演讲稿

大家好，我们小组本次project所选择的方向是电子商务数据分析，研究题目是基于用户数据画像的淘宝用户行为数据分析。我是汇报人袁恒宸，以下是我们小组成员。

那么这是我们这次汇报的一个大题结构，首先我们来进入第一个模块，我们本次项目的实践意义。众所周知互联网中带来便携的一个根本原因就是庞大的数据集，那么用户数据又是这其中最最机密的一环，其重要程度不言而喻。通过对用户的数据分析，可以让商家或者企业更好的拟定自己的合适发展方向，同时又能针对用户提供更加具有针对性的服务。这对电商行业未来的发展具有重要意义。

接下来简单介绍一下我们本次项目使用到的数据，我们所用到的数据来自于阿里云天池提供的数据集，原始的数据类型十分简单，只有五个字段，且其中有三个都是用于唯一确定对应种类的ID字段，其余的就是具体行为与时间戳。下面是我们对数据字段的命名（源数据是没有列名的），其中Betype这一列一共会有四种用户行为，分别是：。。。。

那么拿到这样一份数据，我们首先会对这个原始数据进行一系列的加工，即数据清洗阶段，才能更方便我们进行数据分析。

简单的去重和去空就不提了，我们这里主要强调一下我们做的几个特殊处理。

首先是数据集的压缩取样，因为我们本次的研究方向是基于用户画像来分析的，那我们就需要不同于传统的采样算法，而是基于用户ID为粒度筛选，保证筛选出的数据集达到我们的预期大小，同时又不丢失某一特定用户的任意操作。

可以看到左边使我们的原始数据，后面是我们采样后的数据，大小压缩了一百倍。为什么要有这样一个过程呢，是因为我们可以看到原先的数据集大小高达一个亿，在源数据集上的任意分析都需要耗费非常庞大的内存，仅仅光是读入数据都要好几分钟。因此我们会分批选择某一特定的用户段进行部分用户画像的分析。（这里我们取得是mod100==0）

接下来是对于时间戳的一个可读性处理，我们把原先计算机记录的时间戳转换为有实际意义的年月日，以及周几，还有一天中的时间段，这对于我们后面进行用户活跃时间的分析是很有必要的。

还有就是在细化了时间之后对时间的合理性进行了又一轮的筛选，我们可以看到，对时间这个字段进行分组后，大部分有效数据都是集中在这个区间的，那我们就可以舍弃其他的离散时间段，这些离散的时间段不仅对我们的分析没有作用，反而会起到干扰。

接下来可以看到我们数据处理后的成果，删除了无用及重复数据后，我们将原始数据集从3.7GB的大小压缩到了70mb，同时提供了更多的字段。

那么接下来就是我们最重要的信息挖掘整理部分了，这部分的代码量和细节处理太过庞杂，我会以一个比较高纬度的角度来概括一下我们的分析成果：那么首先呢这时我们基于用户和商品两个抽象层去构建的一个图网络。在商品部分我们简单的使用其基数（、、。）和特征（）作为分类标准，那么用户呢，我们细分了四个方面的指标，分别是。。。。。。

基于以上的网络，我们进一步的处理思路是这样的：对于用户的四个指标，我们进行了一次用户画像的构建与整理，同时基于RFM分类指标，使用K-MEans方法对用户进行了一个分类。对于商品部分，我们对其指标进行综合评分，这一部分的分类简单理解是可以找出我们所理解的（爆款商品，以及针对特殊类人群的特殊商品），作为我们最后商品推荐的一个补充。

我们之前所一直提到的用户画像是什么东西呢，就是根据用户的行为一层层的进行深度剖析，来描绘出他一个整体的特征性向，供电商平台进行参考。

一下就是我们经过对数据层层剖析， 得出来的一个用户数据画像，因为数据量实在太大，这里展示一小部分结果：

然后是我们对商品部分其中一个部分做的一个排名图，因为我们只能获取到ID但不知道对应的具体商品，所以这部分大家看个乐呵就行。

接着呢我们会对购物的一个整体情况做一个overall的分析，我们可以看到这个是随日期的一个用户行为的一个变化曲线，可以看到这个浏览量是断层的高，我们对下面一部分进行这个这个分析，可以看见在12-02和12-03这两部分有明显的提高，当然这个就是周六和周日，那么对应的11-25和11-26。。 数据原因。

然后是对一天之内的时间段操作做了一个曲线图，这部分的偏好也是十分明显了，在早晨的购物量是最少的，随着时间向晚上推移，这个活跃量也是逐渐达到顶峰。

其实对于像用户这种前后关系非常强的操作（比如浏览到加购到购买这一流程），我们对其进行了漏斗分析，这是一种基于业务流程的一种分析模型，具体可以看到这样，那么我们可以看见这个流程过程中，究竟是哪一部分造成最大的用户流失，这里可以清楚的看到，从浏览到加购这一过程中少了百分之90的用户。这也可以使得对应的商家/平台进行更有针对性的分析。

最后呢使我们根据之前的用户画像所进行的RFM分析，这个RFM简单来说就是对应的价值指标，比如。。。。我们对其进行简单的划分，同时对三个指标综合进行3维的K-means聚类，但是由于M在这里并没有在数据中体现，因此我们进行的是一个二维的聚类。那么最终的结果也可以从图中看到。

那么这一个环节也结束了，针对这有限的仅仅五列数据集，我们不仅对用户进行了一个丰富的数据画像，也对商品进行了一个简单的排序，同时又对该平台的一个整体情况进行了评估。

针对我们以上的成果，我们计算出通过用户画像获得的用户偏好以及对应商品所属类别（这一部分的处理我们是做了一个网络流图，那么同属一个商品类别的我们就认为其距离为0，其余的情况就设置为默认值，同时又对特定的指标进行了一系列的修正，如是否为复购用户等等，因为能参考的数据相关度有限），这样来我们计算出商品与对应用户的一个pearson相关系数，并乘上对应的特定性为偏好对应的权重，计算出来具体item与某一用户的相似度，并进行排序。

因为这部分的实现还稍有粗糙，我们选取了用户ID为xxx的用户进行了以上的分析，得出来的结果如下：