[赛题背景： 2](#_Toc1036394458)

[赛题任务： 2](#_Toc774344481)

[字段表： 2](#_Toc2102488040)

[数据探索： 3](#_Toc522062787)

[数据集划分： 8](#_Toc1155853197)

[特征工程 8](#_Toc1800560674)

**生活大实惠：O2O优惠券使用预测**

**赛题背景：**

**随着移动设备的完善和普及，移动互联网+各行各业进入了高速发展阶段，这其中以O2O（Online to Offline）消费最为吸引眼球。据不完全统计，O2O行业估值上亿的创业公司至少有10家，也不乏百亿巨头的身影。O2O行业天然关联数亿消费者，各类APP每天记录了超过百亿条用户行为和位置记录，因而成为大数据科研和商业化运营的最佳结合点之一。 以优惠券盘活老用户或吸引新客户进店消费是O2O的一种重要营销方式。然而随机投放的优惠券对多数用户造成无意义的干扰。对商家而言，滥发的优惠券可能降低品牌声誉，同时难以估算营销成本。 个性化投放是提高优惠券核销率的重要技术，它可以让具有一定偏好的消费者得到真正的实惠，同时赋予商家更强的营销能力。本次大赛为参赛选手提供了O2O场景相关的丰富数据，希望参赛选手通过分析建模，精准预测用户是否会在规定时间内使用相应优惠券。**

**赛题任务：**

**通过分析线上和线下的数据分析，预测用户在2016年7月领取优惠券后15天以内的使用情况。**

**提供数据：**ccf\_offline\_stage1\_test\_revised.csv，ccf\_offline\_stage1\_train.csv，ccf\_online\_stage1\_train.csv，sample\_submission.csv****

**字段表：**

Table 1: 用户线下消费和优惠券领取行为(ccf\_offline\_stage1\_train.csv)

|  |  |
| --- | --- |
| **Field** | **Description** |
| User\_id | 用户ID |
| Merchant\_id | 商户ID |
| Coupon\_id | 优惠券ID：null表示无优惠券消费，此时Discount\_rate和Date\_received字段无意义 |
| Discount\_rate | 优惠率：x \in [0,1]代表折扣率；x:y表示满x减y。单位是元 |
| Distance | user经常活动的地点离该merchant的最近门店距离是x\*500米（如果是连锁店，则取最近的一家门店），x\in[0,10]；null表示无此信息，0表示低于500米，10表示大于5公里； |
| Date\_received | 领取优惠券日期 |
| Date | 消费日期：如果Date=null & Coupon\_id != null，该记录表示领取优惠券但没有使用，即负样本；如果Date!=null & Coupon\_id = null，则表示普通消费日期；如果Date!=null & Coupon\_id != null，则表示用优惠券消费日期，即正样本； |

Table 2: 用户线上点击/消费和优惠券领取行为(cf\_online\_stage1\_train.csv)

|  |  |
| --- | --- |
| **Field** | **Description** |
| User\_id | 用户ID |
| Merchant\_id | 商户ID |
| Action | 0 点击， 1购买，2领取优惠券 |
| Coupon\_id | 优惠券ID：null表示无优惠券消费，此时Discount\_rate和Date\_received字段无意义。“fixed”表示该交易是限时低价活动。 |
| Discount\_rate | 优惠率：x \in [0,1]代表折扣率；x:y表示满x减y；“fixed”表示低价限时优惠； |
| Date\_received | 领取优惠券日期 |
| Date | 消费日期：如果Date=null & Coupon\_id != null，该记录表示领取优惠券但没有使用；如果Date!=null & Coupon\_id = null，则表示普通消费日期；如果Date!=null & Coupon\_id != null，则表示用优惠券消费日期； |

Table 3：用户O2O线下优惠券使用预测样本(ccf\_offline\_stage1\_test\_revised.csv)

|  |  |
| --- | --- |
| **Field** | **Description** |
| User\_id | 用户ID |
| Merchant\_id | 商户ID |
| Coupon\_id | 优惠券ID：null表示无优惠券消费，此时Discount\_rate和Date\_received字段无意义 |
| Discount\_rate | 优惠率：x \in [0,1]代表折扣率；x:y表示满x减y。单位是元 |
| Distance | user经常活动的地点离该merchant的最近门店距离是x\*500米（如果是连锁店，则取最近的一家门店），x\in[0,10]；null表示无此信息，0表示低于500米，10表示大于5公里； |
| Date\_received | 领取优惠券日期 |

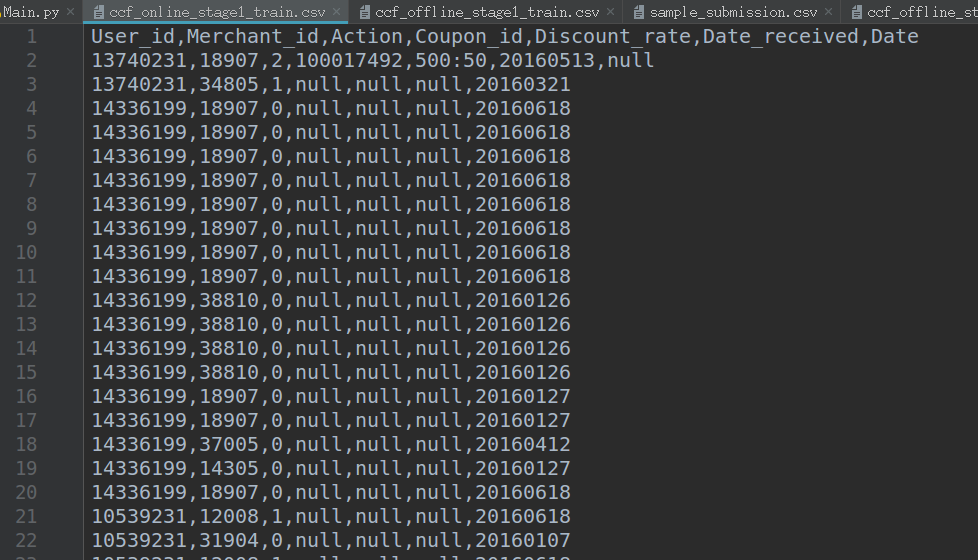
Table 4：选手提交文件字段，其中user\_id,coupon\_id和date\_received均来自Table 3,而Probability为预测值

|  |  |
| --- | --- |
| **Field** | **Description** |
| User\_id | 用户ID |
| Merchant\_id | 商户ID |
| Coupon\_id | 优惠券ID：null表示无优惠券消费，此时Discount\_rate和Date\_received字段无意义 |
| Date\_received | 领取优惠券日期 |
| Probability | 15天内用券概率，由参赛选手给出 |

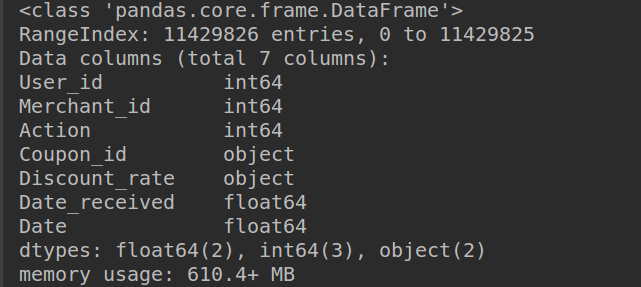
**数据探索：**

**首先对所给的数据进行预处理分析，对给予的几个基本数据集进行分析。**

1. **对于ccf\_online\_stage1\_train.csv文件中给出的用户线上消费用户券的情况，如下：**



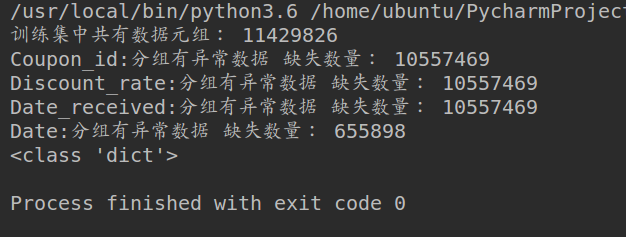
**从文件的title可知该数据集共有8个属性:User\_id, Merchant\_id, Action, Coupon\_id, Discount\_rate, Date\_received, Date。其中数据类型如下:**



发现on\_line训练集中共有11429826个样本，现对样本中的八个属性做缺省分析。通过groupby语句，对八个属性按属性值分类处理(类似决策树),因为groupby过滤了空值，所以可以间接得到每个属性的缺失数量。代码如下：

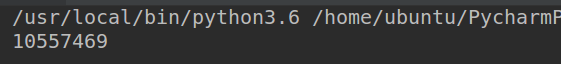


结果：

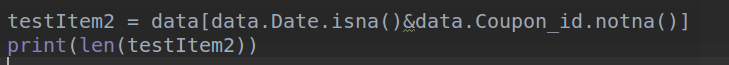


分组有异常数据即表示，该在table中该属性下有为null的属性值，缺失数量便是该属性中null属性值的数量。通过分析得知Coupon\_id,Distcount\_rate,Date\_received的null值相同，现猜想，三者是否同时为null。验证代码如下：





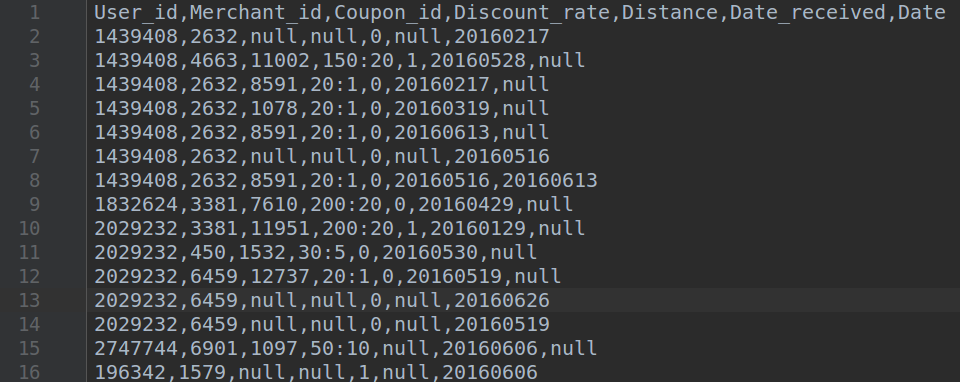
这与上面的缺失数量相吻合，说明三个属性是同时为null，严格来说是Coupon\_id为null时，则Discount\_rate和Date\_received也为null，这证明了Table栏中所言，“Coupon\_id为null表示无用户券消费时，Discount\_rate和Date\_received字段无意义”。除此外，发现Date项有65万之多数据为空。通过分析得知，Date为空时Coupon\_id必然不为空，这与Table栏中所言“如果Date=null & Coupon\_id != null，该记录表示领取优惠券但没有使用”。相吻合。



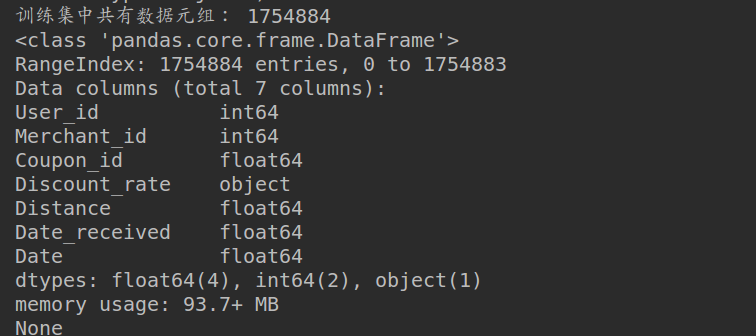


通过上面的分析，发现online数据集中的数据都是正常的，虽然也有不少项为null，但是其依然蕴含了某种特殊的信息，并没有真正意义上的异常数据，都是有用的数据，也就不需要了对online的数据预处理（例如对缺失项按照均值，众数进行填充，或则直接删掉）。

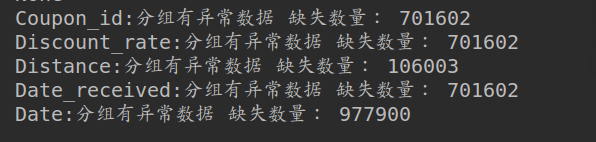
1. **对于ccf\_offline\_stage1\_train.csv文件中给出的用户线下消费用户券的情况，如下：**

户

**从文件的title可知该数据集共有7个属性性:User\_id,Merchant\_id,Coupon\_id,Discount\_rate,Distance,Date\_received,Date。其中数据类型如下:**

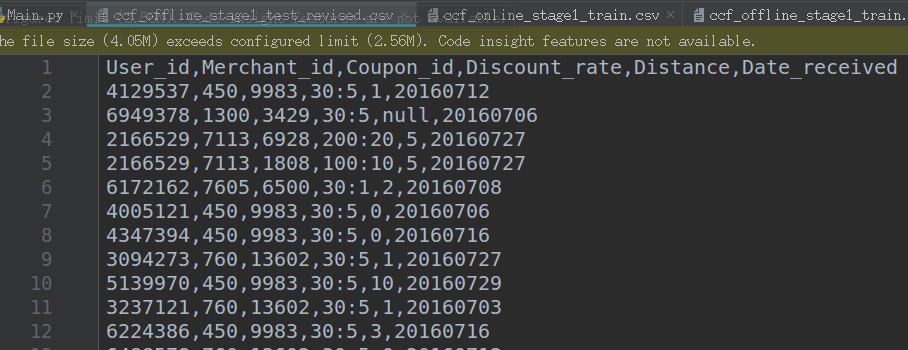


通过和上面处理online数据集同样的方式，可得



发现和online数据集一样，Coupon\_id为null时，则Discount\_rate和Date\_received也为null。即没有领取到优惠券时，折扣率和优惠券领取时间数据也就不存在，符合现实逻辑。同样offline数据集中也存在只要Date为空时Coupon\_id必然不为空的情况。这表明存在不少领取了优惠券但是没有使用的情况。结合online数据集，发现这种现象很是奇怪，即只要用户没有消费，那他一定领取了优惠券，从概率上看，样本数据有点不合逻辑。除此外，发现offline数据集中distance项有null，即缺少了对应的user经常活动的地点离该merchant的最近门店的距离，通过分析发现这些缺失的distance对应的用户领取的消费券都不为空，有点匪夷所思。难不成时门店的员工的内部消费？但是从Distance属性值上来分析x in [0,10]覆盖了所有大于等于0的距离，那么员工应该也在这个范围内啊，不是很能理解。

1. **对于ccf\_offline\_stage1\_train.csv文件中给出的用户线下消费用户券的情况，如下：**



ccf\_offline\_stage1\_test\_revised是要预测的用户在2016年7月领取优惠券的测试集。其中有6个属性

User\_id,Merchant\_id,Coupon\_id,Discount\_rate,Distance,Date\_received。显然这些与offline数据集的6集基本属性是一致的，我们的目标便是根据这6个属性来预测该用户在领取优惠券15天内使用该券的概率。分析得知除了少部分Distance存在null缺失量，其他数据正常。

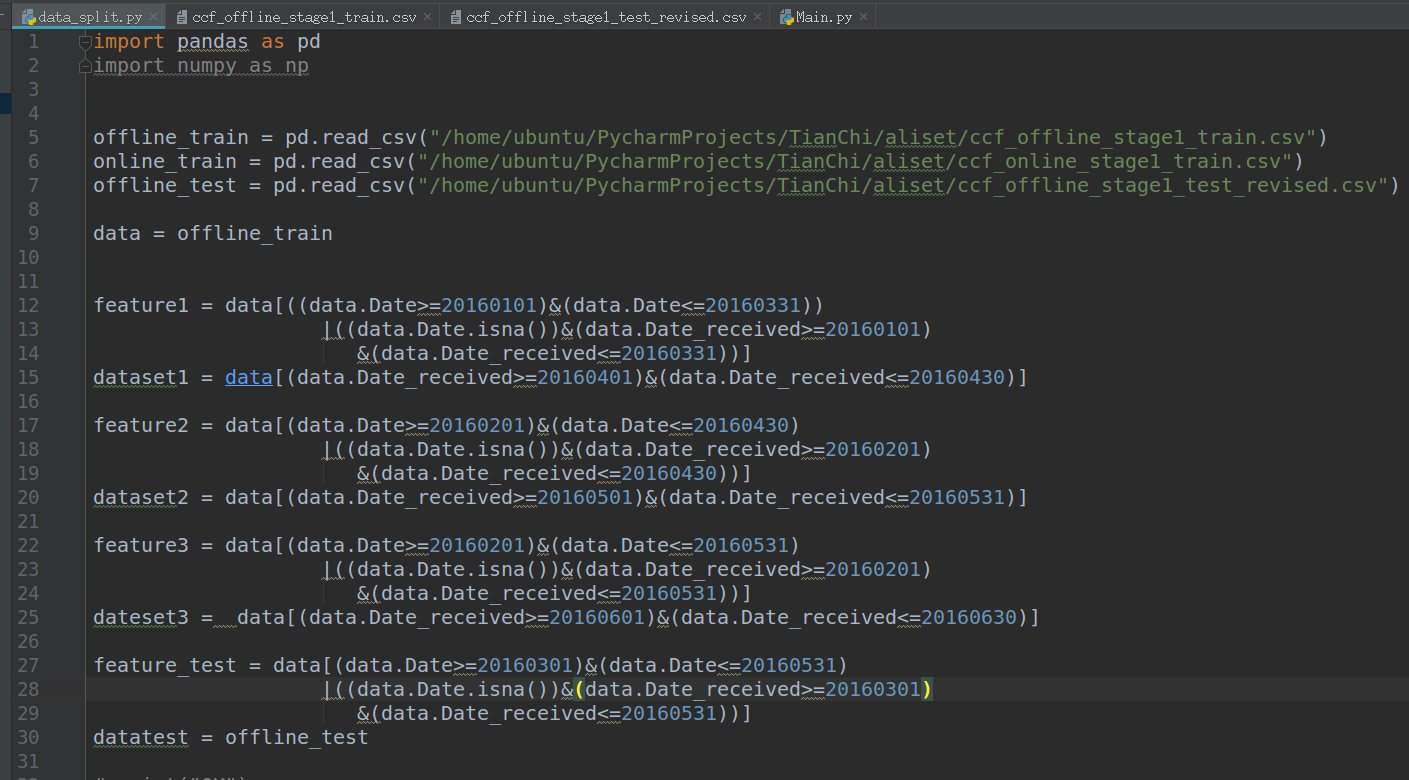
总结：以上通过上述简单的分析，发现只有在线下的数据集中Distance部分有缺失，而在实际给出的测试集中也有Distance为Null的情况，也就是说这个Distance属性无论缺失与否都对最终模型的建立有着重要的作用，至于其他属性诸如Coupon\_id之类的其为null蕴含这特殊的信息(没有领取到优惠券等等)。所以对于给定的数据集并不需要在缺失项上的填充或删除等处理，这些数据都是有意义的。以上主要针对了给定的数据集数据项是否缺失那一块进行了分析，实际在根据个属性groupby中也发现了不少有用的信息，比如说在Online数据集中关于Discount\_rate那一项，发现只有”x：y“型（满x减y）和fixed（低价限时优惠），并不存在折扣率。不少用户(User\_id)同时出现在online和offline数据集中(这一点在用户行为分析上十分重要)，而且通过进一步分析发现在test数据集中的绝大多数用户（User\_id）和绝大部分商家（Merchant\_id）都有出现在online和offline数据集中，优惠券虽然没有出现在online和offline中，但是其对应的折扣率（Discount\_rate）有出现过。综上，从主要的两份原始数据集(online和offline)中，我们可以从online数据集中提取到与用户相关的线上特征（网上浏览习惯(点击率，领取率等等)），而offline数据集则可以提取到更加丰富的特征：用户的线下特征（网上浏览习惯，用户领取优惠券并使用的概率，用户平均核销优惠券次数，用户最喜欢消费的工作日等等），商家的相关特征（商家优惠券被领取的次数，商家优惠券领取后核销的次数，商家优惠券被核销的平均时间，商家分发的最受欢迎的优惠券……），用户和商家的交互特征（商家的回头客比例，用户对商家的惠顾率，用户对折扣优惠券的使用比例等等），优惠券的相关特征（优惠券类型，优惠券的领取时间，领取次数最多的优惠券类型，核销率最高的优惠券类型……）。

数据集划分：

对于这种给定一段时间的数据集要预测未来n天的情况的问题，选择使用滑窗法来对原始训练集的划分，这样可以充分利用到原始数据，尤其是在测试集上，可以利用已知的Label进行验证模型。特征区间划分的越小，得到的训练数据集越多。我的划分方案如下：

|  |  |  |
| --- | --- | --- |
|  | 预测区间（提取Label） | 特征区间（提取Feature） |
| 训练集1 | 20160401~20160430 | 20160101~20160331 |
| 训练集2 | 20160501~20160531 | 20160201~20160430 |
| 验证集 | 20160601~20160630 | 20160201~20160531 |
| 测试集 | 20160701~20160731 | 20160301~20160531 |

代码如下：(详见data\_split.py文件)



特征工程

基于以上的数据探索，将特征分组为用户特征，商家特征，用户商家特征，优惠券特征。

一．用户特征（描述每个用户的消费喜好，可进一步分为线上，线下，线上-线下特征）

* 线下特征
* 用户领取优惠券次数
* 用户线下核销率
* 用户核销过的商家数量
* 用户使用所有优惠券的平均时间
* 用户核销优惠券的平均折扣
* 用户核销率最高的优惠券折扣率
* 用户核销的时间是否为节假日
* 用户核销的平均距离
* 等等户
* 线上特征
* 用户的记录条数
* 用户点击率
* 用户领取率
* 用户购买率
* 等等
* 线上-线下特征
* 用户线下记录的占比数户
* 用户用户线下核销次数的占比数
* 用户对同一折扣率优惠券的线下使用占比数
* 等等

1. 商家特征（描述商家的受欢迎程度机器商品的被消费规律）

* 商家优惠券被核销率
* 商家优惠券被领取的次数
* 商家所被消费的优惠券的平均消费折扣率
* 商家优惠券被核销的平均时间
* 商家所分发的优惠券的种类数量
* 等等

1. 优惠券特征（描述优惠券自身的特征及其类别的历史消费规律）

* 优惠券折扣率
* 优惠券核销率
* 优惠券类型
* 优惠券被核销的时间
* 等等

1. 用户-商家特征（描述用户对某些商家的消费偏好）

* 用户领取该商家的优惠券次数
* 用户在该商家领取的优惠券的占比
* 用户在该商家的核销量的总占比
* 用户在该商家的核销率
* 等等

部分代码如下：（详见feature\_extract.py文件)

