南開大學
Nankai University

《机器学习》

期末课程项目报告

学院：人工智能学院

姓名：刘崇轩

学号：2312801

2025 年 6 月 27 日

金融风险预测实验报告

一、 问题描述

金融风险评估是金融科技领域中的关键问题，尤其是在银行信贷、消费金融和网贷场景中，对于客户是否违约的预测（违约概率）直接关系到资产安全和收益控制。本实验基于某金融机构提供的用户贷款数据，使用 XGBoost 模型预测用户是否存在违约风险（isDefault），并输出预测概率和重要特征排序结果。

二、 实验配置

1.语言环境：Python 3.12
2.开发平台：Windows 11 + Anaconda
3.核心库版本：
- xgboost = 3.0.2
- pandas = 2.2.2
- scikit-learn = 1.5.1
- category_encoders = 2.8.1
- matplotlib = 3.9.2

4.输入数据：
- train.csv: 训练集，含字段 id,isDefault（目标），若干特征；
- testA.csv: 测试集，含字段 id，无标签。
- 预测目标：isDefault（是否违约，1 表示违约）

三、 算法原理

1. XGBoost 简介

XGBoost（eXtreme Gradient Boosting）是提升树（Boosting Trees）的一种高效实现，适用于结构化数据的分类与回归任务。其核心思想是通过加权集成多个弱分类器（通常是 CART 决策树），不断优化目标损失函数。

本实验使用二分类目标函数（binary:logistic）和 AUC 评价指标（eval_metric=auc），以区分是否违约的风险。

2. 特征编码与预处理

（1）数值特征：用中位数填补缺失；

（2）类别特征：用 TargetEncoder 编码（即用目标变量的平滑期望值替代类别）；

（3）数值标准化：统一分布以加快模型收敛。

3. 模型训练与验证

- 使用 StratifiedKFold 分层交叉验证（5 折）；

- 每折构建训练集与验证集，分别训练和评估模型；
- 启用 early_stopping_rounds=50 避免过拟合；
- 使用 tree_method='hist' 以提升训练效率；
- 使用 scale_pos_weight 处理类别不均衡。

四、 具体实现过程

1. 数据读取与字段划分

（1）使用 pandas 加载 CSV；

（2）提取标签列 isDefault，并分离 id 列；

2. 特征类型判断与缺失值处理

（1）利用 pandas 类型判断识别数值和类别特征；

（2）数值特征用中位数填补，类别特征填充 missing 常量。

3. 编码与标准化

（1）类别特征通过 TargetEncoder 编码；

（2）所有数值列经 StandardScaler 标准化处理。

4. 模型训练

（1）使用 StratifiedKFold 分层交叉验证（5 折）；

（2）每一折中构造 xgb.DMatrix 数据结构；

（3）使用 xgb.train()对每一折进行训练；

（4）设置参数并调用 xgb.train 训练模型，启用 early_stopping_ rounds=50；

（5）累加每一折的验证 AUC、预测值与特征重要性。

5. 预测与输出

（1）所有测试集预测结果做均值融合，保存为 Result.csv；

（2）计算平均特征重要性，绘制前 20 个重要特征的横向柱状图，保存为 feature_importance.png。

五、 实验结果与分析

1.验证集表现（5 折交叉验证 AUC）

折数	验证集 AUC
Fold 1	0.7357
Fold 2	0.7351
Fold 3	0.7340
Fold 4	0.7367
Fold 5	0.7341
平均	0.7351

说明模型具有良好的区分能力，AUC 稳定在 0.735 附近，泛化能力较强，无明显过拟合或欠拟合。

2.拟合过程观察

（1）训练集 AUC 从 0.70 提升至 0.77+，验证集 AUC 在 0.73 附近趋于平稳；

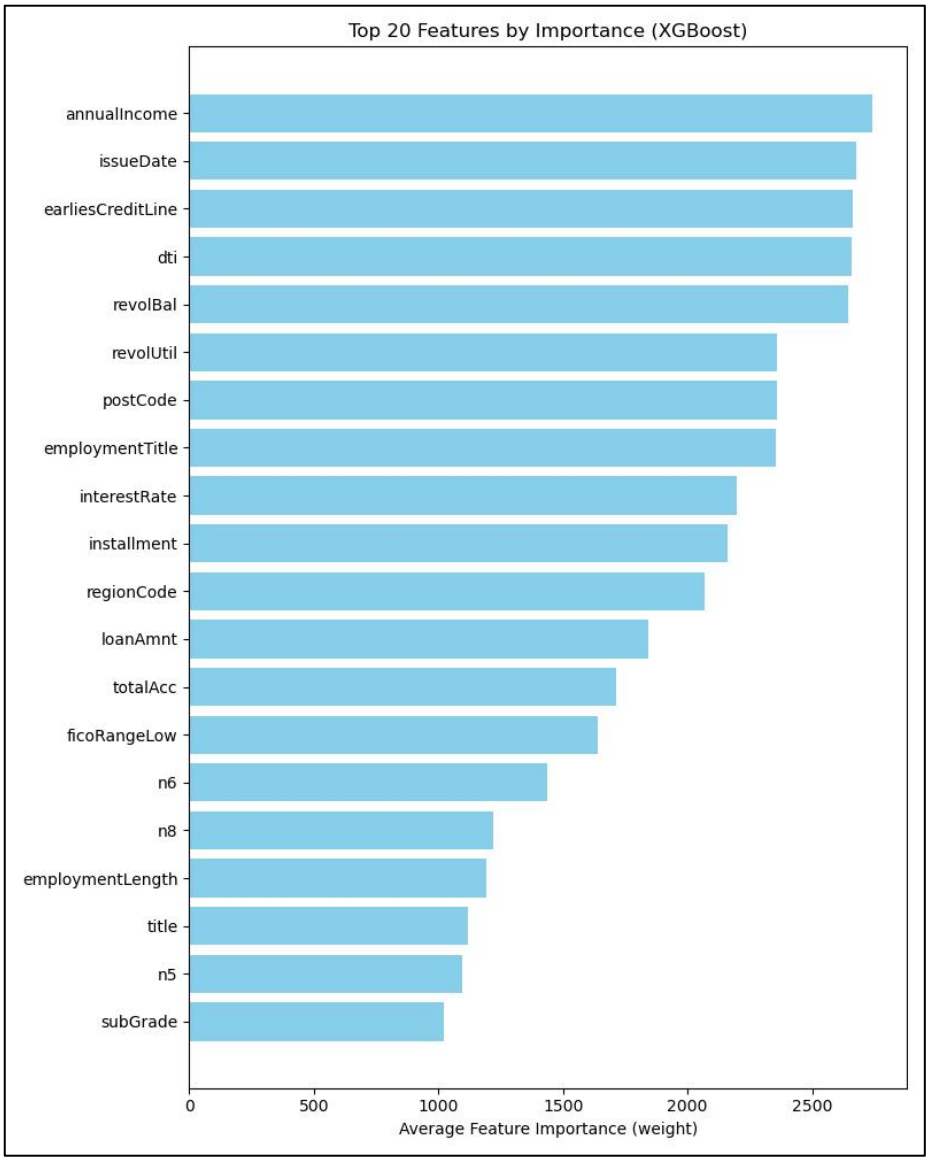
（2）使用 early_stopping_rounds=50 时，多在 800~900 轮停止训练，避免拟合；

（3）参数设置合理，梯度更新平稳，模型收敛过程可控。

3.特征重要性分析

（1）生成的 feature_importance.png 显示了前 20 个最重要特征（按出现频次）；

（2）可作为后续解释模型或优化特征工程的依据。



六、 总结

本实验通过合理的数据预处理、特征工程和模型调参，成功构建了一个基于 XGBoost 的信用违约预测模型。模型在 5 折交叉验证中表现稳定，平均 AUC 达 0.7351，具有实际应用的可行性。

同时，通过特征重要性图，我们可以进一步理解模型的决策机制，有助于提升业务可解释性。未来可进一步引入 LightGBM、CatBoost 或深度学习模型以提升表现。

参考文献

[1] Chen, T. & Guestrin, C. (2016). *XGBoost: A Scalable Tree Boosting System*. In *KDD 2016*. https://arxiv.org/abs/1603.02754

[2] Friedman, J.H. (2001). Greedy Function Approximation: A Gradient Boosting Machine. Annals of Statistics.

[3] XGBoost 官方文档: https://xgboost.readthedocs.io/

[4] category_encoders 文档: https://contrib.scikit-learn.org/category_encoders/