**task1**

**1. 特征**

(1)特征选择

初选：app-cat: 211 dimensions，删除空值多、训练集与测试集的分布差异大的特征

精选：ignore\_column = name:app\_cat2\_43,app\_cat2\_191,app\_cat2\_143,app\_cat2\_14

每次尝试删除feature importances中非“不能删除标记”的首个特征，如果删除后，性能提升，则记为可以删除的特征。

不稳定特征：手机型号、厂商、OS\_release。

(2)特征构造

**ip特征：0/1的特征**

背景：task1 train数据和test 数据中都有ip 相同的数据， 比如train 数据中 有 某ip 为 192.168.0.1 的数据有3条，把train 数据中有相同ip的数据筛选出来，1的比例并不低。

处理：筛选出来的ip相同数据中，手机型号也完全相同的数据有 1400多条，1的比例只有0.03 左右，纯度比较高，所以将 train 数据中是否有 ip相同并且手机型号也相同作为特征，test 数据同样找出这个特征。认为同一数据中出现ip 相同并且手机型号也相同的数据是作弊数据，作弊者不会购买理财产品。

后续改进：虽然ip 相同和手机型号也完全相同的数据中 纯度较高，但还是存在标签为1的数据，所以经过分析，标签为1的数据部分原因在于 虽然ip 和手机型号相同，但是app 特征不一致，证明两个数据并非同一个用户，所以将 ip相同并且手机型号相同 的条件改为 ip相同并且 app特征完全一致。1的比例下降了一半左右。

通过分析生于标签为1的数据，发现标签为1 的数据，只要有经纬度，大部分数据经纬度不一致所以这个特征最终定义为 ip相同并且app特征完全相同，并且只要有经纬度，经纬度要相同。

**2. 分类器融合——**gbm+rf+lbt

(1)classifier1: lightgbm (microsoft github, c++)

(2)classifier2: random forest (scikit learn, python)

(3)classifier3: loss boosted trees (c++)

(4)combination of classifiers: rank average, by excel

rank => 0~1:rank/test\_num => average: (rank1+rank2+rank3)/3

没有任何后处理！