作业2总结

目标

为Maydup Bank plc开发一个预测客户违约的模型。
具体要求包括：

- 根据数据集中的变量决定是否接受或拒绝新申请人。
- 可选择分类模型或概率模型（需指定分类阈值）。
- 项目需覆盖端到端的机器学习流程，包括：
- 生成高质量的预测结果。
- 向客户反馈数据质量及实用性的见解。

数据

- **训练数据**：`P2data????.csv`（9,000行）。
- **保留数据**：约4,500条（用于外部验证）。
- **数据特点**：稀疏，含缺失值，且`GOOD`/`BAD`标签不完整。

关键变量

- **目标变量**：`GOOD`（好客户）/`BAD`（坏客户）。
- **特征变量**：
  - **人口统计**：`cust_age`（年龄）、`occ_code`（职业代码）、`time_emp`（工作年限）、`res_indicator`（居住状态）。
  - **财务状况**：`disp_income`（可支配收入）、`CA_01`-`CA_03`（当前账户状态）、`D_01`-`D_02`（负债余额）。
  - **信用历史**：`S_01`-`S_02`（信用查询记录）、`I_01`-`I_06`（账户开立情况）、`P_01`（信用卡持有率）。

分类变量说明

- **职业代码**：SA（自雇）到FT（失业/学生）。
- **居住状态**：H（自有住房）、P（与父母同住）、R（租房）。
- **CA_01（账户状态）**：

- 1：逾期3-6个月
- 2：逾期1-2个月
- 3：正常/未激活
- 4-5：数据缺失

建模选项

1. **二分类模型**：直接预测`GOOD`/`BAD`，需指定分类阈值。
2. **三分类模型**：增加`PASS`（数据不足无法分类）类别。
3. **概率输出**：返回违约概率及置信度。
4. **客户要求**：必须明确分类阈值或决策规则。

任务要求

1. **数据处理**：

- 处理缺失值和稀疏数据。
- 代码需可复现（随机种子使用学号后四位`????`）。

2. **模型验证**：

- 使用内部验证（如交叉验证、训练集-测试集划分）。
- 比较性能指标（如准确率、AUC-ROC）。

1. **文档说明**：

- 代码需详细注释，便于客户复用。
- 报告需在5页PDF内总结方法和结论。

关键注意事项

- **数据挑战**：数据为模拟生成，含噪声和人为相关性。

- **验证一致性**：内部验证与保留数据验证结果应接近。
- **成果展示**：
- **Jupyter笔记本**：结构清晰，注释完整。
- **报告**：重点说明重要变量及模型推荐理由。

提交内容

1. **Jupyter笔记本**：

- 包含数据清洗、特征工程、模型训练与验证代码。
- 客户可替换为保留数据直接运行。

1. **PDF报告（≤5页）**：

- 方法概述、模型比较、最终建议。
- 突出重要/不重要变量。

其他提示

- **协作**：可讨论方法，但避免共享结果（数据差异影响可比性）。
- **优先级**：模型解释性 > 复杂度，确保客户可理解。