1. 导入数据集
2. 合并数据集
3. 特征建立

###短信风险识别

1. 统计不同的 对端号码前N位 的联系频次
2. 用户收发短信的数量比例
3. 最近短信的时间距离——转化成时间（秒）

最近一次发送短信时间

最近一次收取短信时间

最近一次发送,接收短信时间

1. 对端号码前N位+对端号码长度 组合后的联系频次
2. (收到的短信)对端号码前N位+对端号码长度 组合后的的联系频次
3. sms全部、发送 及 收取的短信 时间间隔 的均值 和 方差
4. 发送sms 时间间隔的均值
5. 接收sms 时间间隔的均值
6. 全部sms 时间间隔的方差
7. 发送sms 时间间隔的方差
8. 接收sms 时间间隔的方差
9. 短信时间间隔最小值

####通话风险识别

1. 每个客户的手机号 通话时间 对端号码前N位 的联系频次
2. 拨打电话 的数量比例
3. 对端号码前N位+对端号码长度 组合后的的联系频次
4. (收听的通话)前N位+长度 组合后的的联系频次
5. 全部 通话时间 的均值
6. 发送 通话时间 的均值
7. 接收 通话时间 的均值
8. 全部 通话时间 的方差
9. 发送 通话时间 的方差
10. 接收 通话时间 的方差
11. 通话时间 最小值
12. 拨打电话 的时长总和 、主动通话时长、接听通话时长、拨打/全部通话的比例、接听时长/全部时长的比例
13. 不同 通话类型 的次数占总通话次数的比例
14. 不同 通话类型 的时长占总通话时长的比例
15. 通话人数的丰富度（通话过多少个不同号码，共多少通电话，不同号码/通话总数的比例）

###上网风险识别

1. 每个客户的 所有的网站和APP ，45天内 日均每个应用 使用频次
2. 每个客户的 所有的网站和APP ，45天内 日均 使用频次
3. 每个客户的 45天内的，上行和下载的总流量；下载/上行 总流程之比
4. 每个客户的 30天内的，上行和下载的总流量；下载/上行 总流程之比
5. 每个客户的 14天内的，上行和下载的总流量；下载/上行 总流程之比
6. 每个客户的 7天内的，上行和下载的总流量；下载/上行 总流程之比
7. 每个客户的 3天内的，上行和下载的总流量；下载/上行 总流程之比
8. 每个客户的 所有的网站和APP ，45天内 每天的 使用频率之和
9. 每个客户的 所有的网站和APP ，45天内 每天的 使用时长之和
10. 3天内，平均使用时长（使用时长之和/使用频率）
11. 7天内，平均使用时长（使用时长之和/使用频率）
12. 14天内，平均使用时长（使用时长之和/使用频率）
13. 30天内，平均使用时长（使用时长之和/使用频率）
14. 45天内，平均使用时长（使用时长之和/使用频率）
15. APP,网站访问的丰富度（多少不同种类，总数，不同种类/总数的比例）

只计算APP（多少不同种类，总数，不同种类/总数的比例）

只计算网站（多少不同种类，总数，不同种类/总数的比例）

all（多少不同种类，总数，不同种类/总数的比例）

1. APP 时间间隔 分析（最大间隔，最小间隔，间隔方差）
2. WEB 时间间隔 分析（最大间隔，最小间隔，间隔方差）

###三个表交叉特征

1. 月使用总金额（短信统计发送条数，通话统计拨打时长的分钟数，流量统计总量；然后按短信条数\*0.1，通话\*0.15，转化成M为单位的流量\*0.1）
2. 各个不同地区通话时长比例
3. 凌晨短信发送总条数，发送的短信占发送总数的比例（23-5点）
4. 凌晨通话总时长，通话总时长占总通话的比例（23-5点）
5. 短信，通话 间隔分析——最长间隔，最短间隔，间隔的方差
6. 通话时长在各个时段的比例
7. 短信条数在各个时段的总比例
8. 通话记录与短信中的号码重合情况
9. 时间间隔的偏度系数（sms,voice,wa分别计算）
10. 时间间隔的峰度系数（sms,voice,wa分别计算）

#########特征建立到此为止###########

1. 通过训练集建立模型，预测test\_a
2. 收集预测test\_a中预测比较有把握的数据
3. 将原始训练集中加入test\_a中预测有把握的数据，然后再次训练合并后的数据集；
4. 最后建立的模型，然后预测数据test-b

计算环境：系统环境win10，8g内存；

软件：R version 3.4.4 (2018-03-15)，R-studio

应用数据包：

data.table

stringr

dplyr

fBasics

xgboost

readr

stringr

caret

car

Matrix