文章编号：2096-1472(2020)-01-27-03
DOI:10.19644/j.cnki.issn2096-1472.2020.01.007

大数据技术在通信运营商异网获客系统的应用

谈俊林

(中国电信股份有限公司上海分公司，上海 200041)

摘　要：电信运营商拥有着庞大的客户群体、海量的数据资源，如何挖掘大数据价值，精准把握客户需求，将成为运营商保持行业领先的关键。为吸引更多优质异网用户在携号转网过程中携入，本次研究通过大数据技术，借助DPI[1,2]解析信息构建异网号码池，运用逻辑回归、决策树、boosting算法[3]等原理搭建转网倾向评估模型，最后通过客户画像、异网策反目标用户分群[4]，并配置针对性营销策略，构建起一个较为完善的“异网获客系统”，提升运营商在携号转网过程中的客户经营服务能力。

关键词：携号转网；DPI；数据挖掘；客户分群；营销标签

中图分类号：TP301.6　**文献标识码**：A

Application of Big Data Technology in Different Network Customer Acquisition System of Telecom Operators

TAN Junlin

(*China Telecom Shanghai Branch,Shanghai* 200041,*China*)

Abstract: Telecom operators have a huge customer base and massive data resources.How to extract the value of big data and grasp the real needs of customers will become the key for operators to maintain the leading position.In order to better attract more high-quality users from different networks,this study establishes the different network number pool by DPI analytic information,and uses logic regression,decision tree,boosting algorithm to build the transfer tendency evaluation model.Finally,with customer portrait,we will output the differentiation of target users in different networks,and configure targeted marketing strategies,so as to build a relatively complete system of “different network customers acquisition” and improve the customer service capability.

Keywords:number portability;DPI;data mining;customer group;marketing label

1 引言(Introduction)

电信运营商拥有着庞大的客户群体，积累了海量的数据资源，如何挖掘沉淀的大数据价值，精准把握客户的真实需求和痛点，提升客户服务体验，将成为运营商保持行业领先地位并保持市场份额的关键。新的竞争形势下，电信产业链和客户行为发生着深刻变化。客户成为新价值链模式下的“核心”，客户体验越来越偏向互联网业务。随着大数据的深入应用，客户画像及精准营销体系建设迫在眉睫，同时如何快速准确识别用户，是业务部门面对的新挑战。为积极应对携号转网带来的不确定竞争风险，运营商急需要系统性构建全方位潜在客户获取系统。

2 异网获客系统的构建(Construction of different network customers acquisition system)

通过深挖用户消费特征，对当前主要关注的携号转网策反进行持续的深入跟踪分析和建模。为更好地吸引更多异网有效用户携入，本次研究方向就是“携号转网异网获客模型的构建”，具体建模思路为：借助构建异网号码池、转网倾向评估模型、异网策反目标用户分群等模型锁定目标用户群，制定针对性策略实施携入引导，异网获客系统搭建思路如图1所示。

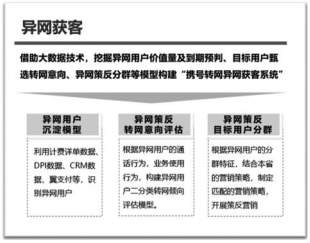

图1 异网获客系统结构设计

Fig.1 Structure design of different network customers acquisition system

www.rjgcz.com

2.1 模型介绍

2.1.1 异网高价值用户识别模型

简介：对异网用户接触本网的各类渠道进行针对性异网用户数据采集设计和开发，包括解析固网DPI[1,2]、人工采集、爬虫采集[5]等。并对已采集的数据进行数据清洗、转换和整合，在已有数据基础上形成异网用户基础标签。

目标：通过各类渠道采集用户号码及相关行为，沉淀异网号码，形成异网用户号码池，为获客系统提供基础数据支撑。

思路与方法：利用计费详单、DPI[1,3]、CRM、翼支付等数据，识别异网用户，同时根据异网用户的通话行为[6]，剔除快递员、外卖送餐员用户，对异网用户进行持续过滤沉淀。如图2所示。

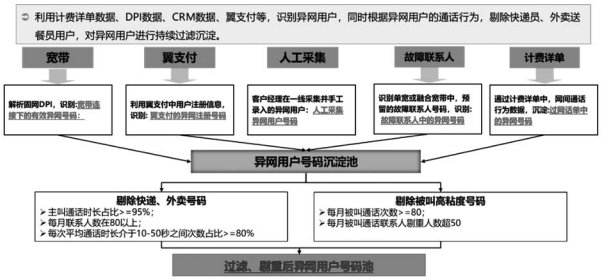

图2 识别异网用户思路方法

Fig.2 Method of identify different network users

通过蒙特卡洛模拟，将沉淀池中的号码的过网通话时长提取出来，模拟估测这批异网用户的实际通话时长，具体如图3所示。

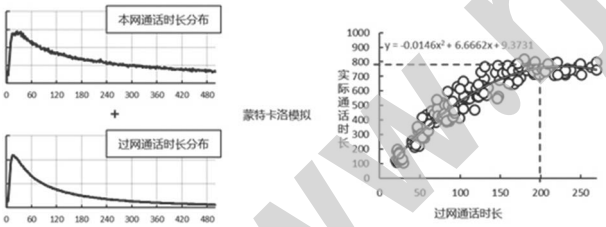

图3 模拟异网用户实际通话时长示例

Fig.3 Example of simulating the actual call duration of a different network user

以同一行业人群行为结构相同的原理，虽然运营商不同，但通话次数、通话时长及其他通话行为相似，推测其ARPU值也相近，因此，通过多元线性回归拟合异网用户ARPU值，并以其交往圈电信用户平均价值量为参考，估测异网用户真实价值量，从而筛选出高价值量用户。

2.1.2 转网倾向评估模型

简介：本模型拟采样logistic模型对异网用户的特征进行训练，计算出每个异网用户的转网概率，并输出标签。

目标：通过评估异网用户的转网倾向，为一线携号转网营销提供名单支撑。

建模设计：将一定周期内转网的异网用户作为正样本，没有转网的用户作为负样本，从业务角度选取异网用户的通话行为，如联系过网号码个数、通话次数、通话时长等，业务使用行为，如翼支付使用数据，电信固网下异网用户行为[7]监测四个维度，通过口径判别，识别历史被成功策反为电信用户，对这些用户进行特征提取，并进行决策树C5.0建模，通过不断调整训练集，反复训练模型，采用boosting算法及交叉验证，直至模型效果较为优良为止，根据模型生成的规则，输出异网用户是否转网及转网概率的标签。

数据准备：结合异网用户转网的影响因素，整理相关的数据字段如详单类信息、宽带连接信息、翼支付信息、终端信息等。

数据加工：将以上与目标分析相关的字段提取出来，得到的数据难免会有缺失值和极值，需要对数据进行质量检查。结合业务经验，对有数据缺失的字段进行处理，填充为-1或者平均值。离群值定义为数据与部分样本不符合的特征，这些样本会影响建模结果，故本次建模对离群值进行丢弃处理。

变量筛选：数据挖掘中数据量非常庞大，特别是特征字段，需要对特征字段进行提取，选取与目标变量最重要且变量之间相关性较弱的变量。若自变量为分类变量，利用卡方检验的方法看是否显著相关；若自变量为连续变量，利用相关性矩阵确定变量的重要性，当连续变量之间相关性大于0.7时，只需选择其中与目标变量相关性较强的变量。本次变量选取通过以上方法进行逐步筛选，最终将筛选后的变量纳入模型。筛选过程如图4所示。

分类	指标	变量贡献率
异网用户基本信息	省份	0.364
	运营商	0.060
整体通话表现	日均过网通话天数趋势	0.022
	日均过网通话人数趋势	0.008
	日均过网通话时长趋势	0.001
	日均过网通话天数	0.011
主叫通话特征	日均主叫过网通话次数趋势	0.030
	日均主叫通话次数	0.045
	前月日均主叫通话次数	0.031
	前前月日均主叫通话次数	0.020
	日均主叫通话占比	0.206
翼支付特征	是否翼支付用户	0.202

```
Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.9382356  0.0364591 -80.590  < 2e-16 ***
PROV_ID      1.0625947  0.0343982  30.891  < 2e-16 ***
OPERATORS    0.1761408  0.0195027   9.032  < 2e-16 ***
CALL_DAYS_TREND  0.0648175  0.0063793  10.161  < 2e-16 ***
BW_USERS_TREND  -0.0223745  0.0093227  -2.400  0.01640 *
CALL_DUR_TREND  0.0017062  0.0006617   2.579  0.00992 **
CALLING_CNT_TREND  0.0876788  0.0058692  14.939  < 2e-16 ***
CALL_DAYS  0.0322593  0.0013282  24.289  < 2e-16 ***
CALLING_CNT  0.1315200  0.0130380  10.087  < 2e-16 ***
CALLING_CNT_1  -0.0910784  0.0154727  -5.886 3.95e-09 ***
CALLING_CNT_2  -0.0576685  0.0139889  -4.122 3.75e-05 ***
CALLING_CNT_RATE  0.6020308  0.0432539  13.919  < 2e-16 ***
PAY_USE_DO  0.5909458  0.0596739   9.903  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

图4 变量筛选方法举例

Fig.4 Examples of variable screening methods

模型建设：使用LR算法，将转网倾向打分模型训练到最优，并验证、测试模型效果。LR模型易于理解，而且便于处理非线性数据的描述数据，能提取数据中隐藏的知识规则。本次模型采用三个月的行为数据，建立异网用户转网模型，不断训练模型，直至达到较好的模型效果为止。

2.1.3 异网策反目标用户分群及策略配置模型

模型简介：利用大数据画像技术，从异网用户的转网倾向、价值度及各基础标签信息对用户进行群体划分，并为每个异网用户进行打标签。

模型目标：针对已在异网号码池中的异网用户目标群体，例如宽带到期的用户，根据宽带用户的使用行为特征进行宽带到期续费新装进行识别，并输出其概率，使得业务人员可以更好地开展营销工作。

www.rjgcz.com

模型构建：根据对电信用户构建的用户画像标签[8](如年龄标签、价值标签、通话使用行为标签、家庭圈-同事圈标签等)，结合对异网用户转网倾向的评估，选择转网倾向高、价值度高的用户优先进行营销分群。

数据准备：从基础信息表、资产表进行筛选、汇总和整理出各个属性的宽表数据信息。

模型算法：此次分群模型采用层次聚类算法[8]，层次聚类是聚类算法的一种，通过计算不同类别数据点间的相似度来创建一棵有层次的嵌套聚类树。在聚类树中，不同类别的原始数据点是树的最低层，树的顶层是一个聚类的根节点。创建聚类树有自下而上合并和自上而下分裂两种方法。

此次通过层次聚类来对异网号码的特征进行挖掘，根据异网号码接触的宽带信息包括访问url的信息和访问app的信息来对这些异网号码进行聚类，根据聚类结果分析每一类异网号码的特征。再匹配分群后的营销策略，主要的匹配思路有：(1)套餐匹配分群：选取家庭圈覆盖下的高价值依附异网用户作为最终策反目标，分场景匹配最佳营销终端及套餐；(2)终端匹配分群：选取家庭圈覆盖下终端到期的异网用户，推荐套餐捆绑进行策反。具体流程如图5所示。

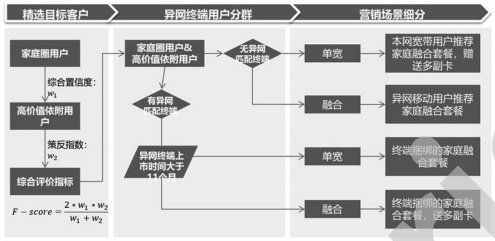

图5 目标客户分群及策略配置举例

Fig.5 Examples of target customer grouping and strategies configuration

2.2 模型评估

通过模型评估一方面对模型进行进一步的优化，使模型性能够更准确；另一方面，通过模型评估看模型实际运行效果，对采取的维系策略的有效性进行评价。评估模型本身的准确性，通过命中率、覆盖率两个重要指标。命中率：描述模型预测准确性性指标，命中率=预测成功转网的用户数/预测转网的用户数；覆盖率：描述模型预测结果与实际结果对比情况指标，覆盖率=预测成功转网的用户数/实际转网的用户数。

本次建模利用当月数据和隔月数据进行验证。通过不断的选取训练集，反复训练模型，最终实现模型命中率和覆盖率符合预期要求。

3 输出成果(The output results)

(1)建立异网号码池：根据识别的异网用户号码池，可开展异网用户的初步画像，并为后续的深度画像提供基础数据资源。

(2)建立转网概率较高的异网高价值用户库：输出异网高价值用户中，预估高概率携号转入的异网策反目标用户，包含：设备号、转网概率、是否转网等。

(3)建立不同群体的异网用户策反策略库：根据异网用户的分群特征，结合匹配本省的营销策略，指导营销人员开展策反营销。

4 结论(Conclusion)

此次研究成果，依托大数据技术筛选出的适合携号转网异网策反的移动用户对象，建立了异网用户号码池，并筛选出了一批高价值客户，还增加了客户画像的标签维度，如价值标签、同事圈标签、家庭圈标签等。有利于推动公司开展精细化营销工作的实施，提高存量、增量经营效能和市场占有率。同时，进一步推进大数据应用，持续提升用户的洞察能力，为后续运营商可针对每一类用户群监测携号转网营销转化率，并根据转化率变化情况及时调整策略。

此次研究挖掘了更深层次的数据价值。强化了既有系统的客户行为洞察能力，又提升了运营商客户经营服务能力。目前，运用大数据技术开展异网用户获取，还处于初步的研究阶段，阶段性的研究成果也不一定很成熟，本文所研究的在通信运营商领域内的数据挖掘模型技术还需要更多的业务应用与实践相结合，不断改进优化，在普及性和深入性上开展更多的研究和应用。

参考文献(References)

[1] SHENG,ZHANG,LI,YANBIAO.Multi-stride Indexing:Improve NFA for Fast and Scalable DPI[J].Chinese Journal of Electronics,2018,27(1):86-92.

[2] 吕锦扬.DPI技术在移动数据网络分析的应用[J].电信技术,2013(6):72-75.

[3] L Miclet,S Bayoudh,A Delhay.Analogical Dissimilarity:Definition,Algorithms and Two Experiments in Machine Learning[J].Journal of Artificial Intelligence Research,2014,32(3):793-824.

[4] YANG,TONG,XIE,GAOGANG,LI,YANBIAO.Guarantee IP Lookup Performance with FIB Explosion[J].Computer Communication Review:A Quarterly Publication of the Special Interest Group on Data Communication,2014,4(4):39-50.

[5] 耿贞伟,保富.网络环境下的大数据采集和处理[J].软件工程,2019,22(6):47-49.

[6] 宋竹,秦志光,罗嘉庆,等.电信数据中用户行为特征测量与分析[J].电子科技大学学报,2015,(6):934-939.

[7] 崇林.基于运营商大数据的互联网海量用户行为分析系统设计与实现[D].南京邮电大学,2016:1-74.

[8] 梁睿博,王思远.基于RAKEL算法的商品评论多标签分类研究与实现[J].软件工程,2019,22(1):8-11.

作者简介：

谈俊林(1982-)，男，硕士，工程师.研究领域：大数据挖掘.

www.jgcz.com