**甜橙金融数据集用户行为相关特征分析**

本文主要是对2018年甜橙金融杯大数据建模大赛的数据集进行用户行为分析，通过对用户操作和交易两种行为依据时间序列进行综合分析，提取用户行为的特征来进行风险预测，同时由于操作和交易类型值是脱敏值，故根据分析对部分风险值进行猜测，最后依据实验进行特征分析验证。

## 1、操作类型编号

操作类型（89种），（op\_type, count）

chroot@ubuntu:~/test data/tiancheng\_data$ cat -b step\_1/op\_type\_count/\*

c8741ce15ceac2a4|001

d25caee90b27fa9b|002

79f8b86938180c3c|003

b501fa4fc58206b9|004

acfaded7e04e7ba0|005

072eee5c88d380df|006

3b02bcd4b685d7a8|007

00094ae2a1d62504|008

963bd8a75ff9ab37|009

20a91b45ef8f8221|010

操作类型由三位十进制数进行编号，第1位“0”表示操作，后两位为操作类型代号，从001到089。

由甜橙金融的翼支付app可知，基本的用户操作包括：余额相关操作、账号相关操作、银行卡操作、代金券相关操作、账单操作、红包金操作、白条操作、借钱操作、信用操作、保险操作、好友分享操作。

余额相关操作有：查询可用余额、充值、提现、转账。账号相关操作有：登录、注销、修改个人信息（头像、昵称、身份信息、手机号、邮箱）、提升账户星级、支付设置、密码管理（修改登陆密码、修改支付密码、忘记支付密码、重置密保问题）、账户安全设置、通用设置（音效、通知、缓存清理）。

由这些操作可推测，比较频繁的操作是登录、查询余额、充值、提现、转账，提现、转账一般属于风险操作，代金券、红包金操作一般是“羊毛党”常用操作。

## 2、交易类型编号

交易类型（15种）,(trans type, count)

chroot@ubuntu: ~/test data/tiancheng\_data$ cat step\_1/trans\_type\_count/\* -b

c2f2023d279665b2|101

6d55c54c8b1056fb|102

61bfb66c928f36ac|103

26bcf43a19df14c8|104

e0d7b8768da99dd4|105

d9c417304a5ae70c|106

9d7dd7b80e806024|107

a19e7a8951e54c06|108

ced62357ad496957|109

eb8d10591677bbe1|110

4adc3de71fe1a83c|111

85bced5214d33ad2|112

e903cf2a79b83d37|113

fd4d2d1006a95637|114

3f469aa3836e71cb|115

交易类型由三位十进制数进行编号，第1位“1”表示交易，后两位为交易类型代号，从101到115

由甜橙金融的翼支付app可知，基本的用户交易包括：支付操作、订单相关操作、退款操作。由此推测，退款操作一般为风险操作。

## 3、用户行为整体分析

用户操作流程（单天） (uid, day, action\_seq, time\_seq)

10000|13|001-031|14:20:56-14:22:59

10000|26|001-101-002-101-002|12:23:48-12:23:56-12:24:01-12:24:17-12:24:21

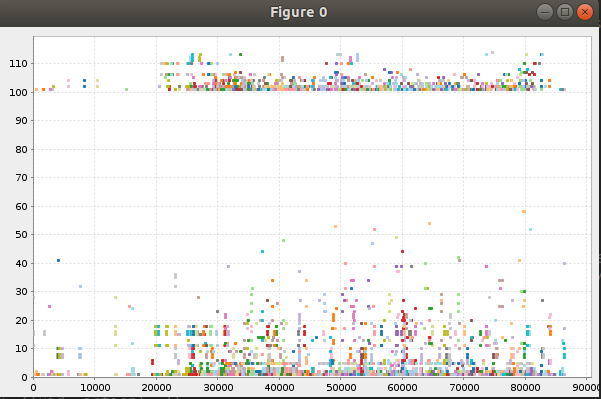
10001|2|001-001-101-101-003-003-002-003-001-001-003-001|08:30:52-08:30:54-08:31:13-08:31:13-08:31:16-08:31:17-08:31:23-08:31:24-08:31:35-18:06:49-18:09:46-18:11:19

10001|3|001-001-001-003-003-002-102-001-001|07:03:58-07:04:02-07:11:07-07:18:05-07:18:06-07:18:16-07:18:17-07:20:09-07:23:09)

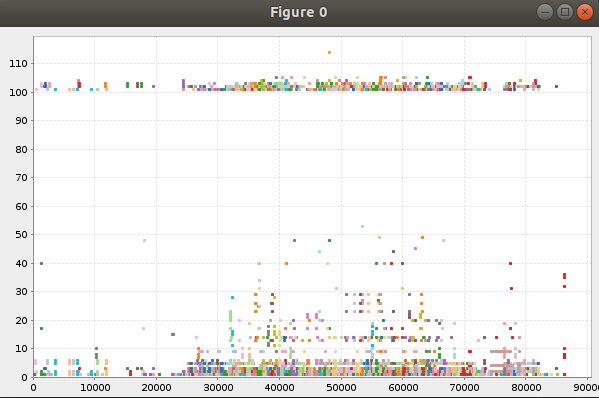
10001|4|001-001-001-003-002-102|22:28:08-22:28:16-22:28:46-22:30:38-22:30:50-22:30:51

将所有用户操作流程、风险用户操作流程、非风险用户操作流程展示在分布图中，如下，横坐标为单天时间（单位s），纵坐标为用户行为类型编号（操作类型和交易类型）。

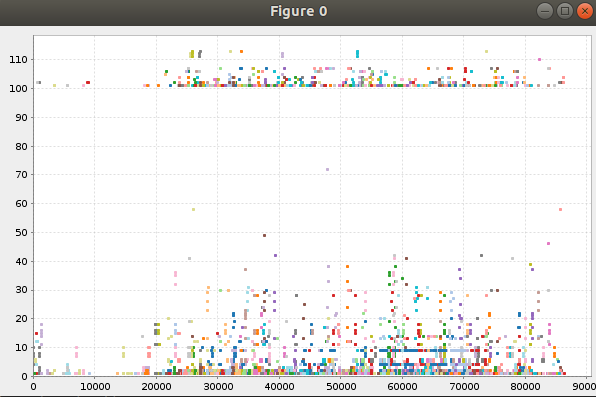
所有用户操作流程分布：



风险用户操作流程分布：



非风险用户操作流程分布：



由分布图知，操作类型集中在050之前，风险用户的交易类型集中在105之前，因此，重点分析001-050的操作类型和101-105的交易类型。

## 4、用户相关行为特征分析

分析用户行为序列，如001-001-101-101-003-003-002-003-001-001-003-001，提取词序相关联的行为对001-101，101-003，003-002，002-003，003-001，001-003，003-001，统计某一行为的后续行为出现次数，转换为比率，提取出现比率top3的行为，作为该行为的相关行为列表。

风险用户行为相关系数：

001,[(101,0.25),(002,0.17),(003,0.10)]

002,[(001,0.19),(102,0.16),(103,0.12)]

003,[(002,0.31),(102,0.30),(101,0.13)]

004,[(009,0.45),(001,0.15),(105,0.14)]

005,[(006,0.30),(101,0.14),(003,0.09)]

006,[(005,0.45),(001,0.11),(013,0.09)]

007,[(008,0.38),(019,0.25),(010,0.24)]

008,[(007,0.35),(019,0.32),(010,0.22)]

009,[(002,0.80),(004,0.08),(003,0.07)]

010,[(007,0.43),(008,0.38),(019,0.11)]

非风险用户行为相关系数：

001,[(101,0.29),(002,0.14),(003,0.07)]

002,[(001,0.35),(101,0.20),(003,0.11)]

003,[(002,0.47),(001,0.26),(101,0.11)]

004,[(001,0.35),(009,0.24),(101,0.08)]

005,[(006,0.41),(001,0.17),(013,0.09)]

006,[(005,0.44),(001,0.15),(017,0.09)]

007,[(008,0.32),(010,0.20),(001,0.16)]

008,[(007,0.32),(010,0.20),(001,0.19)]

009,[(002,0.79),(004,0.10),(105,0.05)]

010,[(007,0.37),(008,0.35),(001,0.13)]

取行为相关系数最高的一列（第一列），将风险用户行为相关矩阵与非风险用户行为相关矩阵进行对比，取差异大的行为即为风险行为，作为重点监控行为以及区分风险用户和非风险用户的特征。操作类型集中在001-050，风险用户行为集中在101-105，故对这个段内的行为类型进行差异比较。

004|[(009,0.45),(001,0.15),(105,0.14)] │004|[(001,0.35),(009,0.24),(101,0.08)]

017|[(001,0.30),(040,0.22),(048,0.19)] │017|[(021,0.53),(038,0.17),(001,0.06)]

018|[(001,0.43),(011,0.35),(039,0.06)] │018|[(011,0.45),(001,0.36),(039,0.04)]

029|[(023,0.66),(002,0.10),(001,0.10)] │029|[(001,0.48),(023,0.39),(101,0.05)]

032|[(036,0.23),(007,0.18),(001,0.18)] │032|[(007,0.20),(035,0.19),(036,0.19)]

036|[(007,0.56),(035,0.19),(032,0.13)] │036|[(035,0.26),(007,0.26),(032,0.26)]

043|[(041,0.29),(007,0.14),(047,0.14)] │043|[(001,0.38),(007,0.13),(010,0.08)]

047|[(001,0.50),(050,0.50)] │047|[(050,0.60),(001,0.26),(010,0.03)]

048|[(006,0.53),(078,0.18),(002,0.18)] │048|[(017,0.44),(006,0.22),(005,0.22)]

049|[(014,0.39),(001,0.30),(003,0.22)] │049|[(001,0.43),(002,0.23),(101,0.08)]

102|[(103,0.38),(001,0.25),(003,0.10)] │102|[(104,0.33),(001,0.29),(101,0.16)]

103|[(001,0.32),(102,0.28),(101,0.16)] │103|[(101,0.52),(001,0.23),(102,0.04)]

进一步分析差异数据，对于004操作，风险用户的相关操作比较多的是009和001，非风险用户的相关操作比较多的是001和009，而且系数差异不大，故特征不够明显。对于017操作，风险用户的相关操作比较多的是001和040，非风险用户的相关操作比较多的是021和038、001，其中001系数比较小，故特征相对明显，可将017作为特征操作。同样的分析可得到043,049,102的差异类型，而由操作分布图可知，043、049处操作分布较少，故不适合作为特征。由结果知，风险用户交易类型集中在101-105，故将106-115的交易操作也可作为特征，如果出现则大概率是非风险用户。

由上面分析知，017、102、106-115适合作为风险特征去分析用户行为。具体是017操作后是多接001或040还是021或038， 102后多接103还是104，以及106-115类的交易操作频次。

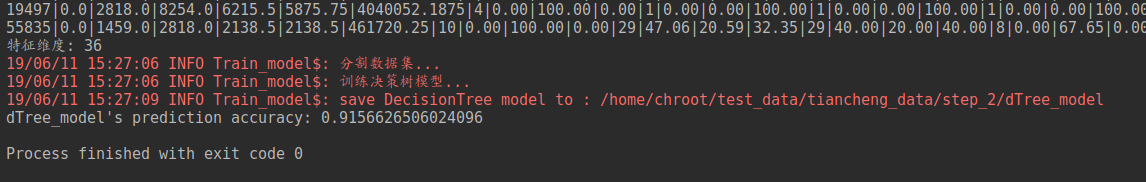
## 5、行为类型猜测

由以上分析，最频繁的操作行为应该是查询余额，故001应该属于查询余额操作。最频繁的交易行为是订单操作，故101应该属于订单操作。017号行为属于特征操作，其后的风险操作040可以推测属于提现操作，043、049等特征不够明显操作估计属于代金券、红包金等相关操作。102属于特征交易行为，其后的风险交易行为103估计属于退款操作，102则属于支付操作。

## 6、特征实验结果

由以上分析，提取对应的三个用户行为特征，使用决策树模型进行训练，与之前的训练结果比对，发现预测准确率略微提升（91.57 -> 92.62）。

提取用户相关行为特征前的结果：



提取用户相关行为特征后的结果：

