****

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：数字媒体技术

学生姓名：宋志强

学 号：15331278

授课教师：郑子彬

2018 年 6 月 14 日

目录

[基于移动网络通讯行为的风险用户识别 2](#_Toc517008248)

[1. 特征工程 2](#_Toc517008249)

[1.1. 用户通话记录数据表 2](#_Toc517008250)

[1.2. 用户短信记录数据表 5](#_Toc517008251)

[1.3. 用户网站/App访问记录数据表 7](#_Toc517008252)

[1.4. 总结 9](#_Toc517008253)

[2. LGB单模型 10](#_Toc517008254)

# 基于移动网络通讯行为的风险用户识别

大赛地址：[基于移动网络通讯行为的风险用户识别](https://jdata.jd.com/html/detail.html?id=3)，该比赛是联通大数据公司发起的，同时作为我们的实训的比赛。下面是自己参加比赛过程的报告：

第一次参加这种比赛，发现自己经验明显不足。初赛0.787（52/624），复赛0.784（71/624），虽然成绩很差，但是感觉对新手来说是一次很好的锻炼机会。简单的介绍一下自己的一些思路。

这次比赛分为初赛和复赛两部分，自己的比赛代码也有好几个版本，这里选取得成绩最好的一次讲讲思路。

## 特征工程

训练数据集一共包括四个文件，用户风险标签，用户通话记录数据，用户短信记录数据，用户网站/App访问记录数据。

测试数据集一共包括三个，用户通话记录数据，用户短信记录数据，用户网站/App访问记录数据。

初赛与复赛分别使用不同的测试数据集，分为test\_A，test\_B，将这些训练集与测试集当正负样本平衡处理。

下面介绍的是自己能想到的一些特征。

### 用户通话记录数据表

通话表如下图 1‑1用户通话记录表所示（下面3张图片来自比赛主页）



图 1‑1用户通话记录表

对通话数据的考虑主要就是判断用户通话是否规律等，还有挖掘出用户接收的通话的特征。

对用户通话数据处理方法：

#### 预处理

对start\_time的时变量做除以2变换，对opp\_len进行分组，避免one-hot变换后分布太稀疏（对线性模型有影响），同理对天变量除以5（或者对用户的生活作息时间进行分组）.

#### 特征提取

#### 统计特征

* 统计每个用户的号码通话的所有与不同的号码数量，以及与均值的差
* 统计用户通话in、out的不同号码数量，in、out的差值，以及所占比例。
* 统计一些特殊号码，如opp\_head为100的，像运营商的号码；170、171虚拟号码段
* 统计不同opp\_head的unique\_count
* 通话时长的相关统计量，如均值， 最大值，中位数，标准差，最小值等
* 统计不同call\_type下的opp\_num
* 统计不同opp\_len下的opp\_num

**一些特征的可视化：**

通话类型统计数均值：

通话出入度

#### One-hot编码特征

* 通话类型做one-hot编码，同时与in、out做特征交叉，统计数量
* 对start\_time的天变量和时变量做one-hot，统计数量
* 对天变量，统计不同日期的通话数量的相关统计量，如均值， 最大值，中位数，标准差，最小值等

### 用户短信记录数据表

用户短信数据表如下图 1‑2用户短信记录数据表



图 1‑2用户短信记录数据表

个人觉得短信记录表比voice表重要一些，主要是被动接收的短信的特征，如opp\_head、opp\_len，接收的不同号码的数量，主要是一些统计特征，还有对可以分类的做one-hot变换。除此之外，找出它们之间的一些相关性特征进行组合。

对用户短信数据处理方法：

* + 1. **预处理**

先统计所有opp\_num，再清洗掉opp\_head为000的系统短信；对opp\_len进行分组，对start\_time的时变量和天变量做类似通话数据的处理。

* + 1. **特征提取**
       1. **统计特征**
* 统计每个opp\_num所有与不同的号码数量，与均值的差
* 统计用户接收短信in、out的不同号码数量，in、out的差值，以及所占比例。
* 分组统计一些特殊号码的所有与不同的数量以及与均值的差，如opp\_head为100的，像运营商的号码；170、171虚拟号码段，106的通知类短信
* 统计不同opp\_len下的opp\_num数量以及与均值的差
* 统计不同opp\_head的数量

**一些特征可视化**

短信接收出入度：

日期分组的短信变化均值：

* + - 1. **One-hot编码特征**
* 对start\_time的天变量和时变量做one-hot，与in、out做交叉，分别求count和unique\_count

### 用户网站/App访问记录数据表

用户网站/App访问记录数据表如下图 1‑3用户网站/App访问记录数据表



图 1‑3用户网站/App访问记录数据表

对用户网站/App访问记录数据主要是考虑访问的app/网站的不同的数量，以及访问时间的特征，上传流量等。

对用户网站/App访问记录数据处理方法：

* + 1. **预处理**

wa数据有很多缺失值，对wa\_date缺失值填充-1做其他类别处理，将天数做除以5变换。求名字长度wa\_len，对wa\_len分组处理。

* + 1. **特征提取**
       1. **统计特征**
* 统计用户访问的wa\_name所有与不同的名字的数量，与均值的差统计用户访问的名字的长度分组的数量。
* 统计用户访问的次数的相关统计量，如均值， 最大值，中位数，标准差，最小值等。
* 统计用户访问时长的相关统计量，如均值， 最大值，中位数，标准差，最小值等。
* 统计用户上传流量的相关统计量，如均值， 最大值，中位数，标准差，最小值等。
* 统计用户下载流量的相关统计量，如均值， 最大值，中位数，标准差，最小值等。
* 统计不同opp\_head的数量

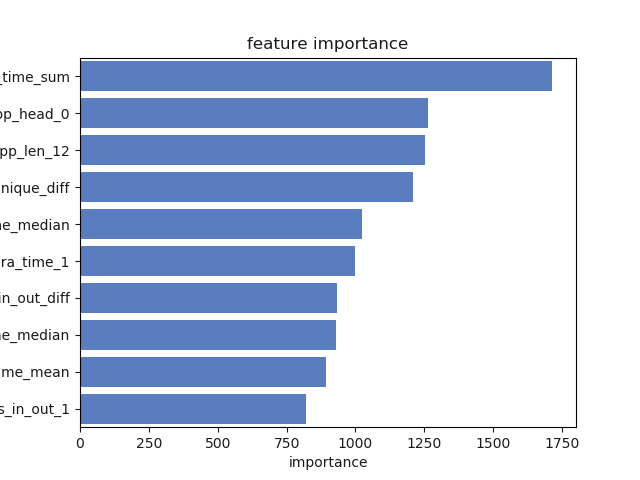
**一些特征可视化**

App/网站访问数量均值：

* + - 1. **One-hot编码特征**
* 对访问的日期进行one-hot编码，与up\_flow、visit\_cnt，down\_flow、做交叉特征，求相关统计量，如均值， 最大值，中位数，标准差，最小值等。
* 对访问的类型进行one-hot编码，与up\_flow、visit\_cnt，down\_flow、做交叉特征，求相关统计量，如均值， 最大值，中位数，标准差，最小值等。

**多项式特征：**

最后根据feature\_importance来选择一些特征进行做多项式组合



* 将访问类型为网站的unique\_count与sms的in类型的opp\_head为106的做二次多项式组合
* 将访问类型为网站的up\_flow与sms的in类型的unique\_count做二次多项式组合
* 将访问类型为网站的unique\_count与sms的in类型的unique\_count为与voice的unique\_count做二次多项式组合

### 总结

赛后发现自己提取的特征明显不够，还有一些重要的特征没有考虑到，如没有对voice，sms，wa出现频次最多的top-k以及对voice和**sms的opp\_head进行one-hot变换**；还有它们两次的最近的一些时间间隔等等。虽然自己也做了很多尝试，如分别对voice,sms,wa每天的一些特征统计量进行处理，但是因为特征方面没有大的提升。到后面复赛赛心态有点崩了，最高的一次分数还是第一次提交时候。赛后才发现对分数提升最大的就是对opp\_head进行one-hot之后，提高了大概0.07的分数，这才是拉开差距的主要特征。

## LGB单模型

**树模型**

使用lgb单模型，选择gdbt，使用GridSearchCV挑选最佳参数。主要调**'min\_data\_in\_leaf'**和**'num\_leaves'**这两个参数

尝试使用stacking将多个模型融合，尝试发现效果并不明显，故最后没有采用。

**参考文献**

（<https://zhuanlan.zhihu.com/p/26820998>）（Kaggle 数据挖掘比赛经验分享）

（<https://www.zhihu.com/question/29316149>）（特征工程到底是什么？）

（<http://lightgbm.apachecn.org/cn/latest/index.html>）（LightGBM 的中文文档）

**比赛相关代码**（<https://github.com/t617/JDATA-UserRiskPre>）