***一.解决的业务逻辑：***

给用户一张优惠券，已知优惠券的信息（店铺所属，优惠率，距离，领券日期）,预测用户是否会在15天内使用发的优惠券

|  |  |
| --- | --- |
| Field | 业务分析 |
| User\_id |  |
| Merchant\_Iid | 用户对商店的感兴趣程度 |
| Coupon\_id | 用户对优惠券是否感兴趣 |
| Discount\_rate | 折扣 |
| Distance | 距离 |
| Date\_received | 领优惠券的日期（周一到周日领的优惠券是否有影响） |

机器学习比较有意思的地方是，它能够对你不曾做出过的行为进行预测

***二.数据分析：***

表一解析：

1数据分布分析

包含539438个用户，正样本75382，负样本701602；共8415个店铺，共有9739个优惠券(id)，46种优惠类型(rate)

2缺失值分析

每列非null的统计

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| User\_id | Merchant\_id | Coupon\_id | Discount\_rate | distant | Date\_received | date |
| 1754883 | 539438 | 1053281 | 1053281 | 1648880 | 1053281 | 776983 |

分析：

1).用户和商铺没有缺少值，优惠券的id和领取时间一一对应；

2).距离的值要大于优惠券的值，但是小于总记录值，说明距离这个属性既记录了非领优惠券的消费情形下的距离，还记录了有优惠券时的距离，不过有缺失

三类行为下的距离缺失数

|  |  |  |  |
| --- | --- | --- | --- |
| 总缺失 | 日常消费 | 负样本 | 正样本 |
| 106003 | 0 | 97786 | 8217 |

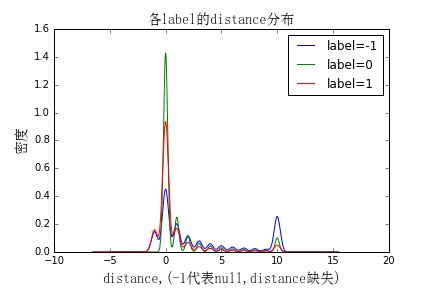
3)由表分析，距离只有在正负样本里面有缺失，即**领了优惠券的用户有距离缺失**；

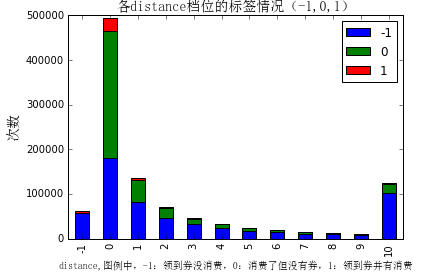
3购买行为分析：

经统计，总共有用户539438，其中有218095（近40%）的用户在半年内没有购买行为；

折扣的分析：

4.距离属性分析



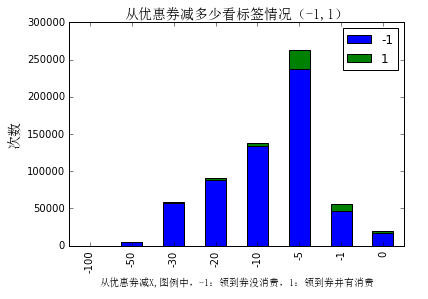
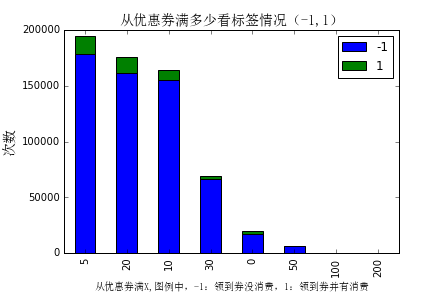


分析：

距离的特性满足一个正太分布，线下的物品越近，购买的概率会越大；

5.优惠力度属性的分析：

把优惠券折成两部分：满多少和减多少来分析其在正负样本中的分布



分析：可以把折扣拆开进行分析；

表二解析（获得用户相似度,，购买能力）

包含762858个用户，共7999个店铺，共有27748个优惠券（id）,65种优惠类型(rate)

1. action=0,1,2的分布

其中0：点击----- 9401780

1：购买----- 1372148，其中优惠券购买216459

2：领优惠券------- 655898

2．Fix的优惠券为131546，而且全都有购买了行为，说明只要用户有fix的优惠券，那么它就一定会购买

3.action = 2的用户就是标记领优惠券的时间；

***三：模型及特征***

使用模型：

特征：

***四：模型评估：***

AUC

实验结果：