**2020“江苏银行杯”金融大数据建模挑战赛**

**初赛方案概要介绍**

|  |  |
| --- | --- |
| 所选赛题 | 赛题1：基于机器学习的贷中风险预测模型 |
| 所使用编程语言 | Python |
| 所使用算法及技术 | 数据预处理，决策树分箱，XGBoost、LightGBM、CatBoost算法，模型融合Stacking |
| 对于赛题场景的理解  （包含对于所使用算法、技术问题场景的基本理解，至少100字） | 题目中的贷中客户分为三种情况：普通贷款，授信，贷中再申请贷款。  这是一个贷中风险控制的评分卡建模过程（二分类问题）,数据分为五张表，由cust\_id进行关联，但并不是每个客户都存在五张表，例如贷款中客户授信信息表，贷中客户申请表，贷中客户申请的缺失值考虑直接补0，授信这种缺失值有其实际意义，考虑进行分箱。  由于该二分类中存在类不平衡问题，我们考虑建立以XGBoost、LightGBM、CatBoost为第一层基模型,逻辑回归为第二模型的Stacking融合模型，为此期待达到更高的模型效果。 |
| 方案概要  （至少包括整体方案实施的步骤；各步骤及各步骤所使用的技术；各类技术如何应用于解决实际问题以及最终解决实际金融问题方式，至少500字） | 流程：1.问题建模-2.数据分析-3.数据清洗-4.特征工程-5.模型训练-6.模型验证（如不理想，则返回2.数据分析）-7.模型预测  1.问题建模\_赛题理解：分析风险来自于哪里，再次申请贷款的原因何在。分为三部分，还款意愿（是否欺诈倾向（历史逾期、虚假信息）），还款能力（收入水平、消费水平、工作变动），其它因素（工资日，节假日，花呗还款日期）。  2.数据探索性分析（EDA）：了解数据（数据类型大小，数据是否干净，缺失值，标签是什么类型），为建模做准备，重点研究《贷中客户授信信息表》与《客户贷款表现表》和《贷中客户流水表》是否存在关联，可否直接代替，即将普通贷款，授信，贷中再申请贷款三种情况经过处理归为一类。  3.数据清洗（数据预处理）：离群点的处理，缺失值处理（是否有特殊的意义），错误值处理。  4.特征工程：统计方式，决策树分箱，特征提取，特征选择（特征交叉，特征衍生）。  5.模型训练：由于该二分类中存在类不平衡问题，我们选择将贝叶斯调参调好的XGBoost、LightGBM、CatBoost模型作为第一层基模型,逻辑回归为第二模型进行Stacking融合模型。  6.模型验证：用训练好的模型进行label标识，提交结果得到反馈，分析问题进行修改。  7.模型预测：用最终训练好的模型进行label标识，提交预测结果Label标识。 |