商户流失率预测-实验报告

梅佳奕 10165300206

项目总目标：

预测抛出门店中，哪些是未来28天内不再有交易（流失）的门店。

阶段一：选取特征值

实验方法：

由于原数据量很大，且属性较多，那么需要担心的是某些属性值较为稀疏、没有实质性有用的信息。为了检测属性值的分布是否稀疏，定义一个量“集中程度”

$$K_{col} = \frac{max\{c_i : \Sigma_i c_i = C\}}{C}$$

在一个共C行的表格中，对每一个attribute（col），count一共出现多少个不同的值，并统计每个值出现的次数，将统计得的结果按降序排序，以作图或表格的形式可视化。

如果该K值很大，超过0.9，或图像呈明显的长尾分布，则可以认为该属性较为稀疏，提供的信息价值不大。

实验过程：

- table throw-point:

  由于目标是从已“抛出”的门店中预测未来是否流失，table throw-point中的信息（store_id, pt）会作为筛选店铺的条件而存在。在后续预测的时候，预测对象就依据table throw-point从所有门店中筛出，从而进行实际预测；不被抛出的门店对项目目标没有贡献。

  所以throw-point是必须信息，但不作为特征值加入训练。

- table merchant:
  - merchant_id:

    是标称数据，不能作为值参与训练，但不同商户有该商户的特征，而门店属于什么商户，经营情况会受到商户影响,所以merchant_id作为将merchant信息与store经营情况join连接的字段而存在。

  - industry_level：

    ```
    industry_level1:

    1    3029
    2    1987
    3     583
    ```

    非稀疏分布，保留。

  - province：

```
province:

1    3742
2    1066
3     791
```

非稀疏分布，保留。

- table transaction：

  首先，通过count数据可以发现：hongbao_channel_mch, hongbao_channel_mch_top_up, card_pre, card_balance, merchant_sn, is_liquidation_next_day, discount_wosai_mch, discount_wosai, hongbao_wosai_mch, hongbao_wosai, wallet_alipay_finance这些属性下其实只有同一个value，他们对不同门店无差别分布，所以直接排除。

  在剩下来的数据中，又通过计算K值的方法发现，K值在0.9以上的属性有： hongbao_channel, discount_channel_mch_top_up, discount_channel_mch, discount_channel, alipay_point, bankcard_credit, type, status；这些属性几乎可以排除，同时考虑到实际情况中一些事实（如退款）发生频率低属正常现象，可以结合属性、数值的实际意义将有效信息转换为tag，一定不按原形式保留。

  还有一些数据，K值在0.8以上，或K值在0.6以上但有较明显的长尾分布，如：alipay_huabei, bankcard_debit, wallet_weixin, sub_pay_way, pay_way；对于这类属性，考虑其现实意义，分组捏合。

  观察到存在某几个属性意义相似、分布相似的情况，那就根据其缺失值情况决定去留，以免造成信息冗余的问题。如：在effective_amount，original_amount，received_amount，paid_amount这一组属性中，由于后两者约有6%的缺失值，且前两者分布更好，最终取前两者。

  另外，在观察到的标称信息中，也并非所有的都有用处，如我们预测商户流失率，那么用户payer_id，terminal_id就可以不必考虑。

实验结果：

标称信息：

- store_id
- merchant_id

特征值：

- original_amount：

  是最主要的特征值，代表了商户的世纪收入，在所有与金额相关的特征中，该特征与门店的关系最为紧密。

- effective_amount:

  向支付通道请求的金额，即用户实际发起的付款。该特征与original_amount有较大重合，但仍存在差别，意指在交易中用户是否得到实质性的优惠。可以认为存在实际优惠，用户将更愿意于该门店进行交易。

- pay_way & sub_pay_way：

  一级、二级主支付方式，K值分别为0.66与0.63，保留。

```
pay_way:          sub_pay_way:

3     1515737     3    1452851
2      789377     1     855294
6        4027     2       1206
17        523     4        314
18          1
```

- ctime & pt：

  ctime是每次交易的具体时间，可通过ctime探索门店交易于一天、一个星期之内集中在什么时间段；pt是粗粒度的日期分区，可探索门店的旺季与淡季。

- 支付方式：
  - bankcard_debit (0.8<K<0.9)
  - wallet_weixin (K=0.60)
  - wallet_alipay (0.8<K<0.9)
  - alipay_huabei (0.8<K<0.9)

    由于以上几个属性意义相关，都属于支付方式，但其分布较稀疏，所以将他们合并、离散化为一个四维向量的tag：该门店提供何种支付方式？四个维度分别对应“储蓄卡”、“微信余额”、“支付宝余额”、“花呗”，如果提供则为1，不提供为0。

- type:

  交易类型，此处我们关心的是“退款”的交易，如果一个门店退款情况时有发生，可推测该门店可能存在售后、质量的问题。

- 折扣：
  - discount_channel
  - discount_channel_mch
  - discount_channel_mch_top_up
  - hongbao_channel

    这四个特征的K值都在0.9以上，分布极其稀疏，所以将它们弱化为一个bool类型的tag：是否提供折扣优惠的活动。注意，此处与上文effective_amount中得到的交易中用户获得实际优惠不同，由于此特征非常弱，所以仅以此特征表示商家是否提供折扣优惠的服务，而不计量折扣优惠的力度、金额。
