**赛题1：基于机器学习的贷中风险预测模型**

|  |  |
| --- | --- |
| **1赛题名称:基于机器学习的贷中风险预测模型** | |
| **赛题简介：介绍整个赛题的实现目标、实用价值、涉及技术** | **实现目标：**网络信用贷款已经成为现阶段较为普及的贷款方式，各式各样的网贷产品也层出不穷，但面向大众的产品同样拥有较高的风险，在通过贷前阶段的风险控制后，贷中风险同样是值得关注的地方。随着时间的推移，客户在贷款后因为工作、生活等因素的影响，所呈现出来的风险等级也会随之变化，因此我们需要预测客户在贷中的风险变化，对出现高风险的客户及时采取防范措施。本赛题的目标是分析网贷客户的交易、行为、信贷表现等数据，设计并实现一种基于机器学习算法的风险预测模型，准确的预测出客户的风险,并对其变化原因给出一定的解释。  **实用价值：**本赛题的价值在于准确、快速的预测出风险变化较大的客户，从而对其进行提前的防范，降低损失。  **涉及技术：**本赛题主要涉及数据分析、数据处理、各类机器学习算法等技术。 |
| **参赛整体要求** | **整体要求：**  1.要求参赛队员掌握相关的基础知识，具有数据分析、处理及模型构建能力；  2.每组成员最多由2人组成；  3.请于 11 月 6 日前提交报名表，并于 11 月21 日前提交初赛最终作品成果； 11 月 25 日前提交决赛作品成果；  4.数据仅供本次比赛期间使用，未经竞赛组委会书面授权许可，任何人不得传播、公开或将数据用于其它科研、教学、商业用途；  5.赛事主办方不鼓励参赛队伍使用侧信道的方式进行解题，为确保竞赛的公平公正，参赛队伍之间不得以任何方式分享解题思路、代码程序或答案；  6.竞赛组委会将认真审核所有队伍的解题报告，如有发现参赛队伍存在抄袭行为，或是未能在规定时间内提交解题报告，则取消该参赛队伍的成绩； |
| **赛题业务场景：描述赛题相关的真实企业业务背景和赛题场景** | **业务背景：**网贷业务中，一个客户的工作或者生活情况的变化往往会对客户产生较大影响，一些负面因素可能导致一个优质客户转变为高风险客户，准确的在贷中场景中识别风险概率变化较大的客户，可以有效的进行防范并降低损失。在实际中，描述一个客户的数据维度复杂且多元，我们需要从万千数据中挖掘，寻找那些可以体现客户风险的特征维度，因此需要引入机器学习模型，通过不同的算法进行大量的运算以确定客户的风险变化情况。  **建模场景：**提供网贷客户的基本数据、交易数据、信用等数据和出险标识，设计并实现贷中风险预测模型，最终预测出高风险的客户。 |
| **基本功能要求** | 1.基于建模数据，设计并实现贷中风险预测模型，有效预测高风险客户； 2.输出建模模型过程和模型结果指标。 |
| **非功能性要求** | 1.算法性能评估；  2.算法实用性、可行性；  3.算法可解释性；  4.算法创新性。 |
| **实现条件：开发环境、实验平台、开发语言、数据库、编译器、涉及硬件等实现条件** | 开发工具自选；  开发语言和机器学习算法等不限。 |
| **测试数据：提供给参赛者的测试数据** | 数据集格式：提供的建模数据集为excel格式；  建模数据分为两部分（测试数据和验证数据），其中：  **1.测试数据集：已知风险标识Label**   1. 样本数据：坏样本（1000个）、好样本（1万个） 2. 维度数据：   ①客户的基本信息（包含年龄、性别、注册时间、所在地等）；  ②行内数据：行内还款行为数据、借款数据、交易流水等数据；  ⑶样本结构：  观察期：2019/12月-2020/3月  表现期：2020/4月-2020/9月  样本均为在表现期末时点未出现逾期的客户  好样本：观察期后6个月未出现逾期的客户  坏样本：观察期后6个月内出现逾期天数超过60天的客户  **2.验证数据集：未知风险标识Label**  （1）样本数据：1万  （2）维度数据：同上 |
| **开发所需设备及设备指标需求说明** | 开发设备不限（个人笔记本电脑、台式机、服务器、公有云等均可）。 |
| **评审要点** | 1.基本要求：F1值达到60%；  2.中等要求：F1值达到80%；  3.非功能性要求完成度；  **备注:**F1=2\*准确率\*召回率/(准确率+召回率)，其中准确率=准确预测客户数/预测的总客户数,召回率=准确预测风险客户数/实际发生风险客户数。 |
| **初赛作品提交要求** | 1.提交模型、验证数据集的预测结果Label标识（以excel形式提交）；  2.模型报告：包含数据分析情况、算法选择、建模过程、参数调整、模型调优、模型效果评估、模型复杂度指标等内容；  3.算法原理说明文档：算法的实现原理和详细细节及源代码。  4.项目演示PPT |
| **初赛评审标准** | 1.模型效果相关指标：验证数据集的F1值、KS、AUC、PSI等；（50分）  2.模型性能：模型时间复杂度和空间复杂度；（25分）  3.模型报告完整性和清晰度；（25分）  初赛根据以上三点评审进入决赛资格。 |

**对比赛内容有疑问可在参赛者咨询群中进行提问**