模型在测试集上整体准确率为83%

其中，

女性标签召回率为85%

男性标签召回率为80%

当前目录下共3个文件夹

1. code： 从input到最终output输出的全过程代码记录

2. input: 包含输入模型的原始数据（行 = 39655，列 = 5）

3. output: 包含最终输出的模型

数据集说明

数据集中包含5个字段： coa.cookieid: cookieid

a.start\_time: 装机首日，首次浏览物品时间

a.end\_time: 装机首日，最后一次离开浏览物品页时间

a.duration: 装机首日，浏览页面总时长

levelsall: 装机首日，物品浏览日志（包含物品的3级类目&goodsid）

每个字段的来源说明如下

1、a. cookieid, a.gender, 来源于ods. appflyer\_install\_log

2、a. start\_time, a. end\_time, a.duration来源于ods.page\_view\_log

3、levelsall 来源于ods.event\_log & dim.dim\_goods

特征工程方面

扩展的特征主要来源于, levels all & a.start\_time这两个字段

1、通过levels all，进一步获得了

'NumProduct’： 浏览物品总数

'FirstCat' : 当日浏览首件物品，一级类目

FirstSubCat' ：当日浏览首件物品，二级类目

'FirstThridCat' ： 当日浏览首件物品，三级类目

'FirstFourthCat' ：当日浏览首件物品，goodsid

'UniqueFirst’: 一级类目（不同&计数）

'UniqueSecond’：二级类目（不同&计数）

'UniqueThrid’：三级类目（不同&计数）

'UniqueFourth' ：goodsid（不同&计数）

'MostFirstCat' ：访问最多的一级类目

'MostSecondCat’：访问最多的二级类目

'MostThirdCat’,：访问最多的三级类目

'MostFourthCat‘： 访问最多的goodsid

2、通过a.stat\_time，获得了字段

'weekday’: 指一周内的第几天

‘startTimeHour’：指开始时间，以小时为单位

‘endTimeHour’：指结束时间，以小时为单位