消费金融场景下的用户购买预测方案与总结

一、比赛介绍

奇点计划——消费金融场景下的用户购买预测

二、最终成绩

队伍名：UAU
成员：Sam、Jason
A 榜分数0.86838，排名 68/1586
B 榜分数0.85744，排名 199/1586

三、问题简介

本题最终目标在于给定用户的部分个人信息、信用卡信息（已脱敏），及部分用户一个月内的APP行为数据，预测用户在未来一周内是否会在掌上生活APP购买优惠券。本方案使用常见的二分类模型进行建模预测。

四、数据描述

1.文件说明

目录	文件	说明
train	train_agg.csv	个人属性与信用卡消费数据
train	train_log.csv	APP操作行为日志
train	train_flag.csv	标注数据
test	train_agg.csv	个人属性与信用卡消费数据
test	train_agg.csv	APP操作行为日志

2.字段说明

个人属性与信用卡消费数据。
USRID表示用户ID。

v1，v2，...，v30表示个人属性和消费数据，包含枚举型特征和数值型，已脱敏。

3.标注数据

USRID表示用户ID。
FLAG表示未来一周是否购买APP上的优惠券，0表示未购买, 1表示购买。

4.APP操作行为日志

USRID	客户号	
EVT_LBL	点击模块名称	已经清洗并编码 ,点击模块名称均为数字编码（形如231-145-18），代表了点击模块的三个级别（如饭票-代金券-门店详情）
OCC_TIM	触发事件	用户触发该事件的精确时间
TCH_TYP	事件类型	0:APP, 1:WEB, 2:H5

五、详细建模步骤

1.数据概览/分析

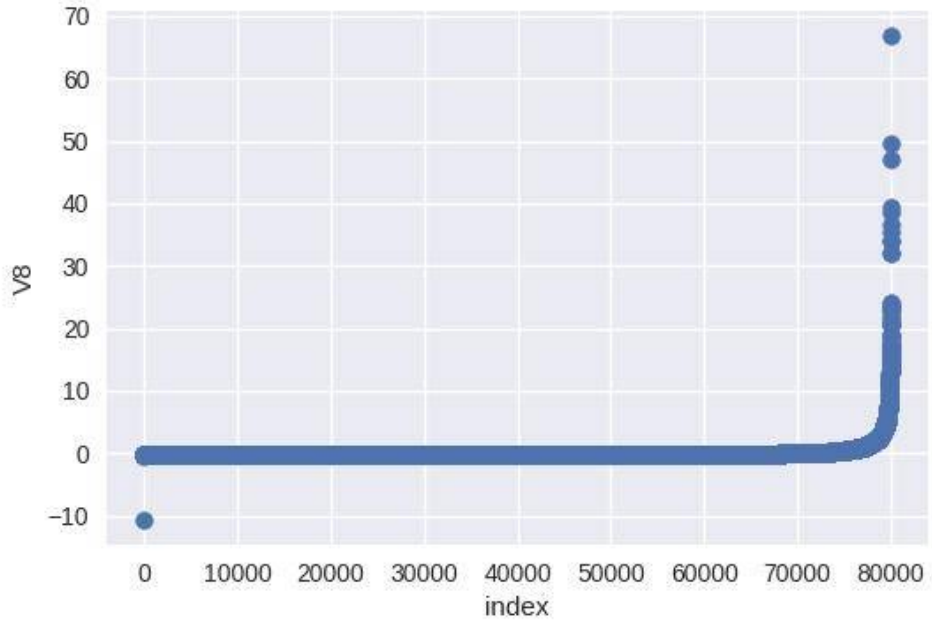
记录数

文件	说明	记录数
train_agg.csv	个人属性与信用卡消费数据	80000
train_log.csv	APP操作行为日志	3533818
train_flag.csv	标注数据 |	80000
test_agg.csv	个人属性与信用卡消费数据	20000
test_agg.csv	APP操作行为日志	891414

agg表各变量的取值个数

取值个数小于30的变量	取值个数大于30的变量
V2、V3、V4、V5	agg中其余变量

对于一些取值个数较小的变量，可以猜测其为离散型变量，其余的视为连续型变量。故本方案仅将V1、V2、V3、V4视为离散型变量，其余的变量视为连续型变量。后面未对其进行onehot编码和特征交叉，应该要做。对于这里的一些连续型变量，观察数据分布。例如，对train_agg中的V8，将变量值升序排序，画出曲线图:



目前观察到的数据分布情况，只在后续的异常值剔除中应用。其他变量进行同样的分析。

FLAG的分布情况

FLAG=1:FLAG=0 = 3176:76824

正负样本比为24.19，正负样本极不均衡。

2.数据预处理

以train_agg的数据作为原始特征，训练Xgboost模型，得到特征重要度。针对特征重要度排名top 5的特征，剔除异常点。异常点可由数据分析步骤得到。

3.特征工程

用户侧特征

基础属性（部分原始特征，部分进行独热编码）

信用消费数据（原始特征）

交互侧特征

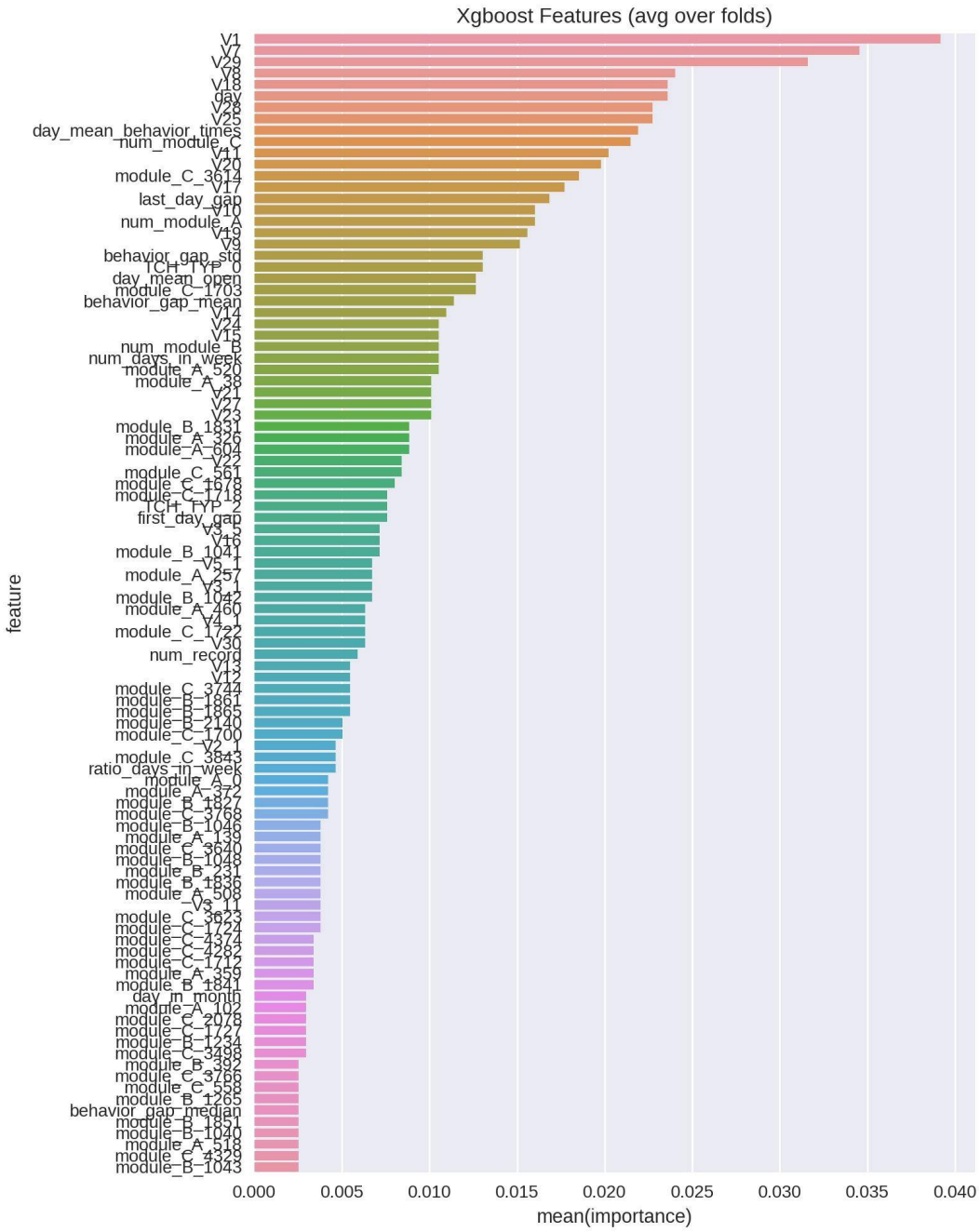
共提取特征255维，详细如下：

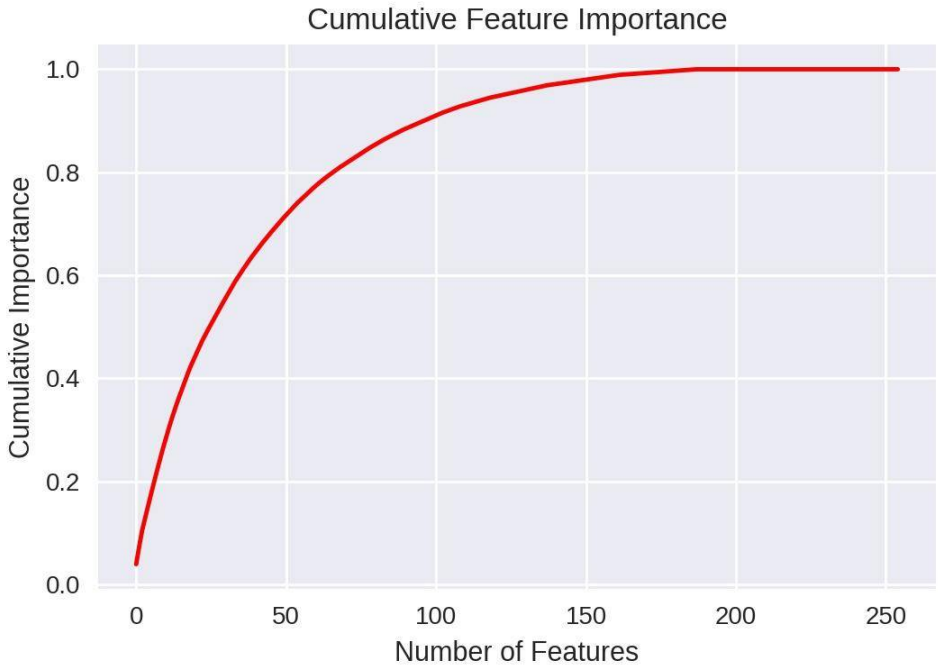
- 对V2 V3 V4 V5做独热编码，V6变量等距划分5个区间后进行独热编码，agg表中其余特征直接使用原始特征
- 用户对点击过的模块A的去重数
- 用户对点击过的模块B的去重数，数量大于训练集中所有用户的数量的平均值+5倍标准差的过滤掉
- 用户对点击过的模块C的去重数，数量大于训练集中所有用户的数量的平均值+5倍标准差的过滤掉
- 用户对模块A集合中每个模块的历史点击次数
- 用户对top100热门的模块B集合中每个模块的历史点击次数
- 用户对top50热门的模块C集合中每个模块的历史点击次数
- 用户月内最多连续n天存在行为
- 月活率，月活跃天数/31

- 用户月内活跃等级，根据月活率划分
- 最近一周的周活率，最近一周的活跃天数/7
- 用户最近一周的周活跃等级，根据最近一周的周活率划分
- 用户首次行为出现的时间距离月末的天数
- 用户最后一次行为出现的时间距离月末的天数
- 用户有行为的日期的间隔天数的均值、中位数、标准差
- TCH_TYP中不同类型的历史使用次数
- 用户月内日均的使用的小时数
- 用户月内日均的行为次数

4.特征选择

本方案使用基于模型的特征选择方法，保留top n个特征。使用全部的训练数据进行一次训练，得到特征的重要性：





选择其中的top 78的特征（根据最终效果选择）进行最终模型的训练，选择部分特征进行降维，达到降低计算量的目的，还可以剔除噪声特征，线下cv分数也有所提高。

5.模型调参

调参步骤:

step 1 通过xgboost自带的cv，设置学习率，设置一个较大的迭代次数，最后根据输出的结果得到最佳的迭代次数。

step 2 利用GridSearchCV，选择一组max_depth和一组min_child_weight进行交叉验证训练，得到最优的max_depth和min_child_weight。max_depth一般不用太大，防止模型过于复杂。

step 3 调整gamma进行剪枝。

step 4 选择效果最优的subsample和colsample_bytree，一般取稍小点，可防止过拟合。

step 5 正则化参数调整，包括L1正则reg_alpha和L2正则reg_lambda。

step 6 降低学习率，同时增加迭代次数，让模型学得更精细。（容易过拟合）

6.模型效果

模型	线下cv分数	线上分数
Xgb_M1	0.86066	未直接提交
Xgb_M2	0.86053	未直接提交
Xgb_M3	0.86088	A榜0.86838
Blending	0.86121	B榜0.85744

Blending 主要将Xgb_M1、Xgb_M2、Xgb_M3的预测结果通过LR再进行一次训练，得到的模型用于预测最终结果。

六、问题汇总及改进建议

1.数据合并

本次比赛在进行数据分析时，仅使用了训练数据，会导致离散型变量分析不到位（例如训练数据和测试数据会有不同的取值），同时后续特征工程时需要分两次进行，增加了冗余的操作和计算量。建议下次将训练数据和测试数据合并用于做数据分析和后续的特征工程（添加训练数据标识和测试数据标识即可），注意训练数据的特征的构造不能引入测试数据的信息。

2.异常数据剔除

本方案分析了多个用户的个人消费数据特征，并利用模型跑出特征的重要性，从中选择了前n个进行剔除，剔除的同时考虑是否对最终分数有提高。建议在后续都进行剔除，或者使用排序特征，增加模型的鲁棒性。如果数据存在缺失值。可以根据缺失值的分布情况来决定是否剔除，缺失特征较多的用户剔除，覆盖用户较少的特征剔除。

3.数据划分

本比赛的正负样本分布极不均衡，正负样本比为1:24。本方案使用了GridSearchCV进行模型的交叉验证训练，但是未指定每轮验证的抽样方式，意味着每次训练时，训练数据和测试数据的分布，随着抽样的不同，可能与原始数据分布产生较大的差异，导致训练出来的模型泛华能力交叉。建议下次使用更好的数据样本抽取方式：

（1）使用GridSearchCV时指定使用分层抽样

只需要在使用GridSearchCV时指定cv=StratifiedKFold(train['FLAG'], n_fold=5, shuffle=True)

（2）easyEnsample

easyEnsample介绍

example

4.特征有效性验证

在模型建立时，一次性的生成数百维的特征，再放入模型进行训练，模型有提升，说明特征有效，否则无效。是否能通过比较有效的方法，在数据分析阶段或特征工程阶段，就能初步判断该特征对auc有提升作用?

5.特征选择

本次比赛的特征选择只使用一次模型训练的特征重要性结果进行选择。不同参数训练出来的模型特征重要性不一样，建议后续使用多组参数进行训练，特征重要度进行平均并排序的结果，作为选择的依据。

6.建立自己的Baseline

我们队在建立模型时没有做这一步。baseline为后面使用更高级的算法进行模型训练、添加新特征后的模型训练提供进行比较的标准建议议根据baseline在线下线上的分数作为标准，判后续的优化是否有效。

7.模型调参

参考上面阐述的模型调参步骤。目前主要存在的问题在于，人工干预过多，时间成本较高，同时调参效果不理想。建议使用BayesianOptimization

8.模型过拟合

问题定位：模型训练误差仍在减小，但测试误差开始增加。在训练集上的训练准确率远大于在验证集上的准确率。我们队伍在建模时，为追求本地更高（提升幅度较小）的cv，没有选择更简单惩罚项更强的模型，导致后面换榜之后的过拟合。
如何解决：
（1）交叉验证，根据测试结果选择最终模型
（2）加数据
（3）提前停止迭代
（4）特征筛选，使用更少的特征
（5）加正则化、选择更简单的模型
（6）融合
（7）训练数据和测试数据分布尽量保持一致，如何操作？

9.多模型对比

建议添加对比多个模型效果的步骤，便于从中选择更好的模型，也为后面模型融合做准备。
参考代码

10.模型融合

本次尝试了使用不同参数效果相近的同一个模型进行Stacking，线下auc有提高，线上b榜也有提升。建议下次训练多个效果相近的不同模型进行融合，尽量使用差异较大的特征，差异较大的最后融合出来效果理论上会更好。

七、关于团队合作

- 关于团队成员数，我们目前成员2名，Jason在江西读研中，本人目前在北京就职，不过是同届，讨论交流没有代沟。
- 第一次合作，最初分工是Jason负责特征工程，本人负责模型部分。后来本人也加入到了特征工程中，提取了不少特征（哈哈，好像把Jason的活干完了）。但由于是第一次合作，默契度已经磨合得不错。不过下次需要提前分好工，计划是在后续的比赛中，每个成员都先独自进行数据探索、特征工程和模型构建（先以个人身份参赛），比赛后期再进行组队，融合各自的模型。
- 我们做比赛的主要目的在于学习。赛中的交流和赛后的总结，希望尽量能使每个队员都参与其中，了解建模的每一个步骤。
- 越到后面，成绩的提升越是越要时间来堆叠。虽然成绩不理想，但是学到的知识还是很多。