**数据集介绍：**

**解压密码：bjlt**

1.用户轨迹链数据，包括用户标识、业务发生时间、业务离开时间、LAC、CI、停留时长。

2.用户对象数据，包括用户标识、性别、年龄、终端品牌。

3.基站对象数据，包括LAC、CI、经度、纬度、方位角、高度、归属行政区。

4.北京中高风险地区数据，包括LAC、CI、街道名称、归属行政区、风险等级。

(1-4是特征分析，做数据清洗,挖掘)

6.用户标签数据，包括是否漫入怀柔、是否漫出怀柔、是否为怀柔常住用户、是否高风险区域返回、是否途径高风险区域。

（标签数据用于预测）

赛题说明：是否漫入怀柔、是否漫出怀柔、是否为怀柔常住用户、是否高风险区域返回、是否途径高风险区域为预测类指标，比赛时，提供含有这些标签的训练数据集，其中验证数据集不提供这些标签，参赛团队需要根据训练数据集训练模型，对验证数据集的这些预测类指标进行预测。用户出行时长、用户出行距离、频繁去向目的地、漫入/漫出怀柔区人群规模等为分析类指标，参赛者可以根据提供的数据集进行数据挖掘，采用算法模型进行个性化分析，丰富重点人群的画像信息。

工具包：1.<https://sklearn.apachecn.org/#/>；

知识点：多任务学习(Multi-task Learning)，机器学习，多任务分类

用户画像：

1. 根据用户轨迹链数据和用户对象数据：可以分析用户年龄段，用户使用的手机，用户在哪些时间内有的流量使用情况；
2. 根据用户轨迹数据和基站对象数据，可以分析用户曾经去过北京哪些行政区
3. 根据用户轨迹数据和北京中高风险地区，可以分析用户去了北京哪些中高风险街道
4. 根据用户轨迹链数据和基站数据，可以分析用户的住所和工作所在的区
5. 根据用户轨迹数据，基站数据，北京中高风险地区，可以分析用户住所，工作所在区的具体街道。
6. 用户出行时长，可以从用户轨迹数据中统计。
7. 用户出行距离：（是否需要从经纬度进行计算，可以再研究看看）
8. 频繁去向目的地，可以从第234的特征中分析出。

根据上述特征做数据分析，筛选特征，预测是否漫入怀柔、是否漫出怀柔、是否为怀柔常住用户、是否高风险区域返回、是否途径高风险区域。