多金融场景下的模型训练

队伍名：吃瓜群众

验证集 AUC：0.7821

测试集 AUC：0.7771

队长：阮琳雄

手机号：13260279076

所属单位：树根互联

解题：

问题：训练集是 2017.4.1 到 2018.5.1 不同金额、不同期限、不同利率的金融产品样本，验证集和测试集是 2018.1.1 到 2018.5.1 机构 A 的产品。
分析：该题与传统的预测问题显著的差异在于训练集和测试集的分布是不同的。因此，首先要找到训练集中与测试集高度相关的样本。
解答：在训练集上，划分训练子集和验证子集，对 tag 特征进行预测(xgboost 模型)，发现训练子集与测试子集的 AUC 均超过 0.95，所以数据可以较准确的预测 tag 特征。运用该模型，对验证集和测试集的 tag 进行预测，使用训练子集上的最优阈值来分割预测结果，发现 99%以上的验证集和测试集 tag 都被预测为 1，所以有理由推断**机构 A 是一个大额分期贷机构**。

特征工程：

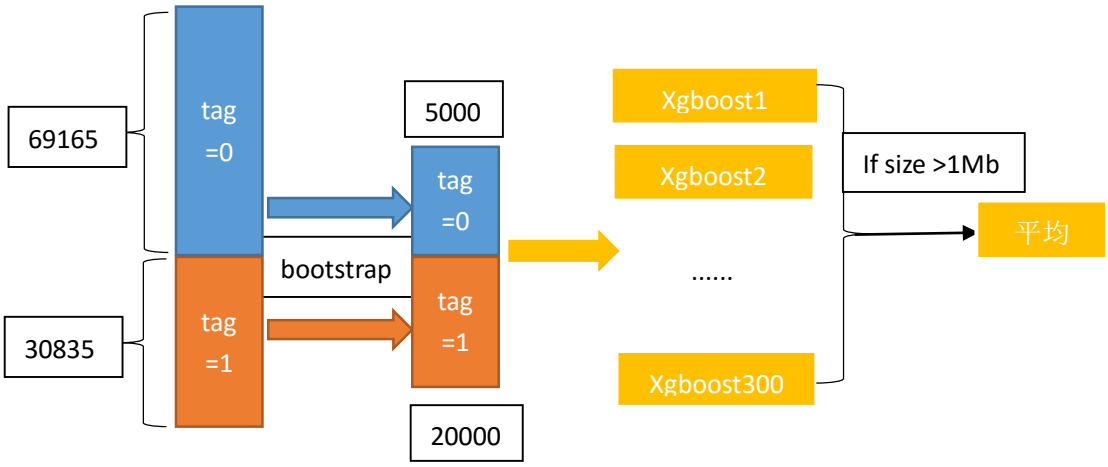
1.统计每个样本的缺失值数量，基于认知：缺失越多，信息越不完善，越有可能逾期。
2.对时间(loan_dt)，提取年、月、日、星期、距离 2017 年 4.1 的天数，五个特征。基于认知：贷款过程往往包含周期性。

数据稀疏化：

数据有大量缺失，xgboost 和 lightgbm 有缺失值处理手段，支持稀疏矩阵输入。所以数据采用 **csc(列)稀疏**。（稀疏矩阵没有列名，所以列名需要额外单独存储，xgboost 模型分析变量重要性时，需要特别注意，python 中 xgboost 自带的变量重要性会按照格式：”f”+列序号,对变量自动命名，所以真实的重要变量需要按照格式反搜索）

模型：

1. 对全样本全特征建立一个 xgboost 模型，验证集 AUC 达到 0.7815；
2. 全样本全特征建立 lightgbm 模型，验证集 AUC 达到 0.77，与 1 中模型相关度较高。
3. 对训练集分层有放回抽样，分层：tag=1 的样本有放回抽取 20000 个，tag=0 的样本有放回抽取 5000 个。基于 bagging 的思想，更换不同的随机种子，多次分层 bootstrap 抽样，建立了 300 个 xgboost 基模型，再从中挑选模型大小大于 1M（xgboost 大约多于 100 棵梯度树）（选入 167 个）。基模型的预测值，取简单平均，得到该模型的预测。验证集 AUC 达到 0.7744。该模型与 1 中模型相关度较低。



该模型基于两个认知：

1.在 tag=1 的样本中增加部分 tag=0 的噪声，不同的噪声在不同层次控制过拟合的程度，同时也允许 tag=0 的样本帮助 tag=1 的样本学习；

2.bootstrap 抽取的不同的样本，使用 xgboost 充分训练样本（size>1Mb），保证了模型之间的差异性从而提升 bagging 的性能。

注：参数选择，第一个模型全样本全特征的 xgboost 使用交叉验证调参，后续模型沿用了该模型的参数。

融合：

上述三个模型的预测结果取简单平均得到最终结果。