**研究方向进展报告**

汇报人：熊俐凯

日期：2025年3月21日

### ****一、本周核心工作****

#### ****1. 文献研读****

* 1. 论文内容：2025年2月发布的一篇论文《IMPACTX: Improving Model Performance by Appropriately predicting CorrecT eXplanations》
  2. 主要内容：该论文讲述了如何将可解释人工智能（XAI）方法不仅用于解释模型决策，还用于自动提升模型性能。现有XAI研究大多着眼于解释，而IMPACTX关注的是如何在无需人工干预的前提下，借助XAI机制作为注意力机制，在训练阶段引导模型聚焦于更重要的输入特征，从而提升分类性能。
  3. 本文提出的创新内容：
     1. 完全自动化的注意力机制：不需要人类提供显著性图或注释，使用XAI方法生成的特征归因图（attribution maps）来训练模型中的一个解释预测模块。
     2. 双分支架构设计（Dual-Branch）上分支：负责分类（特征提取器 M + 分类器 C）；下分支：负责生成解释（Latent Explanation Predictor + Decoder）
     3. 模型无关性（Model-agnostic）：可适用于任何黑盒模型，不需要访问内部结构。
     4. 在推理阶段无需外部XAI方法：一旦训练完成，IMPACTX可以直接输出模型预测及其对应的解释，无需额外运行SHAP等后处理方法。
     5. 联合优化目标函数：结合了分类损失（Cross Entropy）与解释重建误差（MSE）：
  4. 对自己方向的想法：

金融欺诈检测+XAI（LIME+SHAP）：

* + 1. 将解释引入训练阶段：目前可能是先用XGBoost训练，再用LIME/SHAP分析。这种方式是“后验解释”。想尝试像IMPACTX一样，将解释引入训练环节，提高模型聚焦能力。
    2. LEP模块启发：可以构建一个类似的LEP模块来预测SHAP图，然后联合优化预测和解释两任务，这样或许提升解释一致性与模型鲁棒性。
    3. 用于非图像数据（如结构化金融数据）：虽然论文是图像分类，但IMPACTX框架本质是通用的，LEP和Decoder也可以用MLP或Transformer结构来适配结构化数据。
    4. 两阶段训练策略适用金融数据：可以先训练XGBoost模型，然后用SHAP生成每条记录的解释图，再训练一个解释预测模型，结合预测结果提升主模型性能

### ****后续计划****

下周进行项目实验。

**不足点：**对本文中所讲述的模型还没有理解清楚（只知道大致的模型思路），对注意力机制不太了解，想跟着此论文的思路进行初步复现，使用金融数据集结合XAI方法提升将传统的模型预测能力（例如XGBOOST+XAI方法）。