**面向高影响性金融风险场景的潜在对齐生成模型（罕见表格事件）**

**项目导师：** He Zhao, Zang Tao

**动机:**

金融机构日益依赖机器学习模型来评估**信誉度**、检测**异常**并预测**违约风险**。然而，这些模型通常在具有强烈**分布偏差**的真实世界数据集上进行训练——例如，**极少数高净值客户**会违约，导致这类边缘案例的代表性严重不足。

这种稀缺性导致了以下问题：

* 模型在关键、高影响性的边缘场景（例如，VIP客户意外违约）上**泛化能力差**，并且由于数据高度不平衡，模型训练困难。
* 下游决策系统的**稳健性和可解释性不足**。
* **缺乏反事实分析工具**来回答高风险金融场景中的“如果……会怎样？”（what-if）问题，例如：
  + “一个高收入客户如果违约，其信用画像会是怎样的？”
  + 作为回应，我们希望有一种方法，能够生成一个交易表格，其嵌入表示**未能**被归入“违约”簇，但同时又具备“高收入”的特征。

**嵌入语言模型（ELMs）能解决这个问题吗？**

嵌入语言模型（ELM，例如 NeurIPS 2024 的 EAGLE）近期已展示出其强大的能力，可以通过一个基于嵌入的效用函数（或奖励函数），来教导大语言模型（LLMs）生成能够达到一个**期望的嵌入区域**的文本输出。这一新发现启发我们提出了一个框架，该框架能够：

1. 在金融记录的表格嵌入中，**识别代表性不足的潜在簇**（例如，违约的VIP客户）。
2. 以表格数据格式，通过基于嵌入空间的效用函数（奖励函数），**生成真实的反事实样本**。
3. **引导模型和分析师**更好地理解和校准金融领域中数据稀疏区域的情况。

**相关工作:**

* **嵌入对齐语言模型 (NeurIPS 2024):** 介绍了EAGLE，这是一个强化学习智能体，它通过编辑LLM的输出来匹配源于行为数据的目标嵌入区域。本次实习将主要依赖EAGLE框架来实现稀有数据合成和反事实样本生成。
* **表格表示学习:** 像TabNet、FT-Transformer和TabPFN等技术在捕捉金融数据结构方面表现出色。本次实习可以利用TabPFN的编码器将EAGLE生成的输出投影到嵌入空间，在其中，一个精心设计的效用函数将提供奖励/指导。
* **反事实解释 (Wachter et al. 2017, Karimi et al. 2021):** 通过生成备选数据实例来探究模型推理，但通常缺乏对全局嵌入属性的控制。
* **表格领域的数据增强:** 与自然语言处理和计算机视觉相比，该领域仍不成熟；近期的工作探索了GANs、SMOTE和少样本Transformer，但通常难以针对数据分布中具有语义意义的空白区域。
* **金融领域的异常风险:** 在可解释人工智能和公平性领域的研究强调，模型需要在罕见但影响巨大的事件上保持稳健，例如富裕客户的策略性违约或画像转变（即“高净值人士违约悖论”）。

**提议的方法:**

**步骤 1: 潜在空间构建**

* 在一个金融数据集（如信贷局数据或贷款组合）上，**训练一个表格基础模型**（例如FT-Transformer或TabLLM编码器）。
* 使用学习到的嵌入表示，将所有数据点映射到一个潜在的语义空间。
* 使用聚类或密度估计（例如，UMAP + DBSCAN, KDE）来**识别潜在空间中的稀疏区域**——这些区域代表了代表性不足或高风险的画像区域。

**步骤 2: 目标簇识别**

* **定义高价值的目标簇**，例如：
  + 高收入 + 低违约可能性（正常）
  + 高收入 + 实际违约（罕见）
  + 低收入 + 未违约（意外的坚韧性）
* 使用监督标签或潜在的行为信号来**标记这些簇**，以用于效用函数。

**步骤 3: 定向剖面生成**

* 使用一个**嵌入对齐生成器**（即EAGLE）来：
  + 从密集区域中现有的数据点开始。
  + 迭代地修改表格特征（例如，使用率、贷款组合、账户年限），以**达到目标嵌入区域**。
  + 使用学习到的约束或领域知识驱动的编辑图来**确保语义和业务规则的有效性**。

**步骤 4: 应用**

* **模型压力测试:** 将合成的罕见案例注入训练集或评估集，以测试模型的稳健性。
* **情景规划:** 为风险团队提供丰富的假设性画像，以改进早期预警系统。
* **反事实解释:** 提供关于画像的微小变化如何可能改变风险预测的深刻见解。

**潜在数据集**

* 台湾信用卡违约数据集 (UCI)
* 德国信贷数据
* 银行营销数据集
* （如果可扩展到时间序列数据）NAB / SKAB 异常基准数据集，用于罕见事件的时间序列建模。

**提议的评估指标**

* **罕见簇的泛化能力** – 在罕见区域的AUC、召回率（Recall）、校准度（Calibration）。
* **潜在空间多样性** – T-SNE / PCA 图、熵、KL散度。
* **增强的基础模型嵌入** – 在下游微调任务上的性能。
* **人工专家验证** – 由人类专家对生成数据的合理性、实用性、有效性进行评分。

**交付成果:**

该项目的预期成果是将研究成果**提交至像ICML 2026这样的顶级会议**。

* **实习开始前:** 针对开放性挑战提出潜在解决方案，并确定一个可行的项目方案。
* **第一个月:** 实习生应搭建候选的开源LLM模型，调研表格数据基准和基础模型嵌入。接下来，实习生应在任意表格数据集上重新实现一个EAGLE框架的最小可用版本，并设计一个能够引导LLM生成目标输出的嵌入空间效用函数（作为奖励函数）。
* **第二个月:** 实习生应搭建从步骤1到步骤3的完整流水线，获得初步实验结果，并建立评估流程。
* **第三个月:** 我们期望实习生继续改进整体方法，并与基线模型进行比较。
* **第四个月:** 实习生将专注于完成实验收尾、撰写论文并清理代码库。