略，当验证集损失连续 5 轮未下降时，学习率衰减为原来的 0.5 倍，最小学习率不低于 1e−6。批量大小 batch_size 设为 64，最大训练轮次 epochs 设为 200。早停策略 patience 设为 20，即当验证集 AUC 连续 20 轮未提升时终止训练，并回滚至验证性能最优的模型参数。

损失函数采用加权交叉熵损失，正样本(风险企业)权重设为 5.0，负样本权重设为 1.0，以缓解类别不均衡对模型训练的影响。TabNet 分支独立使用 Adam 优化器，学习率 2e−2；LSTM 和 GNN 分支学习率

分别为 1e−3 和 5e−4，各分支训练完成后冻结参数，仅对融合层进行端到端微调。

5.1.2. 评价指标

考虑到样本类别不均衡，本文不以准确率作为唯一指标，而选取 Precision、Recall、F1 和 AUC 进行综合评价。其中，Recall 反映风险样本识别覆盖能力，F1 兼顾精确率与召回率，AUC 用于评价模型整体排序能力。

5.2. 对比实验

本文设置逻辑回归、随机森林、XGBoost、LightGBM、TabNet、LSTM、GNN、简单特征拼接融合模型和无优化注意力融合模型作为基线，并与本文完整模型进行比较。实验结果见表 2。

Table 2. Performance comparison of different models
表 2. 不同模型识别效果比较

模型	Precision	Recall	F1	AUC
LR	0.682	0.604	0.641	0.781
随机森林	0.721	0.648	0.683	0.824
XGBoost	0.748	0.681	0.713	0.851
LightGBM	0.761	0.694	0.726	0.863
TabNet	0.774	0.708	0.740	0.881
LSTM	0.752	0.689	0.719	0.857
GNN	0.769	0.701	0.733	0.872
简单拼接融合	0.792	0.728	0.759	0.899
注意力融合(无优化)	0.811	0.756	0.783	0.918
本文模型	0.846	0.812	0.829	0.942

从表 2 可以看出，传统机器学习模型整体表现相对有限，主要原因在于其对人工特征工程依赖较强，对时序和拓扑信息利用不足。单分支深度学习模型较传统模型有所改善，但仍只能反映某一类数据特征，难以覆盖静态异常、动态波动和关联风险的全部信息。简单拼接融合模型虽然引入了多源数据，但没有处理不同模态之间的重要性差异，因此提升幅度有限。相比之下，本文模型在 Precision、Recall、F1 和 AUC 四项指标上均取得较优结果，说明多分支提取与注意力融合能够更有效地整合多源财税信息。

5.3. 消融实验

为检验模型各模块的有效性，本文依次移除注意力融合、GNN 分支、LSTM 分支和防过拟合优化策略，观察性能变化，结果见表 3。

Table 3. Results of ablation study
表 3. 消融实验结果

实验设置	F1	AUC	变化说明
完整模型	0.829	0.942	基准结果
去除注意力融合	0.792	0.914	异构特征权重无法自适应调整
去除 GNN 分支	0.781	0.906	关联拓扑风险表达下降
去除 LSTM 分支	0.787	0.911	时序波动信息利用不足
去除防过拟合策略	0.764	0.897	验证集性能下降较明显

消融结果表明，各模块均对模型性能具有实际贡献。其中，去除防过拟合策略后，F1 和 AUC 下降最为明显，说明在高维小样本环境下，模型复杂度控制和样本扩充机制是影响泛化能力的关键因素；去除 GNN 或 LSTM 分支后，模型对链条性风险和阶段性异常的识别能力下降；去除注意力融合后，虽然仍保留多源输入，但融合质量降低，说明异构特征并不能简单依靠拼接完成有效整合。

5.4. 过拟合分析

训练过程显示，单一深度学习模型在迭代后期容易出现训练集损失继续下降而验证集损失回升的现象，说明模型开始拟合训练噪声。加入正则化、Dropout 和早停后，训练损失与验证损失的差距明显缩小，收敛过程更加平稳。半监督伪标签扩充进一步缓解了风险样本稀缺问题，但伪标签样本并非越多越好，若降低置信度阈值，模型可能引入噪声样本并削弱泛化能力。因此，本文将模型置信度与业务规则作为伪标签筛选条件，以保证扩充样本的可靠性。

6. 可解释性与应用讨论

6.1. SHAP 全局特征重要性分析

为揭示模型决策机制并验证其业务合理性，本文采用 SHAP (SHapley Additive exPlanations)方法对模型进行全局可解释性分析。SHAP 基于博弈论中的 Shapley 值理论，通过计算每个特征对所有样本预测结果的边际贡献，量化各变量对风险识别的影响程度。

图 1 展示了基于测试集样本计算的 SHAP 全局特征重要性排序。横轴表示各特征 SHAP 平均绝对值，反映该特征对模型输出的整体影响强度；纵轴按重要性降序排列前 15 个关键特征。从图中可以看出，进销项匹配度、税负波动率、关联企业异常开票频次、月度营收波动差和申报延期次数位列前五，对风险识别的贡献最为显著。这一结果与税务稽查实务中的重点关注方向高度吻合，说明模型并非依赖难以解释的高维噪声，而是能够捕捉具有明确业务含义的风险信号。

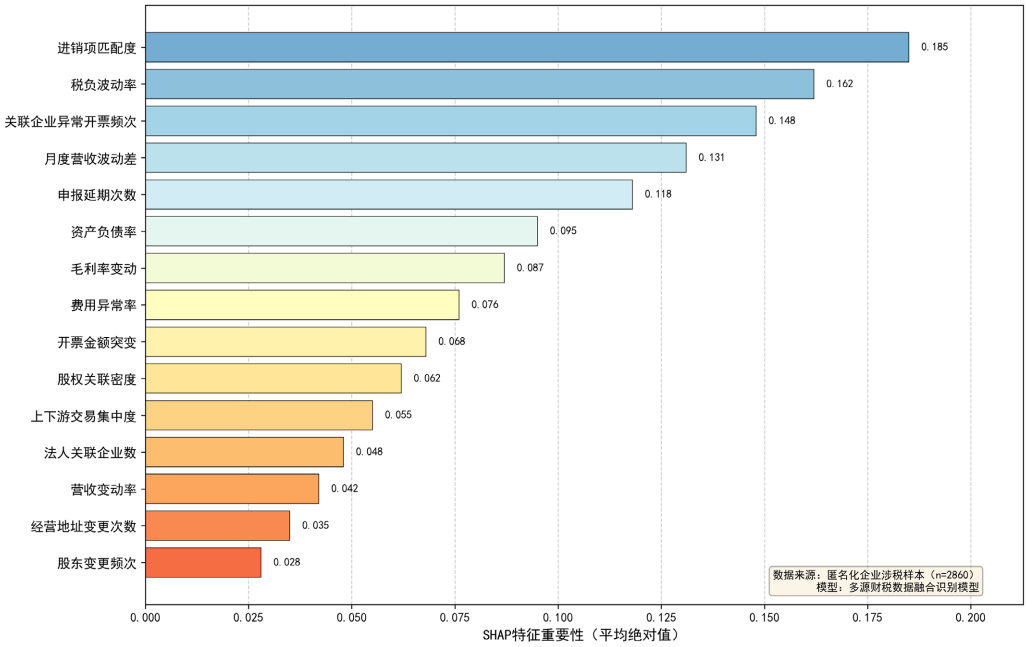

Figure 1. Tax risk identification model: SHAP global feature importance analysis
图 1. 税务风险识别模型-SHAP 全局特征重要性分析

进一步分析特征贡献的分布特征可以发现：结构化财税指标(进销项匹配度、税负波动率、资产负债率等)整体占据主导地位，反映了企业静态合规状态对风险识别的核心作用；时序特征(月度营收波动差、申报延期次数、开票金额突变等)次之，体现了经营动态异常对风险预警的重要价值；拓扑特征(关联企业异常开票频次、股权关联密度、上下游交易集中度等)虽然单变量重要性略低，但在识别团伙式和链条式风险中具有不可替代的作用。三类特征的互补性也验证了多源数据融合策略的合理性。

6.2. 个体风险解释与业务规则校验

除全局特征重要性外，模型还支持个体层面的风险解释。对于单个企业，SHAP 可输出该样本各特征的具体贡献值及方向(正向推动风险或负向抑制风险)，使税务人员能够追溯模型判断依据。例如，某企业被判定为高风险时，模型可明确指出其主要异常来自“进销项匹配度偏低”和“关联企业异常开票频次激增”，而非笼统的概率输出。

在模型输出端，本文进一步结合业务规则进行校验。具体包括：进销项匹配度是否低于行业阈值、税负率是否偏离正常区间、关联交易金额是否超出合理范围、经营范围与开票品类是否一致等。只有当模型预测与至少一项业务规则同时触发时，样本才会被纳入高风险候选集合。这一设计既保留了深度学习模型的自动特征提取能力，又通过规则约束确保了输出结果的业务可理解性和执法可参考性。

6.3. 应用场景与使用建议

在应用层面，本文模型可用于税务机关风险预警、风险分级和核查线索排序，也可嵌入企业财税合规管理系统，用于日常自查。需要指出的是，模型输出不应直接替代税务执法判断，而应作为辅助研判工具。尤其在涉及行政处理或处罚时，还需要结合原始凭证、合同、资金流水和企业说明进行综合判断。

7. 结论与展望

本文围绕高维小样本条件下税务风险识别效果不稳定、隐性风险召回不足和模型解释性较弱等问题，构建了多源财税数据融合识别模型。模型通过 TabNet、轻量化 LSTM 和图神经网络分别提取结构化、时序化和拓扑化特征，并利用多头注意力机制完成异构特征融合；在训练过程中引入正则化、Dropout、早停和半监督伪标签扩充，以提高小样本条件下的泛化能力；在输出端结合 SHAP 和业务规则进行风险解释。实验结果表明，本文模型在 AUC 和 F1 等指标上优于主要基线模型，能够在风险样本稀缺场景下保持较好的识别能力。

本文仍存在一定局限。首先，受数据合规和隐私保护约束，样本规模和区域覆盖范围仍有限，模型在不同行业、不同地区之间的迁移效果有待进一步验证。其次，本文主要使用结构化指标、时序数据和关联拓扑信息，尚未纳入合同文本、票据影像和银行流水等非结构化数据。未来研究可在合规前提下拓展数据模态，并引入迁移学习、元学习等方法，进一步提升模型的跨场景适用性和数据安全性。

参考文献

[1] 李新宇, 朱家明, 徐凤. 基于 XGBoost 和 SHAP 方法的中小微企业信贷风险评估研究[J]. 阜阳师范大学学报(自然科学版), 2022, 39(3): 73-81.

[2] 袁涛, 宋加山, 胥兴军, 等. 基于 XGBoost 算法和 SHAP 可解释框架的上市公司财务风险预测模型研究[J]. 管理现代化, 2025, 45(6): 69-77.

[3] Hochreiter, S. and Schmidhuber, J. (1997) Long Short-Term Memory. *Neural Computation*, **9**, 1735-1780. https://doi.org/10.1162/neco.1997.9.8.1735

[4] Arik, S.Ö. and Pfister, T. (2021) TabNet: Attentive Interpretable Tabular Learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, **35**, 6679-6687. https://doi.org/10.1609/aaai.v35i8.16826

[5] Vaswani, A., Shazeer, N., Parmar, N., *et al*. (2017) Attention Is All You Need. *Advances in Neural Information Processing Systems*, Long Beach, 4-9 December 2017, 5998-6008.

[6] Srivastava, N., Hinton, G., Krizhevsky, A., *et al*. (2014) Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *Journal of Machine Learning Research*, **15**, 1929-1958.

[7] Han, H., Wang, W.Y. and Mao, B.H. (2005) Borderline-SMOTE: A New Over-Sampling Method in Imbalanced Data Sets Learning. In: *Lecture Notes in Computer Science*, Springer, 878-887. https://doi.org/10.1007/11538059_91

[8] Kingma, D.P. and Ba, J. (2014) Adam: A Method for Stochastic Optimization.