付费提升评估报告

# ⼀、项目介绍

* 1. 项目背景

重庆市弹窗运营目前使用热门推荐，即将近期热门节目推给用户，用户感兴趣就会点击触达，最终选择付费。而在用户高触达率的情况下订购转化率仅为触达率的千分之几，用户粘性差，消费意愿低。

* 1. 项目目标

本项目通过向用户推荐可能喜欢的产品包，增强用户的付费意愿，最终达到提升触达用户的付费转化率的目的。

# ⼆、整体流程

### *数据获取*

数据来源为省公司，主要为用户收视行为数据和订购数据，用户行为数据包括但不限于用户的设备信息、观看的节目信息、观看时间、观看次数、节目类型等等，用户的订购行为数据包括订购产品包、产品包类型、产品包订购时间等。多维度的数据能够使模型效果更好。

### *数据清洗*

清洗用户的观看行为，去除掉无用字段和空指标，同时对用户的设备标识去重，处理掉部分缺失值。对订购行为数据清洗，主要为订购时间的格式清洗统一。最终目标是建立两张宽表，然后通过用户的设备标识关联，得到一张大宽表。

### *模型建立*

采用ALS算法模型，该算法属于协同过滤算法的一种，混合了User-ItemCF,并使用隐式反馈在其基础上拓展。同时确定了将喂入模型的各维度指标，确定对应关系，导入到模型中。

* 1. *模型训练*

按照8:2的样本比例切分出训练集和验证集，筛出无效验证集，最后将清洗好的数据喂入模型中。开始训练。

### 模型调优

通过诸项指标确定模型的效果好坏，然后调整模型里的各项参数，以使效果更好。

### 结果验证

通过交叉验证、精准率、召回率、f1score等方式指标验证结果。

# 三、数据基本情况

3.1 数据情况

选取10天数据共1114312名开机用户，86950个观看过的节目，120个产品包，数据集中用户包括用户id，用户观看时间，观看节目名，订购产品包，产品包类型等相关信息。

同时为了数据的准确性，我们将拿到的订购数据与历史观看数据的用户设备标识作匹配，大约为60%的匹配率，即全量订购的用户设备信息有60%可以在收视行为中找到。

# 四、效果验证

4.1 实验结果

为了测量算法的准确度，本文主要基于F1-Score指标评估方法，针对处理后的数据集分为训练集和验证集，周期内全量数据按照8:2的样本比例作为训练集和验证集构建推荐模型。

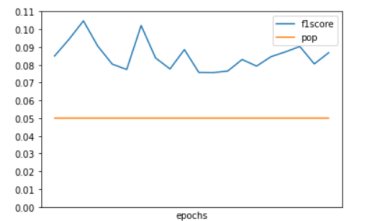
同时本文选择上述热门推荐和ALS两种算法作为比较方法，热门推荐只考虑训练集内播放次数最多的节目，并针对每个用户过滤掉其看过的节目作推荐。

图1 完整数据集下两种模型F1Score分数

该模型由20200801～20200810十天数据量经过清洗 提取全量用户行为数据和订购数据，通过数据集样本的走势大致推测两种模型的优劣

上图是热门推荐算法和ALS推荐随着迭代次数的增加其f1score走向图，有图可知，ALS推荐效果约为热门推荐的两倍。同时，该指标的含义为每推荐10个就有一个命中。

## 4.2 总结

本文提出了以ALS算法为基础、在其上拓展开来的混合模型，该模型针对用户收视行为数据和订购行为数据做清洗处理并建模，周期内的原始数据需要通过ETL处理成两个文件，一个是用户对节目的收视行为数据（包括用户设备标识、节目标识、观看次数、观看时长、记录时间），另一个是订购行为数据（包括用户的订购设备标识、产品包标识以及产品包类型）

准备完数据集后就开始编写对应的模型算法和评价指标，这期间碰到了一些问题，主要是数据量的过大引起的模型迭代时间过长的问题，办法是去除了一些不重要的维度，减少计算量，提升模型的精确性。

本文中的付费提升模型经过多次迭代训练的情况下，其效果约能够提升为之前热门推荐的1.5-2倍效果，其对所有用户订购产品包进行打分评估，分数区间在0~1，分数越接近0，订购该产品包的可能性越小；分数越接近1，订购该产品包的可能性越大。

在后续的版本迭代中，会尝试利用用户画像标签数据来训练模型，同时会采用其他的一些深度学习模型来综合评价各个模型之间的优劣，找到更好效果的模型应用进来。