**研究方向进展报告**

汇报人：熊俐凯

日期：2025年3月14日

### ****一、本周核心工作****

#### ****1. 文献研读****

* 1. 论文内容：深入研读SHAP框架的论文《A Unified Approach to Interpreting Model Predictions》

主要内容：SHAP 方法：基于 Shapley 值的解释方法

•SHAP 方法基于合作博弈论中的 Shapley 值，为每个输入特征分配一个归因值，表示该特征对模型输出的贡献。

•SHAP 统一了现有的特征归因方法，如 LIME（Local Interpretable Model-agnostic Explanations） 和 DeepLIFT。

* 1. 初步阅读了两篇论文（乳腺肿瘤和诈骗信息），使用可解释性AI主要是用在了可解释的辅助诊断系统来提升医生的诊断可信度。用了四种可解释性方法局部自解释（基于反向传播和动态路由算法）：识别影响预测的关键文本区域。全局自解释（基于词频统计和多头向量分析）：提供全局预测依据。LIME（局部模型无关方法）：分析单个预测的影响特征。SHAP（基于博弈论的解释方法）：计算各特征的贡献度。诈骗信息也是使用LIME和SHAP方法来增强模型透明度。后续工作可以模仿这两篇论文的思想进行实践

#### ****2. 了解其他可解释性方法****

* 1. DeepLIFT方法：用于理解深度神经网络的决策过程。他通过计算输入相对参考输入的贡献值，来衡量每个特征对模型预测的影响。能够比较好的解决以下问题：
     1. 梯度消失问题
     2. ReLU非线形问题
     3. 基线比较
  2. 逐层相关性传播方法：该方法其实与DeepLIFT方法等价，就是将所有神经元的参考激活值被固定为零。主要也是为了识别输入特征对模型预测对贡献，然后从输出层反向传播相关性到输入层，以理解哪些输入特征对预测最重要。

1. 数据集准备

从kaggle中获取到了Lending Club Loan Data数据集，后续针对这大量的数据进行XAI方向的初步实验。

### ****后续计划****

**使用下载好的数据集进行初步实验。**