腾讯实习报告总结

Zirui Wang
Halıcıoğlu Data Science Institute
University of California, San Diego
La Jolla, CA 92093
zwcolin@ucsd.Edu

目录

1. 实习主要内容

- 用户分群
  - 给定用户的某些信息和偏好，如何对用户进行分群？
  - 用户的哪些信息可以进一步挖掘被用作分群指标？
  - 分群后如何对目标用户群做出精准定位和营销？
- 用户路径归因分析
  - 用户为什么会打开一个特定的页面、执行一个特定的动作？
  - 导致一个用户打开一个特定的页面、执行一个特定的动作的原因一定会每次都引导该用户打开相同的页面、执行相同的动作吗？

2.1目标定义

- 根据这次实习的主要内容和参考到的实习背景，我对于这次实习调研的目标定义如下：给定某个平台一段时间内所有用户的Log记录(包含用户的页面轨迹和行为轨迹)，以用户的页面轨迹和行为轨迹对用户进行分群，分群后每个群组内的每个路径点可以进行归因解释，并使分群后每个用户群体的路径归因有更好的同源性
- 关键词定义
  - 用户：以一个人或一个独立的个体为单位，在实验过程中通过每个人的ID体现
  - 页面轨迹：某个用户根据时间轴前进的访问页面顺序(例如：9点在主页，9点01分在某个商品的介绍界面)
  - 行为轨迹：某个用户根据时间轴前进的具体行为顺序(例如：9点在预览，9点01分在加入购物车)
  - 路径点：某个用户在某个时间点在某个页面做出了某个行为(因实验数据原因，每个用户在每个页面一定会对应一个行为。例如：用户A在9点01分的时候在编号为1001的页面执行了加入购物车操作)
  - 同源性：给定一个路径点A及一个归因长度N，该路径点A的所有归因路径点之后的N个长度内包含A本身的概率
- 同源性(动机单一性)是该模型中评价根据用户路径分群前后归因质量好坏的一个重要指标

2.2同源性的定义&在归因质量中的意义

- 同源性：给定一个路径点A及一个归因长度N，该路径点A的所有归因路径点之后的N个长度内包含A本身的概率。概率越大，同源性越好，该群组的动机单一性越强，宏观想法和动机越相似
- 因此，这里衍生出了两个验证用户路径分群模型显著性的指标：不加权归因回溯系数、加权归因回溯系数（此处的定义为末次互动的定义，即归因长度N = 1）：
  - 不加权归因回溯系数是指：对于某一个路径点A的所有归因路径点(B, C, …, D)，计算所有归因路径点中每个路径点之后的路径点为A的比例，最后不对每个归因路径点总出现数量的进行加权的情况下计算平均值。该平均值为不加权归因回溯系数
  - 加权归因回溯系数是指：对于某一个路径点A的所有归因路径点(B, C, …, D)，计算所有归因路径点中每个路径点之后的路径点为A的比例，最后对每个归因路径点总出现数量的进行加权的情况下计算平均值。该平均值为加权归因回溯系数

3. 背景论文参考

- The Broad View of Task Type Using Path Analysis (https://dl.acm.org/doi/pdf/10.1145/3234944.3234951)
  - 这篇论文提到了用户的路径行为与用户基本画像之间存在的一些潜在关联
  - 通过这篇论文验证路径分群的有效性：根据路径分群后每个用户组别的基本画像不是随机分布的，而会有一个规律，这里通过①展示每个组的年龄、性别分布和②将组标作为特征让一个XGB模型对用户年龄、性别进行预测，达到更高的准确率进行验证
- Your click decides your fate: Inferring Information Processing and Attrition Behavior from MOOC Video Clickstream Interactions (https://arxiv.org/pdf/1407.7131.pdf)
  - 这篇论文提到了可以根据用户的微观行为顺序进行建模反映用户的宏观想法
  - 这里由于在访问归因分析中，路径的种类过多，无法手动对所有的微观行为顺序进行建模，但此处因为宏观想法相近的微观行为顺序在NLP中会通过相似的矢量表示，我决定此处通过Doc2Vec得出用户的宏观想法

4. 研究目标

- CIKM 2019电商AI数据集
  - user.csv
    - User ID: 用户ID，每个ID代表一个用户
    - Gender: 性别，0代表男，1代表女，2代表未知
    - Age: 年龄，从18到90岁不等
    - Purchasing Power: 购买力， 从1 – 9不等
  - Item.csv
    - Item ID: 一个商品的ID
    - Category ID: 一个商品对应的商品类型的ID
    - Shop ID: 售卖这个商品的店铺ID
    - Brand ID: 售卖这个商品的品牌ID
  - user_behavior.csv
    - User ID: 用户ID，可用于映射user.csv中的用户信息
    - Item ID: 一个商品的ID，可用于映射item.csv中的商品信息
    - Behavior Type: 一个用户ID对一个商品ID所做的行为
      - P: 预览
      - C: 加入购物车
      - F: 收藏
      - B: 购买
    - Timestamp: 时间戳，即一个用户ID对一个商品ID所做的行为发生的时间

5.1 用户路径构建思路

1. 将商品ID、该商品对应的商铺ID、品牌ID、和产品类型ID当成四种不同的页面ID
2. 将用户对某商品做出的行为当做用户在该商品ID页面、商铺ID页面、品牌ID页面、产品类型ID页面所做出的的行为
3. 对于同一个用户，两次行为发生的时间大于10000个单位视为该用户的第二个行为是新session的首个行为。两次行为发生的时间小于等于10000个单位适于该用户的前后行为在同一个session中
4. 根据同一个用户所有行为的时间戳将用户的页面ID+对应行为顺序排列，生成分session的用户路径轨迹
5. 对于连续且重复的同页面ID同行为路径点进行合并，将所有大于2次连续重复的路径点删除，并在第二次路径点做重复标记

- 样例：{279: [[p1001, p1001, b1001], [p1001, p1001x, c1002]]}的意思是用户ID为279的用户在**第一个session**中连续**浏览**了**两次**页面ID为**1001**的页面，随后**购买**了**一次**页面ID为**1001**的页面所对应的产品，在**第二个session**中连续**浏览**了**多次(大于等于三次)**页面ID为**1001**的页面，随后将页面ID为**1002**的页面所对应的产品**加入了一次购物车**。

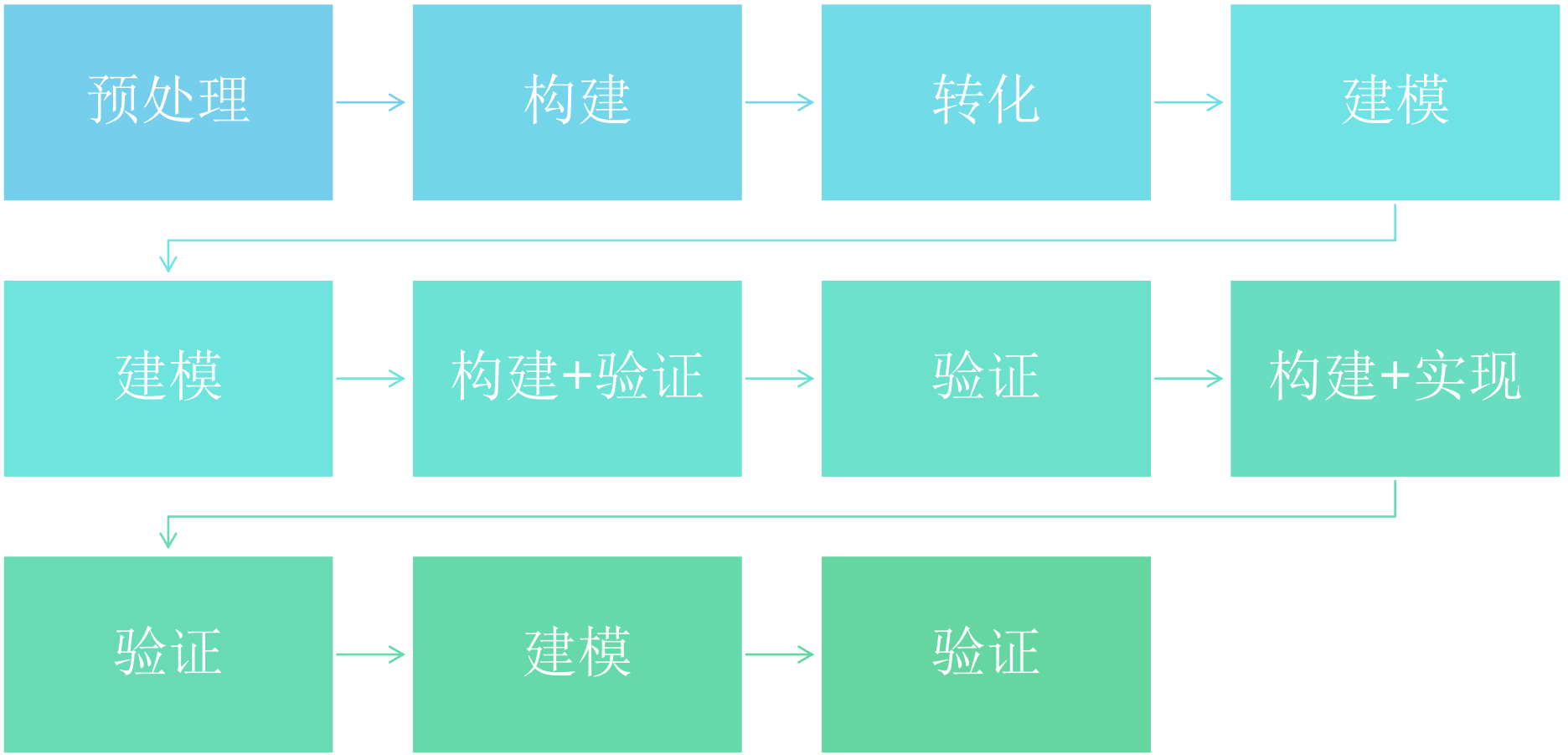
通过以上规则生成每个用户的路径轨迹，对每一种页面ID分别进行后续的用户分群和归因分析

5.2 拆分页面、行为构建ngrams动机

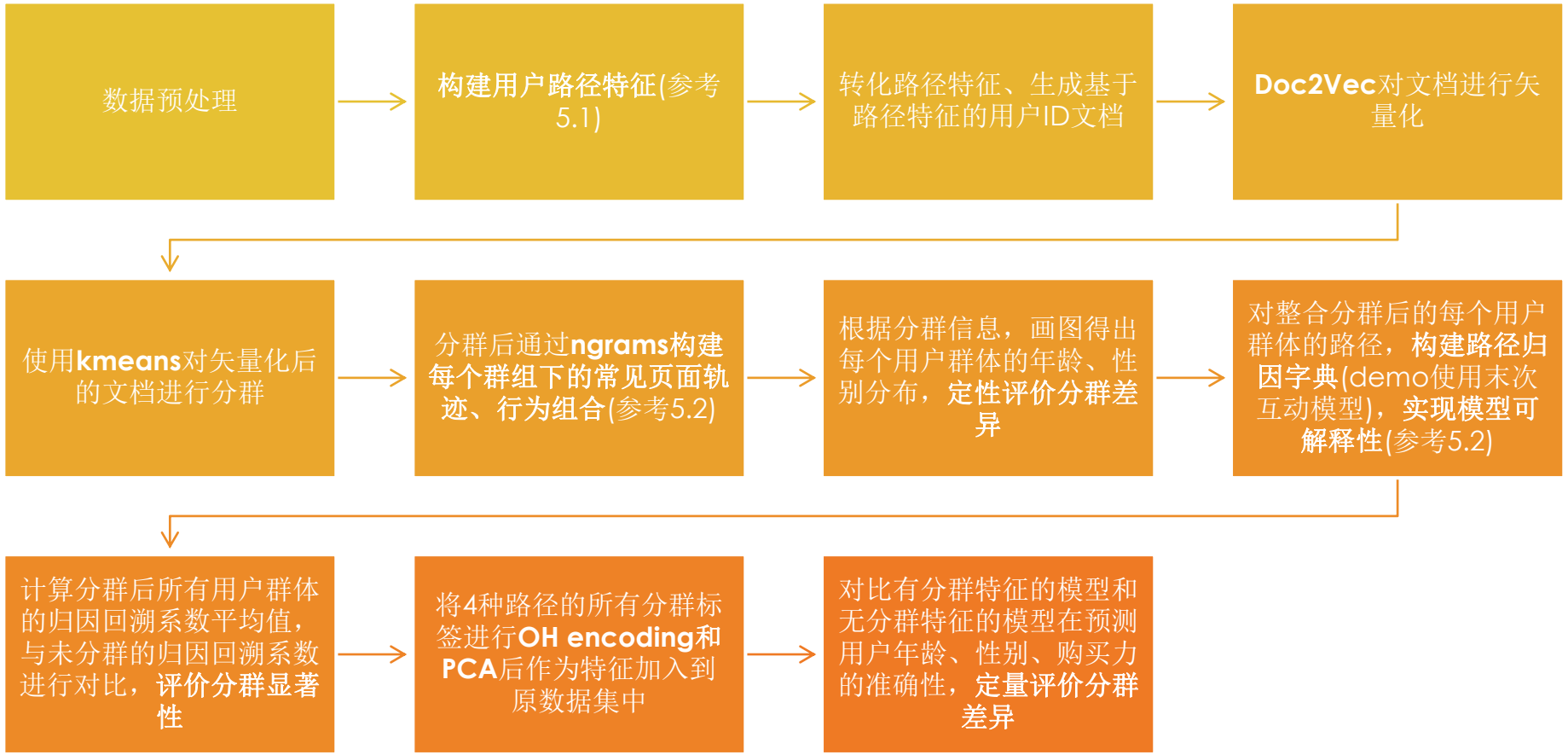
- 动机: 在通过结合用户行为和页面访问的情况下进行用户分群后，很难对每个用户群组的长度为N(demo中N = 4)的频繁页面访问顺序或动作执行顺序进行ngrams统计
- 例如: [p1001, b1001, p1002, p1001]和[p1001, c1001, p1002, p1001]存在一样的页面访问顺序和不同的动作执行顺序，因此此处ngrams会把两个顺序当成两种情况处理。在拆分页面、行为后，该顺序变成了[1001, 1001, 1002, 1001], [p, b, p, p]和[p, c, p, p]，更便于ngrams进行统计。在对相同动作执行顺序、不同页面访问顺序的ngrams中，拆分动作顺序和页面顺序同理也可以帮助ngrams更好的进行频繁项统计

通过以上规则生成的路径归因可使用户路径分群模型实现可解释化，即通过不同权重模型了解每个路径点背后的归因点，反映用户动机

6.1 Pipeline实现流程 – 简化



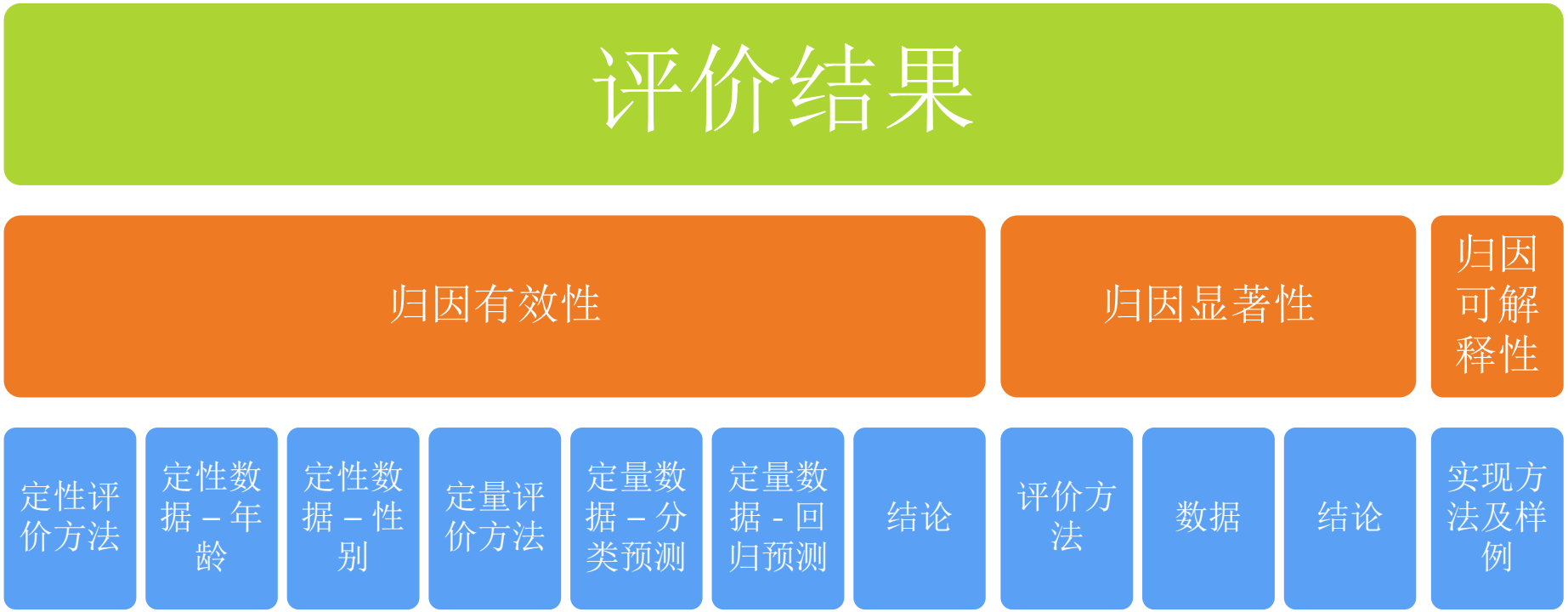
6.2 Pipeline实现流程 – 详细



7. 评价方式

- 1. 评价归因模型的有效性：根据第一篇参考论文的观点，一个有效的根据用户访问路径的用户分群模型在分群后，每个用户群体的年龄、性别等基本信息分布是非随机的，即每个组的年龄、性别是有一个宏观趋势的。
  - 在定性方面，如果分群后每个群组的年龄、性别分布不一样，那么路径归因是有效的
  - 在定量方面，如果分群后每个群组的Label能使一个预测用户性别、年龄的分类器模型中起到明显的重要性并提升了模型预测的准确性，那么路径归因是有效的
- 2. 评价归因模型的显著性：根据第二篇参考论文的观点，一个有效的根据用户访问路径的用户分群模型在分群后，能够更好地反映用户的宏观想法。
  - 在定量方面，如果分群后同一个用户群体的不加权平均归因回溯系数和加权平均归因回溯系数大于不分群状态下的不加权平均归因回溯系数和加权平均归因回溯系数(即分群后用户组路径同源性更高)，那么归因模型的效果是有显著性的
- 3. 评价归因模型的可解释性：根据https://www.jianshu.com/p/a1fa42c5cc42 的常用模型对每个路径点进行归因
  - 通过这个权重可以解释每一个路径点的归因信息

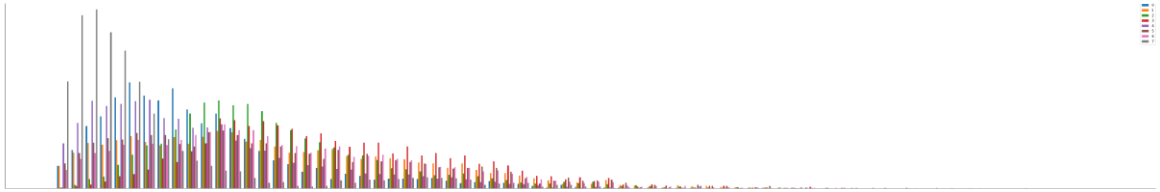
8. 评价结果


评价结果
归因有效性
归因显著性
归因可解释性
定性评价方法
定性数据 – 年龄
定性数据 – 性别
定量评价方法
定量数据 – 分类预测
定量数据 - 回归预测
结论
评价方法
数据
结论
实现方法及样例

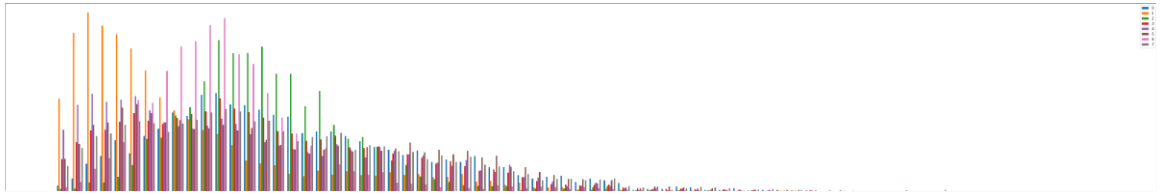
8.1.1 评价结果 – 归因有效性 – 定性评价方法

1. 通过使用基于Doc2Vec模型以用户的产品浏览路径、产品类型浏览路径、店铺浏览路径、品牌浏览路径(共4种不同路径)分别为语料对用户进行矢量化
2. 使用K = 8将所有的用户矢量分为8组进行聚类，聚类后每一类的即为以某类型路径分群后的一类用户。
3. 通过统计路径分群后每一类用户的年龄和性别分布，用图表画后后，通过观察分布是否具有随机性来评价归因有效性

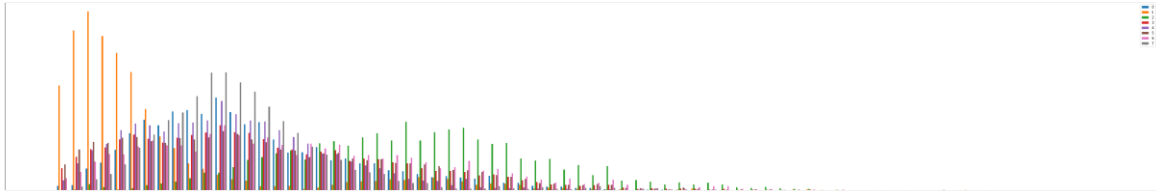
8.1.2 评价结果 – 归因有效性 – 定性数据 – 年龄分布



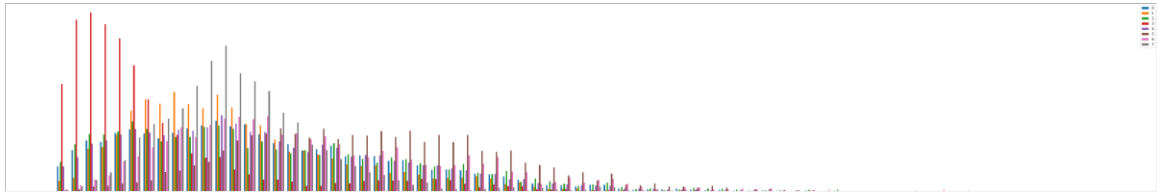
图一：依据用户产品浏览路径的分群后每组(以颜色区分)年龄分布



图二：依据用户产品类型浏览路径的分群后每组(以颜色区分)年龄分布



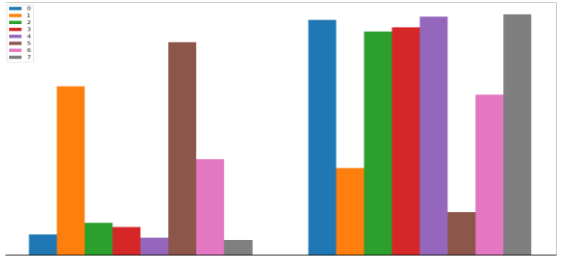
图三：依据店铺浏览路径的分群后每组(以颜色区分)年龄分布



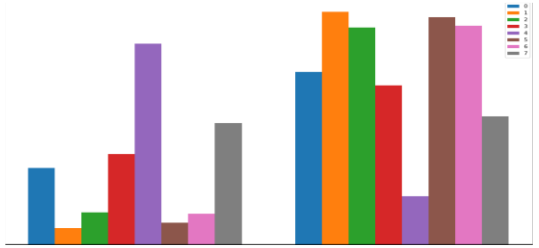
图四：依据品牌浏览路径的分群后每组(以颜色区分)年龄分布

由四张图可以看出，根据不同路径类型进行用户分群后，同一路径类型下每个用户群体的年龄分布均有明显特征，且不同路径类型的分组年龄分布不同

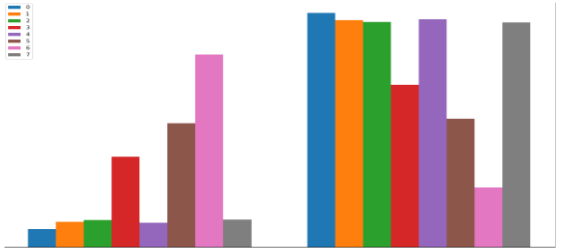
8.1.3 评价结果 – 归因有效性 – 定性数据 – 性别分布



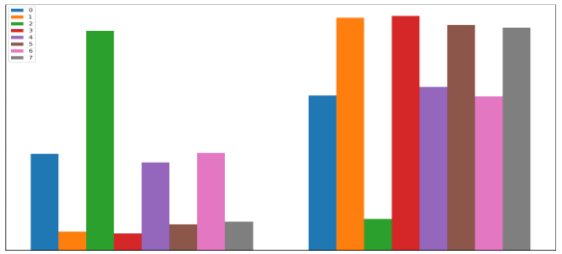
依据用户产品浏览路径的分群后每组(以颜色区分)性别分布(左侧为男性比例，右侧为女性比例)



依据用户产品类型浏览路径的分群后每组(以颜色区分)性别分布(左侧为男性比例，右侧为女性比例)



依据用户店铺浏览路径的分群后每组(以颜色区分)性别分布(左侧为男性比例，右侧为女性比例)



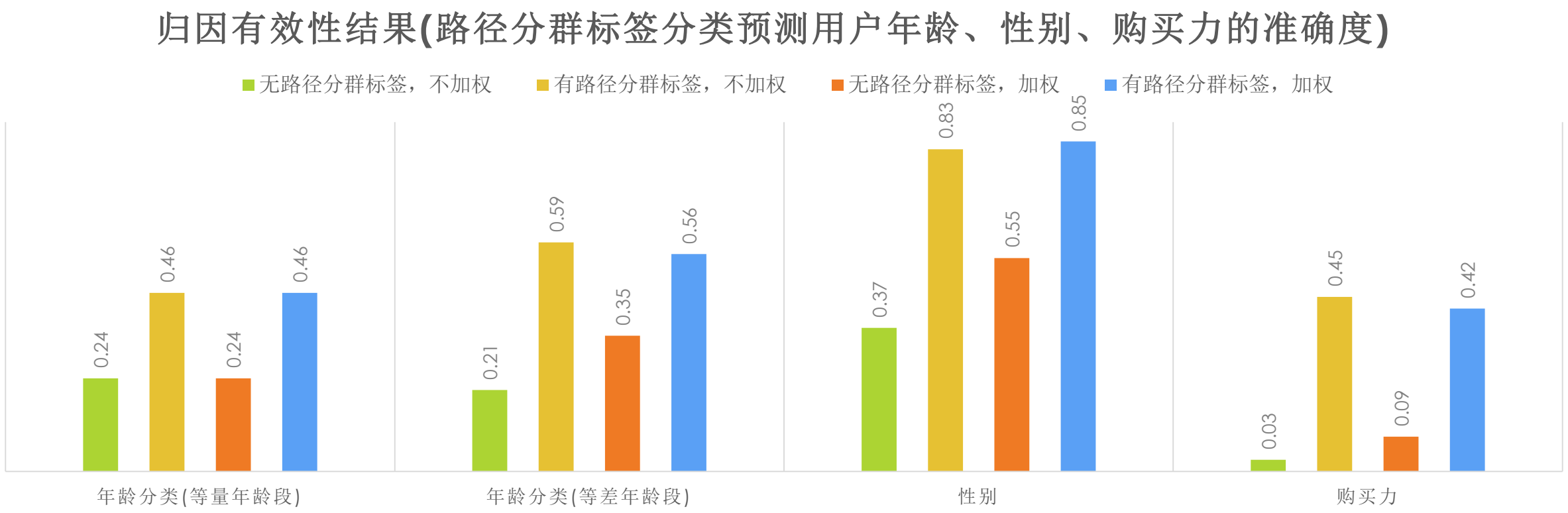
依据品牌浏览路径的分群后每组(以颜色区分)性别分布(左侧为男性比例，右侧为女性比例)

由四张图可以看出，根据不同路径类型进行用户分群后，同一路径类型下每个用户群体的性别分布均有明显特征，且不同路径类型的分组性别分布不同

8.1.4 评价结果 – 归因有效性 – 定量评价方法

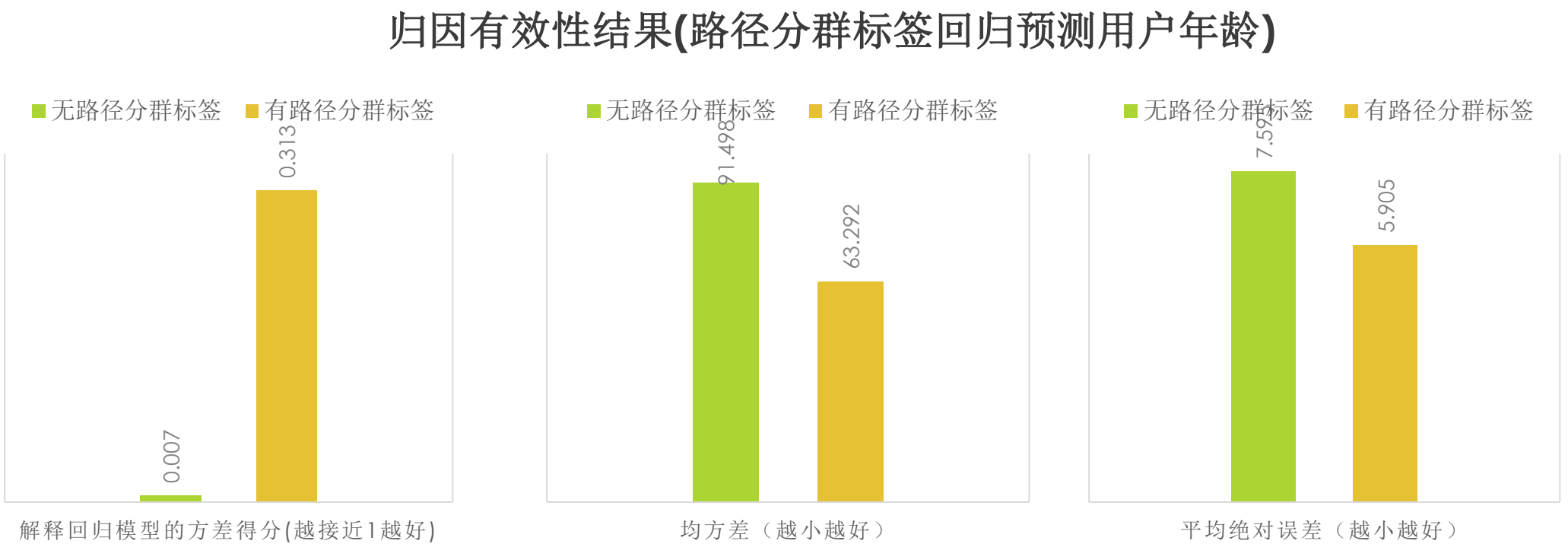
1. 通过使用基于Doc2Vec模型以用户的产品浏览路径、产品类型浏览路径、店铺浏览路径、品牌浏览路径(共4种不同路径)分别为语料对用户进行矢量化，生成4个基于不同用户路径类型的用户矢量模型
2. 每个矢量模型中使用K = 8，通过KMeans将所有的用户矢量分为8组进行分群，生成32个用户组别特征
3. 通过PCA将32个组别特征降维到8个特征
4. 对比加入这些特征前后分类和回归器模型预测用户年龄、性别、和购买力的准确性来评价归因有效性
   - 此处对年龄进行分类预测分两类：
     - 等量年龄段：每一个类的样本数量大致相同，但同类中最大和最小年龄差不同
     - 等差年龄段：每一个类的样本数量不同，但同类中最大和最小年龄年龄差大致相同（20岁以下为同一类，及50岁以上为同一类，其余以10岁为一类进行分类）
   - 等量年龄段：5分类
   - 等差年龄段：5分类
   - 性别：2分类
   - 购买力：10分类

8.1.5 评价结果 – 归因有效性 –定量数据



由四张图可以看出，有路径分群标签的模型在预测用户的年龄段、性别和购买力的表现均好于没有路径分群标签的模型，体现在更高的加权和不加权后的分类器平均准确度

8.1.6 评价结果 – 归因有效性 –定量数据



由三张图可以看出，有路径分群标签的模型在预测用户的准确年龄时的表现均好于没有路径分群标签的模型，体现在更高的解释回归模型的方差的分，更低的均方差和更低的平均绝对误差

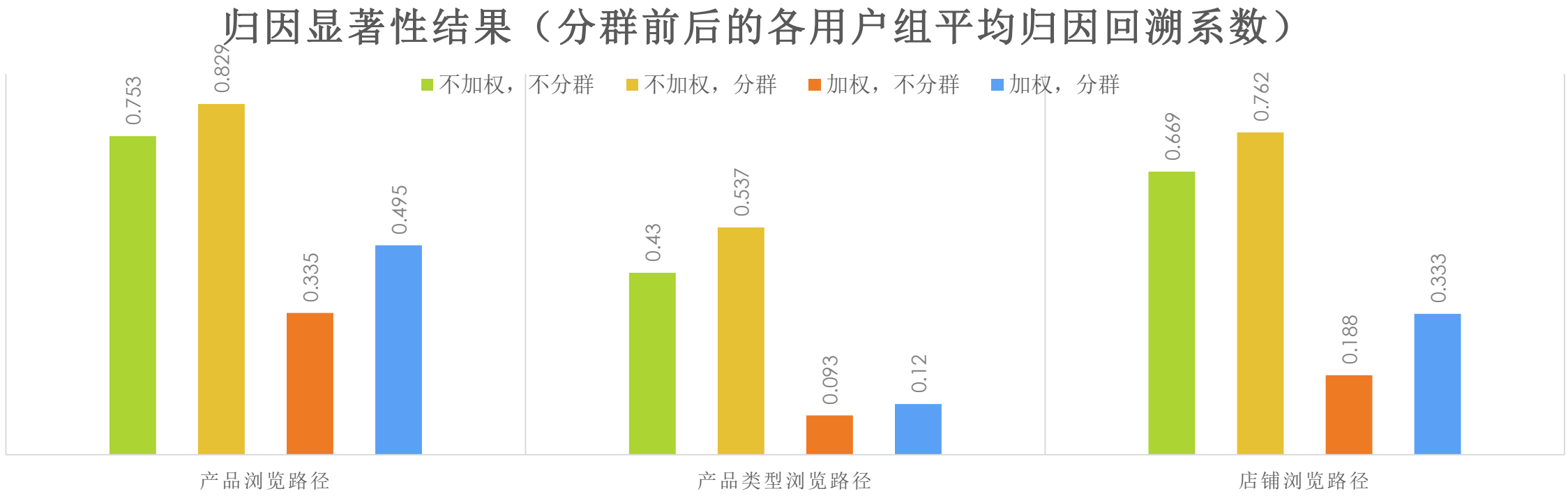
8.1.7 评价结果 – 归因有效性 – 结论

- 根据第一篇参考论文的观点，一个有效的根据用户访问路径的用户分群模型在分群后，每个用户群体的年龄、性别等基本信息分布是非随机的，即每个组的年龄、性别是有一个宏观趋势的
  - 由定性实验数据可以看出，在以不同类型的路径为基准的路径分群模型中，分群后每一组用户的年龄和性别分布是非随机的，每组的年龄会集中在某一个段上，每组的性别分布也是有明显差异的
  - 由实验数据可以看出，有路径分群标签的所有分类和回归用户群体年龄、性别、购买力预测任务的表现均好于没有路径分群标签对应的的分类和回归任务
- 因此，结合第一篇参考论文的观点和路径归因模型的实验数据可以得出该路径归因模型具有有效性

8.2.1 评价结果 – 归因显著性 – 评价方法

1. 通过使用基于Doc2Vec模型以用户的产品浏览路径、产品类型浏览路径、店铺浏览路径、品牌浏览路径(共4种不同路径)分别为语料对用户进行矢量化，生成4个基于不同用户路径类型的用户矢量模型
2. 每个矢量模型中使用K = 8，通过KMeans将所有的用户矢量分为8组进行分群
3. 在同一个路径类型下，计算分群后每个用户组中每个路径点的归因点和归因系数
4. 计算分群后每个路径点的归因点对于路径点本身的不加权归因回溯系数和归因回溯系数的平均值，并与其将不分群情况下的不加权归因回溯系数和归因回溯系数的值进行对比，通过分群和不分群后归因回溯系数值的差异验证归因显著性。

8.2.2 评价结果 – 归因显著性 – 数据



由三张图可以看出，在所有路径类型下，分群后的不加权平均归因回溯系数大于不分群后的不加权归因回溯系数、分群后的加权平均归因回溯系数大于不分群后的加权归因回溯系数。在加权表现中，粒度越细的路径在根据用户路径分群后得出的归因回溯系数与不分群后得出的系数差距越明显

8.2.3 评价结果 – 归因显著性 – 结论

- 由于实验数据集中品牌浏览路径的无效数据过多(即某路径点的页面ID为-1的情况)，该组数据没有纳入评价显著性的一部分
- 在通过产品浏览路径为基准的情况下评价归因显著性，经过路径分群的不加权归因回溯系数和加权归因回溯系数相较于不经过群分的系数分别提升了0.07和0.16
- 在通过产品类型浏览路径为基准的情况下评价归因显著性，经过路径分群的不加权归因回溯系数和加权归因回溯系数相较于不经过群分的系数分别提升了0.11和0.02
- 在通过店铺浏览路径为基准的情况下评价归因显著性，经过路径分群的不加权归因回溯系数和加权归因回溯系数相较于不经过群分的系数分别提升了0.10和0.15
- 由此可以得出以下两点结论
  - 根据用户路径进行用户分群后每组用户的路径同源性提升，代表归因显著性存在
  - 路径同源性的提升幅度随着路径复杂度的增加而增加(即每个路径点的粒度越细，同源性提升效果越好，在此处可以体现在加权归因回溯系数上产品浏览路径的同源提升效果 > 店铺浏览路径的同源提升效果 >产品类型路径的同源提升效果)

8.2.4 评价结果 – 归因可解释性 – 实现方法及样例

- 给定一个用户群组的路径集合（区分不同session），通过不同归因模型对该集合中每一个路径点找出其归因点以及它们的归因系数
  - 末次互动模型(最常用/demo演示模型)：目标路径点之前的一个路径点为它的归因点，该点记(1)分
  - 首次互动模型：目标路径点所在的整条路径的起始路径点为它的归因点，该点记(1)分
  - 线形归因模型：目标路径点之前在该条路径的所有路径点为它的归因点，每个点占(1/该条路径所有归因点的数量)分
  - 线形时间衰减模型：目标路径点之前在该条路径的所有路径点为它的归因点，每个点占(该点到整条路径起始点的距离/所有路径点每个点到整条路径起始点的总和)分
  - U形归因模型：目标路径点之前的一个路径点及目标路径点所在路径的第一个点各占(0.4)分，其余点占(0.2/该条路径所有归因点的数量-2)分。若该路径只有一个归因点，该归因点占(1)分，若该路径只有2个归因点，两个归因点各占(0.5)分
- 将所有归因点的得分进行求和并标准化
- 样例：{'p26991039': [['p22830135', 'b12579266', ''], [0.3333333333333333, 0.3333333333333333, 0.3333333333333333]]}的意思是**路径点p26991039**的**归因点为p22830135**， **b12579266**和**无**(代表该**路径点是整条路径的第一个点**)，这三个点的对**p26991039**的归因系数**均为0.3333333333333333**

通过以上规则生成的路径归因可使用户路径分群模型实现可解释化，即通过不同权重模型了解每个路径点背后的归因点，反映用户动机

9. 附录

- 基于用户路径分析进行用户分群改善用户组内路径归因同源性的demo：
  - Attribution Improvement by User Path & Behavior Clustering_EN_US.pdf