贷款违约预测模型构建

**一、目标**

给定数据集包含1091条数据，每个样本包含20个字段，数据已进行脱敏处理，每一行代表一个样本，且对于每个样本已标注出其是否违约。要求根据给定数据构建模型来预测未来用户的贷款违约情况。

**二、数据集概况**

数据变量说明：

|  |  |
| --- | --- |
| 性别 | sex |
| 最高学历 | degree |
| 大学类型 | uni\_level |
| 借款用途 | u\_loan |
| 担保 | guarantee |
| 项目金额 | amount |
| 收益 | income |
| 周期\_月 | period\_m |
| 月还本息 | repayment\_m |
| 总还款本息 | repayment\_T |
| 借款描述字数 | count |
| 累积借款金额 | c\_borrowing |
| 累积还款金额 | c\_repayment |
| 累积逾期金额 | c\_overdue |
| 累积已还次数 | c\_fre |
| 累积逾期次数 | c\_over\_fre |
| 累积逾借金额比 | c\_overrate |
| 累积逾期次数比 | c\_over\_frerate |
| 履约 | performance |

**三、案例分析要求**

1. 贷款违约预测模型需求理解和目标分析（5分）

目标描述：理解需求产生的背景，梳理出分析的目标，定位数据挖掘模式（分类，聚类还是关联分析）。

1. 认识数据（15分）
2. 阐述每一步数据探索操作的目的（5分）
3. 采用Python matplotlib等数据可视化包完成数据探索代码实现及可视化效果。（10分）
4. 数据预处理（15分）

（1）阐述每一步数据预处理操作的目的（5分）

（2）采用Python完成数据预处理代码实现及执行效果。（10分）

1. 特征选取（特征工程）（15分）
2. 简单描述拟采用特征选取方法的原理（5分）
3. 采用Python完成特征选取代码实现及特征重要程度展示。（10分）
4. 模型构建（30分）
5. 阐述所选择两种及两种以上算法构建相应的数据挖掘模型的理由？（10分）
6. Python实现算法1构建模型及效果（10分）
7. Python实现算法2构建模型及效果（10分）
8. 模型评估与选择（15分）
9. 阐述选择模型评估指标的原因（5分）
10. 模型评估代码实现及最终选择模型的效果展示（10分）
11. 贷款违约模型构建过程的完整性和逻辑性（5分）