****

课 程：2018软件工程综合实训

项 目： 数据挖掘实训项目

院 系：数据科学与计算机学院

专 业： 软件工程 \_

学生姓名： 郑品

学 号： 15331422

授课教师： 郑子彬 \_

2018 年 6 月 20 日

**目 录**

[一、 实验内容 2](#_Toc517253831)

[二、 实验主要思想与方法 2](#_Toc517253832)

[2.1数据集训练与可视化 3](#_Toc517253833)

[2.1.1用户通话记录 3](#_Toc517253834)

[2.1.2用户短信记录 8](#_Toc517253835)

[2.1.3用户网站/App访问记录。 11](#_Toc517253836)

[2.2特征提取 12](#_Toc517253837)

[2.3LightGBM 17](#_Toc517253838)

[三、 实验结果 18](#_Toc517253839)

[四、心得体会 18](#_Toc517253840)

# 实验内容

参加“基于移动网络通讯行为的用户风险识别”大赛，本次大赛以模拟的语音通话、短信收发、网站及App访问记录等移动网络使用行为为基础，参赛队伍需要通过数据挖掘技术和机器学习算法，构建识别风险用户的预测模型，判别用户属于风险用户的可能性。从而为各行业提供风控保障，助力新时代大数字生态的健康有序发展。

# 实验主要思想与方法

本实验主要根据用户的通话记录、短信记录、网站和APP访问记录的信息，对用户进行分类和预测。

在实验方法介绍分为三部分，第一部分为数据集训练与可视化，第二部分介绍特征提取的主要策略，在第三部分将介绍此次实验使用的LightGBM模型。

## 2.1数据集训练与可视化

### 2.1.1用户通话记录



图1 用户通话记录数据。

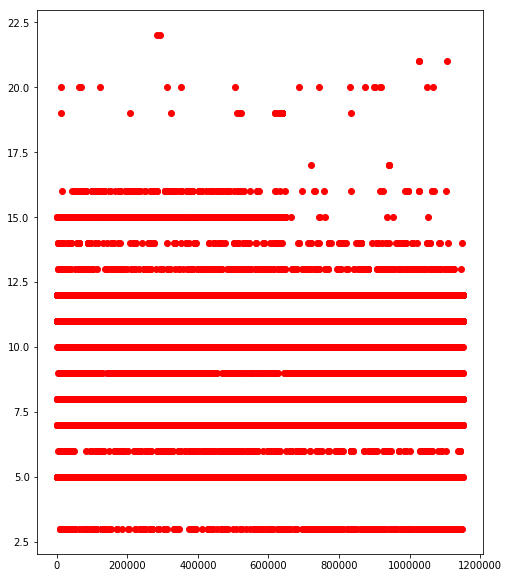


图2 通话号码长度分布。

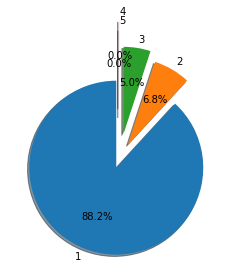


图3 不同通话类型所占比例。

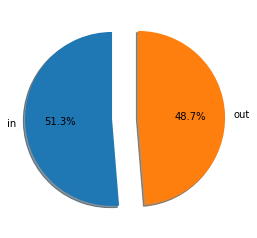


图4 打入和打出电话所占比例。

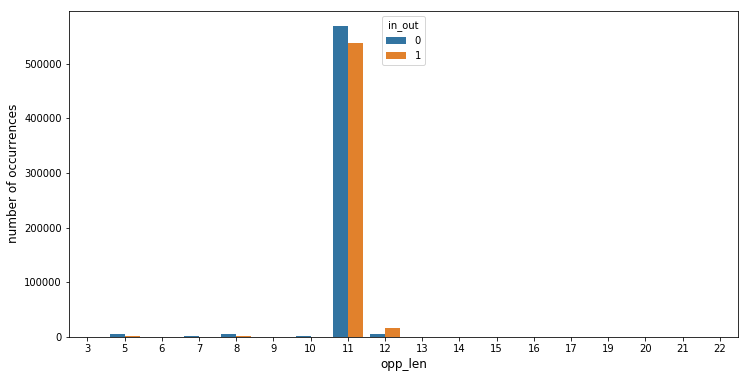


图5 打入和打出电话不同长度出现次数。

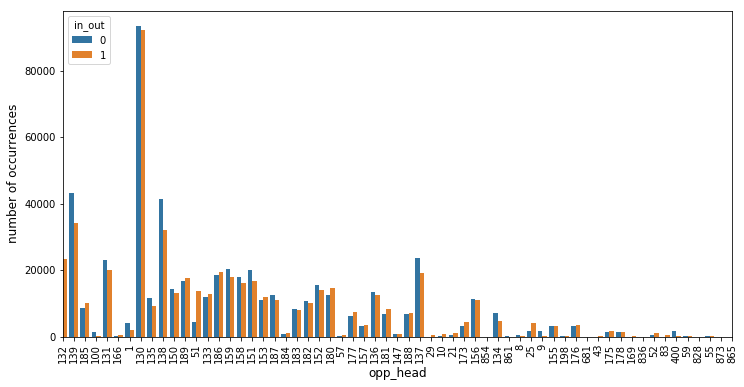


图6 打入和打出电话号码不同头部出现次数。

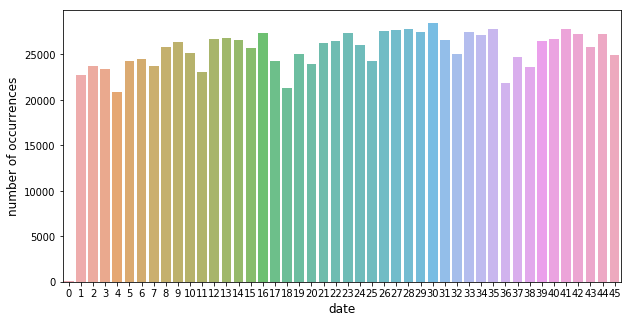


图7 每日活跃次数。

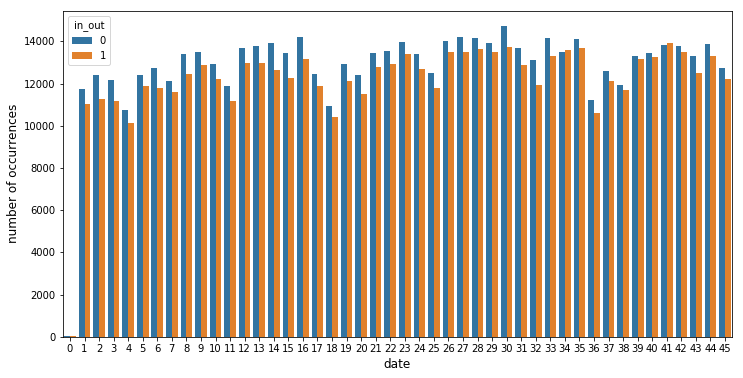


图8 打入和打出情况下每天的活跃次数。

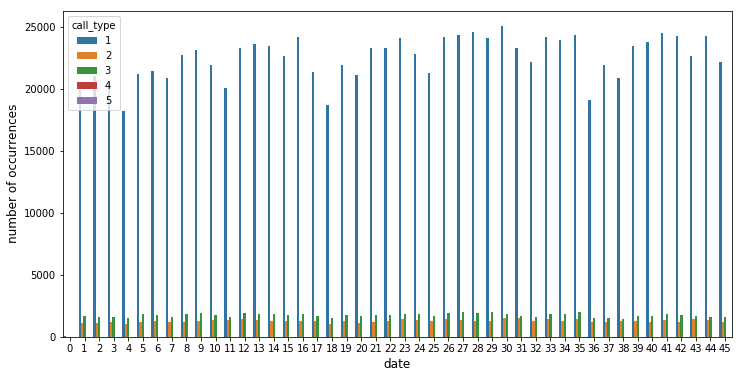


图9 5种不同类型下每天的活跃次数。

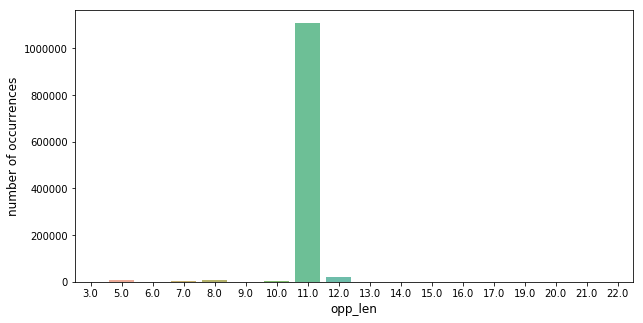


图10 通话号码不同长度出现次数。

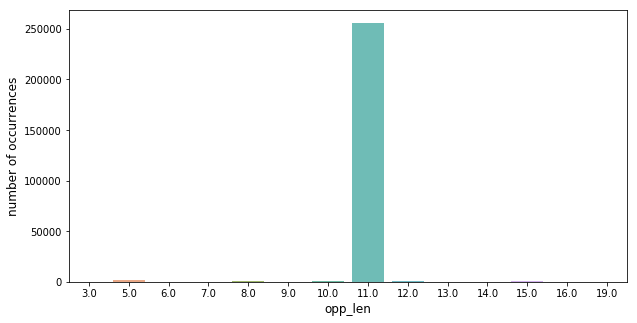


图11 已给风险用户的通话号码长度出现次数。

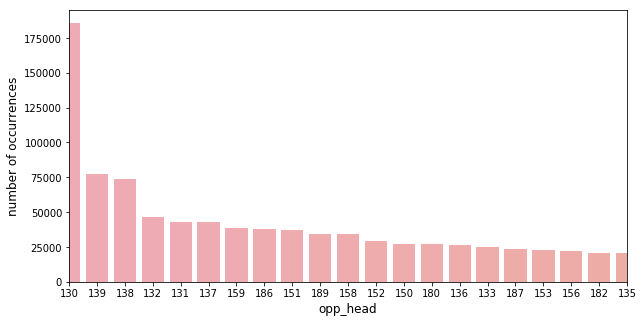


图12 通话号码头部出现次数（仅截取主要部分）。

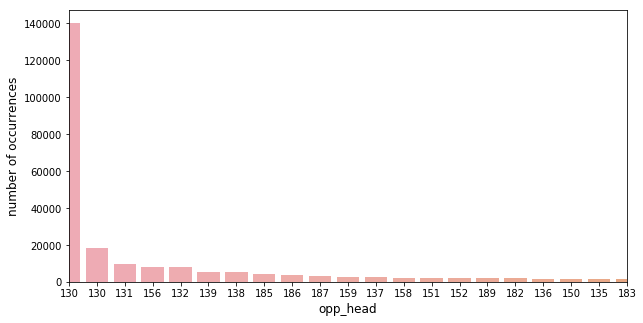


图13 已给风险用户通话号码头部出现次数（仅截取主要部分）。

### 2.1.2用户短信记录



图14 用户短信记录数据。

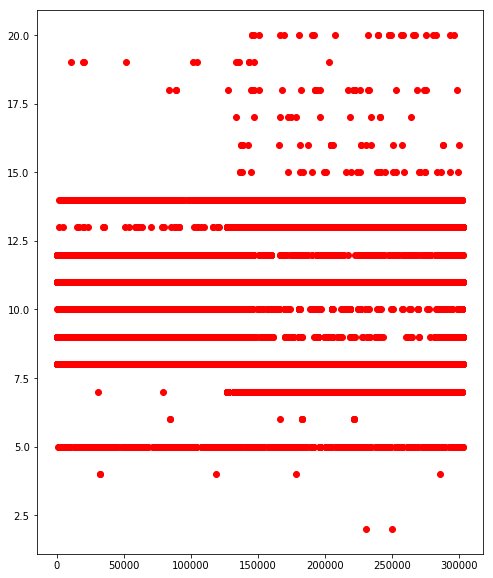


图15 短信号码长度分布。

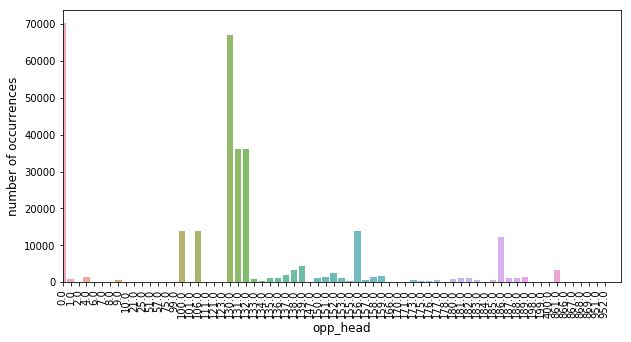


图16 短信号码不同头部出现次数。

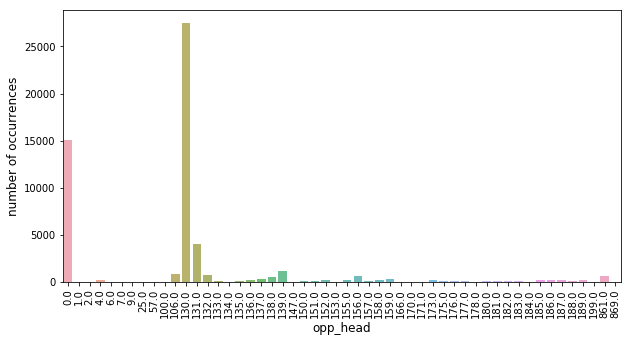


图17 已给风险用户的短信号码头部出现次数。

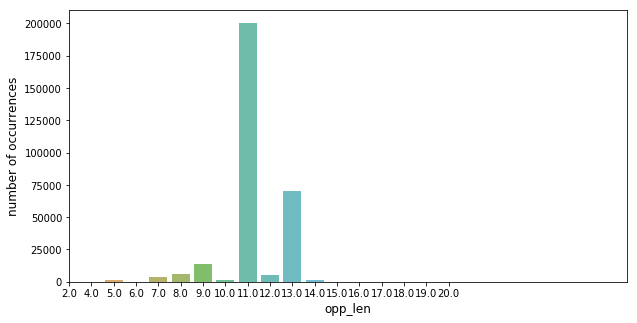


图18 短信号码不同长度出现次数。

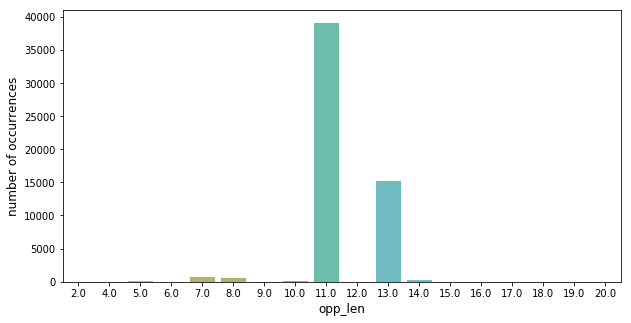


图19 已给风险用户的短信号码长度出现次数。

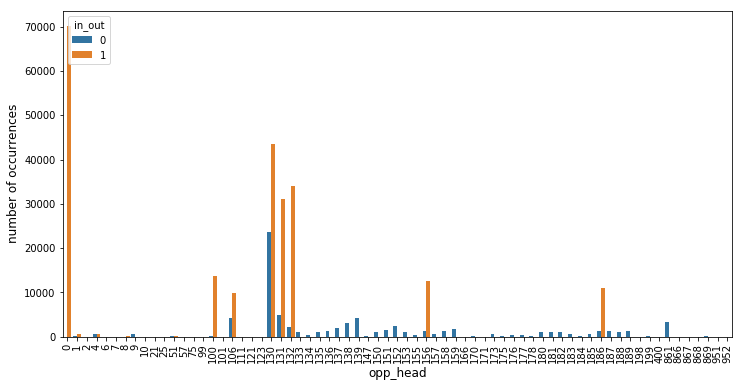


图20 接收/发送短信情况下不同短信头部出现次数。

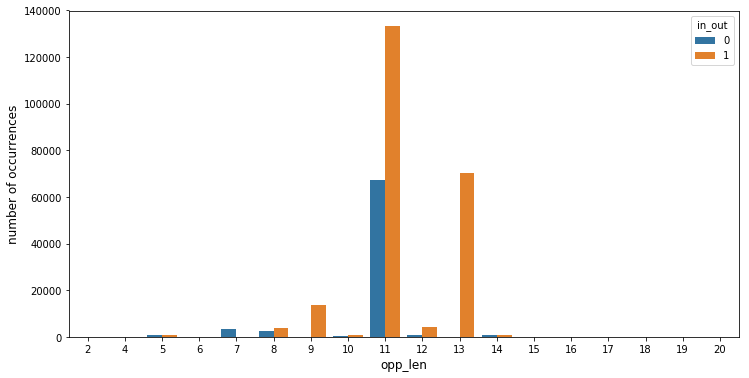


图21 接收/发送短信情况下不同短信长度出现次数。

### 2.1.3用户网站/App访问记录。



图22 用户网站/App访问记录数据。

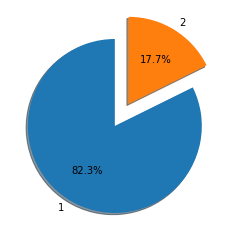


图23 用户访问网站/APP分别占的比例

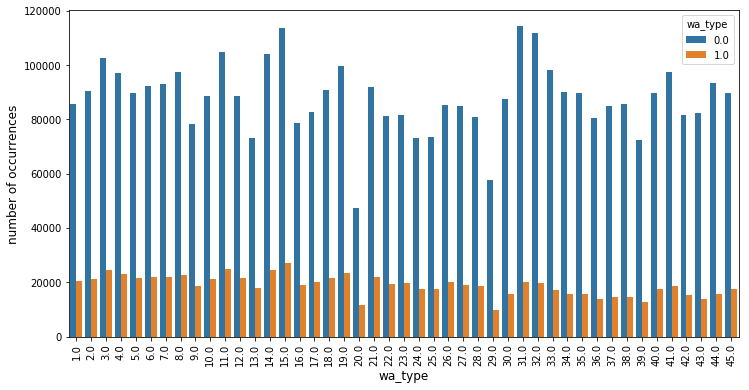


图24 两种类型每日活跃次数。

## 2.2特征提取

经过对可视化结果的分析，对特征进行如下处理：

1. 通话记录：
2. voice\_time:将end\_time、start\_time转为时间格式并计算通话时长(end\_time-start\_time)；
3. voice\_date:取start\_time的前两位
4. 短信记录：sms\_date:取start\_time的前两位。
5. 对已给特征使用常规的提取方法，如取:sum、mean、min、max、median等。
6. 将提取的特征与in\_out、call\_type、wa\_type等进行结合。
7. 多变量进行结合，如：

voice\_feature[‘voice\_cnt\_peruniquecall’]= voice\_feature[‘voice\_opp\_num\_unique’]/voice\_feature[‘voice\_date\_cnt’]，求每天有通话的号码数量等。

经过以上特征处理，最终得到如下特征：

feature,importance

sms\_opp\_head\_max\_in,92

voice\_len\_mean\_in,70

sms\_opp\_len\_mean\_in,57

voice\_cnt\_peruniquecall,56

wa\_type1\_date\_mean,54

wa\_dura\_sum\_percnt,50

wa\_type0\_namecnt\_perday,50

sms\_opp\_head\_min\_in,49

voice\_head\_mean\_in,47

sms\_opp\_head\_mean\_in,41

sms\_opp\_len\_mean,39

voice\_opp\_head\_min,39

voice\_headinunique\_perday,37

voice\_head\_mean\_type1,36

wa\_type0\_dura\_perday,36

wa\_down\_flow\_median,35

voice\_time\_median,35

voice\_date\_mean\_type1,34

voice\_timesum\_peruniquecall,34

voice\_head\_mean\_out,33

sms\_start\_time\_mean,33

voice\_time\_min,33

voice\_start\_mean\_in,33

wa\_type1\_dura\_perday,33

voice\_head\_max\_type1,32

wa\_namecnt\_perday,32

sms\_callunique\_perday,31

voice\_len\_max\_in,31

voice\_start\_mean\_out,30

sms\_calluniquein\_mean\_perday,30

sms\_calluniquein\_perday,30

sms\_len\_perday,30

voice\_headoutunique\_perday,29

voice\_time\_mean\_out,29

voice\_len\_mean\_type1,29

wa\_type1\_date\_cnt,29

wa\_up\_flow\_median,28

wa\_type1\_namecnt\_perday,28

voice\_head\_unique\_in,28

wa\_type0\_namedura\_perday,28

wa\_type0\_date\_mean,27

voice\_headout\_perday,27

voice\_timeoutmean\_percall,26

voice\_headin\_perday,26

voice\_opp\_head\_mean,26

sms\_callcnt\_perday,26

wa\_visit\_date\_cnt,26

wa\_type1\_namevicnt\_perday,26

wa\_type0\_dura\_mean,25

voice\_opp\_head\_max,25

voice\_date\_cnt\_in,25

wa\_type1\_namedura\_perday,25

wa\_name\_count\_uinque,24

wa\_type0\_visitcnt\_mean,24

voice\_time\_max,24

sms\_opp\_num\_count\_in,23

wa\_type0\_date\_cnt,23

voice\_headmean\_peruniquecall,23

wa\_type0\_namevicnt\_perday,23

wa\_type1\_namevicntall\_perday,23

voice\_time1mean\_percall,23

wa\_visit\_cnt\_mean,23

wa\_type1\_dura\_mean,23

sms\_opp\_head\_max,22

voice\_opp\_len\_max,22

wa\_visit\_dura\_median,22

sms\_head\_unique\_perday,22

voice\_cntoutsum\_perday,22

voice\_opp\_unique\_in,22

wa\_down\_flow\_mean,22

wa\_type1\_downflow\_mean,22

sms\_start\_time\_mean\_in,21

voice\_time\_mean\_type1,21

wa\_visitdura\_perday,21

wa\_up\_flow\_std,21

voice\_time\_mean\_in,21

sms\_opp\_len\_sum,21

wa\_namedura\_perday,20

sms\_start\_time\_mean\_out,20

wa\_type0\_up\_persec,20

wa\_type1\_visitcnt\_all,20

wa\_type0\_visitcnt\_perday,20

voice\_timeoutunique\_percall,20

voice\_len\_mean\_out,20

wa\_up\_flow\_max,20

wa\_type0\_flow\_mean,19

voice\_cntoutmean\_perday,19

voice\_timeoutsum\_perday,19

voice\_head\_peruniquecall,19

wa\_type1\_upflow\_mean,19

sms\_calluniquemeanout\_perday,19

sms\_calluniquemeanin\_perday,19

wa\_type1\_visitcnt\_perday,19

voice\_time\_std,19

wa\_visitcnt\_perday,19

wa\_type1\_down\_percnt,19

wa\_flow\_sum\_perday,19

sms\_opp\_head\_min,19

wa\_type0\_downflow\_all,18

voice\_cntinsum\_perday,18

wa\_type1\_flow\_mean,18

sms\_calluniqueout\_mean\_perday,18

voice\_numcnt3mean\_perday,18

wa\_type1\_name\_unique,18

sms\_opp\_head\_mean\_out,17

wa\_down\_flow\_std,17

voice\_timemean\_peruniquecall,17

voice\_lenmean\_peruniquecall,17

voice\_cnt\_perday,17

wa\_type0\_name\_unique,17

sms\_opp\_head\_mean,17

voice\_timeout\_percall,16

voice\_opp\_unique\_out,16

wa\_type0\_down\_perday,16

wa\_type0\_namevicntall\_perday,16

wa\_flow,16

sms\_date\_cnt\_in,16

wa\_upflowcnt\_perday,16

voice\_timeinunique\_percall,16

sms\_opp\_len\_max,15

wa\_up\_flow\_mean,15

wa\_visit\_dura\_std,15

sms\_date\_cnt,15

wa\_visit\_dura\_mean,15

voice\_numcnt1mean\_perday,15

voice\_opp\_len\_mean,15

voice\_date\_mean\_type2,15

voice\_len\_max\_out,15

wa\_namevicnt\_perday,15

sms\_date\_mean\_in,14

voice\_lensum\_peruniquecall,14

voice\_timeoutsum\_percall,14

voice\_time\_sum\_out,14

wa\_upflow\_perday,14

voice\_timeoutmean\_perday,14

voice\_time\_mean\_type3,14

wa\_type1\_visitcnt\_mean,14

wa\_type0\_upflow\_mean,14

voice\_time\_sum\_type3,13

wa\_downflowcnt\_perday,13

wa\_type1\_up\_perday,13

wa\_flow\_sum\_percnt,13

sms\_date\_cnt\_out,13

wa\_down\_flow\_max,13

voice\_timemean\_percall,13

sms\_calluniquecntin\_perday,13

voice\_timein\_percall,13

voice\_date\_mean\_out,12

voice\_time\_sum\_in,12

voice\_date\_count\_type1,12

voice\_time\_sum\_type1,12

voice\_date\_count\_type3,12

wa\_type1\_up\_persec,12

voice\_head\_mean\_type2,12

voice\_time1\_percall,12

voice\_time2\_percall,12

voice\_time3\_percall,12

voice\_len\_peruniquecall,12

wa\_type0\_upflow\_all,12

wa\_flow\_sum\_persecond,12

wa\_type0\_down\_percnt,12

voice\_opp\_len\_unique,11

wa\_type0\_visitcnt\_all,11

wa\_type0\_up\_perday,11

sms\_calluniqueout\_perday,11

voice\_time3mean\_percall,11

wa\_down\_flow\_sum,11

voice\_timeinmean\_percall,11

wa\_visit\_cnt\_std,11

voice\_date\_mean\_type3,11

voice\_date\_cnt\_out,10

wa\_up\_flow\_sum,10

wa\_upflowdura\_persec,10

voice\_time2mean\_percall,10

voice\_cntinmean\_perday,10

voice\_head\_mean\_type3,10

sms\_opp\_head\_min\_out,10

wa\_visit\_cnt\_median,10

wa\_visit\_cnt\_sum,10

voice\_timeinsum\_perday,10

wa\_type0\_downflow\_mean,9

wa\_type1\_down\_perday,9

sms\_opp\_num\_unique\_in,9

wa\_visit\_dura\_sum,9

sms\_opp\_num\_count\_unique,8

voice\_opp\_count\_all,8

sms\_calluniquecntout\_perday,8

voice\_date\_count\_type2,8

sms\_date\_mean\_out,8

voice\_date\_cnt,8

voice\_time\_mean\_type2,8

voice\_time\_mean,8

voice\_opp\_num\_unique,8

voice\_numcnt1\_perday,8

voice\_numcnt2mean\_perday,8

sms\_opp\_num\_count\_out,8

wa\_visit\_cnt\_max,8

wa\_type1\_upflow\_all,7

voice\_opp\_count\_out,7

voice\_numcnt2\_perday,7

wa\_downflowdura\_persec,7

voice\_headsum\_peruniquecall,7

wa\_visit\_dura\_max,7

voice\_timeinsum\_percall,7

voice\_date\_mean\_in,6

wa\_downflow\_perday,6

sms\_opp\_len\_mean\_out,6

voice\_time\_sum,6

sms\_opp\_num\_count\_all,6

voice\_time\_percall,6

voice\_head\_unique\_out,6

voice\_num\_cnt\_type1,6

voice\_num\_cnt\_type3,6

voice\_opp\_head\_unique,5

sms\_opp\_num\_unique\_out,5

wa\_type1\_downflow\_all,4

voice\_len\_sum\_out,4

sms\_opp\_head\_max\_out,4

voice\_numcnt3\_perday,4

sms\_opp\_head\_unique,3

voice\_timeinmean\_perday,3

voice\_head\_max\_type3,3

voice\_time\_sum\_type2,2

voice\_head\_max\_type2,2

voice\_len\_sum\_in,2

voice\_opp\_count\_in,2

voice\_num\_cnt\_type2,0

在特征工程中，删除了一些低分的特征，最终确定维数为228。

## 2.3LightGBM

LightGBM与其他流行算法采用depth-wise的叶子生长策略不同，其使用的是带深度限制的leaf-wise的叶子生长策略。与depth-wise相比，leaf-wise算法收敛得更快。但leaf-wise的一个缺点就是：参数选择不当可能会产生比较深的决策树从而产生过拟合。因此LightGBM在Leaf-wise上增加了一个最大深度限制，在保证高效率的同时防止过拟合。

在调参方面，由于LightGBM使用的是leaf-wise算法，因此调节使用的是num\_leaves控制模型复杂度，大致设置为num\_leaves=2^max\_depth。另外，min\_data\_in\_leaf参数用于处理过拟合的问题，设置大可以防止过拟合，但需要注意可能会导致欠拟合。

由于特征过多，必须要考虑多过拟合的处理，所以设置较小的num\_leavea和max\_depth以避免过深的树，同时要设置合适的min\_data\_in\_leaf避免过拟合或欠拟合。

以下为参数设置:

lgb\_params = {

'boosting\_type': 'gbdt',

'objective': 'binary',

'is\_training\_metric': False,

'learning\_rate': 0.08,

'num\_leaves':16,

'max\_depth':6,

'min\_data\_in\_leaf':40,

'verbosity':-1,

'is\_training\_metric':False,

'feature\_fraction': 0.8,

'bagging\_fraction': 0.8,

'bagging\_freq':1,

'is\_unbalance':True

}

# 实验结果

通过特征提取和调整模型参数，提交结果前最终线下测试结果为0.809，线上测试结果为0.760，最终排名142/211。

比赛结果截图：



# 四、心得体会

本次实训对我来说实际上是一次很大的考验。因为先前既没有python的基础，也没有上过数据挖掘理论课，所以一开始对相关的理论概念一无所知。

在这样的情况下，我从网上的教程一步步开始学习，包括学习python的网站以及在同学推荐下看数据挖掘和机器学习的网课，慢慢摸索。尽管这样，在初赛阶段，写出来的代码和测试结果仍然惨不忍睹（线上测试分数只有0.39）。

后来，在同学的帮助下（包括模型讲解和编码的指导）和参考了助教提供的baseline，逐渐把分数提高，最终有这样的成绩。

可以说现在我依然对数据挖掘相关的理论知识一知半解，仅仅停留在照着葫芦画瓢的阶段。但至少在这个过程中我逐渐熟悉了python的使用，以及一些很实用的python库如numpy、pandas等，还算是有些收获。

**参考文献**

[1]lightGBM原理、改进简述.

<https://blog.csdn.net/niaolianjiulin/article/details/76584785>

[2]助教提供的baseline.

<https://github.com/Hanszhuang/mobileRiskUser>

[3]个人代码.

<https://github.com/MidnightDJPin/JDATAdm>