第三届融360天机-智能金融算法挑战赛多场景金融赛题--TOP1方案

已关注

7 人赞了该文章

经‘永恒分母’团队同意，融360赛题三PPT及解题思路开源于‘菜鸡的数据科学进阶之路’专栏。

永恒分母团队由清华大学、北京大学和武汉大学三位大佬组成，本赛题用简单的两个5折即获取AB榜的TOP1(其实一个5折也位于TOP1），充分说明模型不在融合而在于数据和特征，队长清华大佬将赛题三代码开源于个人GitHub，欢迎关注

答辩目录：

一.赛题分析：

在实际金融场景中，我们常常会上线一些新产品，但对于新产品来说，建模样本必然是不足的，那么我们想要尽快的完善新产品的模型会考虑尝试利用其他产品的样本与新产品样本的共性的部分，来加强模型的预测能力。于是多场景金融赛题应运而生。

赛题分析

- 赛题任务：对用户各类信用相关数据进行分析处理，挖掘数据价值，形成建模特征，预测用户逾期的概率。
- 训练样本：包括从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，并给出每个样本的类型（属于大额分期贷或小额现金贷产品）是否逾期。（约10万样本）
- 验证样本：2018.1.1到2018.5.1机构A的产品，验证集不提供样本是否逾期，参赛选手自行完成是否逾期预测后，可以提交至比赛平台评估结果。（约2万样本）
- 测试样本：与验证样本来源相同且同分布。测试集不提供样本是否逾期，参赛选手只能在比赛最后的评比阶段将预测结果提交至比赛平台评估，且只能提交一次。（约2万样本）
- 赛题特征：时间特征和大量的匿名特征。
- 评价指标：AUC

二.数据处理：

数据特征的选择相对也比较常规，剔除缺失比例大于70%的特征，剔除值唯一的特征，保留高相关性特征中的一个，剔除feature_importance为0的特征。

数据处理

以下是针对特征的缺失比例做的统计

02 数据处理——缺失值

表1 列缺失值比例下的特征数

列缺失值所占比例	特征个数	特征总数
0.6	1137	6812
0.7	696	6812
0.8	201	6812
0.9	90	6812
0.95	42	6812

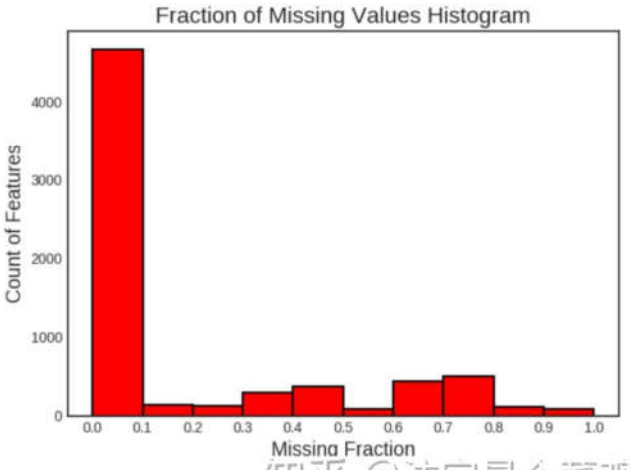


图1 不同缺失比例下的特征个数

以下是对特征为单一值得统计

02 数据处理——单一值

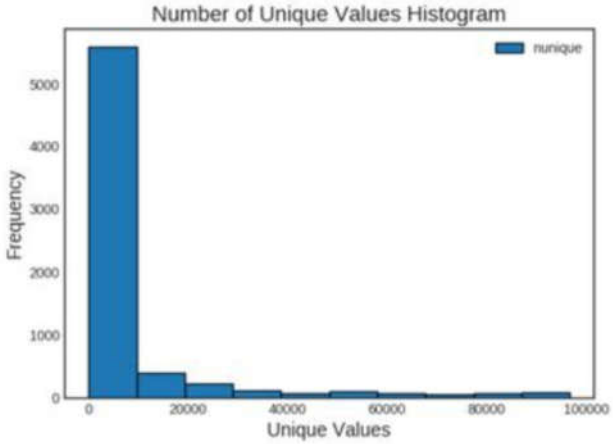


图2 不同独特值个数下对应的特征数

经过统计发现存在26个特征只有一种值，

这种特征不存在任何的数据价值应该剔除。

特征相关性图

02 数据处理——去除高相关性列

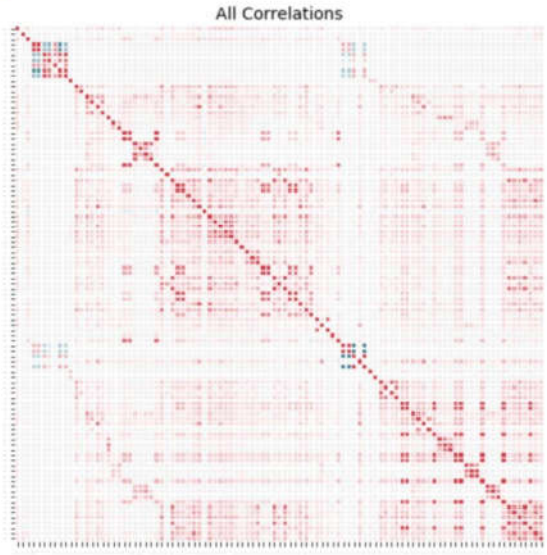


图3 特征相关性图

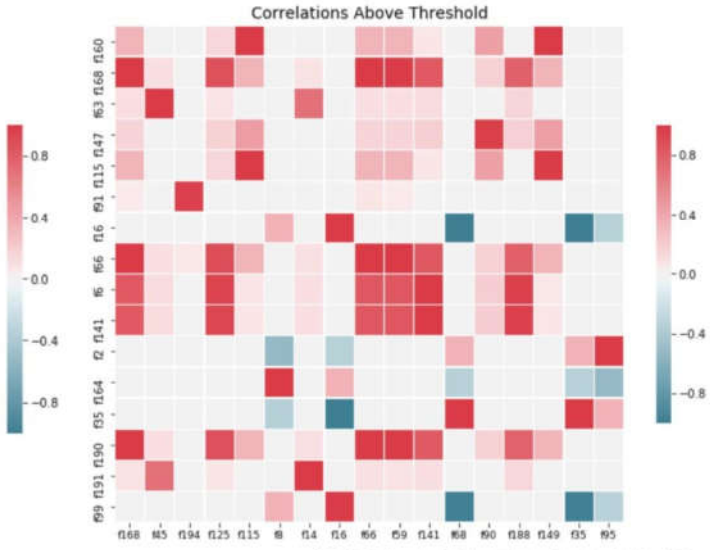


图4 高相关性特征图

特征重要性图

02 数据处理——去除零重要性特征

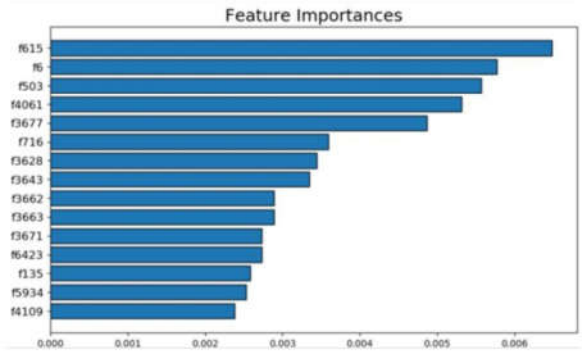
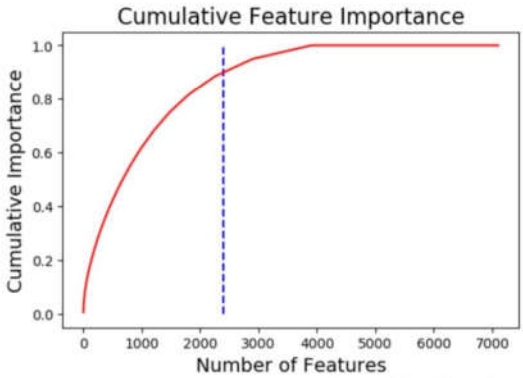


图5 特征重要性(top15)

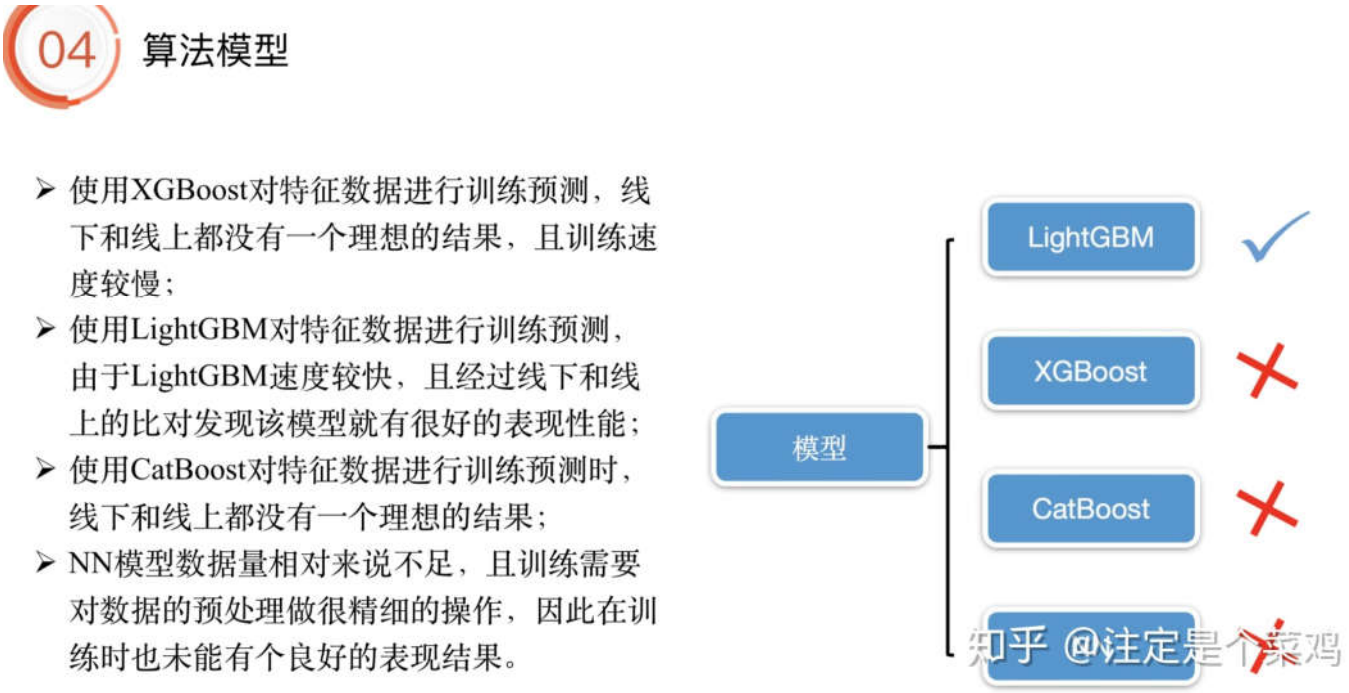


要想达到90%的分类性能只需要大概2400个特征

图6 特征数累计重要性

三.特征工程：针对匿名特征，采用启发式的特征交互，以及对时间的周期做分解。

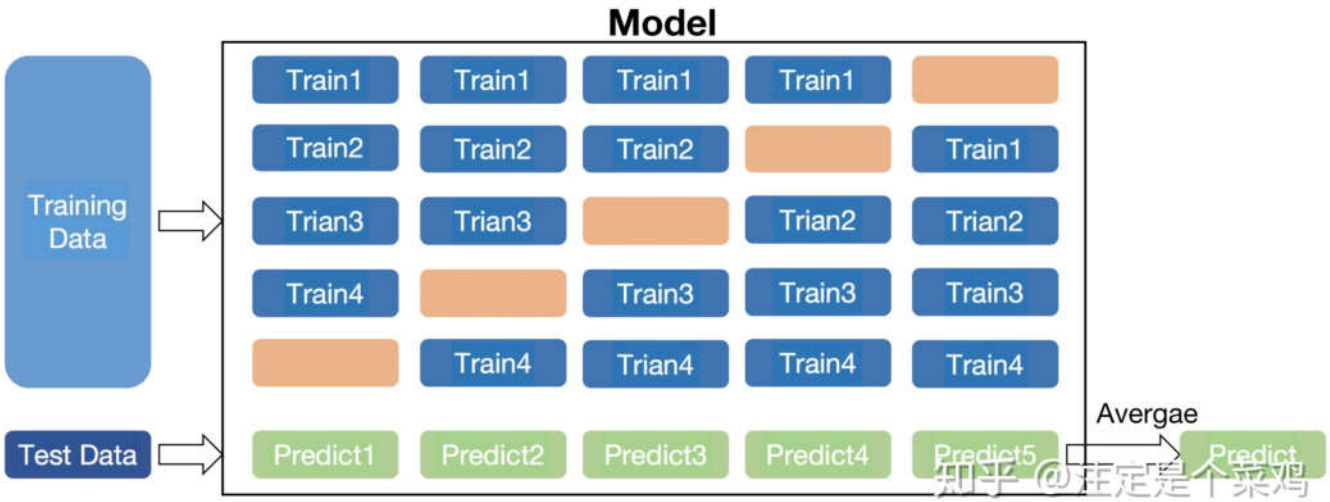
四.算法模型：多模型尝试，最终选择最优模型



融合：采用单次K折，最后分别将有二次特征工程和没有二次特征的两模型进行融合

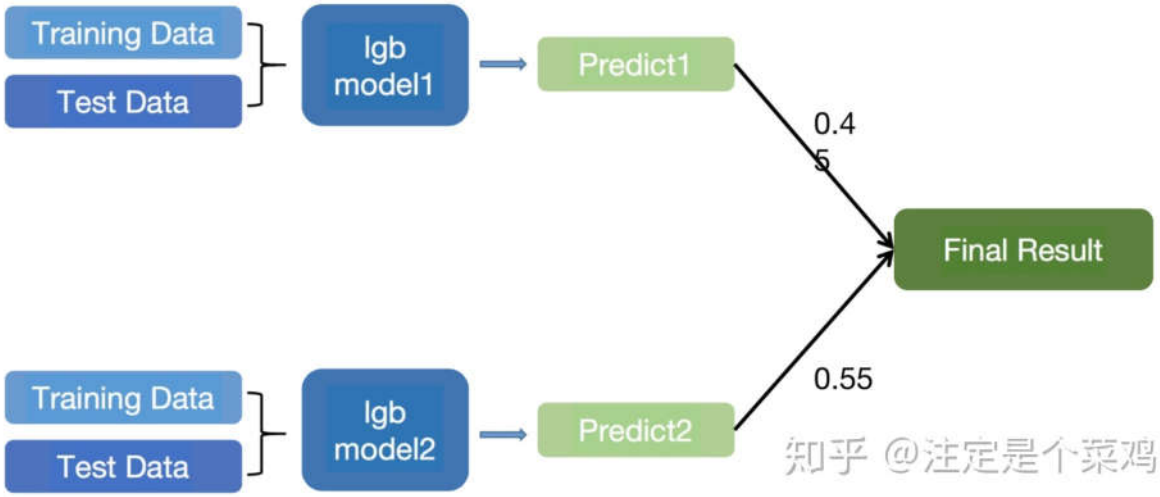
04 模型训练

使用LightGBM 模型并用5折交叉验证构造出5个模型，将预测出来的结果进行加权平均作为单模型最终结果，保证模型训练结果的稳定性



04 模型融合

- **lgb model1** 表示使用数据预处理得到后的特征直接进行训练(验证集单模可到第3)。
- **lgb model2** 表示使用数据预处理得到后的特征+特征工程构造的特征进行训练(验证集单模第1)。



本赛题均为常规操作，用两个5折即取得了AB榜同时第一，也侧面证明了数据的预处理在模型中的重要性。所以在玩数据挖掘的时候应该更多把时间花在数据处理和特征工程上面，这样才能更加稳健的构建出最优模型，且对于落地也较为容易。

编辑于 2018-12-09