# 说明

* 文档下半部分使用report.md查看更为清晰
* 题目说明见introduce.md/introduce.docx
* 文件说明见file\_introduce.md
* 行为数据的可视化可查看webapps下的demo.html

# 步骤详解

### 数据统计

------------用户表数据------------  
+-------+--------+---+----------+-----------+  
|user\_id| age|sex|user\_lv\_cd|user\_reg\_tm|  
+-------+--------+---+----------+-----------+  
| 200001|56������| 2| 5| 2016-01-26|  
| 200002| -1| 0| 1| 2016-01-26|  
| 200003| 36-45��| 1| 4| 2016-01-26|  
| 200004| -1| 2| 1| 2016-01-26|  
| 200005| 16-25��| 0| 4| 2016-01-26|  
| 200006| 36-45��| 2| 2| 2013-04-10|  
| 200007| 36-45��| 2| 3| 2016-01-26|  
| 200008| -1| 2| 3| 2016-01-26|  
| 200009| 36-45��| 2| 2| 2016-01-26|  
| 200010| 36-45��| 2| 3| 2016-01-26|  
| 200011| -1| 0| 5| 2013-04-10|  
| 200012| 36-45��| 2| 3| 2013-04-10|  
| 200013| 26-35��| 2| 4| 2013-04-10|  
| 200014| 26-35��| 2| 4| 2013-04-10|  
| 200015| 26-35��| 1| 3| 2016-01-26|  
| 200016| 26-35��| 2| 3| 2016-01-26|  
| 200017| 36-45��| 2| 4| 2016-01-26|  
| 200018| 16-25��| 2| 2| 2016-01-26|  
| 200019| 16-25��| 2| 2| 2016-01-26|  
| 200020| -1| 2| 2| 2016-01-26|  
+-------+--------+---+----------+-----------+  
only showing top 20 rows  
  
用户表年龄统计  
+--------+--------+  
| age|count(1)|  
+--------+--------+  
| 36-45��| 30336|  
| -1| 14412|  
| 46-55��| 3325|  
| 26-35��| 46570|  
|56������| 1871|  
|15������| 7|  
| 16-25��| 8797|  
| NULL| 3|  
+--------+--------+  
分析：用户年龄存在空值和乱码问题  
  
用户表性别统计  
+----+--------+  
| sex|count(1)|  
+----+--------+  
| 0| 42846|  
| 1| 7737|  
| 2| 54735|  
|NULL| 3|  
+----+--------+  
分析：用户性别存在空值问题  
  
用户表年龄不为空值  
+-------+----+----+----------+-----------+  
|user\_id| age| sex|user\_lv\_cd|user\_reg\_tm|  
+-------+----+----+----------+-----------+  
| 234073|NULL|NULL| 1| NULL|  
| 238906|NULL|NULL| 1| NULL|  
| 267705|NULL|NULL| 1| NULL|  
+-------+----+----+----------+-----------+  
  
用户表用户等级统计  
+----------+--------+  
|user\_lv\_cd|count(1)|  
+----------+--------+  
| 3| 24563|  
| 5| 36088|  
| 1| 2666|  
| 4| 32343|  
| 2| 9661|  
+----------+--------+  
  
  
------------产品表数据------------  
数据  
+------+---+---+---+----+-----+  
|sku\_id| a1| a2| a3|cate|brand|  
+------+---+---+---+----+-----+  
| 10| 3| 1| 1| 8| 489|  
|100002| 3| 2| 2| 8| 489|  
|100003| 1| -1| -1| 8| 30|  
|100006| 1| 2| 1| 8| 545|  
| 10001| -1| 1| 2| 8| 244|  
|100016| 3| -1| 2| 8| 214|  
|100029| 3| 2| 2| 8| 214|  
| 10003| 3| 1| 2| 8| 214|  
|100045| 2| 2| 2| 8| 124|  
|100057| 3| 1| 2| 8| 306|  
|100069| 1| 2| -1| 8| 403|  
|100076| 3| 2| 1| 8| 214|  
|100097| 1| 1| 2| 8| 545|  
| 10011| 3| 1| 1| 8| 214|  
|100112| 3| 1| 2| 8| 489|  
|100117| -1| 1| 2| 8| 174|  
|100119| 2| 1| 2| 8| 812|  
|100129| 3| 1| 2| 8| 214|  
|100134| 3| 1| 2| 8| 214|  
|100137| 1| -1| 1| 8| 403|  
+------+---+---+---+----+-----+  
only showing top 20 rows  
  
a1统计  
+---+--------+  
| a1|count(1)|  
+---+--------+  
| -1| 1701|  
| 3| 14144|  
| 1| 4760|  
| 2| 3582|  
+---+--------+  
  
a2统计  
+---+--------+  
| a2|count(1)|  
+---+--------+  
| -1| 4050|  
| 1| 13513|  
| 2| 6624|  
+---+--------+  
  
a3统计  
+---+--------+  
| a3|count(1)|  
+---+--------+  
| -1| 3815|  
| 1| 8394|  
| 2| 11978|  
+---+--------+  
  
cate品类统计  
+----+--------+  
|cate|count(1)|  
+----+--------+  
| 8| 24187|  
+----+--------+  
  
brand品牌统计  
+-----+--------+  
|brand|count(1)|  
+-----+--------+  
| 800| 1015|  
| 124| 932|  
| 51| 7|  
| 574| 4|  
| 383| 3|  
| 635| 199|  
| 200| 72|  
| 101| 42|  
| 541| 10|  
| 3| 19|  
| 30| 659|  
| 479| 1|  
| 556| 34|  
| 594| 10|  
| 554| 1|  
| 214| 6444|  
| 484| 29|  
| 804| 47|  
| 70| 40|  
| 174| 50|  
+-----+--------+  
only showing top 20 rows  
  
  
------------评论表数据------------  
数据  
+----------+------+-----------+---------------+----------------+  
| dt|sku\_id|comment\_num|has\_bad\_comment|bad\_comment\_rate|  
+----------+------+-----------+---------------+----------------+  
|2016-02-01| 1000| 3| 1| 0.0417|  
|2016-02-01| 10000| 2| 0| 0.0|  
|2016-02-01|100011| 4| 1| 0.0376|  
|2016-02-01|100018| 3| 0| 0.0|  
|2016-02-01|100020| 3| 0| 0.0|  
|2016-02-01|100021| 1| 0| 0.0|  
|2016-02-01|100028| 1| 0| 0.0|  
|2016-02-01|100031| 3| 0| 0.0|  
|2016-02-01|100033| 2| 0| 0.0|  
|2016-02-01|100035| 1| 1| 1.0|  
|2016-02-01|100037| 1| 0| 0.0|  
|2016-02-01|100039| 4| 0| 0.0|  
|2016-02-01|100046| 0| 0| 0.0|  
|2016-02-01|100047| 1| 0| 0.0|  
|2016-02-01| 10005| 1| 1| 1.0|  
|2016-02-01|100051| 1| 0| 0.0|  
|2016-02-01|100056| 2| 0| 0.0|  
|2016-02-01|100068| 0| 0| 0.0|  
|2016-02-01|100076| 3| 1| 0.0227|  
|2016-02-01|100086| 4| 1| 0.0896|  
+----------+------+-----------+---------------+----------------+  
only showing top 20 rows  
  
日期统计  
+----------+--------+  
| dt|count(1)|  
+----------+--------+  
|2016-04-15| 46546|  
|2016-04-11| 46546|  
|2016-03-14| 46546|  
|2016-02-01| 46546|  
|2016-02-15| 46546|  
|2016-03-21| 46546|  
|2016-02-22| 46546|  
|2016-02-29| 46546|  
|2016-04-04| 46546|  
|2016-02-08| 46546|  
|2016-03-07| 46546|  
|2016-03-28| 46546|  
+----------+--------+  
分析：评论时间都发生在这12天  
  
数量统计  
+--------+  
|count(1)|  
+--------+  
| 558552|  
+--------+  
  
有效评论数量  
+--------+  
|count(1)|  
+--------+  
| 558552|  
+--------+  
  
商品评论数量统计  
+------+--------+  
|sku\_id|count(1)|  
+------+--------+  
|100140| 12|  
|100553| 12|  
|100704| 12|  
|100964| 12|  
|102113| 12|  
|102952| 12|  
|103432| 12|  
| 10351| 12|  
|104870| 12|  
|106090| 12|  
|106510| 12|  
|107930| 12|  
|108231| 12|  
|108800| 12|  
| 1090| 12|  
|109818| 12|  
|110071| 12|  
|112261| 12|  
|113417| 12|  
|115770| 12|  
+------+--------+  
only showing top 20 rows  
分析：每件商品都有12个评价  
  
评论数统计  
+-----------+--------+  
|comment\_num|count(1)|  
+-----------+--------+  
| 3| 119642|  
| 0| 19993|  
| 1| 85430|  
| 4| 164789|  
| 2| 168698|  
+-----------+--------+  
分析：对商品的评论数量只有上述的几个数值  
  
是否有差评统计  
+---------------+--------+  
|has\_bad\_comment|count(1)|  
+---------------+--------+  
| 0| 292978|  
| 1| 265574|  
+---------------+--------+  
  
差评率统计  
+----------------+-----+  
|bad\_comment\_rate|count|  
+----------------+-----+  
| 0.124| 1|  
| 0.1687| 1|  
| 0.2525| 1|  
| 0.1097| 1|  
| 0.168| 1|  
| 0.2676| 1|  
| 0.1674| 1|  
| 0.1201| 1|  
| 0.1082| 1|  
| 0.1118| 1|  
| 0.1285| 1|  
| 0.1584| 1|  
| 0.32| 1|  
| 0.11| 1|  
| 0.1021| 1|  
| 0.147| 1|  
| 0.1295| 1|  
| 0.1513| 1|  
| 0.2239| 1|  
| 0.2688| 1|  
+----------------+-----+  
only showing top 20 rows  
  
  
  
------------用户行为表数据------------  
数据  
+--------+------+-------------------+--------+----+----+-----+  
| user\_id|sku\_id| time|model\_id|type|cate|brand|  
+--------+------+-------------------+--------+----+----+-----+  
|266079.0|138778|2016-01-31 23:59:02| | 1| 8| 403|  
|266079.0|138778|2016-01-31 23:59:03| 0| 6| 8| 403|  
|200719.0| 61226|2016-01-31 23:59:07| | 1| 8| 30|  
|200719.0| 61226|2016-01-31 23:59:08| 0| 6| 8| 30|  
|263587.0| 72348|2016-01-31 23:59:08| | 1| 5| 159|  
|296130.0|103126|2016-01-31 23:59:11| | 1| 4| 174|  
|296130.0|103126|2016-01-31 23:59:11| 0| 6| 4| 174|  
|296130.0|103126|2016-01-31 23:59:11| 0| 6| 4| 174|  
|217892.0|137328|2016-01-31 23:59:14| 111| 6| 6| 159|  
|283139.0| 4732|2016-01-31 23:59:14| | 1| 6| 375|  
|217892.0|137328|2016-01-31 23:59:16| 111| 6| 6| 159|  
|243907.0| 27659|2016-01-31 23:59:22| 0| 6| 4| 752|  
|243907.0| 27659|2016-01-31 23:59:24| 17| 6| 4| 752|  
|209278.0| 6998|2016-01-31 23:59:25| 0| 6| 8| 545|  
|209278.0| 6998|2016-01-31 23:59:25| | 1| 8| 545|  
|266079.0|138778|2016-01-31 23:59:40| 0| 6| 8| 403|  
|263587.0|127189|2016-01-31 23:59:48| | 1| 5| 159|  
|200719.0| 61226|2016-01-31 23:59:50| 11| 6| 8| 30|  
|209278.0| 24679|2016-01-31 23:59:52| | 1| 8| 658|  
|209278.0| 24679|2016-01-31 23:59:53| 0| 6| 8| 658|  
+--------+------+-------------------+--------+----+----+-----+  
only showing top 20 rows  
  
数据总数  
+--------+  
|count(1)|  
+--------+  
|50601736|  
+--------+  
  
用户存在的行为总数  
+--------+  
|count(1)|  
+--------+  
|50601736|  
+--------+  
  
产品存在的行为总数  
+--------+  
|count(1)|  
+--------+  
|18128055|  
+--------+  
分析：只有18128055条可用数据  
  
model\_id统计  
+--------+--------+  
|model\_id|count(1)|  
+--------+--------+  
| |20655896|  
| 0|12112574|  
| 216| 4842945|  
| 217| 4335430|  
| 27| 1468903|  
| 26| 1388148|  
| 218| 1203986|  
| 211| 651555|  
| 24| 568860|  
| 29| 486169|  
| 21| 349610|  
| 111| 228742|  
| 17| 220852|  
| 210| 201725|  
| 219| 182587|  
| 222| 125944|  
| 23| 123880|  
| 19| 116766|  
| 220| 115865|  
| 31| 112592|  
+--------+--------+  
only showing top 20 rows  
分析：近半数的用户行为没有model\_id  
  
type统计  
+----+--------+  
|type|count(1)|  
+----+--------+  
| 3| 256053|  
| 5| 109896|  
| 6|30630744|  
| 1|18981373|  
| 4| 48252|  
| 2| 575418|  
+----+--------+  
  
  
  
所有用户行为  
+---+--------+  
|day|count(1)|  
+---+--------+  
| 31| 15|  
| 32| 93746|  
| 33| 84508|  
| 34| 81019|  
| 35| 73341|  
| 36| 58926|  
| 37| 52289|  
| 38| 42265|  
| 39| 52225|  
| 40| 57202|  
| 41| 60712|  
| 42| 70187|  
| 43| 83257|  
| 44| 91591|  
| 45| 107038|  
| 46| 127085|  
| 47| 125318|  
| 48| 134550|  
| 49| 127415|  
| 50| 113448|  
+---+--------+  
only showing top 20 rows  
  
浏览  
+---+--------+  
|day|count(1)|  
+---+--------+  
| 31| 6|  
| 32| 35420|  
| 33| 32035|  
| 34| 31063|  
| 35| 28301|  
| 36| 22585|  
| 37| 20402|  
| 38| 16123|  
| 39| 20214|  
| 40| 21961|  
| 41| 23229|  
| 42| 26937|  
| 43| 31898|  
| 44| 34773|  
| 45| 40963|  
| 46| 48031|  
| 47| 51905|  
| 48| 55515|  
| 49| 47797|  
| 50| 42962|  
+---+--------+  
only showing top 20 rows  
  
加购  
+---+--------+  
|day|count(1)|  
+---+--------+  
| 32| 1388|  
| 33| 1152|  
| 34| 1027|  
| 35| 851|  
| 36| 721|  
| 37| 641|  
| 38| 560|  
| 39| 642|  
| 40| 705|  
| 41| 677|  
| 42| 939|  
| 43| 1081|  
| 44| 1213|  
| 45| 1507|  
| 46| 1660|  
| 47| 1782|  
| 48| 1858|  
| 49| 1936|  
| 50| 1530|  
| 51| 189|  
+---+--------+  
only showing top 20 rows  
分析：存在有些天没有加购行为  
  
删除  
+---+--------+  
|day|count(1)|  
+---+--------+  
| 32| 490|  
| 33| 468|  
| 34| 333|  
| 35| 331|  
| 36| 280|  
| 37| 218|  
| 38| 175|  
| 39| 203|  
| 40| 203|  
| 41| 235|  
| 42| 352|  
| 43| 351|  
| 44| 443|  
| 45| 505|  
| 46| 600|  
| 47| 589|  
| 48| 625|  
| 49| 642|  
| 50| 592|  
| 51| 78|  
+---+--------+  
only showing top 20 rows  
  
下单  
+---+--------+  
|day|count(1)|  
+---+--------+  
| 32| 90|  
| 33| 84|  
| 34| 78|  
| 35| 47|  
| 36| 26|  
| 37| 17|  
| 38| 11|  
| 39| 25|  
| 40| 16|  
| 41| 18|  
| 42| 47|  
| 43| 41|  
| 44| 46|  
| 45| 65|  
| 46| 58|  
| 47| 103|  
| 48| 108|  
| 49| 92|  
| 50| 85|  
| 51| 79|  
+---+--------+  
only showing top 20 rows  
分析：下单数量极少，说明订单转化率低  
  
关注  
+---+--------+  
|day|count(1)|  
+---+--------+  
| 32| 209|  
| 33| 248|  
| 34| 169|  
| 35| 218|  
| 36| 151|  
| 37| 146|  
| 38| 183|  
| 39| 172|  
| 40| 162|  
| 41| 155|  
| 42| 208|  
| 43| 201|  
| 44| 262|  
| 45| 286|  
| 46| 402|  
| 47| 332|  
| 48| 357|  
| 49| 322|  
| 50| 295|  
| 51| 426|  
+---+--------+  
only showing top 20 rows  
  
点击  
+---+--------+  
|day|count(1)|  
+---+--------+  
| 31| 9|  
| 32| 56149|  
| 33| 50521|  
| 34| 48349|  
| 35| 43593|  
| 36| 35163|  
| 37| 30865|  
| 38| 25213|  
| 39| 30969|  
| 40| 34155|  
| 41| 36398|  
| 42| 41704|  
| 43| 49685|  
| 44| 54854|  
| 45| 63712|  
| 46| 76334|  
| 47| 70607|  
| 48| 76087|  
| 49| 76626|  
| 50| 67984|  
+---+--------+  
only showing top 20 rows  
  
分析：每一天的具体行为的统计，具体请看demo.html的图形展示  
  
每一天的用户行为数量统计(31为1月31日，往后累加1)  
+---+------------+------------+------------+------------+------------+------------+  
|day|sum(utype\_1)|sum(utype\_2)|sum(utype\_3)|sum(utype\_4)|sum(utype\_5)|sum(utype\_6)|  
+---+------------+------------+------------+------------+------------+------------+  
| 31| 6| 0| 0| 0| 0| 9|  
| 32| 35420| 1388| 490| 90| 209| 56149|  
| 33| 32035| 1152| 468| 84| 248| 50521|  
| 34| 31063| 1027| 333| 78| 169| 48349|  
| 35| 28301| 851| 331| 47| 218| 43593|  
| 36| 22585| 721| 280| 26| 151| 35163|  
| 37| 20402| 641| 218| 17| 146| 30865|  
| 38| 16123| 560| 175| 11| 183| 25213|  
| 39| 20214| 642| 203| 25| 172| 30969|  
| 40| 21961| 705| 203| 16| 162| 34155|  
| 41| 23229| 677| 235| 18| 155| 36398|  
| 42| 26937| 939| 352| 47| 208| 41704|  
| 43| 31898| 1081| 351| 41| 201| 49685|  
| 44| 34773| 1213| 443| 46| 262| 54854|  
| 45| 40963| 1507| 505| 65| 286| 63712|  
| 46| 48031| 1660| 600| 58| 402| 76334|  
| 47| 51905| 1782| 589| 103| 332| 70607|  
| 48| 55515| 1858| 625| 108| 357| 76087|  
| 49| 47797| 1936| 642| 92| 322| 76626|  
| 50| 42962| 1530| 592| 85| 295| 67984|  
| 51| 53364| 189| 78| 79| 426| 82228|  
| 52| 56502| 1785| 655| 82| 374| 87196|  
| 53| 59569| 2323| 717| 146| 430| 95601|  
| 54| 46914| 1867| 683| 79| 354| 73479|  
| 55| 51567| 1943| 711| 98| 385| 80995|  
| 56| 62932| 2194| 671| 113| 445| 99291|  
| 57| 60963| 2042| 669| 134| 428| 96386|  
| 58| 66850| 554| 157| 106| 482| 101163|  
| 59| 69143| 2005| 673| 108| 505| 107382|  
| 60| 69191| 2288| 806| 120| 499| 109167|  
| 61| 100030| 4236| 1217| 284| 669| 162234|  
| 62| 97132| 3670| 1204| 186| 680| 150099|  
| 63| 108937| 3820| 1327| 199| 714| 168482|  
| 64| 106663| 3396| 1096| 175| 684| 168466|  
| 65| 126211| 1109| 333| 195| 775| 195420|  
| 66| 133272| 0| 0| 174| 860| 203095|  
| 67| 139028| 5598| 1840| 431| 806| 223294|  
| 68| 108066| 3837| 1346| 165| 723| 173434|  
| 69| 103682| 3831| 1298| 140| 683| 164091|  
| 70| 135537| 5365| 1652| 239| 789| 215958|  
| 71| 129322| 2269| 742| 223| 737| 206582|  
| 72| 146609| 0| 0| 268| 826| 222634|  
| 73| 150507| 0| 0| 150| 872| 229886|  
| 74| 197769| 6906| 2725| 325| 1192| 323315|  
| 75| 332352| 13936| 4835| 1297| 1423| 511895|  
| 76| 136023| 4865| 2273| 264| 764| 205174|  
| 77| 102595| 3135| 1829| 156| 625| 152845|  
| 78| 93084| 3009| 1346| 164| 630| 151107|  
| 79| 108946| 3020| 1551| 144| 610| 181132|  
| 80| 104044| 2761| 1327| 135| 642| 167688|  
| 81| 94540| 3171| 1589| 168| 637| 154499|  
| 82| 88334| 2890| 1387| 147| 599| 145628|  
| 83| 77735| 2118| 1092| 115| 549| 126816|  
| 84| 81635| 2869| 1340| 145| 493| 134266|  
| 85| 88554| 2917| 1300| 211| 510| 149030|  
| 86| 89217| 2695| 1176| 116| 581| 145831|  
| 87| 104352| 2706| 1235| 138| 636| 174613|  
| 88| 98074| 3231| 1329| 138| 598| 165312|  
| 89| 93469| 4075| 1581| 297| 618| 157511|  
| 90| 100305| 3738| 1494| 158| 654| 165992|  
| 91| 90258| 3274| 1313| 171| 571| 150755|  
| 92| 94411| 3104| 1321| 208| 553| 159236|  
| 93| 121696| 3670| 1523| 199| 721| 204638|  
| 94| 111963| 3424| 1386| 174| 609| 190451|  
| 95| 112156| 3280| 1396| 175| 782| 191220|  
| 96| 100138| 3473| 1543| 181| 643| 170494|  
| 97| 105346| 3509| 1530| 199| 658| 179234|  
| 98| 149258| 6397| 2211| 446| 863| 254245|  
| 99| 175708| 7095| 2566| 493| 982| 294470|  
|100| 161347| 5595| 2242| 331| 835| 271234|  
|101| 138242| 4374| 1969| 249| 754| 228440|  
|102| 123194| 4328| 1926| 216| 694| 208022|  
|103| 114159| 3976| 1811| 213| 661| 195789|  
|104| 105914| 3621| 1687| 203| 621| 181329|  
|105| 133934| 5169| 2093| 379| 776| 247565|  
|106| 145111| 4964| 2379| 375| 850| 262225|  
+---+------------+------------+------------+------------+------------+------------+

### 数据清洗

------------用户表数据------------  
  
清洗后的用户表  
1. 填补确实值  
2. 去除了无用的字段和无效的用户  
3. 将年龄和性别进行one hot编码  
+------+---+-----+-----+-----+------+-----+-----+-----+-----+-----+-----+----------+-----------+  
|userId|age|sex\_0|sex\_1|sex\_2|age\_d1|age\_1|age\_2|age\_3|age\_4|age\_5|age\_6|user\_lv\_cd|user\_reg\_tm|  
+------+---+-----+-----+-----+------+-----+-----+-----+-----+-----+-----+----------+-----------+  
|200001| 6| 0| 0| 1| 0| 0| 0| 0| 0| 0| 1| 5| 2016-01-26|  
|200002| -1| 1| 0| 0| 1| 0| 0| 0| 0| 0| 0| 1| 2016-01-26|  
|200003| 4| 0| 1| 0| 0| 0| 0| 0| 1| 0| 0| 4| 2016-01-26|  
|200004| -1| 0| 0| 1| 1| 0| 0| 0| 0| 0| 0| 1| 2016-01-26|  
|200005| 2| 1| 0| 0| 0| 0| 1| 0| 0| 0| 0| 4| 2016-01-26|  
|200006| 4| 0| 0| 1| 0| 0| 0| 0| 1| 0| 0| 2| 2013-04-10|  
|200007| 4| 0| 0| 1| 0| 0| 0| 0| 1| 0| 0| 3| 2016-01-26|  
|200008| -1| 0| 0| 1| 1| 0| 0| 0| 0| 0| 0| 3| 2016-01-26|  
|200009| 4| 0| 0| 1| 0| 0| 0| 0| 1| 0| 0| 2| 2016-01-26|  
|200010| 4| 0| 0| 1| 0| 0| 0| 0| 1| 0| 0| 3| 2016-01-26|  
|200011| -1| 1| 0| 0| 1| 0| 0| 0| 0| 0| 0| 5| 2013-04-10|  
|200012| 4| 0| 0| 1| 0| 0| 0| 0| 1| 0| 0| 3| 2013-04-10|  
|200013| 3| 0| 0| 1| 0| 0| 0| 1| 0| 0| 0| 4| 2013-04-10|  
|200014| 3| 0| 0| 1| 0| 0| 0| 1| 0| 0| 0| 4| 2013-04-10|  
|200015| 3| 0| 1| 0| 0| 0| 0| 1| 0| 0| 0| 3| 2016-01-26|  
|200016| 3| 0| 0| 1| 0| 0| 0| 1| 0| 0| 0| 3| 2016-01-26|  
|200017| 4| 0| 0| 1| 0| 0| 0| 0| 1| 0| 0| 4| 2016-01-26|  
|200018| 2| 0| 0| 1| 0| 0| 1| 0| 0| 0| 0| 2| 2016-01-26|  
|200019| 2| 0| 0| 1| 0| 0| 1| 0| 0| 0| 0| 2| 2016-01-26|  
|200020| -1| 0| 0| 1| 1| 0| 0| 0| 0| 0| 0| 2| 2016-01-26|  
+------+---+-----+-----+-----+------+-----+-----+-----+-----+-----+-----+----------+-----------+  
only showing top 20 rows  
  
------------产品表数据------------  
  
  
清洗后的产品表  
1. 填补缺失值  
  
+---------+---+---+---+-----+  
|productId| a1| a2| a3|brand|  
+---------+---+---+---+-----+  
| 10| 3| 1| 1| 489|  
| 100002| 3| 2| 2| 489|  
| 100003| 1| -1| -1| 30|  
| 100006| 1| 2| 1| 545|  
| 10001| -1| 1| 2| 244|  
| 100016| 3| -1| 2| 214|  
| 100029| 3| 2| 2| 214|  
| 10003| 3| 1| 2| 214|  
| 100045| 2| 2| 2| 124|  
| 100057| 3| 1| 2| 306|  
| 100069| 1| 2| -1| 403|  
| 100076| 3| 2| 1| 214|  
| 100097| 1| 1| 2| 545|  
| 10011| 3| 1| 1| 214|  
| 100112| 3| 1| 2| 489|  
| 100117| -1| 1| 2| 174|  
| 100119| 2| 1| 2| 812|  
| 100129| 3| 1| 2| 214|  
| 100134| 3| 1| 2| 214|  
| 100137| 1| -1| 1| 403|  
+---------+---+---+---+-----+  
only showing top 20 rows  
  
------------评论表数据------------  
  
  
清洗后的评论表  
1. 填补缺失值  
  
+----------+---------+-----------+---------------+----------------+  
| dt|productId|comment\_num|has\_bad\_comment|bad\_comment\_rate|  
+----------+---------+-----------+---------------+----------------+  
|2016-02-01| 1000| 3| 1| 0.0417|  
|2016-02-01| 10000| 2| 0| 0.0|  
|2016-02-01| 100011| 4| 1| 0.0376|  
|2016-02-01| 100018| 3| 0| 0.0|  
|2016-02-01| 100020| 3| 0| 0.0|  
|2016-02-01| 100021| 1| 0| 0.0|  
|2016-02-01| 100028| 1| 0| 0.0|  
|2016-02-01| 100031| 3| 0| 0.0|  
|2016-02-01| 100033| 2| 0| 0.0|  
|2016-02-01| 100035| 1| 1| 1.0|  
|2016-02-01| 100037| 1| 0| 0.0|  
|2016-02-01| 100039| 4| 0| 0.0|  
|2016-02-01| 100046| 0| 0| 0.0|  
|2016-02-01| 100047| 1| 0| 0.0|  
|2016-02-01| 10005| 1| 1| 1.0|  
|2016-02-01| 100051| 1| 0| 0.0|  
|2016-02-01| 100056| 2| 0| 0.0|  
|2016-02-01| 100068| 0| 0| 0.0|  
|2016-02-01| 100076| 3| 1| 0.0227|  
|2016-02-01| 100086| 4| 1| 0.0896|  
+----------+---------+-----------+---------------+----------------+  
only showing top 20 rows  
  
------------用户行为表数据------------  
  
清洗后的用户行为表  
1. 填补缺失值  
2. 过滤掉商品不存在的数据  
+------+---------+---+--------+-------+-------+-------+-------+-------+-------+----+-----+  
|userId|productId|day|model\_id|utype\_1|utype\_2|utype\_3|utype\_4|utype\_5|utype\_6|cate|brand|  
+------+---------+---+--------+-------+-------+-------+-------+-------+-------+----+-----+  
|266079| 138778| 31| -2| 1| 0| 0| 0| 0| 0| 1| 8|  
|266079| 138778| 31| 0| 0| 0| 0| 0| 0| 1| 6| 8|  
|200719| 61226| 31| -2| 1| 0| 0| 0| 0| 0| 1| 8|  
|200719| 61226| 31| 0| 0| 0| 0| 0| 0| 1| 6| 8|  
|209278| 6998| 31| 0| 0| 0| 0| 0| 0| 1| 6| 8|  
|209278| 6998| 31| -2| 1| 0| 0| 0| 0| 0| 1| 8|  
|266079| 138778| 31| 0| 0| 0| 0| 0| 0| 1| 6| 8|  
|200719| 61226| 31| 11| 0| 0| 0| 0| 0| 1| 6| 8|  
|209278| 24679| 31| -2| 1| 0| 0| 0| 0| 0| 1| 8|  
|209278| 24679| 31| 0| 0| 0| 0| 0| 0| 1| 6| 8|  
|217567| 22387| 31| -2| 1| 0| 0| 0| 0| 0| 1| 8|  
|217567| 22387| 31| 0| 0| 0| 0| 0| 0| 1| 6| 8|  
|245959| 145895| 31| -2| 1| 0| 0| 0| 0| 0| 1| 8|  
|245959| 145895| 31| -2| 0| 0| 0| 0| 0| 1| 6| 8|  
|245959| 145895| 31| -2| 0| 0| 0| 0| 0| 1| 6| 8|  
|200719| 123773| 32| 0| 0| 0| 0| 0| 0| 1| 6| 8|  
|200719| 123773| 32| -2| 1| 0| 0| 0| 0| 0| 1| 8|  
|200719| 123773| 32| 0| 0| 0| 0| 0| 0| 1| 6| 8|  
|266079| 151327| 32| -2| 1| 0| 0| 0| 0| 0| 1| 8|  
|266079| 151327| 32| 0| 0| 0| 0| 0| 0| 1| 6| 8|  
+------+---------+---+--------+-------+-------+-------+-------+-------+-------+----+-----+  
only showing top 20 rows

### 特征抽取

+------+---------+--------+-------+-------+-------+-------+-------+-------+---+---+---+-----+-----+-----+-----+------+-----+-----+-----+-----+-----+-----+----------+-----------+---------------+----------------+-----+  
|userId|productId|model\_id|utype\_1|utype\_2|utype\_3|utype\_4|utype\_5|utype\_6| a1| a2| a3|brand|sex\_0|sex\_1|sex\_2|age\_d1|age\_1|age\_2|age\_3|age\_4|age\_5|age\_6|user\_lv\_cd|comment\_num|has\_bad\_comment|bad\_comment\_rate|label|  
+------+---------+--------+-------+-------+-------+-------+-------+-------+---+---+---+-----+-----+-----+-----+------+-----+-----+-----+-----+-----+-----+----------+-----------+---------------+----------------+-----+  
|211158| 9852| -2| 2| 1| 0| 0| 0| 3| 2| 2| 2| 812| 1| 0| 0| 0| 0| 0| 1| 0| 0| 0| 4| 24| 0| 0.0| 0|  
|275205| 9852| 218| 2| 0| 0| 0| 0| 2| 2| 2| 2| 812| 1| 0| 0| 0| 0| 0| 0| 1| 0| 0| 4| 24| 0| 0.0| 0|  
|231611| 9852| -2| 1| 0| 0| 0| 0| 1| 2| 2| 2| 812| 0| 0| 1| 0| 0| 0| 1| 0| 0| 0| 5| 24| 0| 0.0| 0|  
|243432| 9852| -2| 2| 0| 0| 0| 0| 4| 2| 2| 2| 812| 1| 0| 0| 0| 0| 0| 0| 1| 0| 0| 5| 24| 0| 0.0| 0|  
|261158| 9852| -2| 2| 0| 0| 0| 0| 3| 2| 2| 2| 812| 0| 0| 1| 0| 0| 0| 0| 1| 0| 0| 3| 24| 0| 0.0| 0|  
|220553| 9852| 0| 2| 0| 0| 0| 0| 3| 2| 2| 2| 812| 0| 0| 1| 0| 0| 0| 1| 0| 0| 0| 3| 24| 0| 0.0| 0|  
|239066| 9852| -2| 2| 0| 0| 0| 0| 3| 2| 2| 2| 812| 0| 0| 1| 0| 0| 0| 1| 0| 0| 0| 5| 24| 0| 0.0| 0|  
|276441| 9852| 216| 0| 0| 0| 0| 0| 7| 2| 2| 2| 812| 0| 0| 1| 0| 0| 0| 1| 0| 0| 0| 5| 24| 0| 0.0| 0|  
|286614| 9852| -2| 2| 0| 0| 0| 0| 3| 2| 2| 2| 812| 0| 0| 1| 0| 0| 0| 1| 0| 0| 0| 4| 24| 0| 0.0| 0|  
|226368| 9852| -2| 1| 0| 0| 0| 0| 1| 2| 2| 2| 812| 1| 0| 0| 1| 0| 0| 0| 0| 0| 0| 4| 24| 0| 0.0| 0|  
|271886| 9852| -2| 1| 0| 0| 0| 0| 1| 2| 2| 2| 812| 0| 0| 1| 1| 0| 0| 0| 0| 0| 0| 4| 24| 0| 0.0| 0|  
|231916| 9852| -2| 0| 0| 1| 0| 1| 0| 2| 2| 2| 812| 1| 0| 0| 0| 0| 0| 1| 0| 0| 0| 5| 24| 0| 0.0| 0|  
|271927| 9852| 0| 2| 0| 0| 0| 0| 2| 2| 2| 2| 812| 0| 0| 1| 0| 0| 0| 1| 0| 0| 0| 4| 24| 0| 0.0| 0|  
|226685| 9852| -2| 2| 0| 0| 0| 0| 3| 2| 2| 2| 812| 1| 0| 0| 0| 0| 0| 1| 0| 0| 0| 5| 24| 0| 0.0| 0|  
|238716| 9852| 0| 1| 0| 0| 0| 0| 2| 2| 2| 2| 812| 1| 0| 0| 0| 0| 0| 0| 1| 0| 0| 5| 24| 0| 0.0| 0|  
|262862| 9852| 0| 2| 0| 0| 0| 0| 4| 2| 2| 2| 812| 1| 0| 0| 0| 0| 0| 1| 0| 0| 0| 4| 24| 0| 0.0| 0|  
|295900| 9852| 216| 10| 0| 0| 0| 0| 6| 2| 2| 2| 812| 0| 0| 1| 0| 0| 0| 0| 1| 0| 0| 3| 24| 0| 0.0| 0|  
|293165| 9852| 216| 2| 0| 0| 0| 0| 3| 2| 2| 2| 812| 1| 0| 0| 0| 0| 0| 0| 1| 0| 0| 3| 24| 0| 0.0| 0|  
|241127| 9852| 217| 2| 0| 0| 0| 0| 3| 2| 2| 2| 812| 1| 0| 0| 0| 0| 0| 1| 0| 0| 0| 5| 24| 0| 0.0| 0|  
|238492| 9852| 0| 4| 0| 0| 0| 0| 4| 2| 2| 2| 812| 1| 0| 0| 0| 0| 0| 1| 0| 0| 0| 5| 24| 0| 0.0| 0|  
+------+---------+--------+-------+-------+-------+-------+-------+-------+---+---+---+-----+-----+-----+-----+------+-----+-----+-----+-----+-----+-----+----------+-----------+---------------+----------------+-----+

1. 提取说明：

* 其中userId、productId、label不属于特征值，其余均为特征值，label是训练标签
* 特征值是将4个表根据userId、productId连接而来
* label值是特征值时间后5天用户的购买行为：0为未购买，1为购买

1. 字段说明：

* 在上述数据分析和数据清洗中已有说明，字段名相同即为同一含义
* 特征数量大概27左右

### 特征细分

+------+---------+---+---+---+-----+-----+-----+-----+------+-----+-----+-----+-----+-----+-----+----------+-----------+---------------+----------------+----+----+----+----+----+----+----+----+----+----+----+----+----+----+----+----+----+----+------+------+------+------+------+------+------+------+------+------+------+------+------+------+------+------+------+------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-----+  
|userId|productId| a1| a2| a3|brand|sex\_0|sex\_1|sex\_2|age\_d1|age\_1|age\_2|age\_3|age\_4|age\_5|age\_6|user\_lv\_cd|comment\_num|has\_bad\_comment|bad\_comment\_rate|t1\_1|t2\_1|t3\_1|t4\_1|t5\_1|t6\_1|t1\_2|t2\_2|t3\_2|t4\_2|t5\_2|t6\_2|t1\_3|t2\_3|t3\_3|t4\_3|t5\_3|t6\_3|t1\_l\_3|t2\_l\_3|t3\_l\_3|t4\_l\_3|t5\_l\_3|t6\_l\_3|t1\_l\_5|t2\_l\_5|t3\_l\_5|t4\_l\_5|t5\_l\_5|t6\_l\_5|t1\_l\_7|t2\_l\_7|t3\_l\_7|t4\_l\_7|t5\_l\_7|t6\_l\_7|t1\_l\_10|t2\_l\_10|t3\_l\_10|t4\_l\_10|t5\_l\_10|t6\_l\_10|t1\_l\_14|t2\_l\_14|t3\_l\_14|t4\_l\_14|t5\_l\_14|t6\_l\_14|t1\_l\_50|t2\_l\_50|t3\_l\_50|t4\_l\_50|t5\_l\_50|t6\_l\_50|label|  
+------+---------+---+---+---+-----+-----+-----+-----+------+-----+-----+-----+-----+-----+-----+----------+-----------+---------------+----------------+----+----+----+----+----+----+----+----+----+----+----+----+----+----+----+----+----+----+------+------+------+------+------+------+------+------+------+------+------+------+------+------+------+------+------+------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-----+  
|211158| 9852| 2| 2| 2| 812| 1| 0| 0| 0| 0| 0| 1| 0| 0| 0| 4| 24| 0| 0.0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 2| 1| 0| 0| 0| 3| 0|  
|275205| 9852| 2| 2| 2| 812| 1| 0| 0| 0| 0| 0| 0| 1| 0| 0| 4| 24| 0| 0.0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 2| 0| 0| 0| 0| 2| 0|  
|231611| 9852| 2| 2| 2| 812| 0| 0| 1| 0| 0| 0| 1| 0| 0| 0| 5| 24| 0| 0.0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 1| 0| 0| 0| 0| 1| 0|  
|243432| 9852| 2| 2| 2| 812| 1| 0| 0| 0| 0| 0| 0| 1| 0| 0| 5| 24| 0| 0.0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 2| 0| 0| 0| 0| 4| 2| 0| 0| 0| 0| 4| 0|  
|261158| 9852| 2| 2| 2| 812| 0| 0| 1| 0| 0| 0| 0| 1| 0| 0| 3| 24| 0| 0.0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 2| 0| 0| 0| 0| 3| 0|  
|220553| 9852| 2| 2| 2| 812| 0| 0| 1| 0| 0| 0| 1| 0| 0| 0| 3| 24| 0| 0.0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 2| 0| 0| 0| 0| 3| 2| 0| 0| 0| 0| 3| 0|  
|239066| 9852| 2| 2| 2| 812| 0| 0| 1| 0| 0| 0| 1| 0| 0| 0| 5| 24| 0| 0.0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 2| 0| 0| 0| 0| 3| 2| 0| 0| 0| 0| 3| 2| 0| 0| 0| 0| 3| 2| 0| 0| 0| 0| 3| 2| 0| 0| 0| 0| 3| 2| 0| 0| 0| 0| 3| 2| 0| 0| 0| 0| 3| 0|  
|276441| 9852| 2| 2| 2| 812| 0| 0| 1| 0| 0| 0| 1| 0| 0| 0| 5| 24| 0| 0.0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 7| 0| 0| 0| 0| 0| 7| 0| 0| 0| 0| 0| 7| 0|  
|286614| 9852| 2| 2| 2| 812| 0| 0| 1| 0| 0| 0| 1| 0| 0| 0| 4| 24| 0| 0.0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 2| 0| 0| 0| 0| 3| 0|  
|226368| 9852| 2| 2| 2| 812| 1| 0| 0| 1| 0| 0| 0| 0| 0| 0| 4| 24| 0| 0.0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 1| 0| 0| 0| 0| 1| 0|  
|271886| 9852| 2| 2| 2| 812| 0| 0| 1| 1| 0| 0| 0| 0| 0| 0| 4| 24| 0| 0.0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 1| 0| 0| 0| 0| 1| 0|  
|231916| 9852| 2| 2| 2| 812| 1| 0| 0| 0| 0| 0| 1| 0| 0| 0| 5| 24| 0| 0.0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 1| 0| 1| 0| 0|  
|271927| 9852| 2| 2| 2| 812| 0| 0| 1| 0| 0| 0| 1| 0| 0| 0| 4| 24| 0| 0.0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 2| 0| 0| 0| 0| 2| 0|  
|226685| 9852| 2| 2| 2| 812| 1| 0| 0| 0| 0| 0| 1| 0| 0| 0| 5| 24| 0| 0.0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 2| 0| 0| 0| 0| 3| 2| 0| 0| 0| 0| 3| 2| 0| 0| 0| 0| 3| 2| 0| 0| 0| 0| 3| 2| 0| 0| 0| 0| 3| 0|  
|238716| 9852| 2| 2| 2| 812| 1| 0| 0| 0| 0| 0| 0| 1| 0| 0| 5| 24| 0| 0.0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 1| 0| 0| 0| 0| 2| 0|  
|262862| 9852| 2| 2| 2| 812| 1| 0| 0| 0| 0| 0| 1| 0| 0| 0| 4| 24| 0| 0.0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 2| 0| 0| 0| 0| 4| 0|  
|295900| 9852| 2| 2| 2| 812| 0| 0| 1| 0| 0| 0| 0| 1| 0| 0| 3| 24| 0| 0.0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 10| 0| 0| 0| 0| 6| 10| 0| 0| 0| 0| 6| 10| 0| 0| 0| 0| 6| 10| 0| 0| 0| 0| 6| 10| 0| 0| 0| 0| 6| 0|  
|293165| 9852| 2| 2| 2| 812| 1| 0| 0| 0| 0| 0| 0| 1| 0| 0| 3| 24| 0| 0.0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 2| 0| 0| 0| 0| 3| 0|  
|241127| 9852| 2| 2| 2| 812| 1| 0| 0| 0| 0| 0| 1| 0| 0| 0| 5| 24| 0| 0.0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 2| 0| 0| 0| 0| 3| 2| 0| 0| 0| 0| 3| 2| 0| 0| 0| 0| 3| 0|  
|238492| 9852| 2| 2| 2| 812| 1| 0| 0| 0| 0| 0| 1| 0| 0| 0| 5| 24| 0| 0.0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 4| 0| 0| 0| 0| 4| 0|  
+------+---------+---+---+---+-----+-----+-----+-----+------+-----+-----+-----+-----+-----+-----+----------+-----------+---------------+----------------+----+----+----+----+----+----+----+----+----+----+----+----+----+----+----+----+----+----+------+------+------+------+------+------+------+------+------+------+------+------+------+------+------+------+------+------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-----+

字段说明:  
- t1\_1：用户在预测日期前第1天进行1号操作的次数  
- t1\_2：用户在预测日期前第2天进行1号操作的次数  
- t2\_2：用户在预测日期前第2天进行2号操作的次数  
- t1\_l\_3：l意为latest(最近)，用户在预测日期前的3天内进行1号操作的次数  
- t1\_l\_5：l意为latest(最近)，用户在预测日期前的5天内进行1号操作的次数  
- t2\_l\_5：l意为latest(最近)，用户在预测日期前的5天内进行2号操作的次数

### 模型训练

仍然使用xgboost进行模型训练，参数如下：

"eta" -> 0.1, //学习率  
 "gamma" -> 0, //用于控制是否后剪枝的参数,越大越保守，一般0.1、0.2这样子。  
 "lambda" -> 2, //控制模型复杂度的权重值的L2正则化项参数，参数越大，模型越不容易过拟合。  
 "subsample" -> 1, //随机采样训练样本  
 "colsample\_bytree" -> 0.8, //生成树时进行的列采样  
 "max\_depth" -> maxDepth, //构建树的深度，越大越容易过拟合  
 "min\_child\_weight" -> 5,  
 "objective" -> "multi:softprob", //定义学习任务及相应的学习目标  
 "eval\_metric" -> "merror",  
 "num\_class" -> 21

使用细化前的特征进行训练：acc：0.08 使用细化后的特征进行训练：acc：0.35

### 总结

在网上看了很多相关的经验介绍，我认为在算法上和模型的构建上大多数人都不会表现出明显的差距，提高准确率最核心的还在于对数据的处理和特征的提取上。 在上述的模型训练中，参数的调节只能略微的提高准确度，而想让模型有质的提升，还是要提高数据的质量和特征的质量。 对于数据和特征的要求应该是求精而不求多。