# 表结构分析

* onehot 编码后生成一个新的字段，名字为{原字段}\_vec

------------prodcut-------------  
打印数据  
+---------+-------+----+-------+-----+-----+  
|productId| price|cate| para1|para2|para3|  
+---------+-------+----+-------+-----+-----+  
| 1| 364.2| 71| 247.95| -1.0| -1.0|  
| 2| 36.7| 46| 6.4| -1.0| -1.0|  
| 3| 953.3| 71| 178.92| -1.0| -1.0|  
| 4| 110.6| 83|6366.88| -1.0| -1.0|  
| 5| 241.5| 71| 31.74| -1.0| -1.0|  
| 6| 368.6| 71| 32.79| -1.0| -1.0|  
| 7|18568.9| 71| 73.0| -1.0| -1.0|  
| 8| 246.9| 101| 222.01| 3.0| 5.0|  
| 9| 886.5| 71| 49.95| -1.0| -1.0|  
| 10| 522.1| 101| 82.89| 3.0| 2.0|  
| 11| 255.6| 101| 81.53| 2.0| 4.0|  
| 12| 398.9| 71| 71.8| -1.0| -1.0|  
| 13| 117.2| 101| 56.43| 2.0| 3.0|  
| 14| 1455.4| 71| 119.94| -1.0| -1.0|  
| 15| 734.2| 71| 71.8| -1.0| -1.0|  
| 16| 908.4| 71| 72.33| -1.0| -1.0|  
| 17| 979.2| 101| 166.57| -1.0| 4.0|  
| 18| 145.8| 71| 15.74| -1.0| -1.0|  
| 19| 1089.3| 1| 24.47| -1.0| -1.0|  
| 20| 450.1| 101| 8.07| 4.0| 2.0|  
+---------+-------+----+-------+-----+-----+  
only showing top 20 rows  
  
统计种类  
+----+-----+  
|cate|count|  
+----+-----+  
| 101|30344|  
| 1| 7333|  
| 83| 7638|  
| 71|41959|  
| 30| 7380|  
| 46| 4758|  
+----+-----+  
  
onehot准备,将cate重新编码  
+---------+-------+----+-------+-----+-----+  
|productId| price|cate| para1|para2|para3|  
+---------+-------+----+-------+-----+-----+  
| 1| 364.2| 3| 247.95| -1.0| -1.0|  
| 2| 36.7| 5| 6.4| -1.0| -1.0|  
| 3| 953.3| 3| 178.92| -1.0| -1.0|  
| 4| 110.6| 2|6366.88| -1.0| -1.0|  
| 5| 241.5| 3| 31.74| -1.0| -1.0|  
| 6| 368.6| 3| 32.79| -1.0| -1.0|  
| 7|18568.9| 3| 73.0| -1.0| -1.0|  
| 8| 246.9| 0| 222.01| 3.0| 5.0|  
| 9| 886.5| 3| 49.95| -1.0| -1.0|  
| 10| 522.1| 0| 82.89| 3.0| 2.0|  
| 11| 255.6| 0| 81.53| 2.0| 4.0|  
| 12| 398.9| 3| 71.8| -1.0| -1.0|  
| 13| 117.2| 0| 56.43| 2.0| 3.0|  
| 14| 1455.4| 3| 119.94| -1.0| -1.0|  
| 15| 734.2| 3| 71.8| -1.0| -1.0|  
| 16| 908.4| 3| 72.33| -1.0| -1.0|  
| 17| 979.2| 0| 166.57| -1.0| 4.0|  
| 18| 145.8| 3| 15.74| -1.0| -1.0|  
| 19| 1089.3| 1| 24.47| -1.0| -1.0|  
| 20| 450.1| 0| 8.07| 4.0| 2.0|  
+---------+-------+----+-------+-----+-----+  
only showing top 20 rows  
  
进行onehot编码  
+---------+-------+----+-------+-----+-----+-------------+  
|productId| price|cate| para1|para2|para3| cate\_vec|  
+---------+-------+----+-------+-----+-----+-------------+  
| 1| 364.2| 3| 247.95| -1.0| -1.0|(5,[3],[1.0])|  
| 2| 36.7| 5| 6.4| -1.0| -1.0| (5,[],[])|  
| 3| 953.3| 3| 178.92| -1.0| -1.0|(5,[3],[1.0])|  
| 4| 110.6| 2|6366.88| -1.0| -1.0|(5,[2],[1.0])|  
| 5| 241.5| 3| 31.74| -1.0| -1.0|(5,[3],[1.0])|  
| 6| 368.6| 3| 32.79| -1.0| -1.0|(5,[3],[1.0])|  
| 7|18568.9| 3| 73.0| -1.0| -1.0|(5,[3],[1.0])|  
| 8| 246.9| 0| 222.01| 3.0| 5.0|(5,[0],[1.0])|  
| 9| 886.5| 3| 49.95| -1.0| -1.0|(5,[3],[1.0])|  
| 10| 522.1| 0| 82.89| 3.0| 2.0|(5,[0],[1.0])|  
| 11| 255.6| 0| 81.53| 2.0| 4.0|(5,[0],[1.0])|  
| 12| 398.9| 3| 71.8| -1.0| -1.0|(5,[3],[1.0])|  
| 13| 117.2| 0| 56.43| 2.0| 3.0|(5,[0],[1.0])|  
| 14| 1455.4| 3| 119.94| -1.0| -1.0|(5,[3],[1.0])|  
| 15| 734.2| 3| 71.8| -1.0| -1.0|(5,[3],[1.0])|  
| 16| 908.4| 3| 72.33| -1.0| -1.0|(5,[3],[1.0])|  
| 17| 979.2| 0| 166.57| -1.0| 4.0|(5,[0],[1.0])|  
| 18| 145.8| 3| 15.74| -1.0| -1.0|(5,[3],[1.0])|  
| 19| 1089.3| 1| 24.47| -1.0| -1.0|(5,[1],[1.0])|  
| 20| 450.1| 0| 8.07| 4.0| 2.0|(5,[0],[1.0])|  
+---------+-------+----+-------+-----+-----+-------------+  
only showing top 20 rows  
  
抽取特征  
+---------+-------+-------+-----+-----+-------------+  
|productId| price| para1|para2|para3| cate\_vec|  
+---------+-------+-------+-----+-----+-------------+  
| 1| 364.2| 247.95| -1.0| -1.0|(5,[3],[1.0])|  
| 2| 36.7| 6.4| -1.0| -1.0| (5,[],[])|  
| 3| 953.3| 178.92| -1.0| -1.0|(5,[3],[1.0])|  
| 4| 110.6|6366.88| -1.0| -1.0|(5,[2],[1.0])|  
| 5| 241.5| 31.74| -1.0| -1.0|(5,[3],[1.0])|  
| 6| 368.6| 32.79| -1.0| -1.0|(5,[3],[1.0])|  
| 7|18568.9| 73.0| -1.0| -1.0|(5,[3],[1.0])|  
| 8| 246.9| 222.01| 3.0| 5.0|(5,[0],[1.0])|  
| 9| 886.5| 49.95| -1.0| -1.0|(5,[3],[1.0])|  
| 10| 522.1| 82.89| 3.0| 2.0|(5,[0],[1.0])|  
| 11| 255.6| 81.53| 2.0| 4.0|(5,[0],[1.0])|  
| 12| 398.9| 71.8| -1.0| -1.0|(5,[3],[1.0])|  
| 13| 117.2| 56.43| 2.0| 3.0|(5,[0],[1.0])|  
| 14| 1455.4| 119.94| -1.0| -1.0|(5,[3],[1.0])|  
| 15| 734.2| 71.8| -1.0| -1.0|(5,[3],[1.0])|  
| 16| 908.4| 72.33| -1.0| -1.0|(5,[3],[1.0])|  
| 17| 979.2| 166.57| -1.0| 4.0|(5,[0],[1.0])|  
| 18| 145.8| 15.74| -1.0| -1.0|(5,[3],[1.0])|  
| 19| 1089.3| 24.47| -1.0| -1.0|(5,[1],[1.0])|  
| 20| 450.1| 8.07| 4.0| 2.0|(5,[0],[1.0])|  
+---------+-------+-------+-----+-----+-------------+  
only showing top 20 rows  
  
------------UA-------------  
打印数据  
+------+---------+----------+---------+----------+------+---+  
|userId|productId|actionDate|actionNum|actionType| month|day|  
+------+---------+----------+---------+----------+------+---+  
| 1| 80036|2017-04-14| 4| 1|201704| 14|  
| 1| 96959|2017-01-12| 2| 1|201701| 12|  
| 1| 8017|2017-03-09| 1| 1|201703| 9|  
| 1| 80036|2017-04-30| 6| 1|201704| 30|  
| 1| 16607|2017-01-12| 3| 1|201701| 12|  
| 1| 16607|2017-02-12| 2| 1|201702| 12|  
| 1| 16607|2017-03-09| 6| 1|201703| 9|  
| 1| 28038|2017-03-09| 1| 1|201703| 9|  
| 1| 45314|2017-03-09| 1| 1|201703| 9|  
| 1| 48100|2017-03-09| 1| 1|201703| 9|  
| 1| 48100|2017-04-30| 1| 1|201704| 30|  
| 1| 80036|2017-03-09| 5| 1|201703| 9|  
| 1| 80036|2017-03-22| 7| 1|201703| 22|  
| 2| 44666|2017-04-25| 2| 1|201704| 25|  
| 2| 44666|2017-04-24| 2| 1|201704| 24|  
| 2| 92546|2017-04-24| 1| 1|201704| 24|  
| 2| 89819|2017-04-24| 1| 1|201704| 24|  
| 2| 83955|2017-04-24| 9| 1|201704| 24|  
| 2| 81082|2017-04-24| 1| 1|201704| 24|  
| 3| 2922|2016-12-08| 4| 1|201612| 8|  
+------+---------+----------+---------+----------+------+---+  
only showing top 20 rows  
  
统计种类  
+----------+-------+  
|actionType| count|  
+----------+-------+  
| 1|6873551|  
| 2| 70590|  
+----------+-------+  
  
统计月份  
+------+------+  
| month| count|  
+------+------+  
|201701|628321|  
|201702|777369|  
|201609|402053|  
|201704|889789|  
|201611|758200|  
|201607|312403|  
|201610|554023|  
|201606|393552|  
|201605|291713|  
|201608|362540|  
|201612|650413|  
|201703|923765|  
+------+------+  
  
对actionType进行onehot编码  
+------+---------+----------+---------+----------+------+---+--------------+  
|userId|productId|actionDate|actionNum|actionType| month|day|actionType\_vec|  
+------+---------+----------+---------+----------+------+---+--------------+  
| 1| 80036|2017-04-14| 4| 1|201704| 14| (2,[1],[1.0])|  
| 1| 96959|2017-01-12| 2| 1|201701| 12| (2,[1],[1.0])|  
| 1| 8017|2017-03-09| 1| 1|201703| 9| (2,[1],[1.0])|  
| 1| 80036|2017-04-30| 6| 1|201704| 30| (2,[1],[1.0])|  
| 1| 16607|2017-01-12| 3| 1|201701| 12| (2,[1],[1.0])|  
| 1| 16607|2017-02-12| 2| 1|201702| 12| (2,[1],[1.0])|  
| 1| 16607|2017-03-09| 6| 1|201703| 9| (2,[1],[1.0])|  
| 1| 28038|2017-03-09| 1| 1|201703| 9| (2,[1],[1.0])|  
| 1| 45314|2017-03-09| 1| 1|201703| 9| (2,[1],[1.0])|  
| 1| 48100|2017-03-09| 1| 1|201703| 9| (2,[1],[1.0])|  
| 1| 48100|2017-04-30| 1| 1|201704| 30| (2,[1],[1.0])|  
| 1| 80036|2017-03-09| 5| 1|201703| 9| (2,[1],[1.0])|  
| 1| 80036|2017-03-22| 7| 1|201703| 22| (2,[1],[1.0])|  
| 2| 44666|2017-04-25| 2| 1|201704| 25| (2,[1],[1.0])|  
| 2| 44666|2017-04-24| 2| 1|201704| 24| (2,[1],[1.0])|  
| 2| 92546|2017-04-24| 1| 1|201704| 24| (2,[1],[1.0])|  
| 2| 89819|2017-04-24| 1| 1|201704| 24| (2,[1],[1.0])|  
| 2| 83955|2017-04-24| 9| 1|201704| 24| (2,[1],[1.0])|  
| 2| 81082|2017-04-24| 1| 1|201704| 24| (2,[1],[1.0])|  
| 3| 2922|2016-12-08| 4| 1|201612| 8| (2,[1],[1.0])|  
+------+---------+----------+---------+----------+------+---+--------------+  
only showing top 20 rows  
  
抽取特征  
+------+---------+----------+---------+------+---+--------------+  
|userId|productId|actionDate|actionNum| month|day|actionType\_vec|  
+------+---------+----------+---------+------+---+--------------+  
| 1| 80036|2017-04-14| 4|201704| 14| (2,[1],[1.0])|  
| 1| 96959|2017-01-12| 2|201701| 12| (2,[1],[1.0])|  
| 1| 8017|2017-03-09| 1|201703| 9| (2,[1],[1.0])|  
| 1| 80036|2017-04-30| 6|201704| 30| (2,[1],[1.0])|  
| 1| 16607|2017-01-12| 3|201701| 12| (2,[1],[1.0])|  
| 1| 16607|2017-02-12| 2|201702| 12| (2,[1],[1.0])|  
| 1| 16607|2017-03-09| 6|201703| 9| (2,[1],[1.0])|  
| 1| 28038|2017-03-09| 1|201703| 9| (2,[1],[1.0])|  
| 1| 45314|2017-03-09| 1|201703| 9| (2,[1],[1.0])|  
| 1| 48100|2017-03-09| 1|201703| 9| (2,[1],[1.0])|  
| 1| 48100|2017-04-30| 1|201704| 30| (2,[1],[1.0])|  
| 1| 80036|2017-03-09| 5|201703| 9| (2,[1],[1.0])|  
| 1| 80036|2017-03-22| 7|201703| 22| (2,[1],[1.0])|  
| 2| 44666|2017-04-25| 2|201704| 25| (2,[1],[1.0])|  
| 2| 44666|2017-04-24| 2|201704| 24| (2,[1],[1.0])|  
| 2| 92546|2017-04-24| 1|201704| 24| (2,[1],[1.0])|  
| 2| 89819|2017-04-24| 1|201704| 24| (2,[1],[1.0])|  
| 2| 83955|2017-04-24| 9|201704| 24| (2,[1],[1.0])|  
| 2| 81082|2017-04-24| 1|201704| 24| (2,[1],[1.0])|  
| 3| 2922|2016-12-08| 4|201612| 8| (2,[1],[1.0])|  
+------+---------+----------+---------+------+---+--------------+  
only showing top 20 rows  
  
------------user-------------  
打印数据  
+------+---+---+---------+  
|userID|age|sex|userGrade|  
+------+---+---+---------+  
| 1| 3| 2| 3|  
| 2| 3| 2| 3|  
| 3| 3| 0| 5|  
| 4| 3| 2| 3|  
| 5| 4| 2| 2|  
| 6| -1| 2| 2|  
| 7| 3| 2| 4|  
| 8| 2| 2| 2|  
| 9| 3| 2| 3|  
| 10| 4| 0| 4|  
| 11| 3| 1| 5|  
| 12| 6| 2| 3|  
| 13| -1| 2| 2|  
| 14| 3| 2| 3|  
| 15| -1| 2| 2|  
| 16| -1| 0| 4|  
| 17| 4| 2| 4|  
| 18| 3| 2| 2|  
| 19| -1| 2| 4|  
| 20| 3| 2| 5|  
+------+---+---+---------+  
only showing top 20 rows  
  
统计年龄  
+---+-----+  
|age|count|  
+---+-----+  
| -1|24636|  
| 1| 18|  
| 6| 840|  
| 3|42486|  
| 5| 1095|  
| 4|21818|  
| 2| 8031|  
+---+-----+  
  
统计性别  
+---+-----+  
|sex|count|  
+---+-----+  
| 1|10146|  
| 2|68577|  
| 0|20201|  
+---+-----+  
  
统计等级  
+---------+-----+  
|userGrade|count|  
+---------+-----+  
| 1| 548|  
| 3|30919|  
| 5|16326|  
| 4|21821|  
| 2|29310|  
+---------+-----+  
  
总数  
98924  
去除负值  
+------+---+---+---------+  
|userID|age|sex|userGrade|  
+------+---+---+---------+  
| 1| 3| 2| 3|  
| 2| 3| 2| 3|  
| 3| 3| 0| 5|  
| 4| 3| 2| 3|  
| 5| 4| 2| 2|  
| 6| 0| 2| 2|  
| 7| 3| 2| 4|  
| 8| 2| 2| 2|  
| 9| 3| 2| 3|  
| 10| 4| 0| 4|  
| 11| 3| 1| 5|  
| 12| 6| 2| 3|  
| 13| 0| 2| 2|  
| 14| 3| 2| 3|  
| 15| 0| 2| 2|  
| 16| 0| 0| 4|  
| 17| 4| 2| 4|  
| 18| 3| 2| 2|  
| 19| 0| 2| 4|  
| 20| 3| 2| 5|  
+------+---+---+---------+  
only showing top 20 rows  
  
进行onehot编码  
+------+---+---+---------+-------------+-------------+-------------+  
|userID|age|sex|userGrade| age\_vec| sex\_vec|userGrade\_vec|  
+------+---+---+---------+-------------+-------------+-------------+  
| 1| 3| 2| 3|(6,[3],[1.0])| (2,[],[])|(5,[3],[1.0])|  
| 2| 3| 2| 3|(6,[3],[1.0])| (2,[],[])|(5,[3],[1.0])|  
| 3| 3| 0| 5|(6,[3],[1.0])|(2,[0],[1.0])| (5,[],[])|  
| 4| 3| 2| 3|(6,[3],[1.0])| (2,[],[])|(5,[3],[1.0])|  
| 5| 4| 2| 2|(6,[4],[1.0])| (2,[],[])|(5,[2],[1.0])|  
| 6| 0| 2| 2|(6,[0],[1.0])| (2,[],[])|(5,[2],[1.0])|  
| 7| 3| 2| 4|(6,[3],[1.0])| (2,[],[])|(5,[4],[1.0])|  
| 8| 2| 2| 2|(6,[2],[1.0])| (2,[],[])|(5,[2],[1.0])|  
| 9| 3| 2| 3|(6,[3],[1.0])| (2,[],[])|(5,[3],[1.0])|  
| 10| 4| 0| 4|(6,[4],[1.0])|(2,[0],[1.0])|(5,[4],[1.0])|  
| 11| 3| 1| 5|(6,[3],[1.0])|(2,[1],[1.0])| (5,[],[])|  
| 12| 6| 2| 3| (6,[],[])| (2,[],[])|(5,[3],[1.0])|  
| 13| 0| 2| 2|(6,[0],[1.0])| (2,[],[])|(5,[2],[1.0])|  
| 14| 3| 2| 3|(6,[3],[1.0])| (2,[],[])|(5,[3],[1.0])|  
| 15| 0| 2| 2|(6,[0],[1.0])| (2,[],[])|(5,[2],[1.0])|  
| 16| 0| 0| 4|(6,[0],[1.0])|(2,[0],[1.0])|(5,[4],[1.0])|  
| 17| 4| 2| 4|(6,[4],[1.0])| (2,[],[])|(5,[4],[1.0])|  
| 18| 3| 2| 2|(6,[3],[1.0])| (2,[],[])|(5,[2],[1.0])|  
| 19| 0| 2| 4|(6,[0],[1.0])| (2,[],[])|(5,[4],[1.0])|  
| 20| 3| 2| 5|(6,[3],[1.0])| (2,[],[])| (5,[],[])|  
+------+---+---+---------+-------------+-------------+-------------+  
only showing top 20 rows  
  
抽取特征  
+------+-------------+-------------+-------------+  
|userID| age\_vec| sex\_vec|userGrade\_vec|  
+------+-------------+-------------+-------------+  
| 1|(6,[3],[1.0])| (2,[],[])|(5,[3],[1.0])|  
| 2|(6,[3],[1.0])| (2,[],[])|(5,[3],[1.0])|  
| 3|(6,[3],[1.0])|(2,[0],[1.0])| (5,[],[])|  
| 4|(6,[3],[1.0])| (2,[],[])|(5,[3],[1.0])|  
| 5|(6,[4],[1.0])| (2,[],[])|(5,[2],[1.0])|  
| 6|(6,[0],[1.0])| (2,[],[])|(5,[2],[1.0])|  
| 7|(6,[3],[1.0])| (2,[],[])|(5,[4],[1.0])|  
| 8|(6,[2],[1.0])| (2,[],[])|(5,[2],[1.0])|  
| 9|(6,[3],[1.0])| (2,[],[])|(5,[3],[1.0])|  
| 10|(6,[4],[1.0])|(2,[0],[1.0])|(5,[4],[1.0])|  
| 11|(6,[3],[1.0])|(2,[1],[1.0])| (5,[],[])|  
| 12| (6,[],[])| (2,[],[])|(5,[3],[1.0])|  
| 13|(6,[0],[1.0])| (2,[],[])|(5,[2],[1.0])|  
| 14|(6,[3],[1.0])| (2,[],[])|(5,[3],[1.0])|  
| 15|(6,[0],[1.0])| (2,[],[])|(5,[2],[1.0])|  
| 16|(6,[0],[1.0])|(2,[0],[1.0])|(5,[4],[1.0])|  
| 17|(6,[4],[1.0])| (2,[],[])|(5,[4],[1.0])|  
| 18|(6,[3],[1.0])| (2,[],[])|(5,[2],[1.0])|  
| 19|(6,[0],[1.0])| (2,[],[])|(5,[4],[1.0])|  
| 20|(6,[3],[1.0])| (2,[],[])| (5,[],[])|  
+------+-------------+-------------+-------------+  
only showing top 20 rows  
  
------------comment-------------  
打印数据  
+------+-------------------+-------+----------+  
|userId| commentTime|orderId|scoreLevel|  
+------+-------------------+-------+----------+  
| 3|2017-03-21 20:38:25| 403068| 1|  
| 7|2016-12-12 09:31:43| 194387| 1|  
| 7|2016-12-17 16:37:23| 197976| 2|  
| 7|2016-12-12 09:30:50| 327504| 1|  
| 8|2017-02-16 22:26:59| 173806| 1|  
| 9|2016-11-05 07:26:59| 319459| 1|  
| 9|2016-11-05 07:24:28| 448441| 1|  
| 9|2016-11-05 07:29:02| 36766| 1|  
| 11|2017-02-10 10:39:55| 612959| 1|  
| 11|2017-02-10 10:38:36| 612959| 1|  
| 11|2017-02-10 10:39:30| 612959| 1|  
| 11|2017-02-10 10:39:44| 612959| 1|  
| 11|2017-01-28 08:59:05| 246666| 1|  
| 11|2017-02-06 21:20:16| 246666| 1|  
| 11|2017-02-06 21:19:30| 641921| 1|  
| 11|2017-02-06 21:28:20| 218972| 1|  
| 11|2017-03-23 08:14:40| 371311| 1|  
| 11|2017-03-23 14:33:21| 424300| 1|  
| 11|2017-03-23 07:34:47| 478342| 1|  
| 14|2017-04-18 13:41:16| 610711| 1|  
+------+-------------------+-------+----------+  
only showing top 20 rows  
  
统计评分等级  
+----------+------+  
|scoreLevel| count|  
+----------+------+  
| 1|218250|  
| 3| 2899|  
| 2| 3135|  
+----------+------+  
  
进行onehot编码  
+------+-------------------+-------+----------+--------------+  
|userId| commentTime|orderId|scoreLevel|scoreLevel\_vec|  
+------+-------------------+-------+----------+--------------+  
| 3|2017-03-21 20:38:25| 403068| 1| (3,[1],[1.0])|  
| 7|2016-12-12 09:31:43| 194387| 1| (3,[1],[1.0])|  
| 7|2016-12-17 16:37:23| 197976| 2| (3,[2],[1.0])|  
| 7|2016-12-12 09:30:50| 327504| 1| (3,[1],[1.0])|  
| 8|2017-02-16 22:26:59| 173806| 1| (3,[1],[1.0])|  
| 9|2016-11-05 07:26:59| 319459| 1| (3,[1],[1.0])|  
| 9|2016-11-05 07:24:28| 448441| 1| (3,[1],[1.0])|  
| 9|2016-11-05 07:29:02| 36766| 1| (3,[1],[1.0])|  
| 11|2017-02-10 10:39:55| 612959| 1| (3,[1],[1.0])|  
| 11|2017-02-10 10:38:36| 612959| 1| (3,[1],[1.0])|  
| 11|2017-02-10 10:39:30| 612959| 1| (3,[1],[1.0])|  
| 11|2017-02-10 10:39:44| 612959| 1| (3,[1],[1.0])|  
| 11|2017-01-28 08:59:05| 246666| 1| (3,[1],[1.0])|  
| 11|2017-02-06 21:20:16| 246666| 1| (3,[1],[1.0])|  
| 11|2017-02-06 21:19:30| 641921| 1| (3,[1],[1.0])|  
| 11|2017-02-06 21:28:20| 218972| 1| (3,[1],[1.0])|  
| 11|2017-03-23 08:14:40| 371311| 1| (3,[1],[1.0])|  
| 11|2017-03-23 14:33:21| 424300| 1| (3,[1],[1.0])|  
| 11|2017-03-23 07:34:47| 478342| 1| (3,[1],[1.0])|  
| 14|2017-04-18 13:41:16| 610711| 1| (3,[1],[1.0])|  
+------+-------------------+-------+----------+--------------+  
only showing top 20 rows  
  
抽取特征  
+------+-------------------+-------+--------------+  
|userId| commentTime|orderId|scoreLevel\_vec|  
+------+-------------------+-------+--------------+  
| 3|2017-03-21 20:38:25| 403068| (3,[1],[1.0])|  
| 7|2016-12-12 09:31:43| 194387| (3,[1],[1.0])|  
| 7|2016-12-17 16:37:23| 197976| (3,[2],[1.0])|  
| 7|2016-12-12 09:30:50| 327504| (3,[1],[1.0])|  
| 8|2017-02-16 22:26:59| 173806| (3,[1],[1.0])|  
| 9|2016-11-05 07:26:59| 319459| (3,[1],[1.0])|  
| 9|2016-11-05 07:24:28| 448441| (3,[1],[1.0])|  
| 9|2016-11-05 07:29:02| 36766| (3,[1],[1.0])|  
| 11|2017-02-10 10:39:55| 612959| (3,[1],[1.0])|  
| 11|2017-02-10 10:38:36| 612959| (3,[1],[1.0])|  
| 11|2017-02-10 10:39:30| 612959| (3,[1],[1.0])|  
| 11|2017-02-10 10:39:44| 612959| (3,[1],[1.0])|  
| 11|2017-01-28 08:59:05| 246666| (3,[1],[1.0])|  
| 11|2017-02-06 21:20:16| 246666| (3,[1],[1.0])|  
| 11|2017-02-06 21:19:30| 641921| (3,[1],[1.0])|  
| 11|2017-02-06 21:28:20| 218972| (3,[1],[1.0])|  
| 11|2017-03-23 08:14:40| 371311| (3,[1],[1.0])|  
| 11|2017-03-23 14:33:21| 424300| (3,[1],[1.0])|  
| 11|2017-03-23 07:34:47| 478342| (3,[1],[1.0])|  
| 14|2017-04-18 13:41:16| 610711| (3,[1],[1.0])|  
+------+-------------------+-------+--------------+  
only showing top 20 rows  
  
------------order-------------  
打印数据  
+------+---------+-------+----------+---------+------+------+---+  
|userId|productId|orderId| orderDate|orderArea|BuyNum| month|day|  
+------+---------+-------+----------+---------+------+------+---+  
| 1| 80036| 480066|2017-03-09| 14| 1|201703| 9|  
| 1| 16607| 106933|2017-01-12| 14| 1|201701| 12|  
| 1| 80036| 542604|2017-04-14| 14| 1|201704| 14|  
| 1| 80036| 196335|2017-04-30| 14| 1|201704| 30|  
| 1| 80036| 42502|2017-03-22| 14| 2|201703| 22|  
| 1| 16607| 235938|2017-02-12| 14| 1|201702| 12|  
| 2| 44666| 554520|2017-04-24| 27| 2|201704| 24|  
| 3| 71817| 403068|2017-03-03| 26| 2|201703| 3|  
| 3| 18217| 403068|2017-03-03| 26| 2|201703| 3|  
| 3| 2816| 403068|2017-03-03| 26| 1|201703| 3|  
| 3| 18217| 519963|2016-11-10| 26| 1|201611| 10|  
| 3| 82091| 519963|2016-11-10| 26| 1|201611| 10|  
| 3| 91286| 18791|2016-05-20| 26| 2|201605| 20|  
| 3| 28001| 22600|2017-01-16| 26| 4|201701| 16|  
| 3| 71817| 519963|2016-11-10| 26| 2|201611| 10|  
| 3| 38635| 519963|2016-11-10| 26| 1|201611| 10|  
| 3| 80409| 25479|2017-01-01| 26| 2|201701| 1|  
| 3| 35919| 157776|2016-11-10| 26| 1|201611| 10|  
| 3| 28001| 110597|2016-07-25| 26| 2|201607| 25|  
| 3| 16586| 519963|2016-11-10| 26| 1|201611| 10|  
+------+---------+-------+----------+---------+------+------+---+  
only showing top 20 rows  
  
统计月份  
+------+------+  
| month| count|  
+------+------+  
|201701| 69797|  
|201702| 91739|  
|201609| 46799|  
|201704|114181|  
|201611| 79291|  
|201607| 34703|  
|201610| 58081|  
|201606| 46146|  
|201605| 35381|  
|201608| 41178|  
|201612| 67054|  
|201703|108373|  
+------+------+  
  
统计4月下单人数  
57658  
统计下单区域  
+---------+------+  
|orderArea| count|  
+---------+------+  
| 31| 6536|  
| 28| 21346|  
| 27| 33483|  
| 26| 22619|  
| 12| 5251|  
| 22| 5480|  
| 1| 5317|  
| 13| 20693|  
| 16| 23521|  
| 6| 133|  
| 3| 7696|  
| 20| 54425|  
| 5| 5503|  
| 19| 8152|  
| 15| 2175|  
| 17| 81484|  
| 9| 17273|  
| 4| 7599|  
| 8|145322|  
| 23| 11048|  
+---------+------+  
only showing top 20 rows  
  
有多少个下单区域  
31  
对orderArea进行onehot编码  
+------+---------+-------+----------+---------+------+------+---+---------------+  
|userId|productId|orderId| orderDate|orderArea|BuyNum| month|day| orderArea\_vec|  
+------+---------+-------+----------+---------+------+------+---+---------------+  
| 1| 80036| 480066|2017-03-09| 14| 1|201703| 9|(31,[14],[1.0])|  
| 1| 16607| 106933|2017-01-12| 14| 1|201701| 12|(31,[14],[1.0])|  
| 1| 80036| 542604|2017-04-14| 14| 1|201704| 14|(31,[14],[1.0])|  
| 1| 80036| 196335|2017-04-30| 14| 1|201704| 30|(31,[14],[1.0])|  
| 1| 80036| 42502|2017-03-22| 14| 2|201703| 22|(31,[14],[1.0])|  
| 1| 16607| 235938|2017-02-12| 14| 1|201702| 12|(31,[14],[1.0])|  
| 2| 44666| 554520|2017-04-24| 27| 2|201704| 24|(31,[27],[1.0])|  
| 3| 71817| 403068|2017-03-03| 26| 2|201703| 3|(31,[26],[1.0])|  
| 3| 18217| 403068|2017-03-03| 26| 2|201703| 3|(31,[26],[1.0])|  
| 3| 2816| 403068|2017-03-03| 26| 1|201703| 3|(31,[26],[1.0])|  
| 3| 18217| 519963|2016-11-10| 26| 1|201611| 10|(31,[26],[1.0])|  
| 3| 82091| 519963|2016-11-10| 26| 1|201611| 10|(31,[26],[1.0])|  
| 3| 91286| 18791|2016-05-20| 26| 2|201605| 20|(31,[26],[1.0])|  
| 3| 28001| 22600|2017-01-16| 26| 4|201701| 16|(31,[26],[1.0])|  
| 3| 71817| 519963|2016-11-10| 26| 2|201611| 10|(31,[26],[1.0])|  
| 3| 38635| 519963|2016-11-10| 26| 1|201611| 10|(31,[26],[1.0])|  
| 3| 80409| 25479|2017-01-01| 26| 2|201701| 1|(31,[26],[1.0])|  
| 3| 35919| 157776|2016-11-10| 26| 1|201611| 10|(31,[26],[1.0])|  
| 3| 28001| 110597|2016-07-25| 26| 2|201607| 25|(31,[26],[1.0])|  
| 3| 16586| 519963|2016-11-10| 26| 1|201611| 10|(31,[26],[1.0])|  
+------+---------+-------+----------+---------+------+------+---+---------------+  
only showing top 20 rows  
  
特征抽取  
+------+---------+-------+----------+---------+------+------+---+---------------+  
|userId|productId|orderId| orderDate|orderArea|BuyNum| month|day| orderArea\_vec|  
+------+---------+-------+----------+---------+------+------+---+---------------+  
| 1| 80036| 480066|2017-03-09| 14| 1|201703| 9|(31,[14],[1.0])|  
| 1| 16607| 106933|2017-01-12| 14| 1|201701| 12|(31,[14],[1.0])|  
| 1| 80036| 542604|2017-04-14| 14| 1|201704| 14|(31,[14],[1.0])|  
| 1| 80036| 196335|2017-04-30| 14| 1|201704| 30|(31,[14],[1.0])|  
| 1| 80036| 42502|2017-03-22| 14| 2|201703| 22|(31,[14],[1.0])|  
| 1| 16607| 235938|2017-02-12| 14| 1|201702| 12|(31,[14],[1.0])|  
| 2| 44666| 554520|2017-04-24| 27| 2|201704| 24|(31,[27],[1.0])|  
| 3| 71817| 403068|2017-03-03| 26| 2|201703| 3|(31,[26],[1.0])|  
| 3| 18217| 403068|2017-03-03| 26| 2|201703| 3|(31,[26],[1.0])|  
| 3| 2816| 403068|2017-03-03| 26| 1|201703| 3|(31,[26],[1.0])|  
| 3| 18217| 519963|2016-11-10| 26| 1|201611| 10|(31,[26],[1.0])|  
| 3| 82091| 519963|2016-11-10| 26| 1|201611| 10|(31,[26],[1.0])|  
| 3| 91286| 18791|2016-05-20| 26| 2|201605| 20|(31,[26],[1.0])|  
| 3| 28001| 22600|2017-01-16| 26| 4|201701| 16|(31,[26],[1.0])|  
| 3| 71817| 519963|2016-11-10| 26| 2|201611| 10|(31,[26],[1.0])|  
| 3| 38635| 519963|2016-11-10| 26| 1|201611| 10|(31,[26],[1.0])|  
| 3| 80409| 25479|2017-01-01| 26| 2|201701| 1|(31,[26],[1.0])|  
| 3| 35919| 157776|2016-11-10| 26| 1|201611| 10|(31,[26],[1.0])|  
| 3| 28001| 110597|2016-07-25| 26| 2|201607| 25|(31,[26],[1.0])|  
| 3| 16586| 519963|2016-11-10| 26| 1|201611| 10|(31,[26],[1.0])|  
+------+---------+-------+----------+---------+------+------+---+---------------+  
only showing top 20 rows

# 提取统计特征

* ua1表示用户浏览行为，ua2代表用户关注行为，s代表用户评分行为，o代表用户下单行为
* 后面的数字代表天数，5代表离预测时间5天内的行为总量，10代表离预测时间10天内的行为总量以此类推

用户行为表  
+------+---------+-----+-----+------+------+------+------+------+------+------+------+------+------+-------+-------+-------+-------+-------+-------+  
|userId|productId|ua1\_5|ua2\_5|ua1\_10|ua2\_10|ua1\_20|ua2\_20|ua1\_40|ua2\_40|ua1\_60|ua2\_60|ua1\_80|ua2\_80|ua1\_100|ua2\_100|ua1\_200|ua2\_200|ua1\_400|ua2\_400|  
+------+---------+-----+-----+------+------+------+------+------+------+------+------+------+------+-------+-------+-------+-------+-------+-------+  
| 39974| 81479| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 1| 0|  
| 37773| 51347| 0| 0| 0| 0| 0| 0| 4| 0| 4| 0| 4| 0| 4| 0| 8| 0| 10| 0|  
| 11980| 45817| 0| 0| 0| 0| 0| 0| 1| 0| 1| 0| 1| 0| 3| 0| 19| 0| 20| 0|  
| 56006| 45817| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 1| 0| 14| 0| 14| 0|  
| 65875| 38710| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 1| 0|  
| 4104| 4908| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 1| 0| 2| 0| 14| 0|  
| 8824| 85363| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 1| 0| 1| 0| 1| 0|  
| 43890| 16585| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 4| 0|  
| 22661| 19403| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 3| 0| 3| 0|  
| 85660| 86925| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 1| 0| 1| 0|  
| 27188| 9915| 0| 0| 0| 0| 0| 0| 0| 0| 2| 0| 5| 0| 6| 0| 6| 0| 6| 0|  
| 38383| 29319| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 4| 0|  
| 48684| 34011| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 1| 0| 2| 0|  
| 86089| 27211| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 1| 0|  
| 35722| 43777| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 1| 0| 24| 0| 29| 1| 30| 1|  
| 82095| 10326| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 1| 0|  
| 15447| 32818| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 2| 0|  
| 34193| 14920| 0| 0| 1| 0| 1| 0| 7| 0| 8| 0| 8| 0| 8| 0| 18| 0| 18| 0|  
| 85556| 95174| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 5| 0|  
| 16170| 2997| 0| 0| 0| 0| 1| 0| 9| 0| 19| 0| 19| 0| 19| 0| 19| 0| 19| 0|  
+------+---------+-----+-----+------+------+------+------+------+------+------+------+------+------+-------+-------+-------+-------+-------+-------+  
only showing top 20 rows  
  
用户评论表  
+------+---------+---+---+---+---+---+---+----+----+----+  
|userId|productId| s5|s10|s20|s40|s60|s80|s100|s200|s400|  
+------+---------+---+---+---+---+---+---+----+----+----+  
| 25917| 22655| 0| 0| 0| 0| 0| 0| 0| 0| 3|  
| 26054| 41706| 0| 0| 0| 0| 0| 0| 0| 3| 3|  
| 29473| 23700| 0| 0| 0| 0| 3| 3| 3| 3| 3|  
| 8039| 16266| 0| 0| 0| 0| 0| 0| 0| 3| 3|  
| 35728| 61085| 0| 0| 0| 0| 0| 0| 3| 9| 9|  
| 21048| 48038| 0| 0| 0| 0| 0| 0| 0| 3| 3|  
| 50661| 36811| 0| 0| 0| 0| 0| 0| 0| -4| -4|  
| 14264| 69167| 0| 0| 0| 3| 3| 3| 3| 3| 3|  
| 23083| 67000| 0| 0| 0| 0| 0| 0| 0| 0| 3|  
| 35722| 43777| 0| 0| 0| 0| 0| 3| 3| 3| 3|  
| 36998| 13684| 0| 0| 0| 0| 0| 0| 3| 3| 3|  
| 94480| 3434| 0| 0| 0| 0| 0| 0| 0| 0| 3|  
| 32199| 93883| 0| 0| 0| 3| 3| 3| 3| 3| 3|  
| 52501| 62178| 0| 0| 0| 0| 3| 6| 6| 6| 9|  
| 96789| 1462| 0| 0| 0| 0| 0| 0| 0| 3| 3|  
| 17181| 12944| 0| 0| 0| 3| 3| 3| 3| 3| 3|  
| 26177| 28038| 0| 0| 0| 0| 0| 0| 0| 0| 9|  
| 86916| 7292| 0| 0| 0| 0| 3| 3| 3| 9| 9|  
| 74593| 36891| 0| 0| 0| 0| 0| 0| 0| 3| 3|  
| 93686| 74385| 0| 0| 0| 0| 0| 0| 0| 0| 3|  
+------+---------+---+---+---+---+---+---+----+----+----+  
only showing top 20 rows  
  
订单表  
+------+---------+---+---+---+---+---+---+----+----+----+  
|userId|productId| o5|o10|o20|o40|o60|o80|o100|o200|o400|  
+------+---------+---+---+---+---+---+---+----+----+----+  
| 31953| 62178| 0| 0| 0| 0| 0| 0| 0| 0| 1|  
| 37773| 51347| 0| 0| 0| 1| 1| 1| 1| 3| 5|  
| 49508| 85938| 0| 0| 0| 0| 0| 0| 0| 1| 1|  
| 59527| 34025| 0| 0| 0| 0| 0| 0| 1| 1| 2|  
| 81506| 32652| 0| 0| 0| 0| 0| 0| 0| 0| 1|  
| 6237| 78567| 0| 0| 0| 1| 2| 2| 2| 4| 4|  
| 48684| 34011| 0| 0| 0| 0| 0| 0| 0| 2| 2|  
| 52281| 24753| 0| 0| 0| 0| 0| 0| 1| 1| 1|  
| 35722| 43777| 0| 0| 0| 0| 0| 0| 1| 1| 1|  
| 87998| 21812| 0| 0| 0| 0| 0| 0| 1| 1| 1|  
| 76183| 24753| 0| 0| 0| 0| 0| 0| 0| 1| 2|  
| 81310| 25543| 0| 0| 0| 0| 0| 0| 0| 0| 1|  
| 16170| 2997| 0| 0| 0| 0| 1| 1| 1| 1| 1|  
| 34193| 14920| 0| 0| 0| 1| 1| 1| 1| 2| 2|  
| 13919| 71891| 0| 0| 0| 0| 0| 0| 0| 1| 2|  
| 37290| 16704| 0| 0| 1| 2| 2| 2| 2| 2| 3|  
| 39093| 28828| 0| 2| 2| 2| 2| 2| 6| 6| 6|  
| 21345| 49591| 0| 0| 0| 0| 0| 0| 0| 0| 4|  
| 89286| 86243| 0| 0| 0| 1| 1| 1| 1| 1| 1|  
| 67503| 94092| 0| 0| 1| 1| 1| 1| 1| 1| 1|  
+------+---------+---+---+---+---+---+---+----+----+----+  
only showing top 20 rows

# 形成训练集

将数据进行聚合  
+------+---------+---+---+---+---+---+---+----+----+----+-----+-----+------+------+------+------+------+------+------+------+------+------+-------+-------+-------+-------+-------+-------+---+---+---+---+---+---+----+----+----+---+---+---------+------+----+------+-----+-----+-----+  
|userId|productId| o5|o10|o20|o40|o60|o80|o100|o200|o400|ua1\_5|ua2\_5|ua1\_10|ua2\_10|ua1\_20|ua2\_20|ua1\_40|ua2\_40|ua1\_60|ua2\_60|ua1\_80|ua2\_80|ua1\_100|ua2\_100|ua1\_200|ua2\_200|ua1\_400|ua2\_400| s5|s10|s20|s40|s60|s80|s100|s200|s400|age|sex|userGrade| price|cate| para1|para2|para3|label|  
+------+---------+---+---+---+---+---+---+----+----+----+-----+-----+------+------+------+------+------+------+------+------+------+------+-------+-------+-------+-------+-------+-------+---+---+---+---+---+---+----+----+----+---+---+---------+------+----+------+-----+-----+-----+  
| 9| 9563| 0| 0| 0| 0| 0| 0| 0| 2| 2| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 3| 0| 3| 0| 0| 0| 0| 0| 0| 0| 0| 3| 3| 3| 2| 3| 736.3| 0|296.21| 4.0| 2.0| 0|  
| 29| 58697| 0| 0| 1| 1| 1| 1| 1| 1| 1| 0| 0| 0| 0| 24| 0| 31| 0| 31| 0| 31| 0| 31| 0| 31| 0| 31| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 4| 2| 3| 299.7| 0|145.39| 6.0| 2.0| 0|  
| 84| 65512| 0| 0| 0| 0| 0| 0| 0| 2| 2| 0| 0| 0| 0| 0| 0| 0| 0| 1| 0| 1| 0| 1| 0| 2| 0| 2| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 3| 1| 3| 492.6| 0|277.33| 4.0| 5.0| 0|  
| 106| 7169| 0| 0| 0| 0| 0| 0| 0| 0| 2| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 6| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 3| 1| 4| 559.4| 4|156.49| -1.0| -1.0| 0|  
| 113| 69167| 0| 0| 0| 0| 0| 0| 0| 0| 1| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 3| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 3| 2| 4| 496.7| 4|223.91| -1.0| -1.0| 0|  
| 124| 27630| 0| 0| 0| 0| 0| 0| 0| 1| 1| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 25| 0| 25| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 4| 2| 3|1588.2| 3| 33.06| -1.0| -1.0| 0|  
| 135| 93713| 0| 0| 0| 0| 0| 0| 0| 0| 1| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 5| 0| 0| 0| 0| 0| 0| 0| 0| 0| 3| 4| 0| 5| 186.0| 1| 23.72| -1.0| -1.0| 0|  
| 182| 80368| 0| 0| 1| 1| 1| 1| 1| 1| 1| 0| 0| 0| 0| 2| 0| 2| 0| 2| 0| 2| 0| 2| 0| 2| 0| 2| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 2| 4|1059.7| 0|272.92| 3.0| 4.0| 0|  
| 223| 76134| 0| 0| 0| 0| 0| 0| 1| 1| 1| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 1| 0| 1| 0| 1| 0| 0| 0| 0| 0| 0| 0| 3| 3| 3| 3| 2| 4| 301.6| 2| 17.85| -1.0| -1.0| 0|  
| 249| 88621| 0| 3| 3| 3| 3| 3| 3| 10| 10| 1| 0| 5| 0| 8| 0| 11| 0| 12| 0| 13| 0| 20| 0| 48| 0| 80| 0| 0| 0| 0| 0| 0| 0| 3| 9| 9| 3| 0| 5| 915.3| 3|107.27| -1.0| -1.0| 0|  
| 327| 35468| 0| 0| 0| 0| 0| 0| 0| 0| 4| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 11| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 3| 1| 5| 130.9| 2| 17.27| -1.0| -1.0| 0|  
| 327| 69217| 0| 0| 3| 3| 3| 9| 9| 9| 9| 6| 0| 6| 0| 28| 0| 40| 0| 46| 0| 55| 0| 59| 0| 67| 0| 67| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 3| 1| 5|1031.5| 3| 86.96| -1.0| -1.0| 0|  
| 365| 81182| 0| 0| 0| 0| 1| 2| 2| 2| 2| 0| 0| 0| 0| 0| 0| 2| 0| 3| 0| 8| 0| 10| 0| 10| 0| 10| 0| 0| 0| 0| 0| 0| 3| 3| 3| 3| 3| 0| 2| 268.6| 0| 117.5| 3.0| 1.0| 0|  
| 401| 54613| 0| 0| 1| 1| 3| 3| 3| 6| 7| 0| 0| 0| 0| 2| 0| 2| 0| 11| 0| 11| 0| 11| 0| 18| 0| 24| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 3| 0| 5| 626.2| 4|139.73| -1.0| -1.0| 0|  
| 417| 86804| 0| 0| 0| 0| 0| 0| 0| 3| 3| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 4| 0| 4| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 4| 0| 5| 733.0| 5| 20.11| -1.0| -1.0| 0|  
| 474| 95559| 0| 0| 0| 1| 3| 3| 5| 5| 5| 0| 0| 0| 0| 5| 0| 13| 0| 28| 0| 29| 0| 44| 0| 53| 0| 54| 0| 0| 0| 3| 3| 9| 9| 9| 9| 9| 3| 2| 3| 625.1| 3| 71.62| -1.0| -1.0| 0|  
| 533| 15968| 0| 0| 0| 0| 0| 4| 4| 4| 4| 0| 0| 0| 0| 0| 0| 2| 0| 3| 0| 7| 0| 7| 0| 7| 0| 7| 0| 0| 0| 0| 0| 0| 3| 3| 3| 3| 4| 2| 3| 184.6| 5| 64.13| -1.0| -1.0| 0|  
| 545| 11149| 0| 0| 0| 0| 4| 4| 4| 4| 4| 0| 0| 0| 0| 0| 0| 0| 0| 2| 0| 2| 0| 2| 0| 2| 0| 2| 0| 0| 0| 0| 0| 3| 3| 3| 3| 3| 6| 2| 5| 314.6| 0|129.75| 3.0| 3.0| 0|  
| 559| 72187| 0| 0| 0| 0| 0| 0| 0| 3| 4| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 15| 1| 16| 1| 0| 0| 0| 0| 0| 0| 0| 3| 3| 4| 2| 4| 543.4| 4|253.88| -1.0| -1.0| 10|  
| 799| 26040| 0| 0| 0| 0| 1| 1| 1| 1| 1| 0| 0| 0| 0| 0| 0| 0| 0| 2| 0| 2| 0| 2| 0| 2| 0| 2| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 2| 2| 616.3| 4| 56.3| -1.0| -1.0| 0|  
+------+---------+---+---+---+---+---+---+----+----+----+-----+-----+------+------+------+------+------+------+------+------+------+------+-------+-------+-------+-------+-------+-------+---+---+---+---+---+---+----+----+----+---+---+---------+------+----+------+-----+-----+-----+  
  
进行one hot编码  
+------+---------+---+---+---+---+---+---+----+----+----+-----+-----+------+------+------+------+------+------+------+------+------+------+-------+-------+-------+-------+-------+-------+---+---+---+---+---+---+----+----+----+------+------+-----+-----+-----+-------------+-------------+-------------+-------------+  
|userId|productId|o5 |o10|o20|o40|o60|o80|o100|o200|o400|ua1\_5|ua2\_5|ua1\_10|ua2\_10|ua1\_20|ua2\_20|ua1\_40|ua2\_40|ua1\_60|ua2\_60|ua1\_80|ua2\_80|ua1\_100|ua2\_100|ua1\_200|ua2\_200|ua1\_400|ua2\_400|s5 |s10|s20|s40|s60|s80|s100|s200|s400|price |para1 |para2|para3|label|age\_vec |sex\_vec |userGrade\_vec|cate\_vec |  
+------+---------+---+---+---+---+---+---+----+----+----+-----+-----+------+------+------+------+------+------+------+------+------+------+-------+-------+-------+-------+-------+-------+---+---+---+---+---+---+----+----+----+------+------+-----+-----+-----+-------------+-------------+-------------+-------------+  
|4155 |92904 |0 |0 |0 |0 |0 |0 |0 |2 |2 |0 |0 |0 |0 |0 |0 |0 |0 |0 |0 |0 |0 |1 |0 |4 |0 |4 |0 |0 |0 |0 |0 |0 |0 |3 |3 |3 |716.5 |80.71 |3.0 |-1.0 |0 |(6,[4],[1.0])|(2,[],[]) |(5,[3],[1.0])|(5,[0],[1.0])|  
|11472 |24220 |0 |0 |0 |0 |0 |0 |0 |0 |1 |0 |0 |0 |0 |0 |0 |0 |0 |0 |0 |0 |0 |0 |0 |0 |0 |8 |0 |0 |0 |0 |0 |0 |0 |0 |0 |0 |197.7 |59.06 |-1.0 |-1.0 |0 |(6,[4],[1.0])|(2,[],[]) |(5,[3],[1.0])|(5,[4],[1.0])|  
|12417 |62813 |0 |0 |0 |0 |0 |0 |0 |0 |1 |0 |0 |0 |0 |0 |0 |0 |0 |0 |0 |0 |0 |0 |0 |0 |0 |8 |0 |0 |0 |0 |0 |0 |0 |0 |0 |3 |1539.6|94.45 |-1.0 |-1.0 |0 |(6,[3],[1.0])|(2,[0],[1.0])|(5,[],[]) |(5,[3],[1.0])|  
|13390 |74500 |0 |0 |0 |0 |0 |0 |0 |1 |1 |0 |0 |0 |0 |0 |0 |0 |0 |0 |0 |0 |0 |0 |0 |1 |0 |1 |0 |0 |0 |0 |0 |0 |0 |0 |0 |0 |58.5 |53.12 |-1.0 |-1.0 |0 |(6,[3],[1.0])|(2,[],[]) |(5,[3],[1.0])|(5,[3],[1.0])|  
|14136 |16266 |0 |0 |0 |0 |0 |0 |0 |1 |1 |0 |0 |0 |0 |0 |0 |0 |0 |0 |0 |0 |0 |0 |0 |5 |0 |5 |0 |0 |0 |0 |0 |0 |0 |0 |0 |0 |577.5 |152.73|3.0 |3.0 |0 |(6,[4],[1.0])|(2,[],[]) |(5,[4],[1.0])|(5,[0],[1.0])|  
|15790 |39033 |0 |0 |0 |0 |0 |1 |1 |1 |1 |0 |0 |0 |0 |0 |0 |0 |0 |0 |0 |0 |0 |0 |0 |0 |0 |0 |0 |0 |0 |0 |0 |0 |0 |0 |0 |0 |93.9 |10.72 |-1.0 |-1.0 |0 |(6,[3],[1.0])|(2,[1],[1.0])|(5,[4],[1.0])|(5,[2],[1.0])|  
|28599 |73753 |0 |0 |0 |0 |0 |0 |0 |1 |1 |0 |0 |0 |0 |0 |0 |0 |0 |0 |0 |0 |0 |0 |0 |1 |0 |1 |0 |0 |0 |0 |0 |0 |0 |0 |0 |0 |193.6 |95.14 |3.0 |2.0 |0 |(6,[0],[1.0])|(2,[],[]) |(5,[4],[1.0])|(5,[0],[1.0])|  
|31753 |95656 |0 |0 |0 |0 |0 |0 |0 |1 |1 |0 |0 |0 |0 |0 |0 |0 |0 |0 |0 |0 |0 |0 |0 |0 |0 |0 |0 |0 |0 |0 |0 |0 |0 |0 |0 |0 |285.2 |133.3 |2.0 |4.0 |0 |(6,[4],[1.0])|(2,[],[]) |(5,[3],[1.0])|(5,[0],[1.0])|  
|38060 |24343 |0 |0 |0 |0 |0 |1 |1 |1 |1 |0 |0 |0 |0 |0 |0 |0 |0 |0 |0 |3 |0 |3 |0 |3 |0 |3 |0 |0 |0 |0 |0 |0 |3 |3 |3 |3 |887.1 |426.68|4.0 |4.0 |0 |(6,[3],[1.0])|(2,[0],[1.0])|(5,[],[]) |(5,[0],[1.0])|  
|47847 |6815 |0 |0 |0 |0 |1 |1 |2 |3 |3 |0 |0 |0 |0 |0 |0 |0 |0 |2 |0 |2 |0 |3 |0 |7 |0 |7 |0 |0 |0 |0 |0 |3 |3 |3 |6 |6 |499.3 |242.31|3.0 |2.0 |0 |(6,[3],[1.0])|(2,[0],[1.0])|(5,[],[]) |(5,[0],[1.0])|  
|48111 |49685 |0 |0 |0 |1 |1 |1 |1 |1 |1 |0 |0 |0 |0 |0 |0 |2 |0 |2 |0 |2 |0 |2 |0 |2 |0 |2 |0 |0 |0 |0 |0 |0 |0 |0 |0 |0 |180.4 |4.08 |-1.0 |-1.0 |0 |(6,[3],[1.0])|(2,[],[]) |(5,[],[]) |(5,[2],[1.0])|  
|51199 |12018 |0 |0 |0 |0 |0 |0 |0 |1 |1 |0 |0 |0 |0 |0 |0 |0 |0 |0 |0 |0 |0 |0 |0 |2 |0 |2 |0 |0 |0 |0 |0 |0 |0 |0 |0 |0 |564.8 |250.05|4.0 |2.0 |0 |(6,[4],[1.0])|(2,[],[]) |(5,[3],[1.0])|(5,[0],[1.0])|  
|53031 |74963 |0 |0 |1 |1 |1 |1 |1 |1 |1 |0 |0 |0 |0 |5 |0 |5 |0 |5 |0 |5 |0 |5 |0 |5 |0 |5 |0 |0 |0 |0 |0 |0 |0 |0 |0 |0 |543.9 |128.77|3.0 |-1.0 |0 |(6,[4],[1.0])|(2,[],[]) |(5,[3],[1.0])|(5,[0],[1.0])|  
|55708 |78567 |0 |0 |0 |1 |2 |2 |2 |3 |5 |0 |0 |0 |0 |0 |0 |4 |0 |6 |0 |7 |0 |8 |0 |10 |0 |27 |0 |0 |0 |0 |0 |0 |0 |0 |0 |0 |955.6 |475.49|5.0 |4.0 |0 |(6,[0],[1.0])|(2,[1],[1.0])|(5,[],[]) |(5,[0],[1.0])|  
|72195 |16266 |0 |0 |0 |0 |0 |0 |0 |1 |1 |0 |0 |0 |0 |0 |0 |0 |0 |0 |0 |0 |0 |0 |0 |11 |0 |11 |0 |0 |0 |0 |0 |0 |0 |0 |0 |0 |577.5 |152.73|3.0 |3.0 |0 |(6,[4],[1.0])|(2,[],[]) |(5,[3],[1.0])|(5,[0],[1.0])|  
|80982 |82014 |0 |0 |0 |0 |0 |0 |0 |1 |1 |0 |0 |0 |0 |0 |0 |0 |0 |0 |0 |0 |0 |0 |0 |8 |0 |9 |0 |0 |0 |0 |0 |0 |0 |0 |0 |0 |594.4 |233.06|-1.0 |-1.0 |0 |(6,[0],[1.0])|(2,[],[]) |(5,[2],[1.0])|(5,[4],[1.0])|  
|82349 |24948 |0 |0 |0 |0 |0 |0 |0 |3 |4 |0 |0 |0 |0 |0 |0 |0 |0 |0 |0 |0 |0 |0 |0 |1 |0 |1 |0 |0 |0 |0 |0 |0 |0 |0 |0 |0 |93.4 |52.02 |-1.0 |-1.0 |0 |(6,[3],[1.0])|(2,[1],[1.0])|(5,[4],[1.0])|(5,[],[]) |  
|82936 |62371 |0 |0 |0 |0 |1 |1 |1 |1 |1 |0 |0 |0 |0 |0 |0 |0 |0 |8 |0 |8 |0 |8 |0 |8 |0 |8 |0 |0 |0 |0 |0 |3 |3 |3 |3 |3 |423.9 |148.17|3.0 |4.0 |0 |(6,[3],[1.0])|(2,[0],[1.0])|(5,[4],[1.0])|(5,[0],[1.0])|  
|84887 |19607 |0 |0 |3 |3 |3 |3 |3 |3 |3 |0 |0 |0 |0 |5 |0 |5 |0 |5 |0 |6 |0 |6 |0 |6 |0 |6 |0 |0 |0 |0 |0 |0 |0 |0 |0 |0 |396.9 |124.35|2.0 |4.0 |0 |(6,[3],[1.0])|(2,[0],[1.0])|(5,[3],[1.0])|(5,[0],[1.0])|  
|85287 |35468 |0 |0 |0 |0 |0 |0 |0 |1 |1 |0 |0 |0 |0 |0 |0 |0 |0 |0 |0 |0 |0 |0 |0 |0 |0 |0 |0 |0 |0 |0 |0 |0 |0 |0 |0 |0 |130.9 |17.27 |-1.0 |-1.0 |0 |(6,[0],[1.0])|(2,[],[]) |(5,[],[]) |(5,[2],[1.0])|  
+------+---------+---+---+---+---+---+---+----+----+----+-----+-----+------+------+------+------+------+------+------+------+------+------+-------+-------+-------+-------+-------+-------+---+---+---+---+---+---+----+----+----+------+------+-----+-----+-----+-------------+-------------+-------------+-------------+  
only showing top 20 rows  
  
特征矩阵化  
# 稀疏矩阵的表示方法，(维数,[位置],[值])  
+------+---------+-------------------------------------------------------------------------------------------------------------------------------------------------------------------------+-----+  
|userId|productId|features |label|  
+------+---------+-------------------------------------------------------------------------------------------------------------------------------------------------------------------------+-----+  
|4155 |92904 |(58,[7,8,21,23,25,33,34,35,36,37,38,39,44,51,53],[2.0,2.0,1.0,4.0,4.0,3.0,3.0,3.0,716.5,80.71,3.0,-1.0,1.0,1.0,1.0]) |0 |  
|11472 |24220 |(58,[8,25,36,37,38,39,44,51,57],[1.0,8.0,197.7,59.06,-1.0,-1.0,1.0,1.0,1.0]) |0 |  
|12417 |62813 |(58,[8,25,35,36,37,38,39,43,46,56],[1.0,8.0,3.0,1539.6,94.45,-1.0,-1.0,1.0,1.0,1.0]) |0 |  
|13390 |74500 |(58,[7,8,23,25,36,37,38,39,43,51,56],[1.0,1.0,1.0,1.0,58.5,53.12,-1.0,-1.0,1.0,1.0,1.0]) |0 |  
|14136 |16266 |(58,[7,8,23,25,36,37,38,39,44,52,53],[1.0,1.0,5.0,5.0,577.5,152.73,3.0,3.0,1.0,1.0,1.0]) |0 |  
|15790 |39033 |(58,[5,6,7,8,36,37,38,39,43,47,52,55],[1.0,1.0,1.0,1.0,93.9,10.72,-1.0,-1.0,1.0,1.0,1.0,1.0]) |0 |  
|28599 |73753 |(58,[7,8,23,25,36,37,38,39,40,52,53],[1.0,1.0,1.0,1.0,193.6,95.14,3.0,2.0,1.0,1.0,1.0]) |0 |  
|31753 |95656 |(58,[7,8,36,37,38,39,44,51,53],[1.0,1.0,285.2,133.3,2.0,4.0,1.0,1.0,1.0]) |0 |  
|38060 |24343 |(58,[5,6,7,8,19,21,23,25,32,33,34,35,36,37,38,39,43,46,53],[1.0,1.0,1.0,1.0,3.0,3.0,3.0,3.0,3.0,3.0,3.0,3.0,887.1,426.68,4.0,4.0,1.0,1.0,1.0]) |0 |  
|47847 |6815 |(58,[4,5,6,7,8,17,19,21,23,25,31,32,33,34,35,36,37,38,39,43,46,53],[1.0,1.0,2.0,3.0,3.0,2.0,2.0,3.0,7.0,7.0,3.0,3.0,3.0,6.0,6.0,499.3,242.31,3.0,2.0,1.0,1.0,1.0]) |0 |  
|48111 |49685 |(58,[3,4,5,6,7,8,15,17,19,21,23,25,36,37,38,39,43,55],[1.0,1.0,1.0,1.0,1.0,1.0,2.0,2.0,2.0,2.0,2.0,2.0,180.4,4.08,-1.0,-1.0,1.0,1.0]) |0 |  
|51199 |12018 |(58,[7,8,23,25,36,37,38,39,44,51,53],[1.0,1.0,2.0,2.0,564.8,250.05,4.0,2.0,1.0,1.0,1.0]) |0 |  
|53031 |74963 |(58,[2,3,4,5,6,7,8,13,15,17,19,21,23,25,36,37,38,39,44,51,53],[1.0,1.0,1.0,1.0,1.0,1.0,1.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0,543.9,128.77,3.0,-1.0,1.0,1.0,1.0]) |0 |  
|55708 |78567 |(58,[3,4,5,6,7,8,15,17,19,21,23,25,36,37,38,39,40,47,53],[1.0,2.0,2.0,2.0,3.0,5.0,4.0,6.0,7.0,8.0,10.0,27.0,955.6,475.49,5.0,4.0,1.0,1.0,1.0]) |0 |  
|72195 |16266 |(58,[7,8,23,25,36,37,38,39,44,51,53],[1.0,1.0,11.0,11.0,577.5,152.73,3.0,3.0,1.0,1.0,1.0]) |0 |  
|80982 |82014 |(58,[7,8,23,25,36,37,38,39,40,50,57],[1.0,1.0,8.0,9.0,594.4,233.06,-1.0,-1.0,1.0,1.0,1.0]) |0 |  
|82349 |24948 |(58,[7,8,23,25,36,37,38,39,43,47,52],[3.0,4.0,1.0,1.0,93.4,52.02,-1.0,-1.0,1.0,1.0,1.0]) |0 |  
|82936 |62371 |(58,[4,5,6,7,8,17,19,21,23,25,31,32,33,34,35,36,37,38,39,43,46,52,53],[1.0,1.0,1.0,1.0,1.0,8.0,8.0,8.0,8.0,8.0,3.0,3.0,3.0,3.0,3.0,423.9,148.17,3.0,4.0,1.0,1.0,1.0,1.0])|0 |  
|84887 |19607 |(58,[2,3,4,5,6,7,8,13,15,17,19,21,23,25,36,37,38,39,43,46,51,53],[3.0,3.0,3.0,3.0,3.0,3.0,3.0,5.0,5.0,5.0,6.0,6.0,6.0,6.0,396.9,124.35,2.0,4.0,1.0,1.0,1.0,1.0]) |0 |  
|85287 |35468 |(58,[7,8,36,37,38,39,40,55],[1.0,1.0,130.9,17.27,-1.0,-1.0,1.0,1.0]) |0 |  
+------+---------+-------------------------------------------------------------------------------------------------------------------------------------------------------------------------+-----+  
only showing top 20 rows  
  
归一化  
# 采用最大最小值归一化方法 计算方法为 x2= (x-MIN)/(MAX-MIN)  
+------+---------+-----+-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------+  
|userId|productId|label|features2 |  
+------+---------+-----+-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------+  
|4155 |92904 |0 |[0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.007380073800738007,0.003703703703703704,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0022172949002217295,0.0,0.006535947712418301,0.0,0.004119464469618949,0.0,0.4,0.4,0.3076923076923077,0.48,0.4444444444444444,0.41935483870967744,0.5714285714285714,0.4444444444444444,0.38095238095238093,0.05140180106665112,7.116613579036948E-4,0.5714285714285714,0.0,0.0,0.0,0.0,0.0,1.0,0.0,0.0,0.0,0.5,0.0,0.0,1.0,0.0,1.0,0.0,0.0,0.0,0.0] |  
|11472 |24220 |0 |[0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.008238928939237899,0.0,0.4,0.4,0.3076923076923077,0.48,0.4444444444444444,0.41935483870967744,0.5,0.3888888888888889,0.3333333333333333,0.013602716171713344,5.200023406302156E-4,0.0,0.0,0.0,0.0,0.0,0.0,1.0,0.0,0.0,0.0,0.5,0.0,0.0,1.0,0.0,0.0,0.0,0.0,0.0,1.0] |  
|12417 |62813 |0 |[0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.008238928939237899,0.0,0.4,0.4,0.3076923076923077,0.48,0.4444444444444444,0.41935483870967744,0.5,0.3888888888888889,0.38095238095238093,0.11137178328913239,8.332962261409976E-4,0.0,0.0,0.0,0.0,0.0,1.0,0.0,0.0,1.0,0.0,0.5,0.0,0.0,0.0,0.0,0.0,0.0,0.0,1.0,0.0] |  
|13390 |74500 |0 |[0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0036900369003690036,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0016339869281045752,0.0,0.0010298661174047373,0.0,0.4,0.4,0.3076923076923077,0.48,0.4444444444444444,0.41935483870967744,0.5,0.3888888888888889,0.3333333333333333,0.0034607874566490835,4.674178342743511E-4,0.0,0.0,0.0,0.0,0.0,1.0,0.0,0.0,0.0,0.0,0.5,0.0,0.0,1.0,0.0,0.0,0.0,0.0,1.0,0.0] |  
|14136 |16266 |0 |[0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0036900369003690036,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.008169934640522876,0.0,0.005149330587023687,0.0,0.4,0.4,0.3076923076923077,0.48,0.4444444444444444,0.41935483870967744,0.5,0.3888888888888889,0.3333333333333333,0.04127444408824644,0.0013492263659423077,0.5714285714285714,0.5714285714285714,0.0,0.0,0.0,0.0,1.0,0.0,0.0,0.0,0.5,0.0,0.0,0.0,1.0,1.0,0.0,0.0,0.0,0.0] |  
|15790 |39033 |0 |[0.0,0.0,0.0,0.0,0.0,0.007518796992481203,0.005,0.0036900369003690036,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.4,0.4,0.3076923076923077,0.48,0.4444444444444444,0.41935483870967744,0.5,0.3888888888888889,0.3333333333333333,0.006039984845393874,9.206714917525098E-5,0.0,0.0,0.0,0.0,0.0,1.0,0.0,0.0,0.0,1.0,0.5,0.0,0.0,0.0,1.0,0.0,0.0,1.0,0.0,0.0] |  
|28599 |73753 |0 |[0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0036900369003690036,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0016339869281045752,0.0,0.0010298661174047373,0.0,0.4,0.4,0.3076923076923077,0.48,0.4444444444444444,0.41935483870967744,0.5,0.3888888888888889,0.3333333333333333,0.013303995570192054,8.394045273843556E-4,0.5714285714285714,0.42857142857142855,1.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.5,0.0,0.0,0.0,1.0,1.0,0.0,0.0,0.0,0.0] |  
|31753 |95656 |0 |[0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0036900369003690036,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.4,0.4,0.3076923076923077,0.48,0.4444444444444444,0.41935483870967744,0.5,0.3888888888888889,0.3333333333333333,0.019977850960277443,0.0011772201439735458,0.42857142857142855,0.7142857142857143,0.0,0.0,0.0,0.0,1.0,0.0,0.0,0.0,0.5,0.0,0.0,1.0,0.0,1.0,0.0,0.0,0.0,0.0] |  
|38060 |24343 |0 |[0.0,0.0,0.0,0.0,0.0,0.007518796992481203,0.005,0.0036900369003690036,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0066518847006651885,0.0,0.0066518847006651885,0.0,0.004901960784313725,0.0,0.003089598352214212,0.0,0.4,0.4,0.3076923076923077,0.48,0.4444444444444444,0.5161290322580645,0.5714285714285714,0.4444444444444444,0.38095238095238093,0.06383149243726867,0.003774399011765385,0.7142857142857143,0.7142857142857143,0.0,0.0,0.0,1.0,0.0,0.0,1.0,0.0,0.5,0.0,0.0,0.0,0.0,1.0,0.0,0.0,0.0,0.0] |  
|47847 |6815 |0 |[0.0,0.0,0.0,0.0,0.008928571428571428,0.007518796992481203,0.01,0.01107011070110701,0.007407407407407408,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.004434589800443459,0.0,0.004434589800443459,0.0,0.0066518847006651885,0.0,0.011437908496732025,0.0,0.007209062821833162,0.0,0.4,0.4,0.3076923076923077,0.48,0.5555555555555556,0.5161290322580645,0.5714285714285714,0.5,0.42857142857142855,0.035576895054352575,0.002142243214319133,0.5714285714285714,0.42857142857142855,0.0,0.0,0.0,1.0,0.0,0.0,1.0,0.0,0.5,0.0,0.0,0.0,0.0,1.0,0.0,0.0,0.0,0.0] |  
|48111 |49685 |0 |[0.0,0.0,0.0,0.008928571428571428,0.008928571428571428,0.007518796992481203,0.005,0.0036900369003690036,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.005649717514124294,0.0,0.004434589800443459,0.0,0.004434589800443459,0.0,0.004434589800443459,0.0,0.0032679738562091504,0.0,0.0020597322348094747,0.0,0.4,0.4,0.3076923076923077,0.48,0.4444444444444444,0.41935483870967744,0.5,0.3888888888888889,0.3333333333333333,0.0123422609506601,3.3285815471052274E-5,0.0,0.0,0.0,0.0,0.0,1.0,0.0,0.0,0.0,0.0,0.5,0.0,0.0,0.0,0.0,0.0,0.0,1.0,0.0,0.0] |  
|51199 |12018 |0 |[0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0036900369003690036,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0032679738562091504,0.0,0.0020597322348094747,0.0,0.4,0.4,0.3076923076923077,0.48,0.4444444444444444,0.41935483870967744,0.5,0.3888888888888889,0.3333333333333333,0.040349138810363415,0.002210762419570714,0.7142857142857143,0.42857142857142855,0.0,0.0,0.0,0.0,1.0,0.0,0.0,0.0,0.5,0.0,0.0,1.0,0.0,1.0,0.0,0.0,0.0,0.0] |  
|53031 |74963 |0 |[0.0,0.0,0.03571428571428571,0.008928571428571428,0.008928571428571428,0.007518796992481203,0.005,0.0036900369003690036,0.0,0.0,0.0,0.0,0.0,0.02403846153846154,0.0,0.014124293785310734,0.0,0.011086474501108648,0.0,0.011086474501108648,0.0,0.011086474501108648,0.0,0.008169934640522876,0.0,0.005149330587023687,0.0,0.4,0.4,0.3076923076923077,0.48,0.4444444444444444,0.41935483870967744,0.5,0.3888888888888889,0.3333333333333333,0.03882639232943782,0.0011371178184193258,0.5714285714285714,0.0,0.0,0.0,0.0,0.0,1.0,0.0,0.0,0.0,0.5,0.0,0.0,1.0,0.0,1.0,0.0,0.0,0.0,0.0] |  
|55708 |78567 |0 |[0.0,0.0,0.0,0.008928571428571428,0.017857142857142856,0.015037593984962405,0.01,0.01107011070110701,0.014814814814814815,0.0,0.0,0.0,0.0,0.0,0.0,0.011299435028248588,0.0,0.013303769401330377,0.0,0.015521064301552107,0.0,0.017738359201773836,0.0,0.016339869281045753,0.0,0.027806385169927908,0.0,0.4,0.4,0.3076923076923077,0.48,0.4444444444444444,0.41935483870967744,0.5,0.3888888888888889,0.3333333333333333,0.06882231224317314,0.004206494930154231,0.8571428571428571,0.7142857142857143,1.0,0.0,0.0,0.0,0.0,0.0,0.0,1.0,0.5,0.0,0.0,0.0,0.0,1.0,0.0,0.0,0.0,0.0] |  
|72195 |16266 |0 |[0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0036900369003690036,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.017973856209150325,0.0,0.01132852729145211,0.0,0.4,0.4,0.3076923076923077,0.48,0.4444444444444444,0.41935483870967744,0.5,0.3888888888888889,0.3333333333333333,0.04127444408824644,0.0013492263659423077,0.5714285714285714,0.5714285714285714,0.0,0.0,0.0,0.0,1.0,0.0,0.0,0.0,0.5,0.0,0.0,1.0,0.0,1.0,0.0,0.0,0.0,0.0] |  
|80982 |82014 |0 |[0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0036900369003690036,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.013071895424836602,0.0,0.009268795056642637,0.0,0.4,0.4,0.3076923076923077,0.48,0.4444444444444444,0.41935483870967744,0.5,0.3888888888888889,0.3333333333333333,0.04250575583598053,0.0020603565672161453,0.0,0.0,1.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.5,0.0,1.0,0.0,0.0,0.0,0.0,0.0,0.0,1.0] |  
|82349 |24948 |0 |[0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.01107011070110701,0.011111111111111112,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0016339869281045752,0.0,0.0010298661174047373,0.0,0.4,0.4,0.3076923076923077,0.48,0.4444444444444444,0.41935483870967744,0.5,0.3888888888888889,0.3333333333333333,0.006003555503744936,4.576799627269688E-4,0.0,0.0,0.0,0.0,0.0,1.0,0.0,0.0,0.0,1.0,0.5,0.0,0.0,0.0,1.0,0.0,0.0,0.0,0.0,0.0] |  
|82936 |62371 |0 |[0.0,0.0,0.0,0.0,0.008928571428571428,0.007518796992481203,0.005,0.0036900369003690036,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.017738359201773836,0.0,0.017738359201773836,0.0,0.017738359201773836,0.0,0.013071895424836602,0.0,0.008238928939237899,0.0,0.4,0.4,0.3076923076923077,0.48,0.5555555555555556,0.5161290322580645,0.5714285714285714,0.4444444444444444,0.38095238095238093,0.030083350333692766,0.0013088584620731592,0.5714285714285714,0.7142857142857143,0.0,0.0,0.0,1.0,0.0,0.0,1.0,0.0,0.5,0.0,0.0,0.0,1.0,1.0,0.0,0.0,0.0,0.0] |  
|84887 |19607 |0 |[0.0,0.0,0.10714285714285714,0.026785714285714284,0.026785714285714284,0.022556390977443608,0.015,0.01107011070110701,0.007407407407407408,0.0,0.0,0.0,0.0,0.02403846153846154,0.0,0.014124293785310734,0.0,0.011086474501108648,0.0,0.013303769401330377,0.0,0.013303769401330377,0.0,0.00980392156862745,0.0,0.006179196704428424,0.0,0.4,0.4,0.3076923076923077,0.48,0.4444444444444444,0.41935483870967744,0.5,0.3888888888888889,0.3333333333333333,0.02811616588465013,0.001097989280019844,0.42857142857142855,0.7142857142857143,0.0,0.0,0.0,1.0,0.0,0.0,1.0,0.0,0.5,0.0,0.0,1.0,0.0,1.0,0.0,0.0,0.0,0.0]|  
|85287 |35468 |0 |[0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0036900369003690036,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.4,0.4,0.3076923076923077,0.48,0.4444444444444444,0.41935483870967744,0.5,0.3888888888888889,0.3333333333333333,0.008735756127415265,1.5005174793466383E-4,0.0,0.0,1.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.5,0.0,0.0,0.0,0.0,0.0,0.0,1.0,0.0,0.0] |  
+------+---------+-----+-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------+  
only showing top 20 rows

# 进行训练

1. 将正负样本进行聚合

val minus = train\_1.filter($"label" === 0).sample(false, 0.01, 4)  
 val plus = train\_1.filter($"label" > 0)  
 val temp = minus  
 .union(plus)

1. 进行训练

//拆分测试训练集  
val Array(train,test) =train\_3.randomSplit(Array(0.8,0.2))  
 //参数设定  
 val(maxDepth , numRound , nworker )=(7,7,7)  
 val paramMap = List(  
 "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" -> 32  
 ).toMap  
  
//模型训练  
 val model:XGBoostModel = XGBoost.trainWithDataFrame(train, paramMap, numRound, nworker,  
 useExternalMemory = true,  
 featureCol = "features2",  
 labelCol = "label",  
 missing = 0.0f)  
//结果预测  
 val reslut = model.transform(test)

1. 训练结果

是否购买的准确度

0.8457095072687508

预测日期的准确度：

不太理想，特征的提取上还需要改进

# 结果总结

有以下几点需要改进：

1. 需要增加时间差的特征来对时间间隔进行更好的预测
2. 可以使用多模型的方式进行预测，一个模型用于预测是否购买，一个模型预测购买行为的具体时间