****

****

**人工智能技术专题论文**

**题    目     用户行为分析**

**学   院       自动化学院**

**专 业 物联网工程**

**年级班别 20级(1)班**

**学 号 3120001384**

**学生姓名 陈铿任**

**2021年 6 月 日**

**目 录**

[1 绪 论 1](#_Toc70949431)

[1.1 了解用户行为分析 1](#_Toc70949432)

[1.2 用户行为分析的目的 1](#_Toc70949433)

[2 用户行为分析的具体内容 1](#_Toc70949440)

[2.1 用户行为分析的指标 1](#_Toc70949441)

[2.2 用户行为分析模型 2](#_Toc70949442)

[2.2.1 漏斗模型 2](#_Toc70949451)

[2.2.2 用户留存分析 2](#_Toc70949452)

[2.2.3 用户消费分析 2](#_Toc70949452)

[2.2.3.1箱型图 2](#_Toc70949452)

[2.2.3.2日ARPPU分析 3](#_Toc70949452)

[2.2.3.3日ARPU情况分析 3](#_Toc70949452)

[2.2.3.4RFM模型分析 3](#_Toc70949452)

[2.2.4 行为事件分析 4](#_Toc70949452)

[3 实例分析—淘宝用户行为分析 4](#_Toc70949449)

[3.1 数据集的介绍 4](#_Toc70949450)

[3.2 数据预处理 5](#_Toc70949453)

[3.2.1 缺失值的处理 5](#_Toc70949454)

[3.2.2 日期标签的处理 5](#_Toc70949455)

[3.3 行为分析 6](#_Toc70949453)

[3.3.1 用户留存分析 6](#_Toc70949454)

[3.2.2 用户的行为数据对比 7](#_Toc70949455)

[3.2.3 模型分析 9](#_Toc70949455)

[3.2.3.1漏斗图模型分析 9](#_Toc70949454)

[3.2.3.2箱型图 10](#_Toc70949455)

[3.2.3.3日ARPPU分析 10](#_Toc70949455)

[3.2.3.4日ARPU情况分析 11](#_Toc70949455)

[3.2.3.5 复购率 12](#_Toc70949455)

[3.2.3.6 RFM模型分析 12](#_Toc70949455)

[4 工作总结和心得体会 13](#_Toc70949469)

[4.1 工作总结 13](#_Toc70949470)

[4.2 心得体会 13](#_Toc70949471)

1. **绪论**

**1.1了解用户行为分析**

随近几年来，大数据不断发展，大数据成为了一个热门的行业，应用在各个领域。

用户行为分析就是其中的一个应用方向。随着智能设备的快速发展，人们的生活已经离不开智能设备了。人们花费在智能设备的大部分时间在购物平台上面，而我们在购物平台上浏览信息会产生不同的信息来反应我们的行为。

用户行为信息在电商平台十分重要，通过用户行为分析，能够了解用户的情况，比如：他们不同的生活习惯、文化背景、生活位置等等这些信息都会影响他们的行为。那么什么是用户行为分析？

用户行为信息指的是在获得网站或APP等平台访问量基本数据的情况下，对其中的数据进行统计、分析，从中发现用户访问网站或APP等平台的规律来建立、进一步修正、或者重新指定网络营销策略等等进而实现精细化运营。

**1.2用户行为分析的目的**

对互联网金融、电商平台、线上教育、证券、短视频等这些行业的产品来说，用户行为数据显得十分重要。用户行为的目的在于自定义留存分析、精细化渠道质量评估、产品分析、精准营销。

1. 自定义留存分析：基于用户行为，我们可以做精细化留存评估。根据产品特性自定义用户留存，根据用户的活跃情况来对产品进一步进行评估、优化等。
2. 精细化渠道质量评估：通过不同的渠道带来的用户，我们需要去了解哪个渠道的用户行为数据最为丰富、最有价值，从而实现降低渠道成本，提升渠道转化。
3. 产品分析：通过漏斗分析或者路径转化，找到产品改进的关键点、找到促进核心转化的相关因素，从而使得产品设计得更人性化。
4. 精准营销：通过用户点击商品详情、搜索行为、浏览历史记录、购买的东西等行为来定位用户的兴趣，从而制定更好的个性化服务。

**2.用户行为分析的具体内容**

**2.1用户行为分析的指标**

根据用户行为表现可以将用户行为分析的指标大体上可以分为三大类，即黏性指标、活跃指标和产出指标。

1. 粘性指标：主要关注用户在某一段时间内持续访问何使用网站的情况，一种持续的状态。比如：新用户数及比例、活跃用户数及比例、用户留存率、用户流失率、用户访问率等等。
2. 活跃指标：指向用户每次访问过程中发生的行为，主要是指用户访问的参与度，比如：活跃用户、新增用户、流失用户、使用频率等等。
3. 产出指标：衡量用户创造的价值输出，比如：页面浏览数(PV)、页面访问人数(UV)、点击次数等等。

**2.2用户行为分析模型**

借助一些模型来对数据(用户特定的行为)进行定性和定量分析分析。

常用的分析模型有：

1. 漏斗模型分析
2. 用户留存分析
3. 用户消费行为分析

1.箱型图

2.日ARPPU分析

3.日ARPU情况分析

4.RFM模型

4) 行为事件分析

**2.2.1漏斗模型分析**

漏斗模型分析是用户在使用APP或者访问网站等多城中，描述各个阶段中关键环节的用户转化和流失率情况，能够科学反映用户行为状态以及从起点到终点各阶段用户转化率情况。

漏斗模型可以分析验证整个流程的设计、决策是否合理、满足客户的需求，通过分析用户行为过程的相关转化率，发现哪个环节的转化率相对低、需要改进，从而发现、决解问题。

**2.2.2用户留存分析**

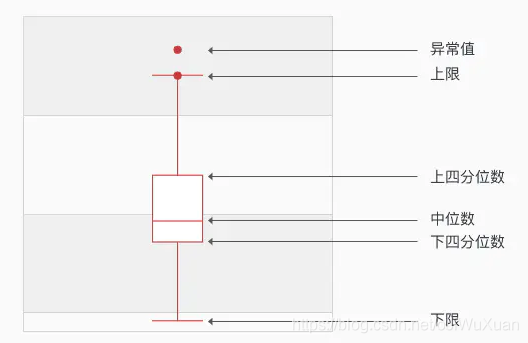
用户留存分析是一种用来分析用户活跃程度的模型。通过留存量和留存率(PV、UV等指标)了解当前用户实际行为从而得知用户的留存和流失情况，发现影响产品可持续增长的关键因素，知道市场决策、产品改进、提升用户价值等等。

**2.2.3用户消费行为分析**

用户消费行为是一种在获得用户的消费信息下，了解用户的消费频率和消费金额和用户的收入情况，从而预测用户复购率(复购行为用户数/消费行为用户总数)、划分用户价值等级的情况，从而进一步做好决策及时调整。

**2.2.3.1箱型图**

箱型图是其中的一种分析图，用作一组数据的分散分布情况的统计图，用于处理异常数据和数据的过滤。(不受异常值的影响，能够准确第描绘出数据离散分布的情况)



其中箱型图包含5个重要的因素：

1.上四分数U：表示在所有样本中只有1/4的数值大于U，即从大到小排序时U处于25%处。

2.下四分位数L：表示在所有样本中只有1/4的数值小于L，即从大到小排序时处于75%处。

3.中位数Q：表示一组数小到大排列处于中间位置的数，若序列数为偶数时，该组的中位数为中间两个数的平均数。

4.上限：表示非异常范围的最大值，四分位距为IQR=U-L，则上限为U+1.5IQR。

5.下限：表示非异常范围内的最小值，即下限为L-1.5IQR。

**2.2.3.2日ARPPU分析**

ARPPU(average revenue per paying user)是指从每位消费用户身上获得的收入，即在统计日期之内消费用户对产品的平均消费。

**ARPPU** = 消费总收入/消费用户数

**2.2.3.3日ARPU情况分析**

ARPU(average revenue per user)是指每位用户的平均收入。ARPU注重的是一个时间段内销售员情况来看，ARPU值越高那么利润越高，效益越好，可以衡量产品的盈利和发展的活力。

**ARPU** = 消费总收入/用户总数

**2.2.3.4RFM模型分析**

RFM模型根据客户活跃度和交易金额的贡献，进行客户价值细分的一种方法。识别优质客户，可以指定个性化的沟通和营销服务，能够衡量客户价值和客户利润创收能力。

1.R(Recency)——最近一次交易时间间隔，基于最近一次交易日期计算得分，举例当前日期越近，得分越高。

2.F(Frequency)——客户在最近一段时间内交易次数，基于交易频率计算得分，交易频率越高，得分越高。

3.M(Monetray)——客户在最近一段时间内交易金额，基于金额计算得分，交易金额越高，得分越高。

RFM模型的假设前提：

1.最近交易过的客户 > 最近没有交易过的

2.交易频率高的客户 > 交易频率低的

3.交易金额打的客户 > 交易金额小的

通过自定义的评分规则，来划分出不同类型的客户，对于哪些客户进行重点关注、维护，哪些客户需要探究流失原因等等。

**2.2.3行为事件分析**

行为事件分析是指根据获取到的指标数据对用户进行行为分析，通过追踪或记录用户行为事件，了解事件的趋势走向和用户的情况，从而挖掘事件的影响因素。

主要研究用户的5W行为，即用户是谁、从什么地方来、发生了什么事件、如何产生的、发生的时间，从而挖掘其中对产品、企业的影响等等。

行为指标：

1) 点击率CTR=点击量/曝光量(商品、咨询内容、视频)\*100%

2) 退出率

3) 消费频率

**3.实例分析**

下面通过一个分析淘宝用户行为来阐述使用过程。

**3.1数据集的介绍**

数据集来源于阿里天池(提供完备的IDE以及丰富的计算资源)，包含有2014-11-18至2014-12-18时期的1048576条用户数据。

|  |  |
| --- | --- |
| Features | Instructions |
| user\_id | The user name |
| item\_id | The commodity name |
| behavior\_type | Behavior(Click,Favorite,Add to cart,Buy corresponds to 1,2,3,4) |
| user\_geohash | The geographical position of user |
| item\_categort | Commodity categories |
| time | The time that the user beahavior takes place |

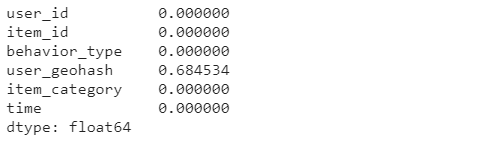
**3.2数据预处理-数据清洗**

**3.2.1缺失值的处理**

通过自定义一个函数来计算各个特征下缺失值的个数占该特征总数的比值。

代码实现：

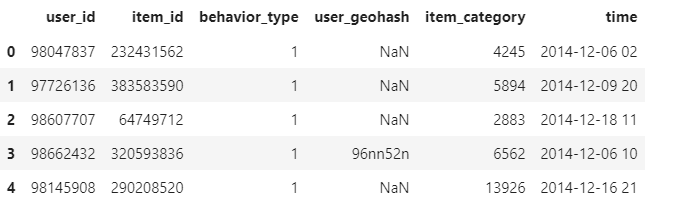
DataFrame.apply(lambda x:sum(x.isnull()/len(x)))



通过观察发现user\_geohash(The location of user while behavior take palce)这个特征缺失值也太多了，那么可以将这特征除去。

DataFrame.drop('user\_geohash',axis=1,inplace=True)

**3.2.2日期标签的处理**

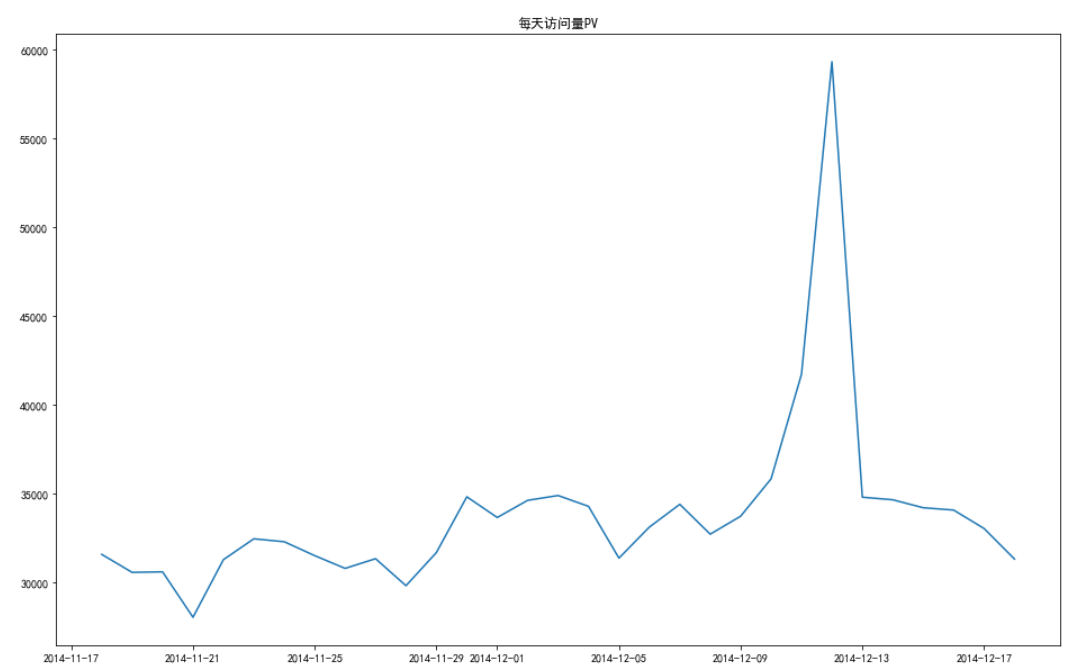
通过DataFrame.head()观察<time>这个标签的数据既含有日期部分也含有时间部分，为了方便分析每日和在相同时间用户行为，将<time>分为两个标签<time>、<hour>。

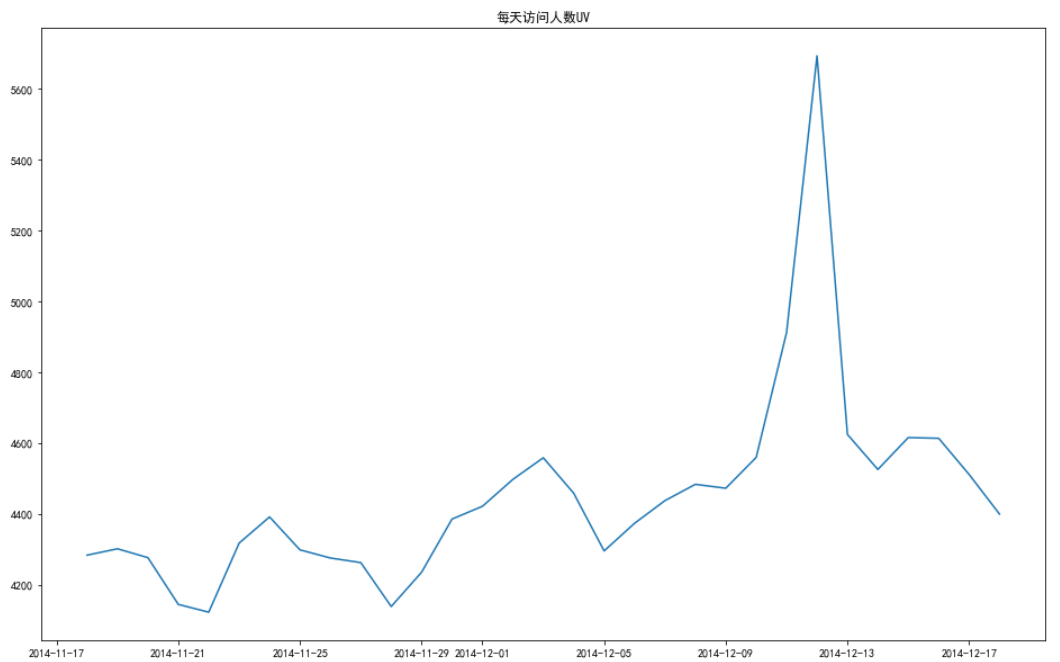
|  |
| --- |
| df['time'] = pd.to\_datetime(df['time'],format="%Y-%m-%d")  df['hour'] = df.time.dt.hour#提取time的具体时间(小时)部分  df['time'] = df.time.dt.normalize() #提取time的日期部分 |

**3.3行为分析**

**3.3.1用户留存分析**

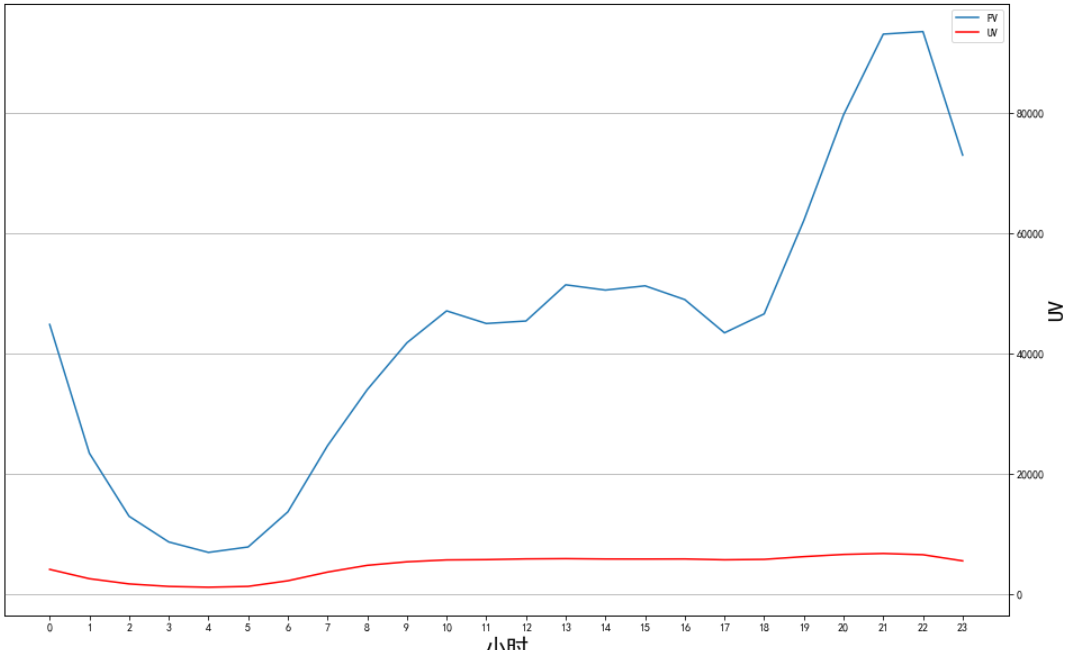
先是对日期维度进行分析：





通过观察发现在双12日访问人数和访问量均达到峰值，而且PV和UV两者在图像形状上很相似。

再对小时维度进行分析：

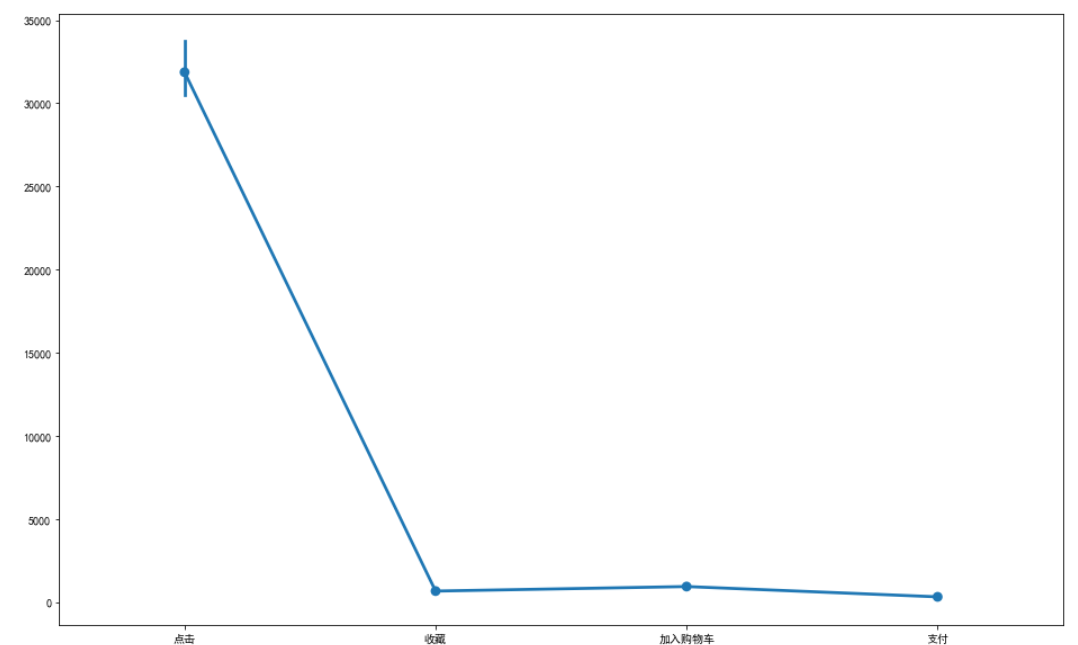


通过图表观察来看，UV整体上没有很明显的波动，但在凌晨0-6点人数会相对下降；PV的20-22点会出现达到访问的峰值，数据较为正常。

**3.3.2用户行为的数据对比**

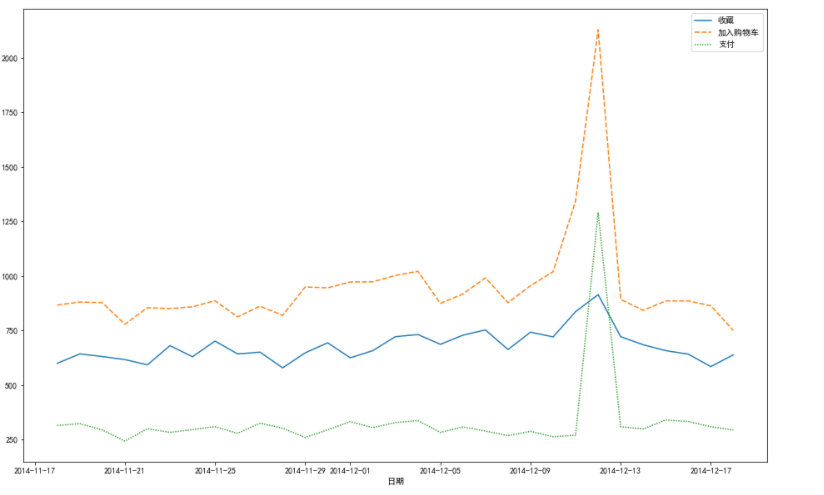
四种行为数据总量对比：

|  |
| --- |
| behavior\_data = pd.pivot\_table(df,index='日期',columns='行为',values='用户名',aggfunc=np.size)  behavior\_data.columns = ["点击","收藏","加入购物车","支付"]  plt.figure(figsize=(16,10))  sns.pointplot(data=behavior\_data[["点击","收藏","加入购物车","支付"]])  plt.show() |

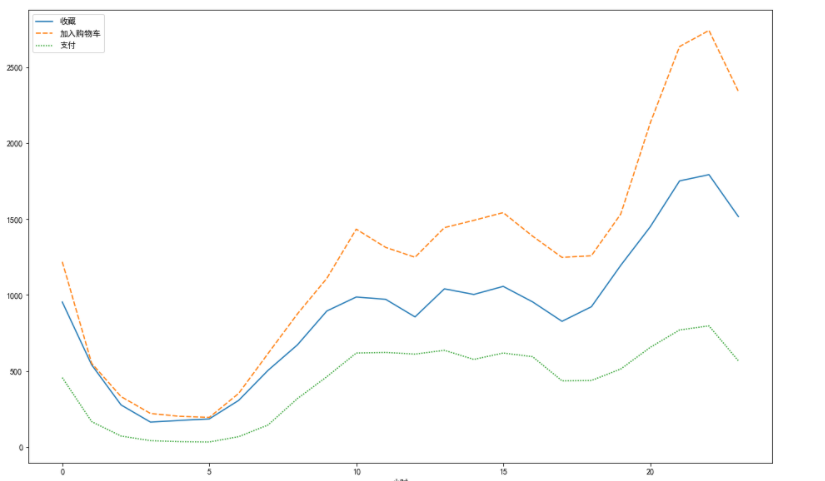


结论：点击的数据量远比其他三种高，所以下面的在日期和小时下来研究行为，只对比三种行为。

以日期维度来分析三种用户行为：



结论：同样，在双12购物节能够引起用户对购物的欲望 ，均在双12期间达到峰值。

以小时维度来分析三种用户行为：

**3.3.3模型分析**

**3.3.3.1漏斗模型分析**

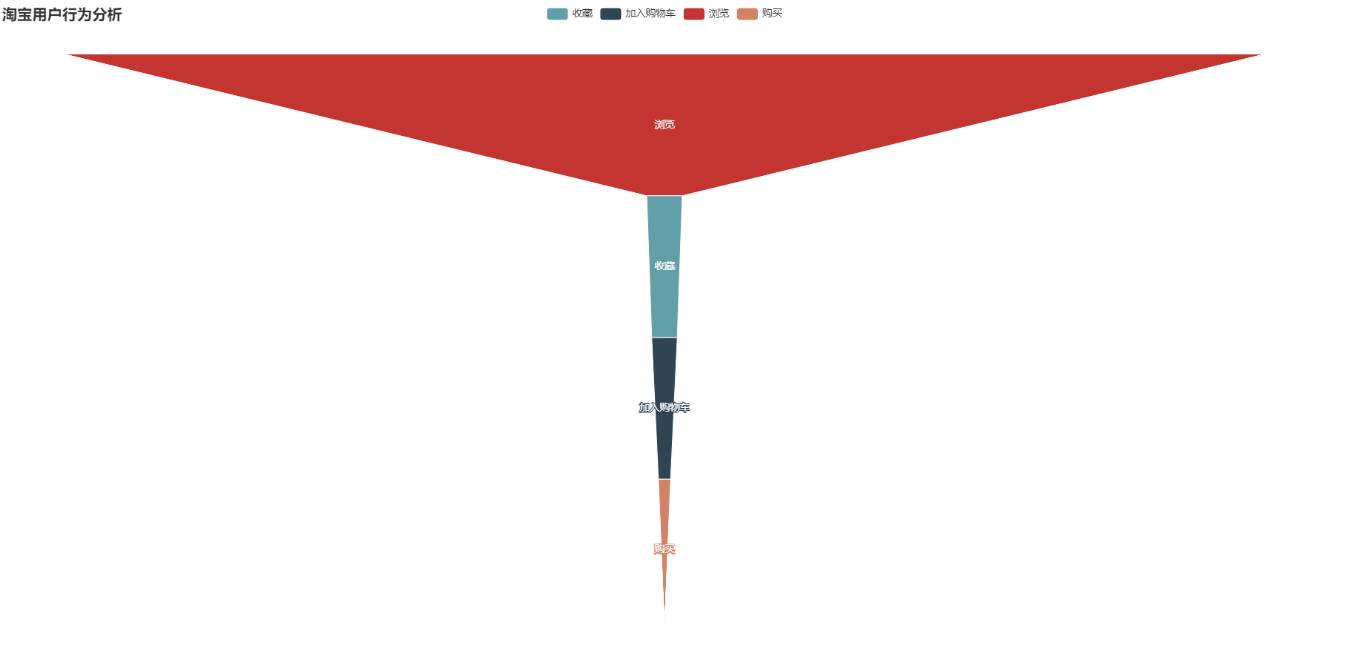
行为的顺序：点击->收藏->加入购物车->购买

单一转化率的计算：就是前一个行为的人数除于下一个行为的人数\*100%(保留两位小数点)，需要将第一个行为的转化率设置为100%。(作为参考系)

|  |
| --- |
| tnum = np.array(df\_behavior["人数"])[:-1]  num = np.array(df\_behavior["人数"])[1:]  conversion\_rate = num/tnum  conversion\_rate.insert(0,1)  print(conversion\_rate)  conversion\_rate = [round(x\*100,2)for x in conversion\_rate] |

总体转化率的计算：就是各个行为的人数分别除于第一个行为的人数。

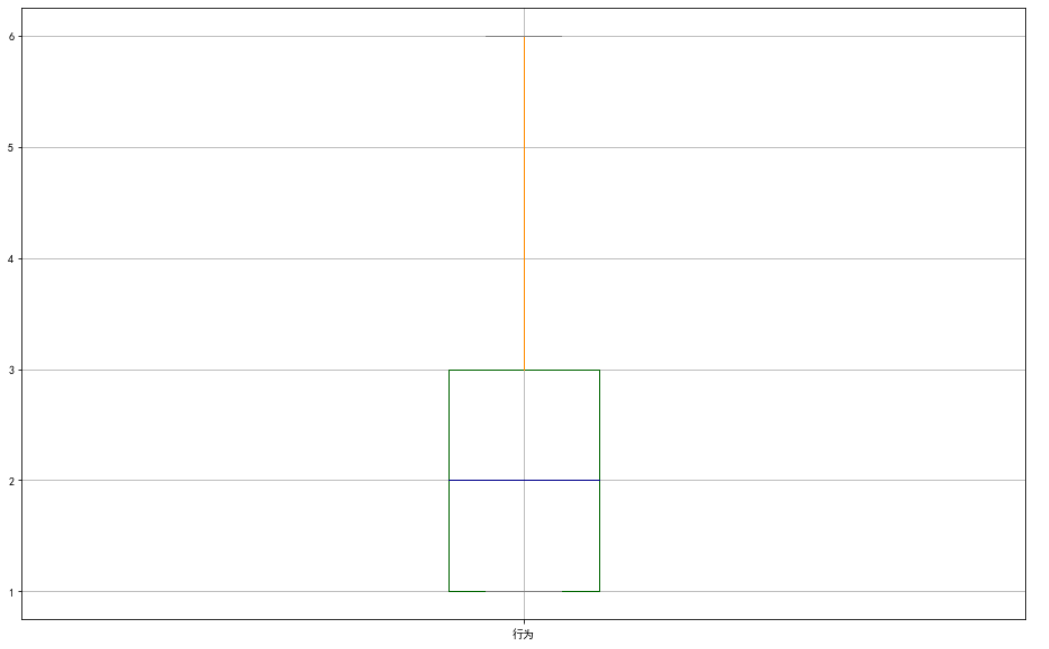
|  |
| --- |
| total\_conversion = df\_behavior["人数"]/df\_behavior["人数"][0]  total\_conversion = [round(x\*100,2)for x in total\_conversion] |



通过观察漏斗图，我们发观从浏览到购买这一过程的转化率十分低，这时候就需要去探究原因。

引发公司的猜想，这说明可能由于用户再浏览过程中，没有找到自己所想要的商品，提高产品个性化推荐或者产品不适合大部分消费者导致产品无法卖出去还是支付的时候出了什么问题还是在购买的时候又看到更让自己满意的商品。

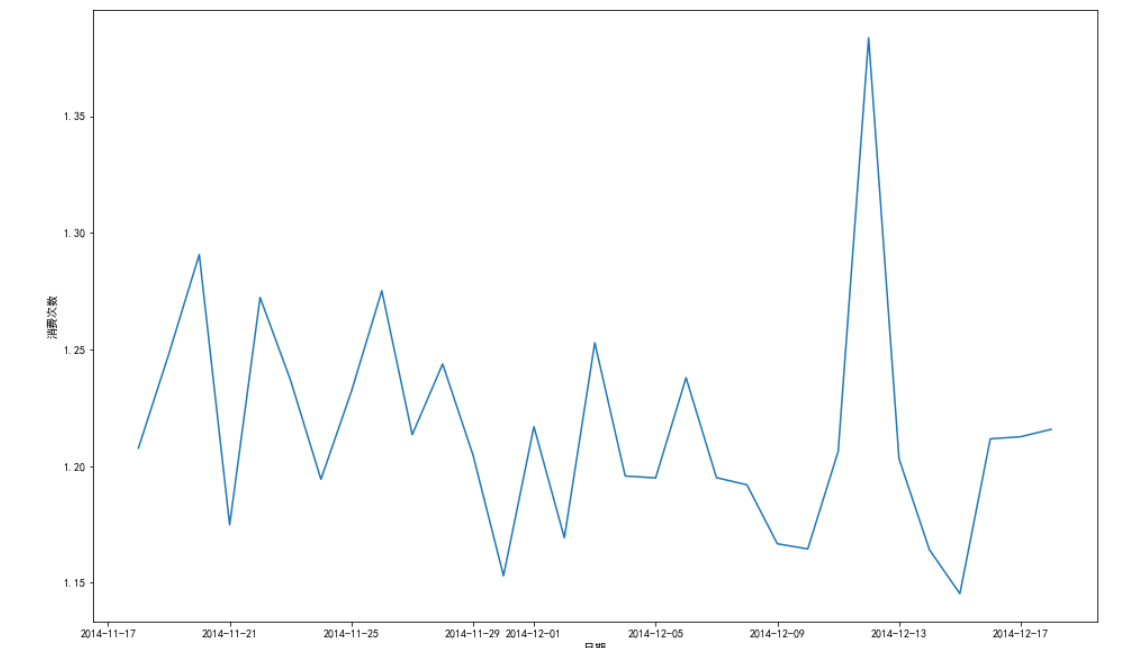
**3.3.3.2箱型图**



通过观察发现25%到50%和50%到75%箱重占比相同，这说明大多数用户处在1-3次消费这个次数；用户的消费次数总体上来说偏少。

**3.3.3.3日ARPPU分析**

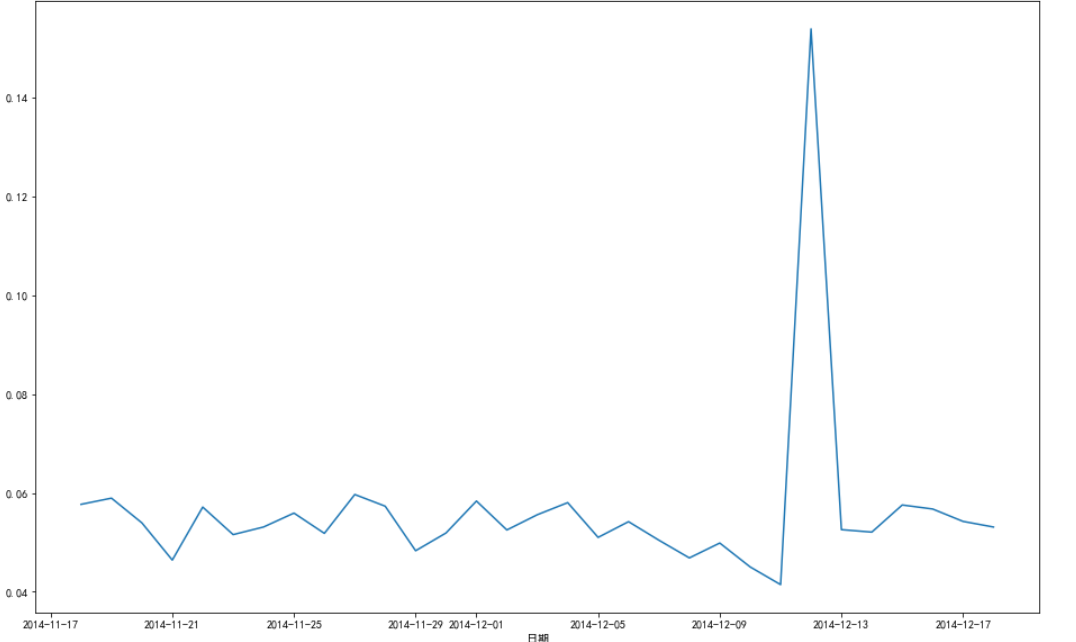
在这个数据集中没有用户的消费金额，我们用消费次数代替消费金额。



通过ARPPU-day统计图我们发现在平时一些日子里面用户消费次数主要在1.20到1.30之间波动，但在双12购物节日的时候剧烈增长，高峰值在1.40左右。说明一些节日会促进用户进行购物。

**3.3.3.4日ARPU情况分析**

同样，因为该数据集中没有消费收入的数据，我们用消费次数来代替收入。



通过ARPU\_day统计图我们发现在这个数据集中用户的消费次数十分的少，每个用户平均消费次数不到0.2。

**3.3.3.5复购率**

|  |
| --- |
| Re\_buy = df[df.行为==4].groupby('用户名')['日期'].apply(lambda x:len(x.unique()))  print(round(Re\_buy[Re\_buy>=2].count()/Re\_buy.count(),2)) |

通过计算复购率，计算结果为0.47，说明产品相当不稳定，不符合人们的需求，导致消费次数、复购率相当低，该公司应该及时做出处理，紧急止损并且做出下一步的营销策略。

**3.3.3.6 RFM模型分析**

由于没有该数据集没有用户的消费金额，在这里只用R(最近一次消费时间)和F(消费频率)来进行评分划分用户。

自定义评分规则：

|  |
| --- |
| def Recency(day):  if day < 5:  return 5  elif day >=5 & day < 10:  return 4  elif day >=10 & day < 15:  return 3  elif day >= 15 & day < 20:  return 2  elif day >= 20 & day <25:  return 1  elif day >= 25:  return 0    def Frequency(time):  if time < 15:  return 0  elif time >= 15 & time < 30:  return 1  elif time >= 30 & time <45:  return 2  elif time >= 45 & time < 60:  return 3  elif time >= 60 & time < 75:  return 4  elif time >= 75:  return 5 |



结论：对于不同的客户给予不同的个性化服务，能够创造更好的价值。

**4.工作总结和心得体会**

**4.1工作总结**

人工智能技术专题这门课让我对人工智能的许多领域，如：医学影像、智能优化、自然语言处理和文本挖掘等等都有了一定的了解，其中我对数据挖掘与分析这个领域产生了极大的 兴趣。其次，这次的选题是老师上课提及到的一个数据挖掘与分析方面的应用，结合我自身的想法确定了这次论文的选题。

在此次工作学习了pyecharts、numpy、pandas、seaborn、matplotlib库的一些使用并了解用户行为分析的整个流程。通过一个淘宝用户行为分析实例，学习并掌握用户行为分析的一些方法，从数据的预处理(数据清洗、特征选择)到行为数据比较分析再到模型分析，一步步地走过来，有了很多收获。

**4.2心得体会**

在学习过程中虽然遇到很多波折，但亦有所获，同时也意识到自己有很多地方的不足之处，自己在技术方面的不足，对刚开始接触数据挖掘与分析不久的自己，还是有很多东西要去学习。同时也要感谢人工智能技术专题帮助我找到自己所热爱的方向。

既然热爱数据挖掘与分析，那就坚持走下去。从一个小白开始，看博客，上官网查文档，动手实践。每次在学习一个知识点的时候，又会涉及到其他新的知识点，虽然过程可能会有点复杂，但在学习到一个个新的东西之后，将其运用到实例上面去。

现在我所学到的知识只不过是数据挖掘与分析的皮毛，想要更深入学习数据挖掘与分析还有很长的一段要走，决不能因为自己在大一就有所松懈，希望自己能够保持学习，坚持走下去能够有所建树。