1）概述

A.本次比赛，采用python3.6+lighgtbm模型。

B/C特征选取

**特征描述：**

本次赛题要求从用户通话/短信/上网记录中识别‘风险’用户，赛题提供用户于过去45天内的行为日志，因此，我们队伍分别通过对用户的短信日志，通话日志和上网日志三个维度描述用户特征，通过特征描述用户，构建用户画像。

特征分析：

1. 通话日志

字段：通话号码（脱敏）

通话号码的长度/区段，通话类型（地区/主叫被叫）

通话起始时间和结束时间

1. 通话时长特征（用户通话时间的统计特征，最大值，最小值，均值等）
2. 通话间隔特征（用户下一次，下两次通话的时间差特征）
3. 尝试对用户通话号码进行LDA/SVD降维度处理（最终结果未采用）

解释：

用户1 号码1 号码2 号码3 号码4

用户2 号码1 号码2 号码4 号码5

用户3 号码1 号码2 号码6

可以采取文本处理的只是，把号码序列当作文本，用NLP的技术处理，得到向量。

这个部分仍可以引入聚类，去给用户贴标签，但是由于个人时间，未能实现。比较遗憾。

1. 构造了类似于协同过滤矩阵的特征

解释：采用pandas的透视功能，构造 号码长度 和 号码区段的特征矩阵，具体可参考：pd.pivot\_table函数的解释。

1. 用户-通话号码的出度与入度特征

解释：将用户和通话号码看作两个节点，分别可以计算与用户相关通话号码的数量关系，同理，计算号码关联的用户的数量关系。

1. 进一步交互特征（经过测试，该问题表现不好）

例子：对通话号码的频次和被叫时常统计，之后在根据用户对通话号码的关系，进一步统计用户都通话号码特征的特征的统计

1. 时间衰减特征

考虑到用户行为的时许行，使用46-当前天数的倒数作为全职，统计用户行为的时候，进行加权处理。也可以直接对权值求和，反馈用户的活跃天数（经过测试，这个问题效果不好）

1. 短信日志

短信字段类型基本与通话部分类似，其特征构造基本一致。

1. 上网日志

其实这个部分提供的特征比较少，我们也只是采用了通话日志特征中的2，4构造了简单的特征。

不足：对于用户与上网记录/通话记录/短信日志构造的协同过滤矩阵，未能进行分解，交互等操作。

最终模型特征维度约 400 维度，线上成绩约0.869附近。其中分析特征重要度，发现存在几个号码段的特征重要度明显高于其他号码。考虑，可以对号码关联的脱敏号码进行聚类，可能分析出更多关系。很多想法，鉴于时间和当前水平有限未能完全实现。例如，借鉴word2vec思想，对号码段进行向量化处理等，文本处理的思路，都可以应到这次的题目。用户历史本身可以理解为行为的文本描述。

2） 说明文档

将文件保存在与code同的data文件夹，并且创建submit文件夹，同时将5.baseline.py放入到code文件夹，直接运行即可得到运行结果。Python3.6运行环境，pandas，numpy，scikit-learn，lighgbm.

3）有趣的发现

A. 使用的小技巧？

借鉴了协同过滤矩阵，可以获得更加详细的用户对不同类型对象的操作描述，但是由于其耗费空间较大，因此预先对号码段进行第一次预测后，将特征重要度较低的字段合并。其实我们借鉴了很多NLP的相关经验，去解决用户行为的特征提取，认为用户的行为历史可以是一段发生的文字故事。

类似的人有类似的历史，类似的历史聚类类似人的特征。

B. 您觉得您最突出的优势是什么？

立足于业务思考，尽可能还原用户每天的操作行为。但是本身作为学生，业务理解比较浅显，特征并没有做深曾挖掘。但是队友们都是工作许久，可以从业务和社会经验给我一些指导和帮助。

C. 每次成绩提升改进说明。

这个部分基本在特征描述部分已经叙述，按照特征描述的顺序，基本是每次尝试的结果。

4）其他您想分享的

多尝试并且立足于业务思考特征和描述特征。组队交流可以进一步扩展思路，不同领域的队友交流过程可以发现更多有意思的东西。