**联通项目目标2流程及数据说明文档**

**目标：预测3G业务什么用户以及在什么情况下会订购流量包；**

**实际实现目标：预测3G未订购流量包的用户在下月是否会订购流量包及预测条件**

***一、流程***

**1）训练数据预处理**

分离样本集，将样本中的用户按2G、3G类型分存为两个文件，分离标准如下：

data[13].equals("A计划") || data[13].equals("B计划") || data[13].equals("C计划") || data[13].equals("IPHONE") || data[13].equals("沃派") || data[13].equals("流量升级套餐") || data[13].equals("校园G卡") || data[13].equals("沃派先锋卡") || data[13].equals("2、3G融合业务") || data[13].equals("无线上网卡") || (data[13].equals("其它") && data[14].contains("3G") )

认定符合以上标准即为3G用户。

**2）生成训练集**

对每个月的数据进行3G用户与2G用户分离，对2013年11月未订购流量包但在2013年12月订购了流量包的3G用户进行标记，标记标准为11月未订购“全国流量”和“省内流量”，而12月订购了至少一种流量包。

对离散的参数进行穷举另存为excel表。连续参数分割后用0-9进行标示。

**3）训练模型**

采用ID3决策树算法对训练集建模，用XML文件存储模型。

**注：**为使算法适应连续属性，并减少运算量，对训练集的连续属性字段编号、及连各续属性的分支分割点进行了固化，各字段连续性及分割点设置详见数据说明2）训练集说明表的连续字段分割点字段。

**4）测试集**

2013年12月未订购流量包但在2014年1月订购了流量包的3G用户。

***二、数据说明***

**1）原始数据：同副本存量需求1。**

**2）训练集**

表1：训练集数据格式。

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 字段索引 | 字段含义 | 是否作为参考 | 字段类型 | 连续字段分割点 |
| 0 | 用户号码 |  |  |  |
| 1 | 用户编号 |  |  |  |
| 2 | 用户身份证号码 |  |  |  |
| 3 | 性别 | 是0 | 离散 | 男、女 |
| 4 | 年龄 | 是1 | 连续 | 无穷小，12,18,24,30,36,42,48,54，无穷大 |
| 5 | 入网时间 |  |  |  |
| 6 | 分公司名称（城市） | 是2 | 离散 |  |
| 7 | 城乡表示（原区县名称） | 是3 | 离散 |  |
| 8 | 渠道类型 | 是4 | 离散 |  |
| 9 | 入网渠道名称 |  |  |  |
| 10 | 入网渠道编码 |  |  |  |
| 11 | 付费类型 | 是5 | 离散 |  |
| 12 | 产品大类 | 是6 | 离散 |  |
| 13 | 套餐档次 |  |  |  |
| 14 | 出账收入 | 是7 | 连续 | 无穷小，20,50,80,110,150,200,250,300，无穷大 |
| 15 | 长途主叫通话分钟数 | 是8 | 连续 | 无穷小，20,50,100,200,400,800，无穷大 |
| 16 | 市话主叫通话分钟数 | 是9 | 连续 | 无穷小，20,50,100,200,400,800，无穷大 |
| 17 | 漫游主叫通话分钟数 | 是10 | 连续 | 无穷小，20,50,100,200,400,800，无穷大 |
| 18 | 被叫通话分钟数 | 是11 | 连续 | 无穷小，20,50,100,200,400,800，无穷大 |
| 19 | 上行点对点短信条数 | 是12 | 连续 | 无穷小，20,50,100,200,400,800，无穷大 |
| 20 | 上行点对点彩信条数 | 是13 | 连续 | 无穷小，20,50,100,200，无穷大 |
| 21 | 是否订购全国流量包 |  |  |  |
| 22 | 是否订购省内流量包 |  |  |  |
| 23 | 是否订购沃派语音包 | 是14 | 离散 |  |
| 24 | 是否订购短信可选包 | 是15 | 离散 |  |
| 25 | 是否订购彩信可选包 | 是16 | 离散 |  |
| 26 | 定向流量 | 是17 | 离散 |  |
| 27 | 非定向流量 | 是18 | 连续 | 无穷小，20,50,100,200，无穷大 |
| 28 | 是否加订流量包 |  |  |  |

表2：参数映射（作为配置文件）。

|  |  |  |  |
| --- | --- | --- | --- |
| 参数名称 | 映射参数名称 | 偏移量 | 取值 |
| 性别 | cateName1 | 3 | 男,女,0 |
| 年龄 | cateName2 | 4 | 3,2,1,0,7,6,5,4,9,8 |
| 分公司名称（城市） | cateName3 | 6 | 遂宁,宜宾,甘孜,乐山,成都,广安,攀枝花,眉山,雅安,资阳,泸州,内江,自贡,阿坝,绵阳,达州,凉山,德阳,广元,南充,巴中 |
| 城乡表示（原区县名称） | cateName4 | 7 | N,Y |
| 渠道类型 | cateName5 | 8 | 电话渠道,集团渠道,自有营业厅,公众直销,社会实体门店,电子渠道 |
| 付费类型 | cateName6 | 11 | 预付费用户,后付费用户 |
| 产品大类 | cateName7 | 12 | B计划,IPHONE,流量升级套餐,无线上网卡,A计划,2、3G融合业务,沃派,C计划 |
| 出账收入 | cateName9 | 14 | 3,2,1,0,7,6,5,4,9,8 |
| 长途主叫通话分钟数 | cateName10 | 15 | 3,2,1,0,7,6,5,4 |
| 市话主叫通话分钟数 | cateName11 | 16 | 3,2,1,0,7,6,5,4 |
| 漫游主叫通话分钟数 | cateName12 | 17 | 3,2,1,0,7,6,5,4 |
| 被叫通话分钟数 | cateName13 | 18 | 3,2,1,0,7,6,5,4 |
| 上行点对点短信条数 | cateName14 | 19 | 3,2,1,0,7,6,5,4 |
| 上行点对点彩信条数 | cateName15 | 20 | 3,2,1,0,5,4 |
| 是否订购沃派语音包 | cateName16 | 23 | yes,no |
| 是否订购短信可选包 | cateName17 | 24 | yes,no |
| 是否订购彩信可选包 | cateName18 | 25 | yes,no |
| 定向流量 | cateName19 | 26 | yes,no |
| 非定向流量 | cateName20 | 27 | 3,2,1,0,5,4 |
|  |  |  |  |

**3）数据输出**

预测条件以json格式给出，所有条件存为一个文本，每一行表示一个条件，格式样例如下（预测条件.txt）：

{"非定向流量":"50-100","产品大类":"B计划","套餐档次":"156","分公司名称（城市）":"乐山","长途主叫通话分钟数":"20-50"}

***三、数据统计结果***

|  |  |  |  |
| --- | --- | --- | --- |
| 年份 | 2013年11月 | 2013年12月 | 2014年1月 |
| 总用户数 | 10323094 | 10323567 | 10322701 |
| 3G用户数 | 3503120 | 3543566 | 3548991 |
| 未订购流量包的3G用户数 | 2634668 | 2647597 | 2620124 |
| 下月订购流量包的用户数 | 59561 | 68075 | Null |

***四、测试结果***

|  |  |
| --- | --- |
| 未订购流量包3G用户数 | 2647597 |
| 次月加订流量包3G用户数 | 68075 |
| 参数：叶子结点最小用户数 | 100 |
| 参数：叶子结点正确数比例 | 0.05 |
| 预测加订流量包3G用户数 | 187292 |
| 预测对的用户数 | 13272 |
| 准确率 | 0.0708626102556436 |
| 覆盖率 | 0.19496143958868895 |