# 观察

## 数据的基本样子

## 训练数据

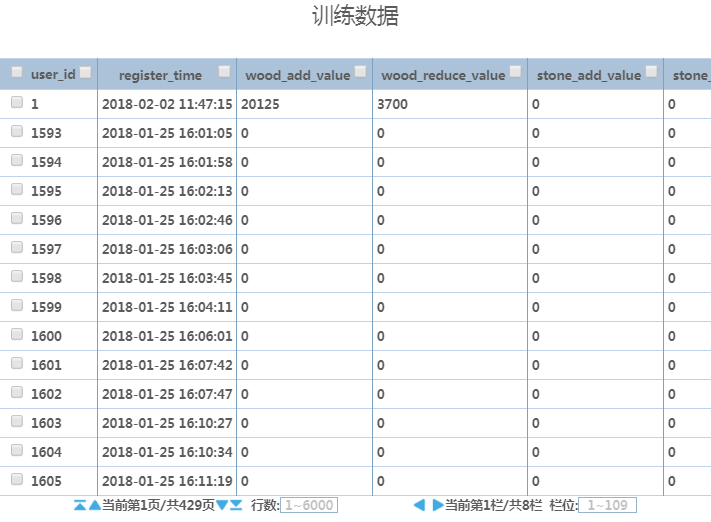


图1-1



图1-2



图1-3

## 预测数据



图1-4



图1-5

## 看一下是否每条记录对应一个id

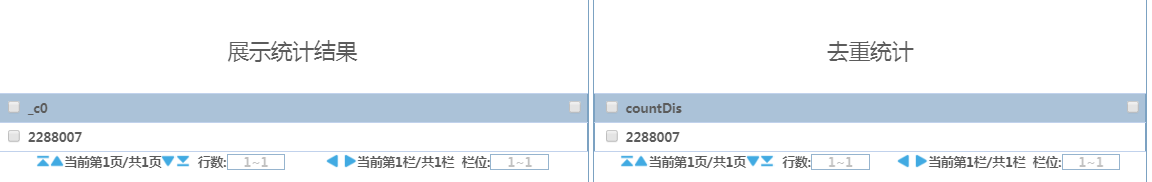


图1-6

是一样的

## 看下付款情况统计

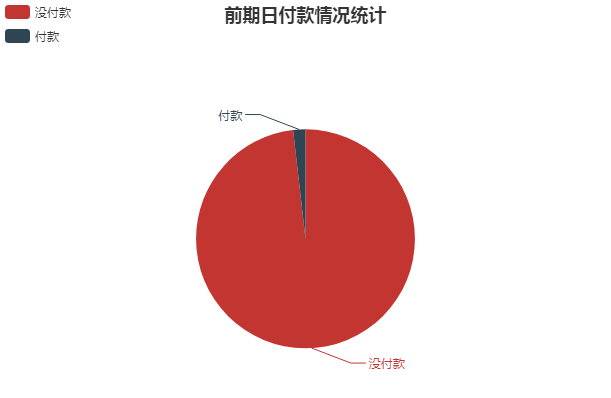


图1-7

其中付款人数41439，占比1.81%

没付款人数2246568，占比98.19%

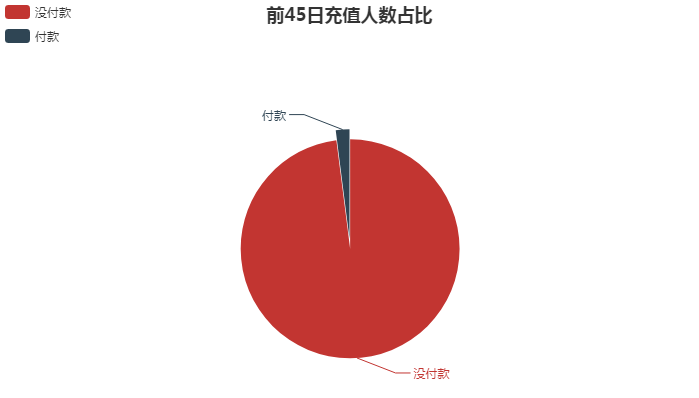


图1-8

其中付款人数45988，2.01%

没付款人数2242019，97.99%

0-7日和7-45日充值轨迹统计



图1-9

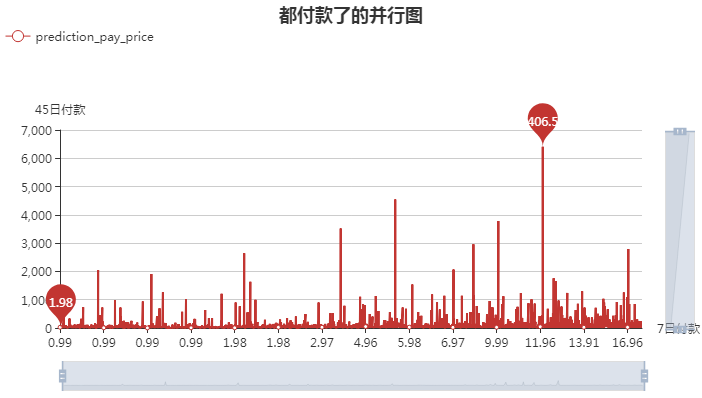


图1-10

发现0.99太多了，而且前7日充值大约都是是0.99的倍数



图1-11

之前充值集中在0.99一线，0.99可能是首次充值的一个集中情况。

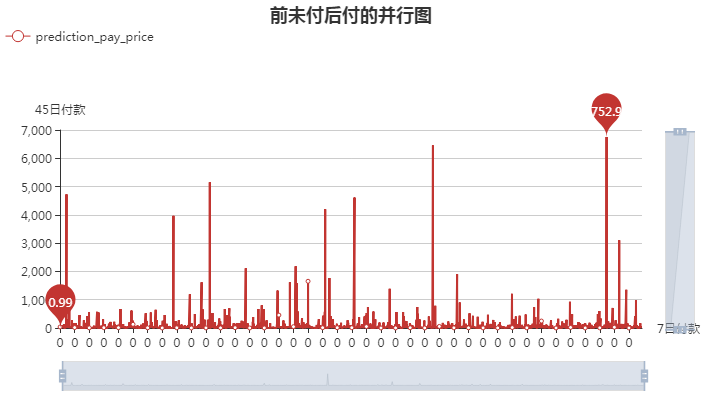


图1-12

观察“都付了”的玩家数据

发现：

1）有前面冲了几十后面不充的

2）有前面充了一两块后面充了一百多的

3）有前面充了1000多后继乏力的

4）大部分前面充的多于10元

## 高充值玩家



图1-13

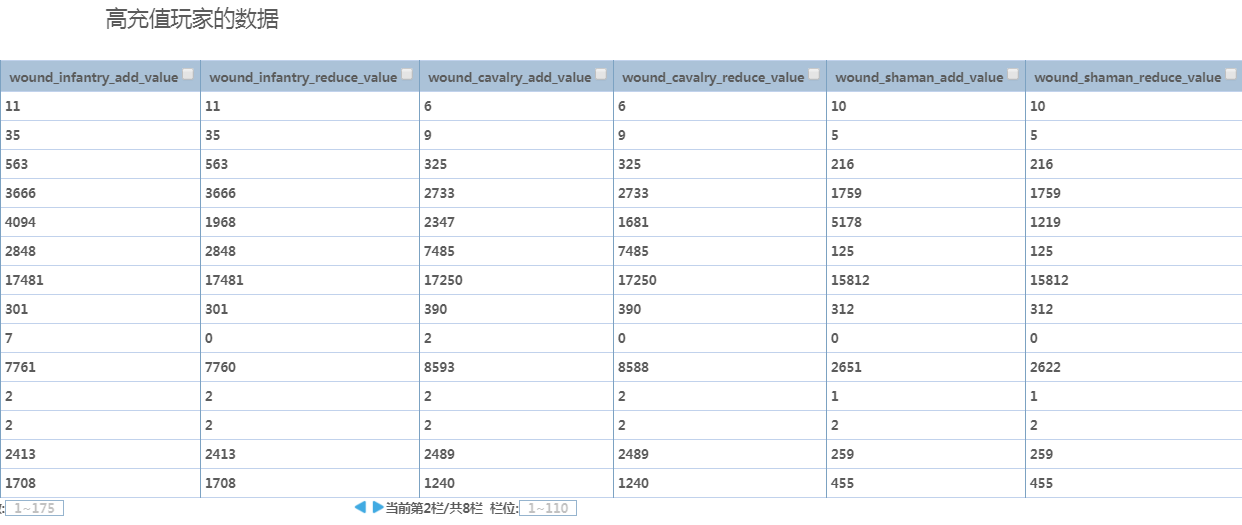


图1-14

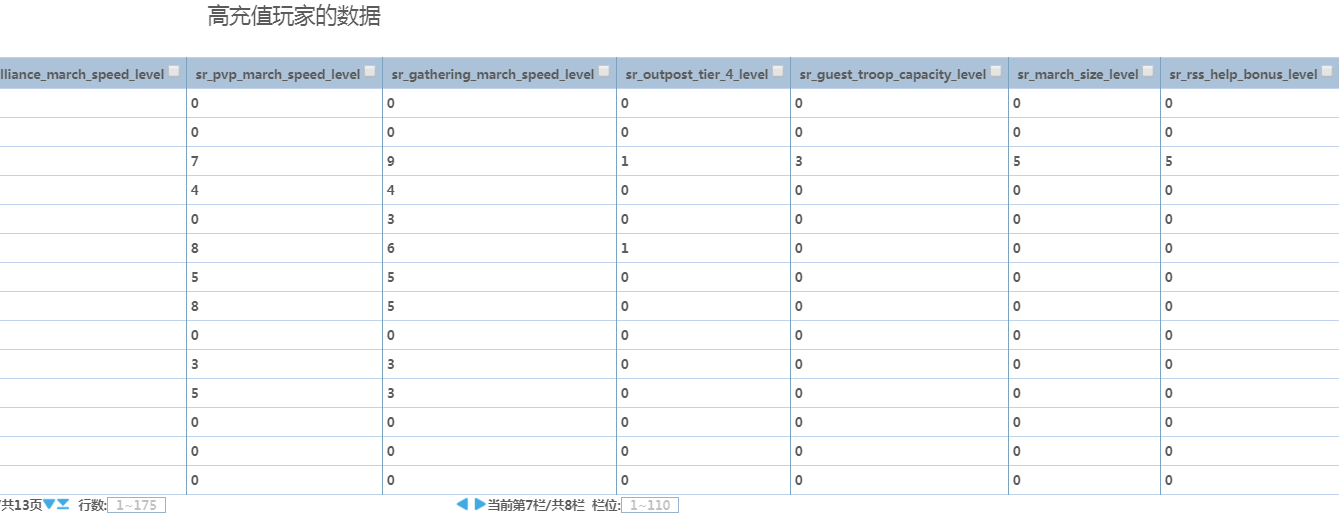


图1-15

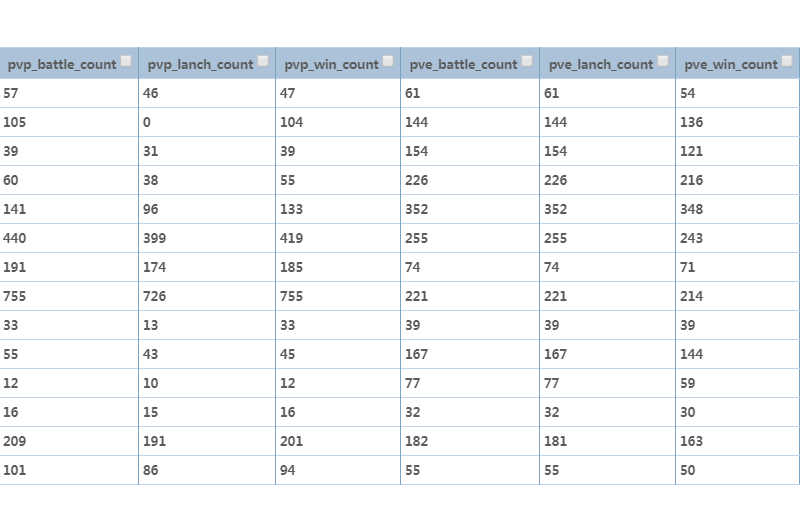


图1-16



图1-17

发现

1）伤病低，部分科技分布不均匀，高充值玩家的各项指标都很好。

2）PVP，PVE乘积很好

3）上线时间长

4）有些前面充了两千多的后面没充，是不是被封号了？

5）注册时间似乎有规律？周六日？上班时间？很晚？

看下时间规律：

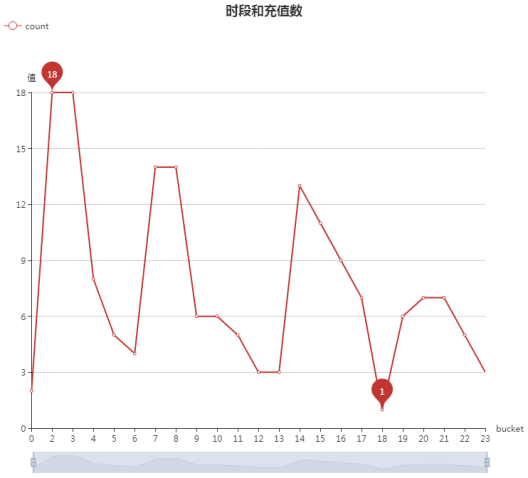


图1-18 图1-19

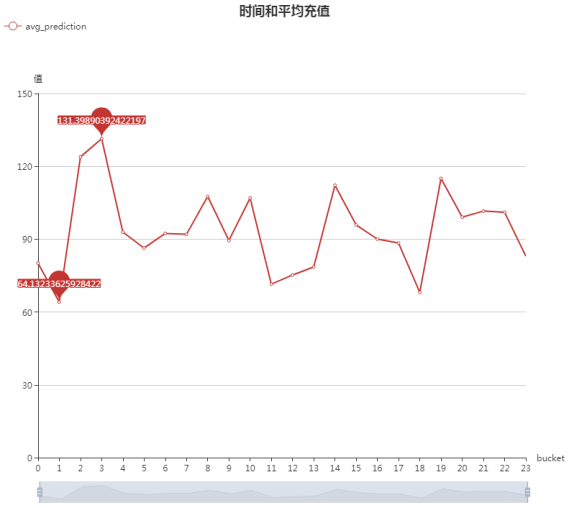
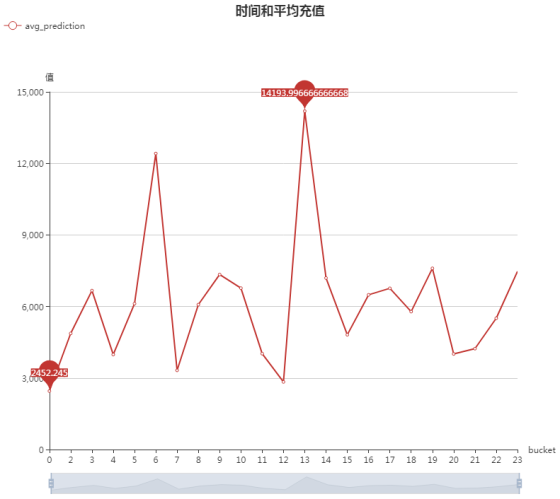


图1-20 图1-21

星期

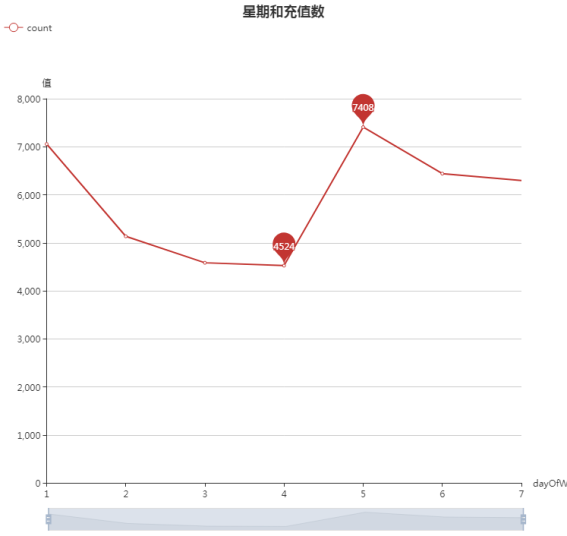
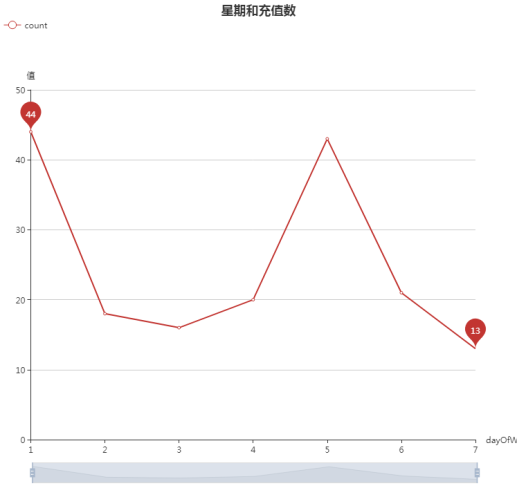


图1-22 图1-23

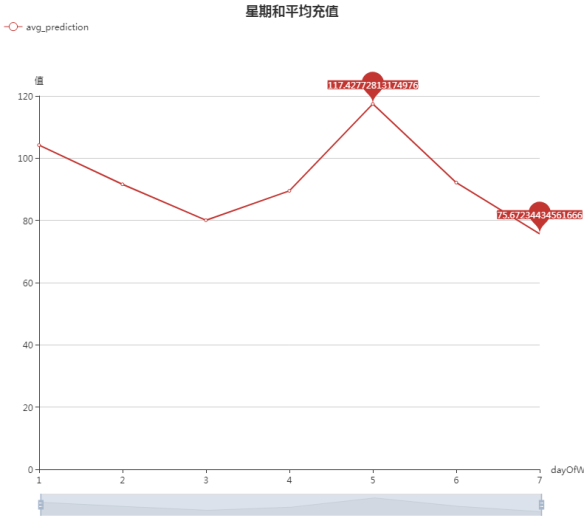
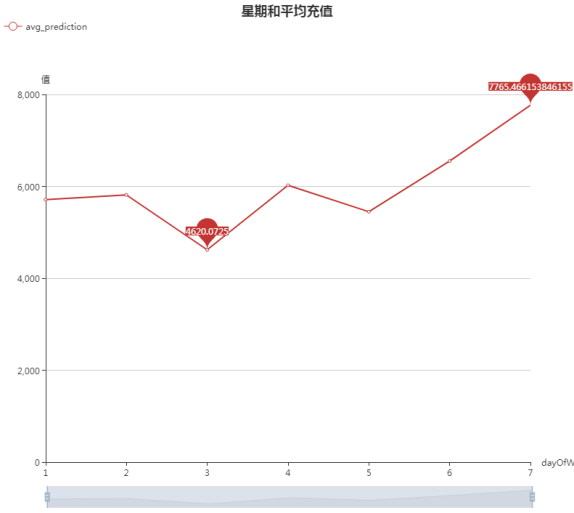


图1-24 图1-25

发现整体趋势和星期以及注册时段有密切关系

在晚间02:00左右注册的人充值潜力最大，14:00注册的人充值潜力最大

在周一和周五注册的充值人群最多，平均充值数也最大。

通过以上说明星期和时段对充值影响是比较显著的。

## 看一下整个充值的分布情况

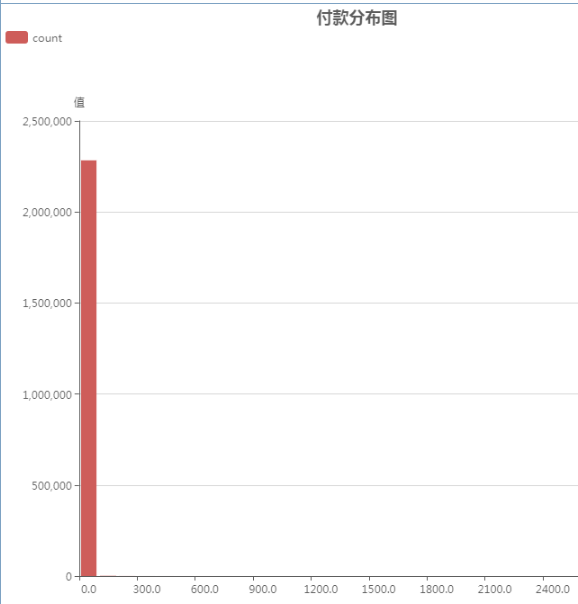


图1-26

发现有严重的结尾特性。0值再的时候其他几乎可以忽略不计。

去掉0值，发现分布如下

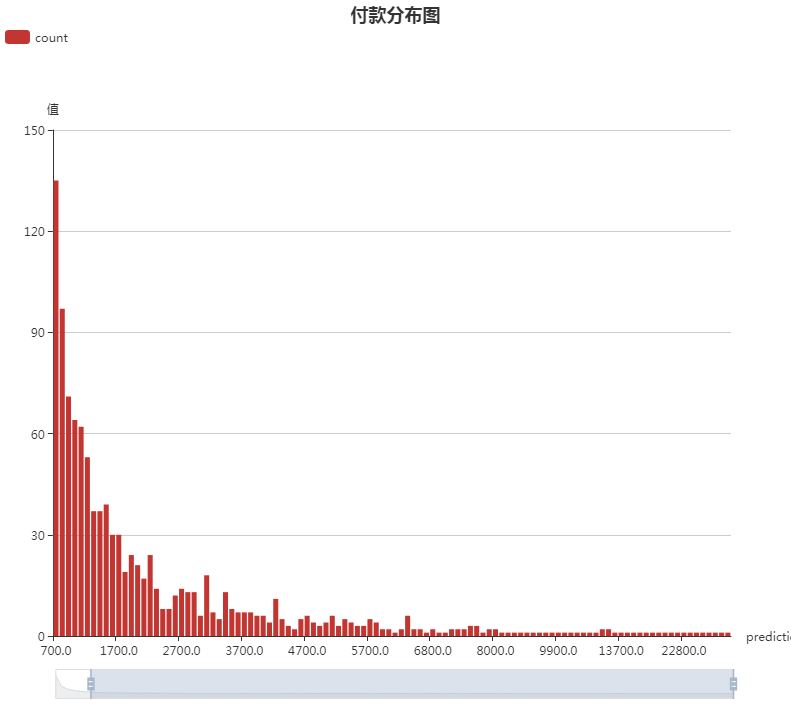


图1-27

发现呈指数式的分布，且有厚尾特性，最后8000+的厚尾玩家需要看一下。

# 发现

1. 时间只是注册时间，数据每条记录代表一个人
2. 根据图1-7/1-8/1-9发现，付款的是2%左右的人，前7日付款人数41439中有约72.7%后期未付款，前7日付款人数41439约有27.3%后期仍然付款。前期没付后期付款的人数约4549占总付款人数45988的9.9%

# 想法

## 大致想法

每个玩家在每段时间内有一个玩家行为的六芒星，这个六芒星在不断的变化，这对玩家充值情况有一定印象

## 以一定的多重共线性

1. 前七日的充值会影响到前45日的充值，这是一个很重要的影响，=> 可以先看一下前七日充值和7-45日充值的xtabs
2. 内部变量也有一定共线性，很多影响是差不多的， => 可以看下隐含变量或者主成分
3. 最终评价结果是RMSE因此着重照顾好搞充值的客户。

## 模型可能有一定的网格特性

不同区域的影响因素效果是不同的，可以使用决策树+监督学习的方法，先用决策树根据特征对对象网格化，每个网格内对应一定的模型，最后组合起来。

当然也可以使用xgbt。

另外模型也有一定的近邻特性

## 特征大致有以下几大类

资源数据，该类和游戏表现成正比

伤病类别

科技类别

PVP、PVE类别

平均在线时长

7日充值

特征处理过程可以通过以下方法：

1. 重要的几个类别的特征分别提取主成分，然后进行处理
2. 可以是通过两三层神经网络，提取隐含层，然后再进行处理

尝试使用集成学习方法将一些机器学习模型结合

## 模型

数据因变量有以下特性：

1. 因变量是严重截尾的
2. 指数分布+长尾

对模型方面的思考可能是这样的：

1. 对付严重截尾，可以先分类，确定不充值人群后在对充值人群进行回归模型。
2. 对付指数分布可以取对数。

# 实现

## 特征选择

分组PCA 进行特征选择

直接先进行特征选择在进行分组PCA

## 嵌套模型