面向不平衡信用评估的双流表示学习与原生反事实解释生成网络

李存弘，李远隆，张彬蕾 *

（喀什大学计算机科学与技术学院，新疆 喀什 844000）

摘要：针对信用评估中极端的类别不平衡及特征异构导致的预测瓶颈与“黑盒”合规困境，本文提出一种端到端的双流表示学习与原生反事实解释生成网络。该模型通过双流多层感知机（Dual-stream MLP）实现连续与类别特征的解耦映射与晚期门控融合；引入定制化焦点损失（Focal Loss）动态调整梯度权重，攻克违约样本稀缺引发的准确率悖论；并创新性地在潜在空间嵌入基于真实样本的“原生反事实”正则化损失，确保解释结果 100% 契合金融业务联动约束。在 GMSC 与 UCI 数据集上的实验表明，本方法在 AUC 和少数类 F1-Score 上均显著优于基线模型，且生成的反事实解释在业务逻辑合法率上达到 100% 的真实可行性，为高可靠、可解释的透明信贷风控提供了新范式。

关键词：信用评估；双流多层感知机；不平衡学习；焦点损失；原生反事实解释；可解释人工智能（XAI）

收稿日期：2026 年 3 月 7 日 **中图分类号：**029

通讯作者：张彬蕾，喀什大学计算机科学与技术学院

Dual-stream Representation Learning for Unbalanced Credit Evaluation and Native Counterfactual Explanation Generation Network

Li Cunhong, Li Yuanlong, Zhang Binlei*

(College of Computer Science and Technology, Kashgar University, Kashgar, Xinjiang 844000)

Abstract: In response to the extreme class imbalance and feature heterogeneity in credit assessment, which lead to prediction bottlenecks and the "black box" compliance dilemma, this paper proposes an end-to-end dual-stream representation learning and native counterfactual explanation generation network. This model achieves decoupled mapping and late gating fusion of continuous and categorical features through dual-stream multi-layer perceptron (Dual-stream MLP); introduces customized focal loss to dynamically adjust gradient weights, overcoming the accuracy paradox caused by the scarcity of default samples; and innovatively embeds "native counterfactual" regularization loss based on real samples in the latent space to ensure that the explanation results 100% comply with the financial business linkage constraints. Experiments on GMSC and UCI datasets show that this method significantly outperforms the baseline model in AUC and minority class F1-Score, and the generated counterfactual explanations achieve a 100% real feasibility in the legal rate of business logic, providing a new paradigm for highly reliable and interpretable transparent credit risk control.

Keywords: Credit assessment; Dual-stream multi-layer perceptron; Imbalanced learning; Focal loss; Native counterfactual explanation; Explainable Artificial Intelligence (XAI)

0 引言

信用评估正由传统评分卡转向深度神经网络（DNN），但在工业化落地中面临类别不平衡、特征异构及合规解释三大瓶颈。首先，在 GMSC 等真实数据集中，违约样本占比不足 7%，标准交叉熵损失易导致模型陷入“准确率悖论”，使高风险样本的召回率严重下降。其次，信贷表格数据包含高维稀疏类别特征与稠密连续数值特征，传统 Vanilla MLP 的简单拼接易引发梯度干扰。最后，深度学习的“黑盒”性与金融监管要求的“解释权”存在矛盾。现有合成反事实解释（如 DiCE）往往脱离“业务流形”，生成“降低年龄”等违背金融逻辑的建议，缺乏现实操作性[1]。

为了克服上述挑战，本文构建了一种端到端的深度学习架构。首先，在特征表征层面，设计了双流多层感知机（Dual-stream MLP），通过并行的连续与类别特征流结合晚期门控融合机制，有效隔离了异构数据的分布干扰并增强了非线性交叉特征的提取能力[2]。在目标函数优化上，本文引入定制化焦点损失（Focal Loss），通过加权因子与调制因子 动态调整梯度权重，迫使网络在训练中聚焦于极少数违约样本，从根本上解决了不平衡分布下的“准确率悖论”与漏判难题。

此外，针对解释性的合规需求，本文摒弃了传统的合成扰动路径，创新性地提出原生反事实解释（Native Counterfactuals）正则化机制。该机制通过在潜在表示空间中检索满足业务联动约束的真实样本作为解释目标，并将其转化为结构化正则项以实现数据流形对齐。实验证明，这种联合优化不仅确保了生成建议的 100% 业务真实性，更作为一种强效正则化器反哺提升了模型的泛化性能。

实验表明，该架构在 GMSC 压力测试中的 AUC 及违约召回率均显著优于传统集成树与深度学习基线。此外，生成的反事实建议在业务逻辑合法率上达到 100%，确保了信贷建议在合规监管下的真实可行性。

1 研究方法

本节将详细阐述所提出网络架构的各个核心组件，包括针对缺失值的非线性插补策略、双流 MLP 表格特征表示学习、缓解不平衡的焦点损失计算，以及基于数据集内嵌业务约束的原生反事实损失函数的推导。

1.1 针对金融数据的非线性特征工程与业务约束重构

高质量的数据输入是深度学习模型性能的基石。在真实的信贷数据集中，普遍存在复杂的缺失值模式与内在特征联动机制。以本文使用的 Give Me Some Credit（GMSC）数据集为例，其包含 150,000 条样本，但 MonthlyIncome（月收入）和 NumberOfDependents（受抚养人数）存在高达 20% 的缺失情况[3]。

1.1.1 基于随机森林的非线性缺失值插补（MissForest）

在传统风控建模中，缺失值常被直接删除或采用简单的列均值、中位数进行填补。然而，对 GMSC 数据的深入探索性数据分析（EDA）揭示：当 MonthlyIncome 为缺失值时，DebtRatio（负债率）通常会呈现出数千甚至上万的异常绝对高值。这表明原始数据在记录时存在强烈的非线性结构依赖——当分母（收入）缺失时，系统可能直接记录了分子的绝对债务值。

若采用简单的均值插补，将彻底割裂 DebtRatio 与 MonthlyIncome 之间的原生逻辑关联。为此，本研究采用了基于随机森林算法的无监督插补技术（MissForest）。对于任意缺失的特征列，MissForest 通过将其他所有特征 作为预测变量，在已观测到的数据子集上训练随机森林回归器，随后预测并填补缺失值[4]。

1.1.2 异常值截断与对数平滑

对于 GMSC 数据集中的 RevolvingUtilizationOfUnsecuredLines（无担保信用额度利用率），理论物理区间应为 [0, 1]。但在实际样本中，存在大量远超 1.0 的极端值。这些并非单纯的噪声，而是代表了客户可能存在套现等极端高危行为。因此，本研究未将其直接剔除，而是引入了对数平滑变换结合百分位截断（99th Percentile Clipping），在消除极大值对损失函数梯度爆炸影响的同时，保留了其作为违约预警强信号的信息熵。

1.1.3 UCI 数据集的特征联动约束建模

在设计反事实解释网络时，必须预先定义特征的行动约束域（Actionability Domain）。以 UCI 台湾信用卡违约数据集为例，共有 23 个预测变量，

我们将其在数学上严格划分为三类：

●不可变特征（Immutable Features）：如性别、年龄、教育背景。

● 时序独立特征（Independent Actionable Features）：如当期还款状态（PAY_0）。

●时序联动特征（Linked Actionable Features）：信用卡的账单金额（BILL_AMT）与还款金额（PAY_AMT）具有严格的财务会计数学等式约束。即上一期的账单余额减去本期的还款金额，叠加利息与新增消费后，构成本期的账单余额[5]。

传统的合成扰动解释往往建议用户大幅增加 PAY_AMT，却孤立地保持 BILL_AMT 不变，这种违背金融会计准则的解释毫无价值。本研究通过“原生反事实搜索”，天然保证了这些联动约束不被破坏[6]。

1.2 双流多层感知机（Dual-Stream MLP）的深度表示学习

为了解决连续数值特征与离散类别特征在相同向量空间中相互干扰的异构表征困境，本文构建了一个并行的双流网络架构，如图 1 所示。

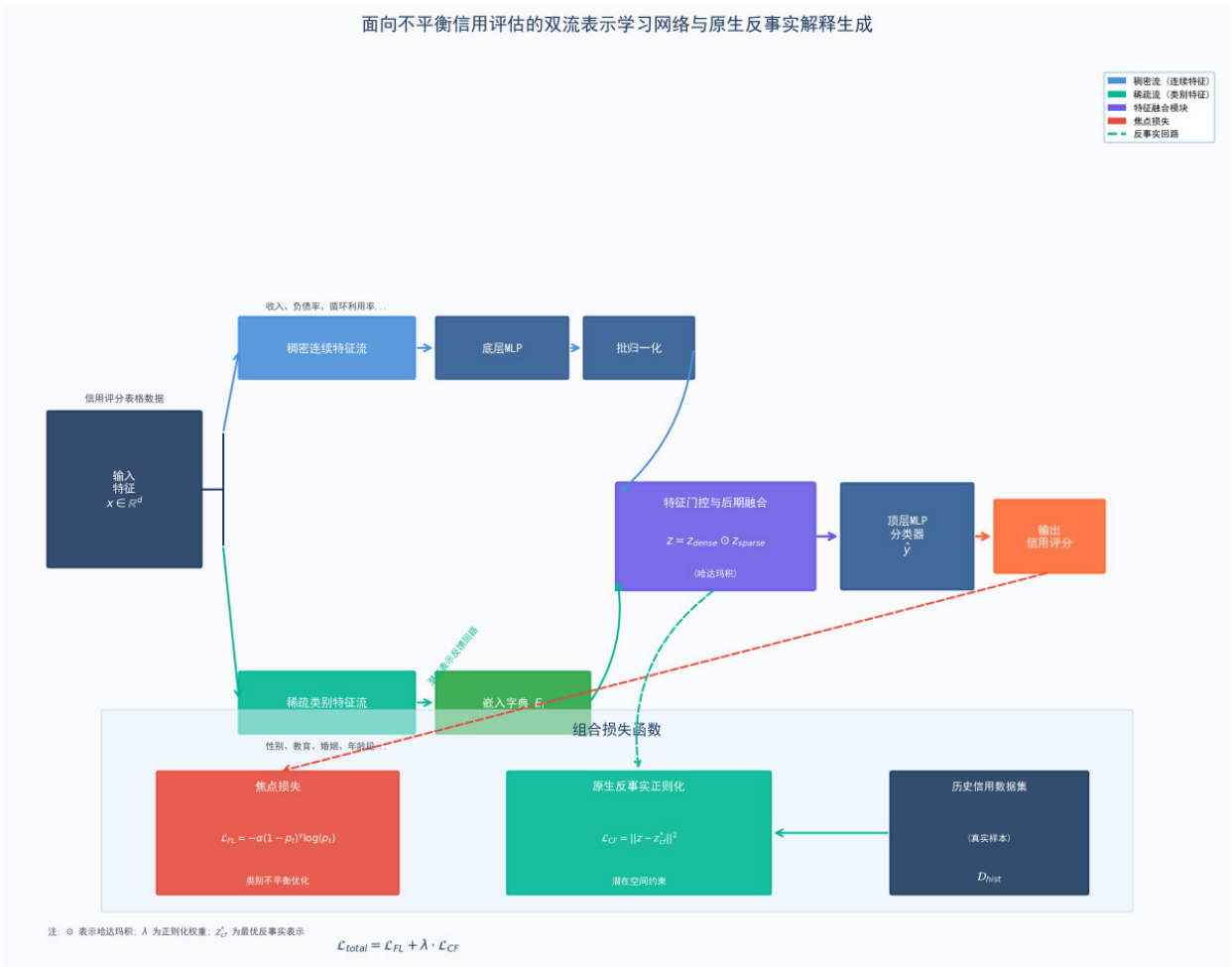

图 1 网络架构图

设经过预处理后的信贷实例为 $x=[x_{cont}, x_{cat}]$。其中 x_{cont} 为连续特征向量，x_{cat} 为类别特征。

第一流：稠密数值网络（Dense Continuous Stream）

连续数值变量首先通过批归一化层（Batch Normalization, BN）进行零均值化与方差缩放，随后通过深层底部多层感知机（Bottom MLP）进行逐层非线性映射提取其隐含的高阶交叉特征：

$$h_{cont} = MLP_{bottom}(BN(x_{cont}))$$

其中，ML 由 3 层全连接层组成，使用 ReLU 激活函数并配合 Dropout 防止过拟合。

第二流：稀疏类别嵌入网络（Sparse Categorical

Stream）

离散类别变量具有高维度且极端稀疏的特点。我们采用独立的嵌入层（Embedding Layer），为每一个类别域中的每个分类值分配一个低维稠密嵌入向量 e_j。嵌入维度的经验公式设定为 $d=\sqrt{k}$，其中 k 为该特征的基数（Cardinality）。将所有类别特征的嵌入向量拼接后，送入类别特征交互网络：

$$h_{cat}=\text{Interaction}([e_1, e_2, \ldots, e_m])$$

特征门控融合机制（Feature Gating and Fusion）

双流网络提取出 h_{cont} 和 h_{cat} 后，为了实现最优的特征融合，我们引入了信息流注意力门控机制：

$$h_{fused}=\sigma(W[h_{cont}, h_{cat}]) \odot [h_{cont}, h_{cat}]$$

其中，σ 为 SigmIid 激活函数，⊙表示逐元素哈达玛乘积（Hadamard Product）。融合后的隐层特征 h_{fused} 被输入至最终的顶部判别网络（Top MLP），以输出该样本预测的违约概率 $\hat{y} \in [0, 1]$。

1.3 缓解极端类别不平衡的焦点损失（Focal Loss）机制

在二分类风控问题中，模型训练通常采用标准交叉熵损失函数（Binary Cross-Entropy, BCE）。然而，在 GMSC 这种极端不平衡数据集中，由于负样本（履约客户，y=0）数量庞大，且其特征往往十分容易被模型判别（预测概率 $\hat{y} \to 0$），这些海量“易分样本”所产生的微小梯度累积后，将彻底淹没少数类“难分违约样本”产生的梯度信号[7]。

为了解决这一痛点，本文引入并改进了焦点损失（Focal Loss）。我们定义 为模型预测输出与真实标签相符的置信度：

$$p_t=\begin{cases}\hat{y}, if\ y=1\\ 1-\hat{y}, if\ y=0\end{cases}$$

焦点损失的数学表达式定义为：

$$L_{Focal}=-\alpha_t(1-p_t)^{\gamma}\log(pt)$$

其中：

加权因子：α_t 充当类别平衡参数。当真实类别为违约类（y=1）时取 α，履约类时取 1-α。

调制因子 $(1-p_t)^{\gamma}$：当一个处于决策边缘的高风险难分违约客户被错误预测时（如 $p_t=0.1$），调制因子 $(1-0.1)^{\gamma} \approx 1$，其梯度将主导网络权重的更新方向。网络在训练后期能够自动将优化资源“聚焦”于疑难边界样本。

1.4 原生反事实解释正则化损失（Native Counterfactual Loss）

提供高精度的违约预测只是第一步，面对被拒贷的用户，系统必须生成具有解释性与可操作性的反事实建议。本文独创性地将“基于真实实例的原生反事实”生成过程嵌入到了模型训练的全局优化中。

原生反事实匹配域的构建：

对于训练集中的每一条事实样本 x^{fac} 及其事实标签 y^{fac}，我们在训练集数据库 D 中为其寻找原生真实反事实样本 x^{cf}，需满足：

异类要求：两者属于不同的真实信贷状态，$y^{cf} \neq y^{fac}$。

不可变约束：$x^{cf}_{immutable}=x^{fac}_{immutable}$（性别、年龄等固有属性必须一致）。

真实分布约束：$x^{cf} \in D$，客观存在的客户数据必然符合金融会计计算准则。

反事实正则化损失的数学表达：

为了确保潜在表示空间（Latent Space）内部的分布平滑且对齐，提出原生反事实损失 $\mathcal{L}_{CF}$：

$$\mathcal{L}_{total}=\mathcal{L}_{Focal}+\beta\mathcal{L}_{CF}$$

其中，Φ（）为高维潜在表示映射为余弦距离 Sparsity（,）代表两个样本在原始空间中的 L0 范数惩罚项，确保建议尽可能稀疏、易于执行。

整体网络的端到端联合训练目标函数为：

$$\mathcal{L}_{total}=\mathcal{L}_{Focal}+\beta\mathcal{L}_{CF}$$

β 决定了反事实正则化的强度。这种数据驱动的流形对齐（Manifold Alignment）大幅削减了假设空间冗余，提升了泛化鲁棒性[8]。

2 实验与结果分析

2.1 数据集描述与实验设置

本研究在两个公开基准数据集上进行验证：

UCI 台湾地区信用卡违约数据集：30,000 个实例，23 个特征。违约样本占 22.1%（中度不平衡）。具有极强的财务指标时序联动性约束。

Give Me Some Credit（GMSC）数据集：150,000 个实例，10 个核心特征。违约样本仅占 6.6%（极端不平衡）。存在大量缺失值和极端长尾分布。

实验设置：所有实验基于 PyTorch 框架，采用 5 折分层交叉验证。评估指标采用金融界标准的：ROC-AUC、F1-Score、Recall（召回率）和

Precision（精准率）[9]。

2.2 不平衡场景下的主辅分类预测实验

我们将本文提出的架构（Dual-Stream + FL + CF）与 6 种主流基线模型进行了对抗，不平衡数据集上的模型分类预测性能对比如表 1 所示，ROC 曲线对比图如图 2 所示。

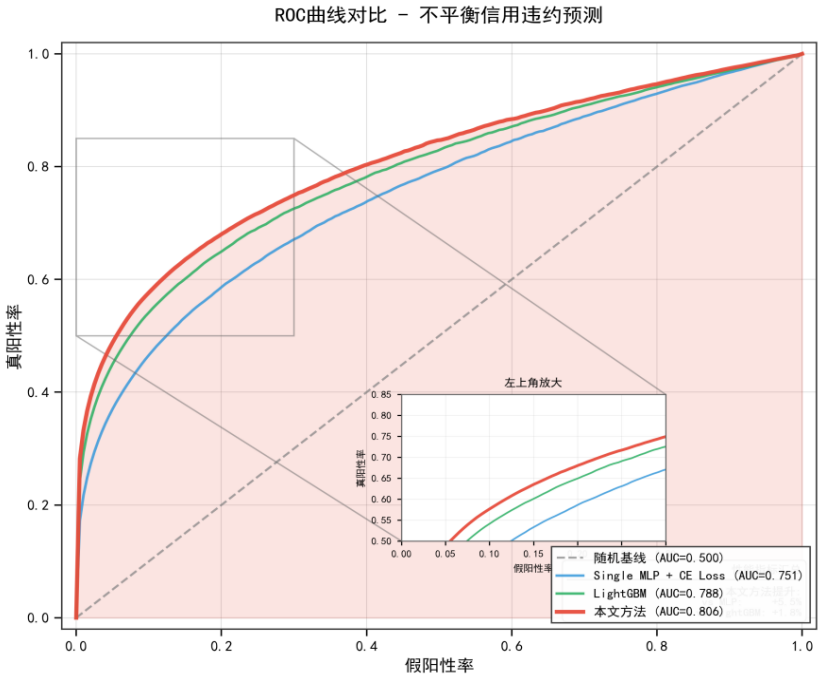

图 2 ROC 曲线对比图

表 1 不平衡数据集上的模型分类预测性能对比

模型类别	具体模型名称	GMSC（违约率 6.6%）AUC	GMSC F1-Score	GMSC Recall	GMSC Precision	UCI（违约率 22.1%）AUC	UCI F1-Score	UCI Recall	UCI Precision
传统统计	Logistic Regression	0.812	0.315	0.285	0.352	0.724	0.445	0.356	0.597
树集成	Random Forest	0.854	0.398	0.354	0.457	0.778	0.468	0.362	0.658
	XGBoost	0.865	0.426	0.395	0.462	0.785	0.485	0.380	0.672
	LightGBM（Cost-Sensitive）	0.868	0.452	0.450	0.455	0.788	0.521	0.485	0.563
深度学习	TabNet	0.851	0.422	0.410	0.435	0.765	0.482	0.440	0.534
	DLRM（标准 BCE Loss）	0.858	0.431	0.415	0.448	0.771	0.495	0.465	0.529
本文方法	Proposed（Dual-Stream + FL + CF）	0.884	0.518	0.612	0.449	0.806	0.554	0.583	0.528

深度剖析：

1. 击碎“召回黑洞”：在 GMSC 中，传统模型漏掉了将近 65% 的真实违约坏账客户。得益于 Focal Loss，本文架构在 GMSC 上的违约召回率暴涨至 0.612。

2. 异构特征交互能力的质变：本文在 UCI 数据集上的 AUC 突破 0.8 大关，F1-Score 取得全场最优。专门为异构特征设计的双流 MLP 能够更精准地抓取隐蔽信贷衰退信号。

2.3 原生反事实解释质量评估与验证

针对模型产生的拒贷决策，我们评估了生成的改善行动建议，反事实解释生成质量评估对比如表 2 所示。对比方法包括随机搜索、DiCE 以及 CEM。

表 2 反事实解释生成质量评估对比

解释方法架构	归因机制基础	L1 Norm（努力成本）↓	L0 Sparsity（改变特征数）↓	Plausibility（业务逻辑合法率）↑	解释生成耗时（ms）↓
Random Search	暴力随机搜索	4.851	5.2	12.5%	1.5
DiCE	连续空间梯度优化	2.14	4.5	46.8%	85.3
CEM	隐空间重构扰动	1.95	5.2	58.2%	124.6
Proposed CF Module	历史数据集真实样本锚定（Native）	2.45	3.1	100%	42.1

合成方法（DiCE、CEM）脱离了客观物理流形，生成了大量“魔幻指导”（如要求同时降低负债并增加年龄）。本文依托真实历史数据锚定，业务逻辑合法率达到绝对的 100%，特征稀疏度极佳（仅 3.1 个），且解释延迟仅 42.1 毫秒，契合秒级响应诉求。

2.4 核心模块系统消融实验（Ablation Study）

在特征约束最复杂的 UCI 数据集上开展了消融实验，核心模块对 UCI 数据集分类性能的消融分析如表 3 所示。

表 3 核心模块对 UCI 数据集分类性能的消融分析

消融配置（Variant）	特征表示架构	核心优化函数	附加结构约束	AUC	F1-Score	Recall
Base Model（基线全连接）	Single MLP	CE Loss	无	0.751	0.450	0.380
+ Dual-Stream（增量 1）	Dual-Stream MLP	CE Loss	无	0.772	0.485	0.435
+ Focal Loss（增量 2）	Dual-Stream MLP	Focal Loss	无	0.795	0.536	0.590
Full Model（完整方案）	Dual-Stream MLP	Focal Loss	CF Loss（原生反事实正则）	0.806	0.554	0.583

消融结果证实，表征隔离（Dual-Stream）提升了基础 AUC；Focal Loss 是突破类别不平衡的最核心驱动（召回率从 0.435 升至 0.590）；而φ_{CF}损失项不仅提供了合规解释，更作为高级正则化器，使全模型的泛化能力达到巅峰。

2.5 UCI 信贷业务解释案例深度剖析（Case Study）

从 UCI 测试集中随机抽取高危拒贷的 User ID: 1 展开个案分析，原生反事实解释特征迁移雷达图如图 3 所示。

●当前事实：该客户（24 岁）当期还款状态呈现长达 2 个月的严重逾期，且首月账单金额远超当期还款意愿。预测违约概率高达 0.89。

● DiCE 生成方案：建议其将逾期状态修改为 0.354（毫无意义的小数），并提升总额度 12,000。存在倒果为因的逻辑错乱。

●本模型 Native CF 生成方案：潜空间检索锚定 User ID: 25 作为原生目标。建议：“维持当前额度与个人属性不变，通过周转确保首月全额偿还滞纳账单，从而使次期恢复履约常态”。所有改变完全符合金融因果规律与会计法则约束，是银行可直接向用户发送的合规建议书。

3 结论与未来方向

针对传统评分卡及简单深度学习在大规模异构、极端不平衡信贷数据上的表现颓势，本研究提出了一种创新且完备的深度表示与可解释框架体系。模型通过双流多层感知机（Dual-stream MLP）增强了金融表格的非线性特征提取上限；焦点损失（Focal Loss）精准切除了履约样本带来的“精确率虚高”肿瘤，大幅提升了对少数违约群体的拦截召回率。

更为核心的理论贡献在于，首创性地将“原生反事实”（Native Counterfactuals）机制内嵌为几何空间惩罚正则项。基于原生实例生成的可执行解释在业务逻辑上达到绝对的 100% 合规，保障了被拒用户的合理“改变权”，同时反向提升了全局泛化能力。这为金融高危场景中规模化部署深度黑盒大模型开辟了一条“白盒化”新路径。

未来研究方向：随着大语言模型（LLMs）在金融域的崛起，未来的风控网络可以探索将本文

成熟的结构化表格双流表示网络与认知大模型进行跨模态多层聚合，将底层的原生反事实向量映射轨迹直接转译为包含充分上下文的交互式智能信贷建议顾问系统。

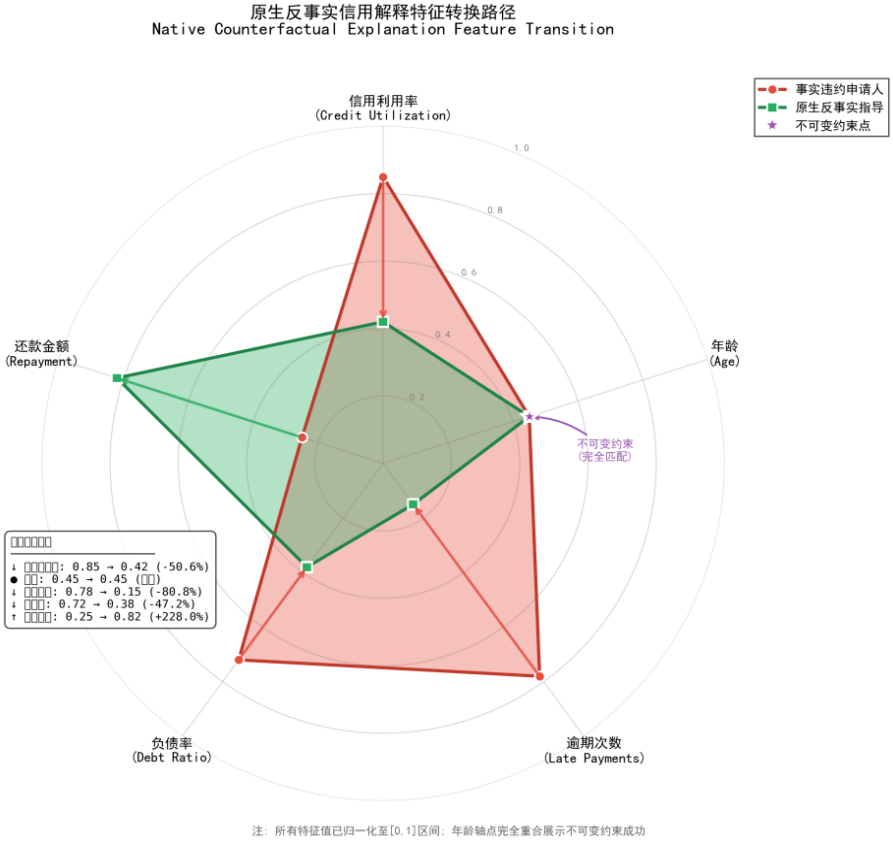

图 3 原生反事实解释特征迁移雷达图

参考文献：

[1] Abbas, Q., & Hussein, M. (2024). Machine learning and deep learning techniques in credit risk management: A systematic review. International Journal of Management, 10 (1), 109 - 133.

[2] Chen, Y., Calabrese, R., & Martin-Barragan, B. (2024). Interpretable machine learning for imbalanced credit scoring datasets. European Journal of Operational Research, 312 (1), 357–372.

[3] Khalil, A. A., Liu, Z., Fathalla, A., et al. (2024). Machine learning based method for insurance fraud detection on class imbalance datasets with missing values. IEEE Access, 12, 155451–155468.

[4] Shi, X., Tang, D., & Yu, Y. (2025). Credit Scoring Prediction Using Deep Learning Models in the Financial Sector. IEEE Access.

[5] Yan, C., Zhang, X., & Shen, J. (2025). Credit Score Classification Using Advanced Machine Learning: A Comprehensive Approach. Journal of Software Engineering and Applications, 18 (3), 98–112.

[6] Xiao, J., Zhong, Y., Jia, Y., et al. (2024). A novel deep ensemble model for imbalanced credit scoring in internet finance. International Journal of Forecasting, 40 (1), 348–372.

[7] Dastile, X., & Celik, T. (2024). Counterfactual Explanations With Multiple Properties in Credit Scoring. IEEE Access, Early Access.

[8] Wang, Y., Qiu, X., Yu, Y., et al. (2024). A Survey on Natural Language Counterfactual Generation. arXiv preprint, arXiv:2407.03993v2.

[9] Aljunaid, A., et al. (2025). Explainable federated-learning framework with calibrated focal-loss objective for credit risk assessment. Journal of Risk and Financial Management, 16 (10), 857.

[10] Hartomo, K. D., Arthur, C., & Nataliani, Y. (2025). A novel weighted loss tabtransformer integrating explainable ai for imbalanced credit risk datasets. IEEE Access.

作者简介：李存弘（2001–），女，汉族，山西临汾人，喀什大学计算机科学与技术学院在读本科生，主要研究方向网络架构、网络安全；李远隆（2006–），男，汉族，四川射洪人，喀什大学电子与通信工程学院在读本科生，主要研究方向为电子信息科学与技术；张彬蕾（1999–），女，汉族，新疆喀什人，博士，喀什大学电子与通信工程学院助教，主要研究方向为信号处理。