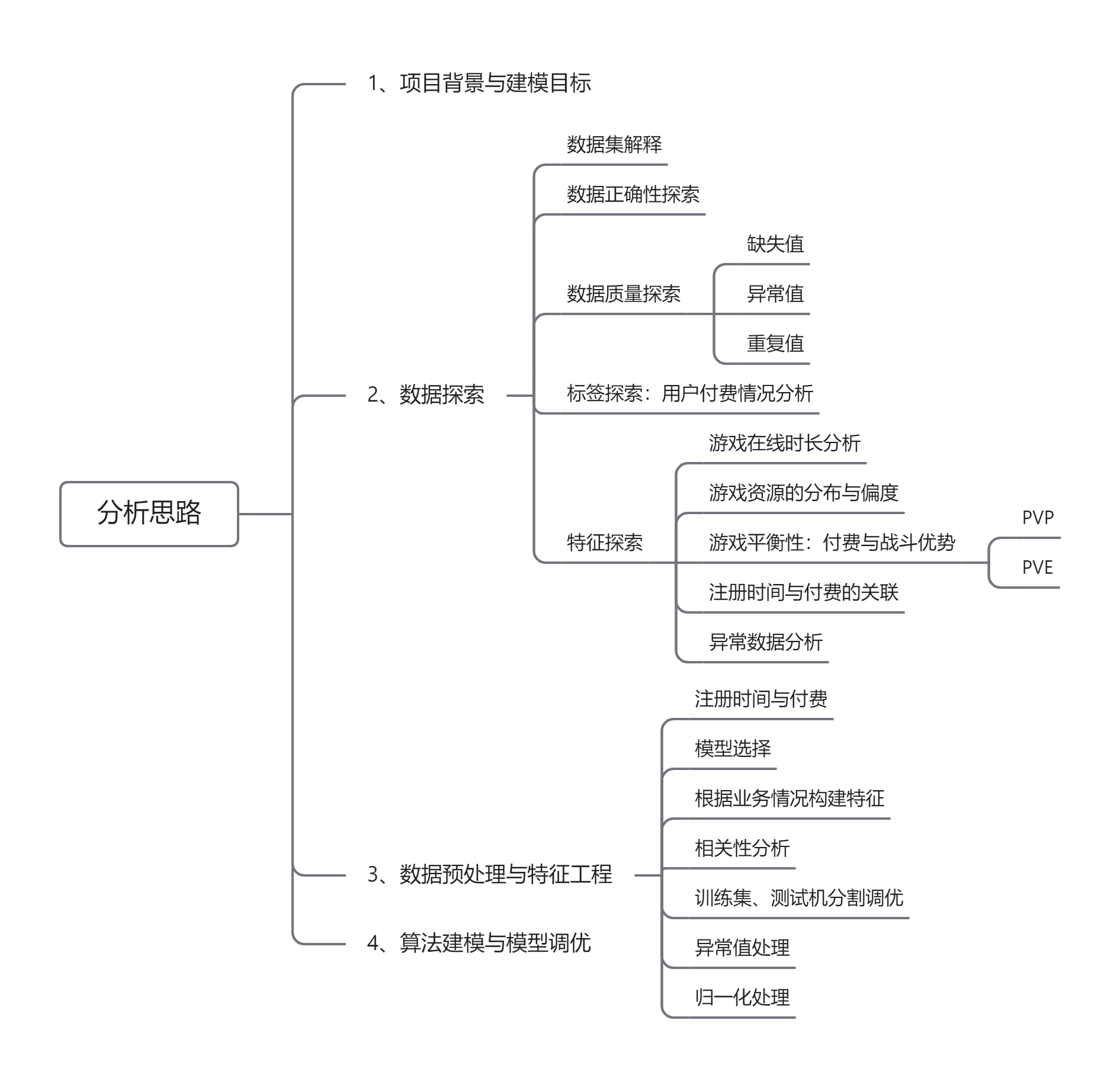
《野蛮时代》付费预测分析

### 一、项目背景与建模目标

本项目是针对SLG手游《野蛮时代》(Brutal Age)用户行为数据的付费预测。该数据集为2018年项目的分析目标是利用玩家在游戏内前7日的行为数据，预测他们每个人在45日内的付费总金额。

分析思路如下：



### 二、数据探索

#### 1、数据集解释

该数据来源于DC竞赛第二届智慧中国杯(ICC)。数据规模为2288007 ×109。数据集的字段主要是：玩家ID，注册时间；资源（石头，木头，象牙，肉，魔法）的获取和消耗数量；角色（勇士，驯兽师，萨满）的招募和损失数量等等。

#### 2、数据正确性探索

数据集‘’玩家唯一ID‘’列无重复

#### 3、数据质量探索

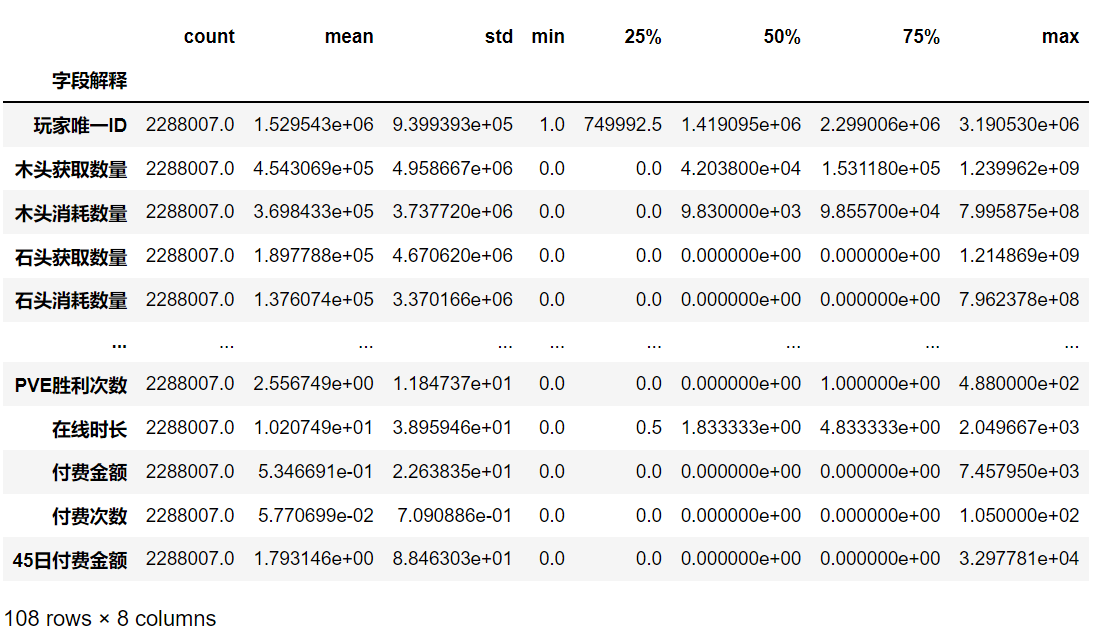
经检验，数据无缺失值和重复值

#### 4、数据类型探索

该数据集有1列数据是对象类型，即注册时间。有95列数据类型都为整数。从游戏业务的独特性、以及特征含义来看，资源数量/等级/发起PVP次数等信息一般不会以浮点数表示，因此显示为整数也十分正常。  
 从特征含义来看，该数据集中没有任何离散型特征。

#### 5、查看分布和统计信息

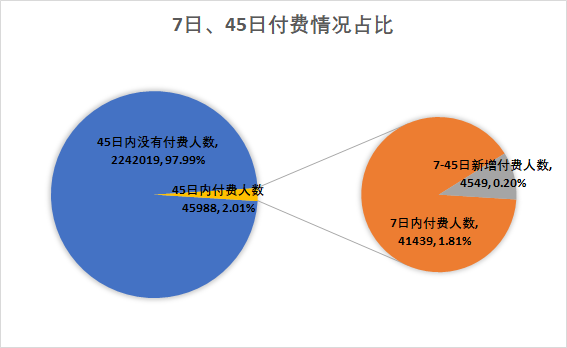
data.describe().T



该数据集右偏程度过高，玩家流失情况可能很严重。高度偏态的信息不利于异常值检测，可以预见大量的长尾数据会被归为异常值，需要谨慎处理异常值。

#### 6、标签探索—付费情况分析

##### （1）7日内与45日内用户付费情况分析



45日内付费率为2.01%，7日内付费率为1.81%。整体来看付费率是偏低的。2016年时全球F2P(free to play)手游的付费率就有2.3%，《野蛮时代》的制作公司Tap4fun具有丰富的SLG手游经验，并且在开服测试阶段就有200w用户数据，转化率应该更高。  
 2020年SLG手游的平均付费率大约在5%左右，顶级手游的付费率能够接近15%。

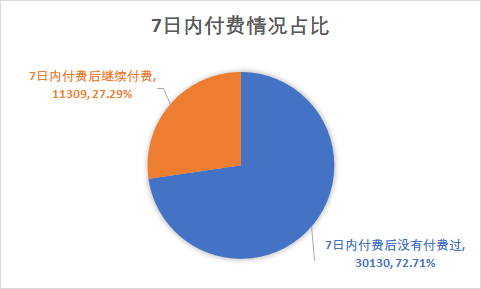
这可能是由于如下原因导致的:

1.大部分用户都留存到了45日，这说明新用户福利过去后，游戏中的转化手段非常贫瘠，转化刺激严重不足；

2.大部分用户在45日之前已经流失，因s此才没有继续转化为付费用户的机会。

##### （2）7日内付过费用户的后续付费情况

###### 7日后的付费情况



7日之前付过费，并且7日之后再也没有付费的用户有3万多人，占到前7天总付费用户的72%。即**一个新玩家在服前七天有付费，但在未来一个月还是有7成的概率不再继续氪金。**

这可能说明：

1、对大部分付费用户而言，最初氪的是新手礼包，后续礼包中的福利以及价格超出了用户对游戏付费的心理预期；

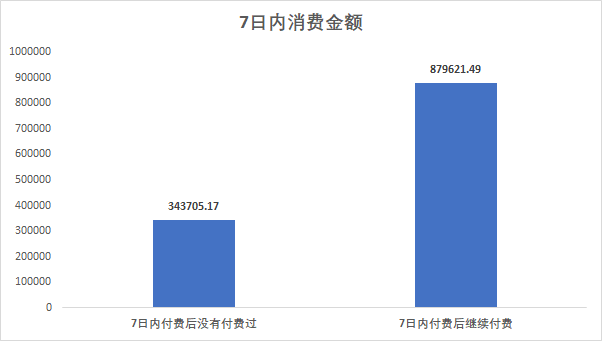
2、新手礼包力度过大，导致后续礼包中的福利看起来不够实惠，用户再付费门槛较高；

3、游戏本身粘性不足，无论用户是否付费，大部分人都在7-45日内自然流失了(甚至在前7日就流失了大部分)；

4、开局需要发育时玩家更愿意氪金解决问题，发育起来之后便不再具有强烈付费需求/发育不起来直接退服了；

5、氪金后体验一般甚至较差，付费后对玩家竞争力无明显提升 。

###### 上述玩家前7日的消费对比



![](data:image/png;base64;base64,)

非常明显，7日后继续付费的玩家在前7日会氪更多的金额，7日后不再付费的玩家中有75%都只氪了6元以下的金额，而7日后继续付费的玩家中有50%以上都氪了16元左右，这极大地印证了第一条理由:**大部分付费用户只购买了新手礼包，而后续礼包的价格在大部分人的心理预期之外。**

同时，7日后不再付费的玩家中也有不少氪过百过千的玩家，最终放弃游戏可能是因为自然流失/氪金后体验不足以满足部分用户的想象。

从7日后继续付费的用户氪金情况来看，这些用户是倾向于越氪越多的，因此这些用户对于氪金后的感受应该是满意的。

高价值用户评判的第一个可能的标准:前7日氪金超过6元的用户是更有价值的用户。

根据历史数据，当用户7日内氪的金额越大，用户对氪金的心理预期越高，继续氪金的可能性就越大，不再氪金的可能性就越小。

因此，必然存在一个金额点，当用户在前7日内付费的金额超过该金额点时，该用户继续氪金的可能性大于不再氪金的可能性。

![](data:image/png;base64;base64,)

在前7日内付费5.7元以下的用户，无论是否自然流失，未来氪金的可能性都是小于继续付费的可能性的。

可以通过设置游戏体验，给与付费5元左右的用户更多的"刺激"，让他们后续继续付费。也可选择直接放弃5.7元以下的付费用户，体验维持现状或减少资源分配，让更多的资源/服务围绕更有可能付费的用户展开。

##### （3）ARPU与ARPPU分析

ARPU = 总付费金额/活跃用户数  
ARPPU = 总付费金额/付费用户数

###### 45日内转化情况



###### 7日内转化情况

  
 目前来看全服的ARPU较低，但ARPPU还不错，这说明虽然付费人数不多，但是高付费玩家的付费能力相对较强。从前5000高氪用户金额占比来看，45日内全服氪金前5000的人(占比小于0.22%)提供了全服89%的收入，贡献巨大。

《野蛮时代》中的氪金行为与玩家竞争力有较为直接的联系，许多大R玩家非常偏爱强PK类游戏，付费可以直接让PVP变得异常简单。

#### 7、游戏在线时长分析

##### 在线时长



75%的人在前7日一共只玩了4.8分钟或以下，这说明75%的用户连新手任务/剧情都没过完就被劝退了。用户以超出常规的方式流失，在这种流失方式下，付费率也很难被拯救。

* **付费用户的在线时长分析**



只有25%以下的付费用户在线时间在半小时以下，游戏时间明显比非付费用户多了很多。但一般来说，投入巨量时间的玩家不会氪太多，游戏时间可以获取需要的资源以及克服游戏中设置的氪金点。

##### 在线时长很少的玩家们贡献的金额

![](data:image/png;base64;base64,)

不难发现，在线时长很少的玩家们贡献的氪金数额都很少。因此**可以将一周在线时长不足20分钟（氪金占比约2%）的用户判断为低价值用户。**

##### 在线时长超长的玩家们贡献的氪金数额

![](data:image/png;base64;base64,)

很多用户虽然深度参与游戏，但是付费占比较低。可以规定，**一周在线时长超出800分钟的用户为(氪金层面)低价值用户。**

#### 8、游戏行为分析

##### （1）游戏资源的分布与偏度

该数据集的数据严重右偏，说明游戏中拥有大量资源人较少。我们已经知道用户在前7天是严重流失的，资源分布会右偏也合理。  
用户严重的流失的原因可能有以下几点：

1、投放到了不太合适的流量渠道；

2、投放的广告效果不好，让用户误以为是其他类型游戏；

3、游戏美工不足、吸引力较低；

4、除了上述理由，用户大量流失还可能是游戏对新手不太友好或者游戏平衡性严重受氪金影响导致的。

在现有数据下，通过分析资源存量来分析游戏中资源获取难度分类的合理性。

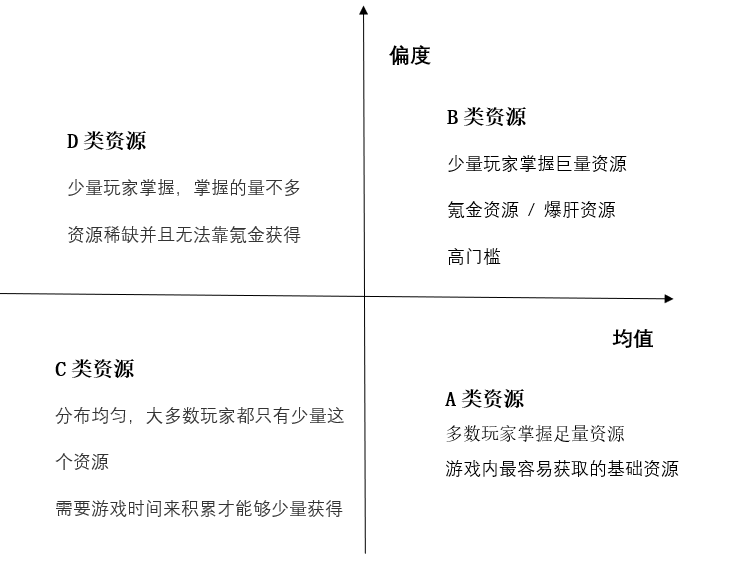
对游戏数据而言，偏度与峰度衡量的内容是类似的:偏度说明分布对称的情况。峰度说明分布集中在少数点的情况，巨大的正偏态和巨大的峰度都指向"大部分玩家只有很少资源/无资源"的状态。  
结合单一资源的均值，我们可以进一步判断资源获取的真实情况。

峰度和偏度似乎呈现高度一致的趋势——当偏度很大时，峰度也倾向于很大。

![](data:image/png;base64;base64,)

因此这里只考虑根据偏度与均值的情况分析游戏资源的合理性。

均值与偏度可以以如下方式分割游戏资源的类型：



在一个健康的游戏中，A类资源应该是足够多的，这类资源决定了大部分玩家的游戏体验，B类资源主要用于付费，收入的主要来源，但可能会严重破坏游戏平衡，D类资源主要在于增加游戏的趣味和公平性，应该是最少的。一个游戏的A类资源要足够多，才算是对新手比较友好。  
 偏度与均值都包括部分极大异常点，因此数据极差非常高，而大部分数据点又集中在较小的值附近，这样的数据在绘图时很容易出现以下情况:

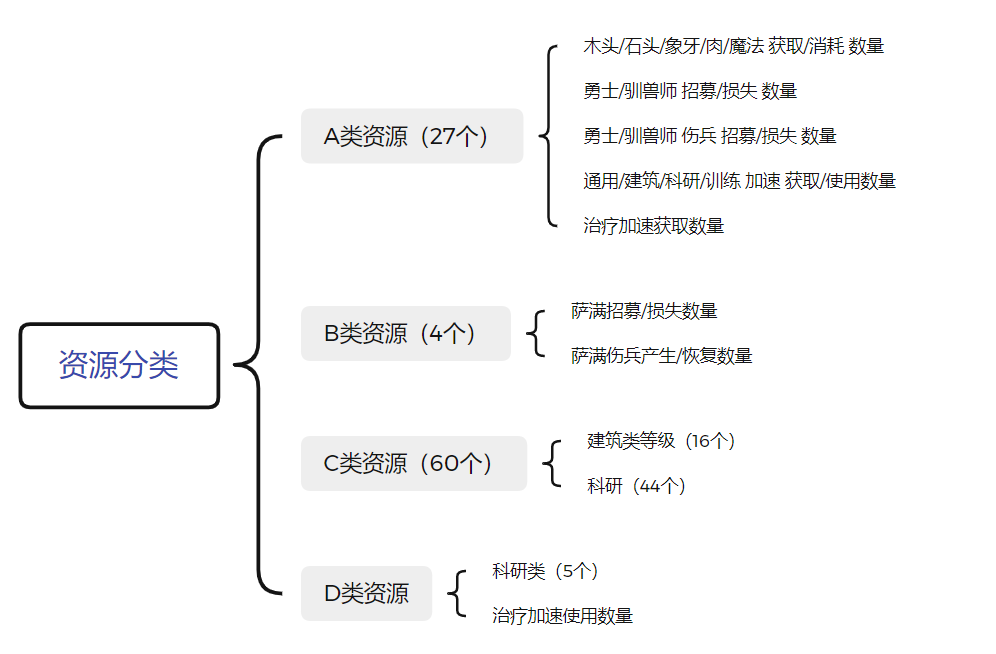
![](data:image/png;base64;base64,)

为了解决这个问题，需要绘制三个纵坐标数值不同的图像：

![](data:image/png;base64;base64,)

规定：偏度>300认为高，资源均值>3认为高。

查看一下四类资源分别有哪些，以及数量是否合理：



从结果来看，各项资源的配比数量还算合理。  
 在游戏早期只有兵种萨满相关的集中资源和技能是需要氪金的，说明游戏依然在致力于为普通玩家提供较好的体验。所以大部分用户应该不是因为新手关太难、剧情无法推进才退游的。  
 而从A类资源的分布来看，只有木头和肉这两种最基础的资源是分布较为正常的，其他资源都是严重右偏，说明"留存率低"是原因，“资源数据的异常"是结果。  
 比较可能的情况就是游戏的新手入门逻辑不顺，使玩家获得感不足、美工不足、系统BUG等让大部分用户在第一天就流失掉了。

##### （2）平衡性影响—付费与战斗优势

分析氪金用户与不氪金用户的战力差异，来判断氪金对游戏平衡性的影响。

##### PVP战斗的参与程度对比分析——付费与非付费

###### 7日内玩家参与PVP情况

* ![](data:image/png;base64;base64,)

###### 7日内玩家的人均PVP次数

![](data:image/png;base64;base64,)

付费玩家人均参与的PVP战斗次数、参与PVP的比例远高于非付费玩家。 当然，不氪金用户中有许多人可能是已经流失的用户，因此不会参与PVP对战。

###### 7日内玩家PVP胜率均值对比

![](data:image/png;base64;base64,)

付费玩家人均PVP胜率是53.3%，对手游付费玩家来说是个可以接受的数字，但并不是特别大的优势。对非付费玩家而言，人均PVP胜率只有25%，并且75%的人的场均胜率不到50%，这个条件对非付费玩家来说的确比较苛刻。

###### 前7日主动发起PVP的比率对比

不付费玩家平均26%的时候主动发起对战，不过大多数时候75%的人都不会发起对战。**PVP状况下是付费玩家占绝对优势。**

##### PVE战斗的参与程度对比分析——付费与非付费

###### 玩家PVE胜率对比

不氪金玩家的PVE平均胜率有92%，是高于氪金玩家的胜率88%的，看来环境怪兽都比较弱小,大部分玩家都能够轻松战胜。  
当PVP上氪金玩家有绝对优势时，不氪金玩家许多会选择在PVE环境中进行战斗,因此我们推断不氪金玩家主动发起PVE战斗的情况会更多:

氪金与不氪金玩家主动发起PVE的可能性都非常高，几乎所有人都是主动发起PVE战斗。  
这可能说明PVE战斗是发展城市必须的环节(比如收集资源肉类需要猎杀野猪等生物、获取魔法道具需要猎杀龙、哥布林等生物)。

从分析的结果来看，**氪金并不会对PVE战斗有影响，主要是影响PVP。**

《野蛮游戏》中玩家可以夺取其他玩家的城池，因此PVP对战的平衡性对玩家流失有重要的影响。

前期留存下来，但后期逐渐流失的非氪金用户有一定的可能是因为PVP胜率太低而离开。

##### （3）异常玩家分析

在特征高度右偏的情况下，数据中必然含有大量的异常值，这些异常值可能代表着某些特殊的人群，但带入异常值进行建模会严重影响模型的稳定性，导致模型容易过拟合。因此在建模之前，我们需要对异常值进行简单的探索，以确定异常值的处理方法。

###### 查看业务上的异常值

从游戏业务角度而言，所有的资源/人物等级/胜利次数/金额等不应该出现负数，因此首先排查负数  
如果存在用户信息，则需要排查年龄不为负，年龄不过高或过低等状况  
查看是否有上线时间较短，但是资源量异常丰富的账号，可能是GM，也可能有人作弊

###### 数据分布上的异常

使用箱线图法则，1123706数据被归为异常，数量太多。经计算，前7日付费用户被100%包含在了异常用户里。付费用户的数据在现有数据的背景下，(大部分用户短时间快速流失掉)下，在统计上是完全异常的，因此建模时想要辨认出付费用户，模型需要很深入的学习，有很高的过拟合风险。异常用户有100w，其中只有4w多付费，即许多与氪金用户行为高度相似的用户都没有氪金，对算法来说这部分用户是很难判断的。  
 在进行算法建模时，用更加灵活的方法来处理这些异常数据。

### 三、数据预处理与特征工程

##### 1.注册时间与付费

将时间分成日期和时刻来考虑，分别绘制横坐标为注册日期/注册时刻、纵坐标为该日期/该时刻注册用户氪金均值的关系图，以此来观察注册日期/时刻与氪金是否有关。

##### 不同注册日期的用户氪金均值

![](data:image/png;base64;base64,)

##### 不同注册日期的用户付费总额与玩家注册数量

![](data:image/png;base64;base64,)

从上图来看，2018-02-19（春节期间）新注册玩家数量激增，且高峰的持续时间很短，可能是因为公司对游戏的推广取得了效果进而使注册用户激增。从游戏运营方的微博来看，春节期间运营方在游戏中开展了一系列活动，包括舞狮活动、酋长参战、跨服战、爆竹活动等，这些活动和春节效应叠加，带来了注册人数的增加，出现峰值。

但大量的新增注册用户并没有给转化情况带来明显改善。从付费均值来看，几乎无法看出日期与氪金数额有直接联系。但是每隔几天就会有一个高峰。  
 将日期转化为一星期中的七天，观察一下星期和氪金数额是否有明显的关系。

##### 不同注册日期（星期）的用户氪金均值

![](data:image/png;base64;base64,)

周五、周一注册的用户有较高的氪金金额，并且在高峰之后会逐渐下降。可能跟游戏投放的节奏和渠道有关。可能每周五、周一可以排到质量较高的用户所在的渠道，或者周五的高峰暗示这两天进入的流量可能是有规律工作、有经济收入的人，玩家付费能力比较高。  
 从特征的角度来看，可以计算一下星期与45日付费金额的相关系数：



相关系数非常非常小，两者几乎不相关，可以不将星期放入矩阵特征。

##### 不同注册时刻（小时）的用户氪金均值

![](data:image/png;base64;base64,)

早上10点-12点之间注册的用户、以及下午4点-5点之间表现出比其他时段注册的用户更高的氪金能力。

相关系数:



看起来注册时刻与付费金额有一点点关系(非常微弱)，可以保留注册时刻为建模用。

##### 特征修改

增加'玩家注册时刻(小时)'列，删除‘’玩家ID列‘’、‘’删除玩家注册时间‘’列，不是有助于建模的特征。

##### 2、模型选择与benchmark

在现有数据量的要求下，首先要考虑计算迅速且简单的线性模型:线性回归。

##### 建立benchmark

###### 训练集上的(Xtrain,Ytrain)

0.5604160157526339

###### 测试集上的(Xtest,Ytest)

0.5581998030190463

模型处于欠拟合状态——在训练集和测试集的分数上都不是很高，并且两者分数相近，简单的线性模型在学习能力上略显不足。

###### RMSE

62.00116183660748

##### 3、根据业务模式新增特征

1、个人PVP胜率=PVP胜利次数/PVP次数，无PVP的用户胜率为0

2、主动发起PVP的概率=主动发起PVP次数/PVP次数，无PVP的用户主动发起为0

3、玩家发育效率=平均资源获取数/在线时长(该指标高的人擅长游戏)

4、玩家升级效率=技能的平均等级/在线时长

5、氪金发育效率=平均资源获取数/7日付费金额（该指标高的人善用氪金资源)

6、氪金升级效率=技能的平均等级/7日付费金额

修改除0错误与极值错误。

在之前分析游戏的经营状况时，得出了一些关于用户价值的关键结论：

1、如果用户**7日内付费超过5.7元**，则在7日后继续付费的概率大于不再付费的概率，是高价值用户。

2、当用户的**在线时间小于20或者大于800分钟**，用户对氪金金额的贡献率较小，是低价值用户。

3、**主动发起PVP概率大于50%**的用户贡献了较多氪金金额，是高价值用户。

基于这些发现，统—让低价值用户被标记为0，高价值用户被标记为1，为特征矩阵创造如下新特征:

**高价值玩家**

潜力玩家 : 7日内付费超过5.70元

好战玩家 : 主动发起PVP概率大于50%

**低价值玩家**

肝帝玩家:7日在线时间过长大于800分钟

流失玩家:7日在线时间小于15分钟

#### 4、相关性分析:筛选特征/特征重要性

现在模型有190多个特征，为了能够增加计算速度，可以简单使用相关系数对特征进行筛选。最简单的方式就是使用皮尔逊相关系数。简化一下:只考虑特征与标签之间的相关性，不考虑特征与特征之间的相关性带来的统计学方面的问题。只有两个特征的相关系数低于0.01，可以考虑都统一保留这些特征。

#### 5、训练集、测试集分割的调优

线性回归是是复杂度较低的模型，学习能力和抗过拟合能力有限，因此训练集过多或者过少都很容易导致过拟合，为此可以尝试几组训练集测试集的分割比例，以确定最适合线性回归的训练集数量。

##### 测试集与训练集上的

![](data:image/png;base64;base64,)

##### 测试集上的RMSE

![](data:image/png;base64;base64,)

testRMSE：[65.40513888232609, 61.25254728964682, 57.409896451713955, 72.69313409149458]

从结果来看，在测试集上RMSE最低的比例是0.4，从R2的表现来看，测试集比例为0.4时，训练集上R2表现出过拟合的倾向，但并不严重。

但考虑到现在模型的核心评估指标是RMSE，因此R2稍有问题，但RMSE明显有优势时，应该优先考虑RMSE上的表现。按照0.4比例分割后，不恢复索引，而让索引跟随Xtrain，Xtest分割后的样子。

#### 6、 异常值处理

根据箱线图法则，所有的付费玩家都被归类到了异常数据里，这与付费玩家往往都是深入参与到游戏的玩家有很大的关系。由于付费玩家都被包括在异常数据中，不能对异常数据进行简单的删除。数据右偏严重，异常值数量众多(超出100w)，删掉50%的数据也不可取。因此在这，可以对异常值进行类似于盖帽的特殊处理。

按照箱线图的规则进行异常检测，并对异常值中7日内没有付费的玩家进行用0覆盖，对于7日内氪金的玩家则不处理，尽量放大付费玩家与非付费玩家特征上的区别。

**注意:异常值处理必须在训练集测试集被分割的情况下进行，因为测试集上的异常值处理是必须依赖对测试集进行异常值处理所产生的中间变量的**

#### 7、归一化处理

使用训练集上的最小值和最大值对训练集/测试集同时进行归一化

### 五、算法建模与模型调优

所有的付费用户都被包括在了异常用户中,这说明有大量非付费的用户表现在游戏资源积累、游戏参与深度上的行为与付费用户极为相似。同时,氪金用户与不氪金用户的用户组成都相对复杂。

在之前的特征处理中，已经对异常值进行过“盖帽"处理，并在特征中增加认为可以让付费和非付费用户区别变大的特征，但直接让单模型对付费用户的氪金数额进行预测是非常困难的。 因此，可以使用2个模型进行融合预测，第1个模型是分类模型逻辑回归，我们让逻辑回归先对用户进行“氪与不氪”的预测，然后将逻辑回归认为会付费的用户放入线性回归，再让线性回归直接对“本来就很可能会付费”的用户进行氪金金额预测。通过这样的方式经过适当调整后,应该能够提升模型表现。

##### 创造逻辑回归使用的y2

付费用户转化为1，不付费用户转化为0



已经有了Xtrain,Xtest,线性回归用的Ytrain,Ytest,逻辑回归的y2, 变成可以和Xtrain,Xtest相匹配的y2train,y2test。

按分割后的Xtrain和Xtest的索引来分割y2，那么y2就是与Ytrain相匹配的。

##### 使用逻辑回归进行建模并使用ROC进行评估

逻辑回归在训练集上的准确率（xtrain,Ytrain2)：0.9972042622253432

逻辑回归在测试集上的准确率（xtest,Ytest2)：0.9972639949825339

ROC：0.9711120160009642

召回率（Recall）：0.8637353069220722

精确度（Precision）：1.0

逻辑回归的效果很好，能够在样本量如此不均衡的情况下达到0.97的ROC，证明数据与逻辑回归很适配。recall和precision结合起来看说明，被预测为1的值都预测正确了，但是还有部分被预测为0的值，实际上真正的值是1。

如果按照现在的结果建模，那RMSE的主要来源就是没有被逻辑回归包括的付费用户的付费金额＋被逻辑回归指向付费、受线性回归预测后的金额差异（状况A)。如果希望逻辑回归能够捕捉到所有真实付费的用户，即便误伤一些没有付费的用户，那么就要努力提升Recall，此时RMSE的主要来源就是被逻辑回归误认为是付费用户的非氪金用户在线性回归中的预测值＋全部真实氪金用户经过线性回归预测后的金额差异（状况B)，该状况下需要线性回归预测的值会更多。

##### 尝试调节阈值提升Recall

经过之后的RMSE判断，阈值=0.02是个比较合适的值。

对于被逻辑回归判断为不付费的样本，预测到此结束，选出应该放在线性回归中的训练集和测试集。开始回归预测。

##### 回归模型的选择

###### 1、线性回归

训练集R2:0.582  
测试集R2:0.537  
测试集RMSE:288.430

###### 2、随机森林

训练集R2:0.515  
测试集R2:0.431  
测试集RMSE:319.896

###### 3、GBR

训练集R2:0.737  
测试集R2:0.531  
测试集RMSE:290.533

在不是很严格的限制下，随机森林表现不如GBR。GBR和线性回归差不多，因为线性回归参数调节性小，因此选择GBR。

##### 模型调优

学习曲线、交叉验证、网格搜索计算成本过高，使用经验调参。确定参数后：

训练集R2:0.792  
测试集R2:0.564  
测试集RMSE:296.464

##### 将逻辑回归与GBR的预测值合并

RMSE=55.113101108167164