****

课 程：2018软件工程综合实训

项 目： 数据挖掘实训项目

院 系： 数据科学与计算机学院

专 业： 软件工程

学生姓名： 杨金辉

学 号： 15331357

授课教师： 郑子彬

2018 年 6 月 16 日

# 目录：

1. 比赛介绍
2. 赛制介绍
3. 评分算法
4. 数据来源与使用
5. 比赛流程
6. 数据预处理
7. 特征工程
8. 模型调参
9. 本地CV
10. 个人总结
11. 参考文献

# 一.比赛介绍

## 1.赛制介绍

基于移动网络通讯行为的风险用户识别

参赛者根据训练集中的用户风险标签及通话、短信、网站/App访问记录数据，开 发风险用户判别模型，对测试集用户进行预测，并提交预测结果。

## 2.评分算法

对于参赛者提交的预测结果，按如下公式计算得分：

score=0.6×auc+0.4×F1

其中：

auc值为在测试集上，根据预测结果按照标准auc定义计算的分值；

F1值为针对测试集中实际标签为1（风险用户）的用户，根据预测结果，按照标准 F-measure定义计算的分值。

## 数据来源与使用

提供45个连续自然日期间，抽样模拟的9999个用户每天的通话、短信、访问网站/App记录的脱敏数据，基于这些用户的移动网络使用行为，判别用户属于风险用户的可能性。

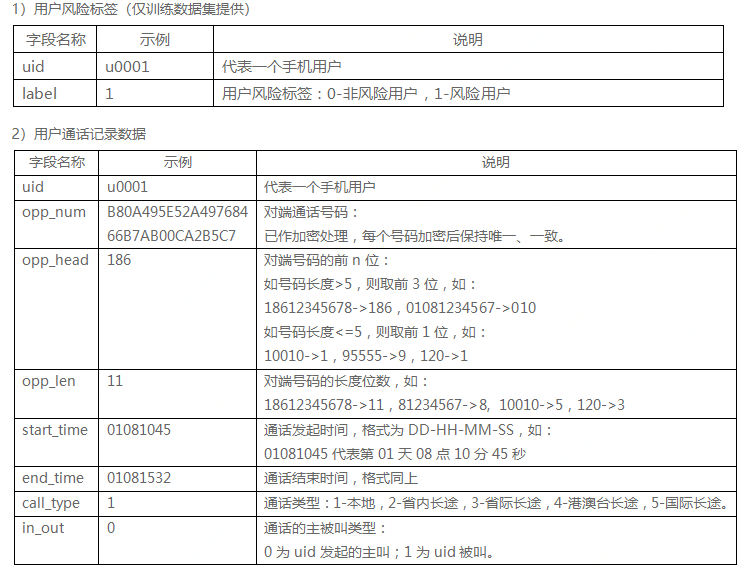
用户id按照u0001～u9999编号。其中：

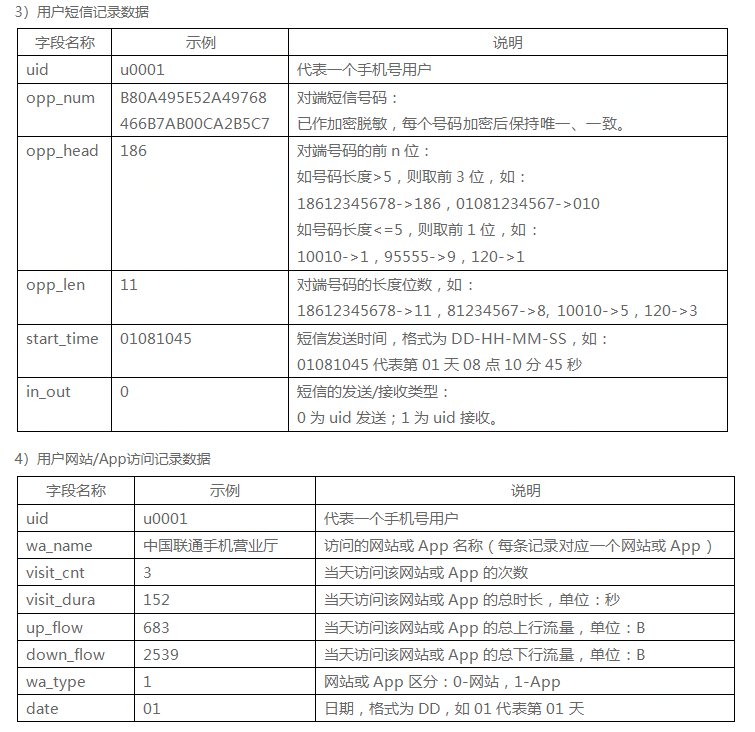
• u0001～u4999用户作为训练集，对每个用户给出是否为风险用户的标签（0-非 风险用户，1-风险用户），在赛事启动时下发；

• u5000～u6999用户作为初赛阶段测试集A，不带用户风险标签，在初赛阶段下发；

• u7000～u9999用户作为复赛阶段测试集B，不带用户风险标签，在复赛阶段下发。

数据表中部分列存在空值或NULL，少量uid可能在部分记录表中未出现（代表该用户在此期间没有相应的通话、短信或上网行为），均属正常情况，请参赛者自行处理。





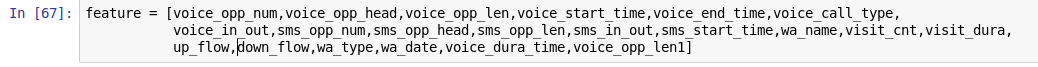
# 比赛流程

## 数据预处理

对比赛所给出的训练集数据进行筛选、排序等

## 特征工程

这里经过筛选选出了以下特征



其中：

一些普通数量特征如voice\_opp\_num：



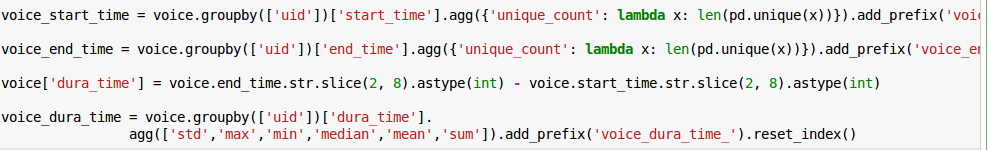
一些数量特征用onehot如voice\_opp\_len：



一些涉及维度的统计特征如voice\_opp\_len1：



除此之外，还有一些无法直接获取的特征需要通过自己操作获 得，如通话时间长短：

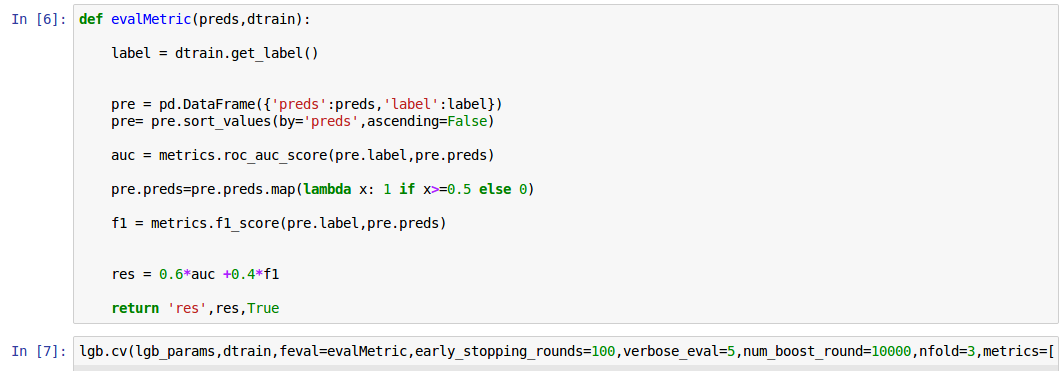


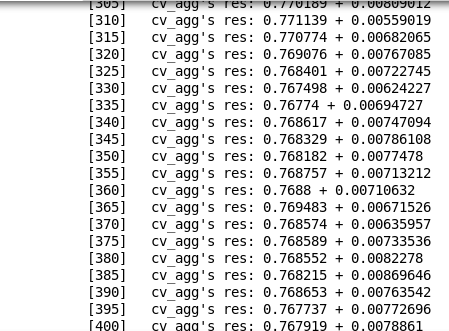
## 模型调参

这里使用了LGB聚类算法，总体来说给出的参数已经算是比较适合的，因为之前微调了好多次都没什么提升，就只有将学习率调到0.09



## 进行本地CV





# 三.个人总结

经过本次比赛，我学到了不少东西，过程中也碰了不少壁。比如在一开始因为刚接触python，很多库都不会用，导致进度一度停滞，后来查阅了相关文献之后发现其实python非常好用，除此之外，一开始不知道怎么下手，后来ta发了baseline之后有了一些头绪，学习了baseline中的方法来构建特征。如对于一些类别的型的可以用onehot，一些涉及维度的可以做统计特征，自己摸索并且筛选可能会有用的特征，可能由于没什么经验，效果一直不太明显，但是总算还是有提高，也发现特征不是越多越好，过拟合会导致分数下降。我觉得这次比赛对我之后学习数据挖掘的课有着很重要的作用，在之后的学习中我还会继续钻研，争取进步。

# 四.参考文献

<https://blog.csdn.net/u014253445/article/details/78512101>（数据预处理总结）

利用Python进行数据分析(Python For Data Analysis中文版)

http://www.sohu.com/a/210721214\_609569（主流机器学习模型模板代码+经验分享）