



专题:大数据

基于大数据挖掘的终端换机模型

王雪琼,熊珺洁,姚晓辉

(中国电信股份有限公司上海研究院,上海 200122)

摘要:目前,移动终端已成为运营商维系用户、拓展市场的战略重心,提升移动终端销量、扩大终端规模是各运营商的工作重点。基于数据挖掘技术,从用户属性、终端使用信息、终端搜索访问信息等维度出发,挖掘海量用户行为数据价值,建立终端换机模型,具体包括基于决策树算法的用户换机倾向识别模型和基于聚类算法的终端推荐模型,助力移动终端精准营销。

关键词:移动终端营销;数据挖掘;决策树;聚类算法

中图分类号:F274

文献标识码:A

doi: 10.11959/j.issn.1000-0801.2016314

Terminal replacement model based on big data mining

WANG Xueqiong, XIONG Junjie, YAO Xiaohui

Shanghai Research Institute of China Telecom Co., Ltd., Shanghai 200122, China

Abstract: In order to incrementally capture, retain and grow the subscriber bases, mobile operators must more effectively maximize the utilization of big data. Promoting the sale of mobile terminals was one of the focus of the operator's business currently. By mining the big data of consumer behaviors, including consumers' attributes, mobile terminal information and DPI data, replacing inclination distinguished model was built based on decision tree and recommending model was built based on clustering algorithm to identify target customers.

Key words: mobile terminal recommendation, data mining, decision tree, clustering algorithm

1 引言

移动互联网时代,OTT业务迅猛发展,中国移动、中国电信和中国联通三大运营商长期赖以依存的语音和短信业务受到OTT应用的严重侵蚀,2015年语音业务收入在移动通信业务收入占比37.97%,比2014年下降12.7个百分点。为了应对移动互联网的冲击,各大运营商将流量经营作为工作重点,力图构建智能管道,寻求以流量和增值服务带动收入增长的模式。另一方面,电信市场出现一种协同竞争的新局面——“端管云”,即手机终端、通信传输终端、云计算中心。可见作为OTT应用载体的移动终端已

成为各大运营商竞相争夺的利润高地。

虽然运营商希望通过移动终端市场突破流量经营的瓶颈,但是通信市场已趋于饱和,工业和信息化部通信运营统计公报显示,2015年移动电话用户总数达13.06亿户,移动电话用户普及率达95.5部/百人,已经不能通过大量增加新用户来拉动电信收入的增长。当前形势下,运营商应该更多地关注存量用户的移动终端使用情况,将终端销售与高价值用户的维系和拓展相结合。

随着用户数据量的激增,传统的数据分析方法和营销方式受到巨大的挑战,急需引入大数据挖掘技术对海量数据进行深度挖掘,探索用户消费行为数据的潜在价值,支



撑终端精准营销。

目前,大数据挖掘技术在电信业务应用方面的研究主要有:

- 参考文献[1]提出将大数据挖掘技术应用到电信运营商终端营销上,但是没有给出具体的建模方法;
- 参考文献[2]基于售前终端营销和售后终端能力分析两大功能模块,构建支撑市场部 and 终端厂商的终端分析体系;
- 参考文献[3]利用大数据挖掘技术剖析用户离网原因,确定目标用户群,进而针对潜在离网用户提出合理的营销政策和建议;
- 参考文献[4]基于 Hadoop 大数据架构采集电信运营商网络侧的数据,并对海量数据进行加工分析,挖掘掩藏于其中的用户行为特征,构建用户行为分析模型,展现了用户行为分析系统的设计思路与实现方法;
- 参考文献[5]基于统计分析和数据挖掘技术,针对手机垃圾短信治理效果不佳等问题设计了垃圾短信过滤系统。该系统根据实时获取的垃圾短信自动生成过滤规则,在短信转发阶段进行过滤。

本文总结了数据处理的主要方法,并基于大数据挖掘技术,多角度综合分析了用户的消费行为数据、移动 DPI 数据和终端使用数据,详细阐述了数据处理、模型构建以及模型评估的具体过程,实现以下功能:

- 基于决策树算法构建用户换机倾向识别模型,有效预测未来两个月内最有可能更换终端的目标用户;
- 基于聚类算法分别构建用户流量—价值九宫格、终端价格—性能九宫格,实现用户业务价值和终端的标签细化;
- 根据业务价值和终端细化标签结果,为目标用户匹配合适的终端,引导营销策略。

2 数据挖掘的关键技术

数据挖掘一般指利用算法搜索隐藏于海量数据中的重要信息的过程。数据挖掘通常与计算机科学有关,并通过统计、在线分析处理、情报检索、机器学习、专家系统(依靠过去的经验法则)和模式识别等诸多方法来实现上述目标。跨行业数据挖掘标准(cross-industry standard process for data mining,CRISP-DM)将一个完整的数据挖掘项目周

期分为业务理解、数据理解、数据准备、建模、评估和部署6个阶段。图1中的箭头指出了各个阶段之间最重要和频繁的关联依赖,图形的外圈表达了数据挖掘本身的循环特性。

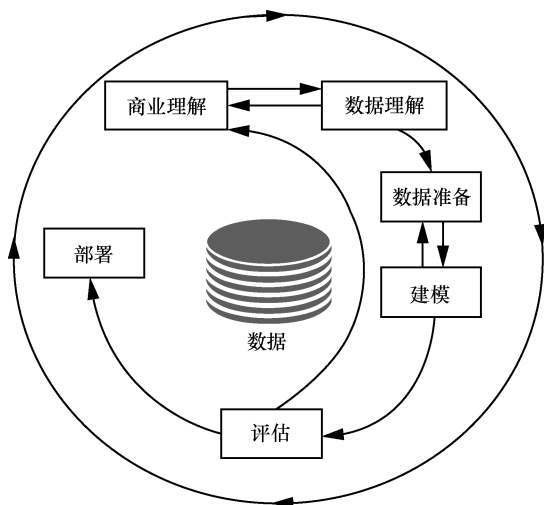


图1 CRISP-DM 标准

2.1 数据处理

数据质量对数据挖掘结果有至关重要的影响,在数据挖掘项目中,接近60%的时间和精力用来熟悉、处理和转换数据,最大限度保证数据的可用性。

(1)数据抽样

若数据全集规模太大,针对数据全集进行分析建模会消耗较多时间,有时过大的数据量会导致软件运行时崩溃。而采用合理的数据抽样,会明显降低这些负面影响且不会影响建模效果。在抽样时要确保输入变量的值域、分布,输出变量的值域、分布与数据全集的保持一致。另一种是预测小概率事件时的抽样,将在第3.6节介绍。

(2)数据审核

在获取数据后对数据进行一致性检查。一致性检查是指根据每个变量的合理取值范围和相互关系,检查数据是否合乎要求,发现超出正常范围、逻辑上不合理或者相互矛盾的数据。进一步,统计缺失值、异常值比例,评估数据可用性。最后处理缺失值和异常值。

(3)数据转换

根据采用转换的逻辑和目的的不同,数据转换可分为4类:产生衍生变量,通过对原始数据进行简单、适当的数学公式推导,产生更加有业务意义的新变量;连续数据的离散化,为了降低变量的复杂性,提升预测能力,可以利用

分箱变换将连续型变量转换成名义型变量 (例如年龄变量,每 10 岁构成一个年龄组,可以有效简化数据);改善变量的分布,针对不对称分布的变量,运用各种数学变换将其分布呈现(或近似)正态分布;数据标准化,该变换的主要目的是将数据按照比例进行缩放,使之落入一个小的区间范围内,消除量纲的影响,使其有平等分析和比较的基础。

(4)筛选有效的输入变量

过多的输入变量会带来干扰和过拟合等问题,导致模型稳定性下降,因此要遵循少而精的原则。主要方法有:结合业务经验进行筛选,很多情况下可以根据业务专家的商业敏感性有效缩小自变量的考察范围;计算变量间的相关系数(如皮尔逊相关系数),若两个变量间的相关系数大于或等于 0.6,说明两个变量间存在中度以上线性关系,建模时只需保留一个。

2.2 样本说明

模型总是在某一样本基础上建立的,将最大限度反映该样本的“核心行为”,但由于样本抽样的随机性,无法确定该模型在其他样本上的表现。为了能更好地评判模型效果,需将建模数据分为训练集和测试集,通常训练集与测试集样本比例为 1:1。训练集用于建立模型,包含实际目标值为 1 的正样本与实际目标值为 0 的负样本。测试集用于评估模型效果,包含实际目标值为 1 的正样本与实际目标值为 0 的负样本。模型应用数据是指用于预测的数据,该部分数据没有目标值,将其作用到构建好的模型上,得到实际预测用户名单。样本说明如图 2 所示。

本文从中国电信某省公司本地网中抽取 50 万户用户的数据进行挖掘分析,首先根据用户历史行为数据构建换机倾向识别模型,识别出 2016 年 9 月和 10 月份有换机倾向的用户;再根据用户现用终端和消费行为信息构建终端推荐模型,为用户推荐合适的终端,实现终端的精准营销。本文结合 SPSSModeler 软件详细阐述模型构建过程。

3 换机倾向识别模型

3.1 模型输入

首先确定宽表的数据窗口、观察月份、目标日期(数据窗口的最后一天)。本文的数据窗口选取 2016 年 4 月、5 月和 6 月,观察月份选取 2016 年 7 月和 8 月。选取的变量维度如下。

(1)用户基本属性维度

用户唯一标识、地市编码、年龄、性别、在网时长、客户渠道、是否合约用户、合约到期时间等。

(2)终端使用信息维度

现用终端上市日期,现用终端主屏幕尺寸,现用和历史终端价格、品牌、类型、终端注册时间,追溯至 $n-3$ 款。

(3)用户消费行为信息维度

用户的 ARPU (average revenue per user,每用户平均收入)值和流量使用量。

(4)终端搜索访问信息(移动 DPI 数据)维度

视频、音乐、游戏、地图、打车和网购 App 使用次数,视频、音乐、游戏、地图、打车和网购 App 流量使用量等。

3.2 数据审核

全面审核样本数据,利用 SPSSModeler 软件中的数据审核节点查看各变量的最小值、最大值、平均值、标准差,异常值、离群值的个数以及缺失值的比例,评估数据的可用性,如图 3 所示。将离群值和缺失值进行强制转换,用最接近的正常值进行替换。图 3 显示合约到期时间缺失值占 70%,这主要是由于只有 29%左右的用户办理合约套餐,在建模时这个字段可以不作为输入变量,但可以计算剩余协议时长,剩余协议时长小于 6 个月的用户优先营销。

3.3 生成衍生变量

(1)换机标识

将在观察月份内更换终端的用户标识为 1,作为建模的正样本,否则标识为 0,作为负样本。根据现用终端注册

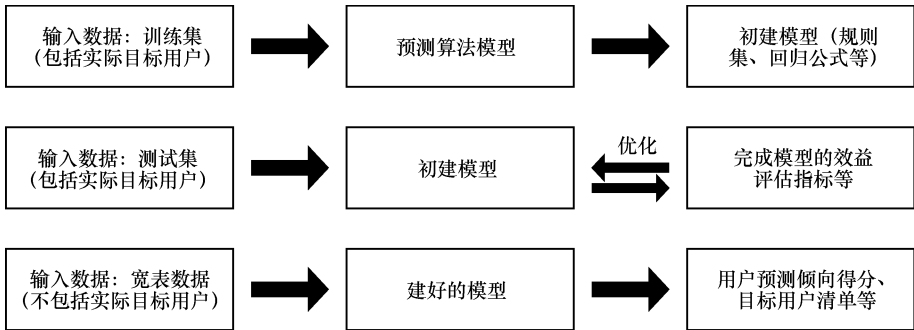


图 2 样本说明

3.6 模型构建

在建模前需用 SPSSModeler 中的分区节点将建模样本分为训练集和测试集,占比为 1:1。数据审核发现本文选取的样本中只有 15%的用户换机,由于模型总是力争使错误率最小化,若直接在这种分布上建立分类模型,所得的模型会偏向占比较高的非换机用户,对该部分用户的预测精度较高,但不能有效识别出换机用户,因此在训练集上要对本样本进行平衡,使得换机用户与非换机用户的比例为 1:1。

用于分类模型的算法主要有决策树 C5.0、决策树 CHAID、神经网络等,如何从众多的算法中选取最合适的构建模型是一个难点。SPSSModeler 中的自动分类器节点利用整体精确性和增益等指标衡量各个算法的优劣。综合考虑增益和总体精确性两个指标发现决策树 CHAID 算法比较理想,如图 5 所示,故选取决策树 CHAID 算法进一步精准建模。CHAID 模型给出预测变量的重要性排序,图 6 显示 ARPU 趋势,流量趋势,离散化的在用终端持机时长、终端价格、应用访问总次数和主屏幕尺寸这些变量对模型

模型 图形 摘要 设置 注解				
排序方式(S): 使用 升序 降序				
图形	模型	增益(前 30%)	总体精确性 (%)	
☒	决策列表 1	1.907	85.194	
☒	C&R 树 1	2.055	84.017	
☒	CHAID 1	2.278	83.698	
☒	C5 1	2.266	77.981	
☒	Quest 1	2.032	71.266	
☒	判别 1	1.721	64.568	
☒	贝叶斯网络 1	1.652	4.715	

图 5 自动分类器结果

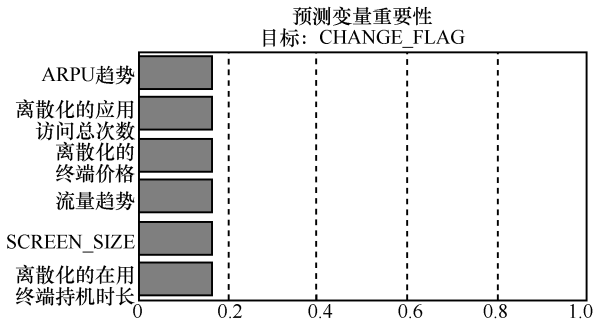


图 6 CHAID 模型运行结果

构建起关键作用。

3.7 模型评估

对建立好的模型做进一步的评估和优化,有两种常用的方法。

(1)SPSSModeler 中的分析节点

该节点分别给出模型在训练集和测试集上的正确率,对模型的准确性给出整体描述。将分析节点作用到 CHAID 模型上,图 7 显示测试集的正确率为 73%,训练集上的正确率为 81%。

分析 注解				
全部折叠(C) 全部展开(E)				
输出字段 CHANGE_FLAG 的结果				
比较 SR-CHANGE_FLAG 与 CHANGE_FLAG				
分区	1 训练	2 测试		
正确	53,990	73.15%	200,958	81.46%
错误	19,816	26.85%	45,747	18.54%
总计	73,806		246,705	

图 7 分析节点结果

(2)在测试集上计算命中率和覆盖率

命中率反映模型输出的预测目标对执行效率的提升情况,计算式为预测成功换机/预测换机;覆盖率反映模型输出的预测目标在实际换机人群中的覆盖情况,计算式为预测成功换机/实际换机。针对本例,预测换机人数为 54 182 人,实际换机人数为 37 099,预测中实际换机人数为 22 767 人,计算命中率为 42.1%,比原始浓度(15%)提升 2.8 倍,覆盖率为 61.3%。

3.8 模型优化

模型测试集的正确率只有 73%,需要对模型做进一步的优化。重新审视建模过程,发现将现用终端持机时长、历史终端持机时长、现用终端价格、在网时长、应用访问总流量、应用访问总次数、年龄、ARPU 均值、流量均值进行离散化时都用的是分位数法,没有考虑各变量与目标变量(换机标识)间的关系,因此在离散化时应选取综合考虑目标变量的最优法进行分箱。调整后继续运行自动分类器节点,如图 8 所示。结合增益和总体精确性两个指标综合考虑,选取决策树 CHAID 算法进一步精准建模,结果如图 9、图 10 所示。

图 11 显示训练集和测试集的正确率均超过 94%,预测换机人数为 45 628 人,实际换机人数为 37 099 人,预测中实际换机人数为 35 002 人,计算命中率为 76.7%,比原始浓度(15%)提升 5 倍,覆盖率为 94.3%,模型效果提升明显。



是否使用?	图形	模型	增益(前 30%)	总体 精确性 (%)
☒		决策列表 1	2.517	95.283
☒		CHAID 1	3.302	94.843
☒		判别 1	2.298	81.948
☒		C&R 树 1	3.108	80.959
☒		C5 1	3.192	76.635
☒		Quest 1	1.997	73.361
☒		贝叶斯网络 1	3.033	71.250
☒		类神经网络 1	3.023	71.168

图 8 自动分类器运行结果

	userid	CHANGE_FLAG	SR-CHANGE_FLAG	SRC-CHANGE_FLAG
1	550112929057	1	0	0.658
2	550113034337	1	1	1.000
3	550113163085	1	1	0.845
4	550113253387	1	1	0.861
5	550113284488	1	1	1.000
6	550113295849	1	1	0.861
7	550113316642	1	1	0.834
8	550113345376	1	1	1.000
9	550113361482	1	1	1.000
10	550113381930	1	1	0.515
11	550113450076	1	0	0.538
12	550113454506	1	1	0.834
13	550113475542	1	1	0.922
14	550113475652	1	1	0.515
15	550113483010	1	1	1.000
16	550113483093	1	1	0.922
17	550113483487	1	0	0.658
18	550113484507	1	1	0.834
19	550113540664	1	1	1.000
20	550113586318	1	1	0.925

图 9 CHAID 模型运行结果

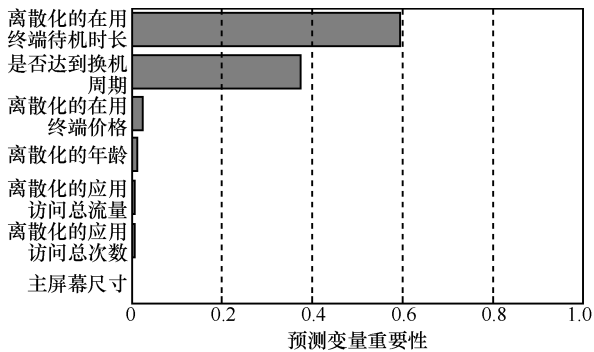


图 10 CHAID 模型运行结果

4 终端推荐模型

本模型利用全网终端价格和性能的聚类细分特征,构建终端价格—性能九宫格,并结合高换机倾向用户的终端

分析

注解

全部折叠(C)

全部展开(E)

■ 输出字段 CHANGE_FLAG 的结果

■ 比较 SR-CHANGE_FLAG 与 CHANGE_FLAG

“分区”	1_训练	2_测试
正确	70,086 94.7%	233,982 94.84%
错误	3,924 5.3%	12,723 5.16%
总计	74,010	246,705

图 11 分析节点结果

九宫格标签和价值标签为其推荐合适的终端。本模型的数据处理与换机倾向识别模型类似,这里不再赘述。

4.1 模型输入

用户基本属性维度:用户唯一标识、地市、是否合约用户、合约到期时间。

终端使用信息维度:现用终端价格,现用终端型号,现用终端上市日期,现用和历史终端品牌、类型,追溯至 $n-3$ 款等。

用户消费行为信息维度:用户的 ARPU 值和流量使用量。

全网终端信息:全网终端主屏幕尺寸(screen_size)、屏幕分辨率(resolution)、电池容量(battery)、主摄像头(camera)、RAM 和 AP 主频(CPU)。

4.2 终端九宫格构建

终端九宫格是将在售和已退市(但仍有人使用)的所有终端按性能和价格分别聚为高中低 3 类,为每款终端赋予九宫格标签。根据终端型号,关联出目标用户终端的九宫格标签。

(1) 构建终端数据库

利用集团终端自注册信息获取全网用户在用终端/历史终端的性能、价格、上市时间等信息。若部分终端的价格和性能信息缺失,可通过网络爬虫工具获取;若终端自注册信息覆盖的终端型号不全,也需通过网络爬虫工具获取该部分终端的价格和性能信息。

(2) 终端价格聚类

先剔除极端值,再利用 k -means 算法将终端聚为 3 类,分别为高价格、中价格和低价格,再将剔除掉的极大值归入高价格类中,极小值归入低价格类中,结果如图 12 所示。

(3) 终端性能聚类

选取最能反映终端性能的主屏幕尺寸、屏幕分辨率、电池容量、主摄像头、RAM 和 AP 主频 6 个指标,利用 k -means 算法将终端聚为高性能、中等性能和低性能 3 类,

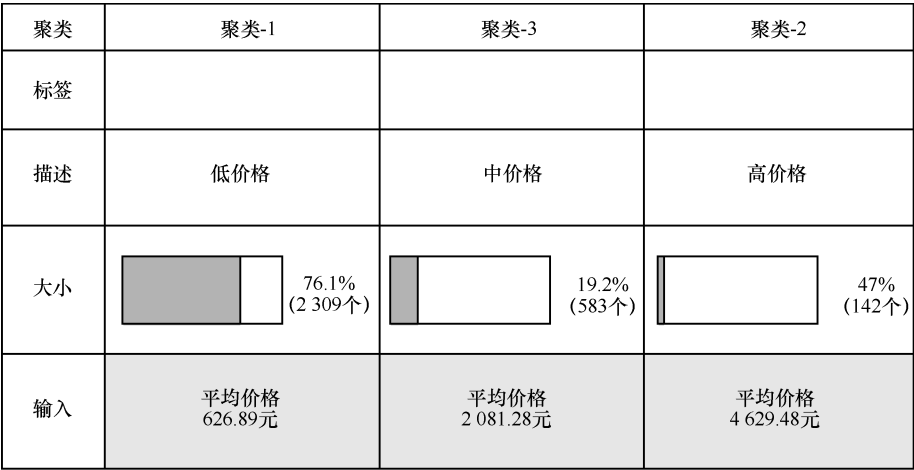


图 12 价格聚类

结果如图 13 所示。

(4)构建终端九宫格

利用 *k*-means 算法, 为每一个终端输出两类标签, 分别是价格标签(高价格、中价格、低价格)和性能标签(高性能、中等性能、低性能)。通过两两组合得到如图 14 所示的 9 个标签, 即每一个终端的最后标签是 9 宫格内的数字。将部分终端归入九宫格内, 结果如图 15 所示。

4.3 用户价值标签

体现用户价值的指标有 ARPU 值、流量使用量(flux)

和语音计费时长等, 但高性能终端对语音计费时长提升不大, 对流量使用量提升较大, 进而提升 ARPU 值。因此在构建用户价值九宫格时, 不考虑语音计费时长, 利用 *k*-means 算法分别对用户 ARPU 值和流量进行聚类, 处理方法与终端价格聚类相同, 结果如图 16 所示。

4.4 用户—终端标签分析

根据终端型号关联目标用户的终端九宫格标签, 分析这些用户的终端和价值信息, 可以看出:

- 大量用户属于低流量、低 ARPU 值, 并且使用低价

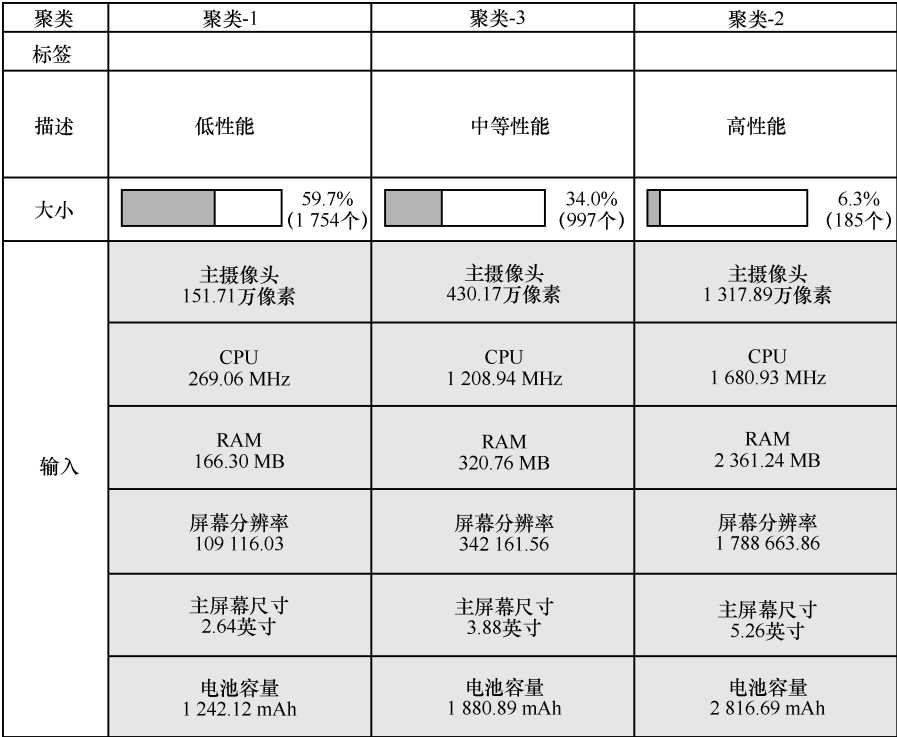


图 13 性能聚类



性能	价格	终端标签
高	高	1
高	中	2
高	低	3
中	高	4
中	中	5
中	低	6
低	高	7
低	中	8
低	低	9

图 14 终端的 9 个标签

格低性能终端(第 9 类);

- 第 5、第 6 类终端用户中,中高流量、中高 ARPU 值的用户可以优先进行终端引导升级;
- 使用第 1 类终端,且高 ARPU 值的用户为高价值用户,需要关怀,可向其推荐高档终端,引导其换机。

4.5 终端推荐迁转路径

基于两个原则设计迁转路径。一是终端性能维度迁移,用户在花销变化不大(同价格段)的基础上,更愿意购

买性能升级的终端;二是终端价格维度迁移,高 ARPU 用户消费水平较高,可向其推荐高价格终端。根据上述原则,结合用户的价值标签设计签转路径如图 17 所示。本文假定要营销的目标终端九宫格标签为 1、2、5、6,故迁转路径的目标只包含这 4 个格子。

4.6 目标终端推荐

首先确定具有换机倾向用户的现用终端九宫格标签,再结合用户的价值标签按照图 17 所示的签转路径确定该用户最适合的终端九宫格标签,应向其优先推荐该格内的目标终端。营销时还要考虑以下两点:

- 品牌忠诚度。若用户有忠诚品牌,则优先推荐该品牌终端;
- 剩余协议时长。优先向剩余协议时长小于 6 个月的用户推荐终端。

5 结束语

本文根据笔者日常工作的实践经验总结了数据预处理的相关方法,并基于中国电信某省公司本地网 50 万用户的终端信息和行为数据,详细阐述了用户换机倾向识别模型和终端推荐模型的构建、评估过程。首先利用换机

性能/价格	高	中	低
高	共54款终端 代表机型: · 华为HW-Huawei MT7 C · HTC-919d · 摩托罗拉Moto-X · 摩托罗拉MOT-XT928 · 三星SCH-I959 · 苹果ACM-A6+ · 小米XMP-2015022 · 宇龙YL-Coolpad 9960 1	共109款终端 代表机型: · HTC-D626d · 华为HW- Honor-7 · 华为HW-GRA · 联想LNV-Lenovo K910 · 魅族MZ-M463C · OPPO OB-OPPO R7c · 三星SCH-SM A5009 · 小米XMP-2013063 2	共28款终端 代表机型: · 华为HW-C2008 · 魅族MZ-M571C · 三星SCH-X979 · 小米XMP-2014910 · 宇龙YL-Coolpad 8297 · 中兴ZTE-B880 · TCL TCL-P606L · 大显DAXIA-M5 3
中	共47款终端 代表机型: · 宇龙Coolpad-N68 · 海信HS-Q9 · HTC-T9 199 · 摩托罗拉MTS-MC659B · 黑莓RIM-BlackBerry 9630 · 三星SCH-B109 · 优派环宇YHY-ViewSonic A · 中兴ZTE-D820 4	共247终端 代表机型: · 华录CHL-EB880 · 创维CW-EG30 · 多普达DPD-S900c+ · 大唐DT-S500C · 长虹GH-ChanghongC20 · 海尔HC-HE N7E · 华为HW-Huawei C8816D · LG-C960 5	共697款终端 代表机型: · 夏新AMO-AMOI E800C · 宇龙Coolpad-219 · 海信HS-EG971 · HTC-D316d · 华为HW-H30 C00 · 华为HW-Huawei Y330 C00 · 联想LNV-i316e · 诺基亚Nokia-2608 6
低	共42款终端 全部为退市老机型,如: · 多普达DPD-S500 · 中光ZTE-C D820 7	共227款终端 几乎全部为退市老机型,如: · 夏新AMO-N7 · 多普达DPD-S505 8	共1 690款终端 几乎全部为退市老机型,如: · 爱宝陵ABL-A6L · 波导BD-Bird AE710 9

图 15 终端九宫格示例

聚类	聚类-1	聚类-3	聚类-2
标签			
描述			
大小	 54.2% (269 891个)	 30.0% (149 159个)	 15.8% (78 900个)
输入	ARPU值 31.23元	ARPU值 116.41元	ARPU值 256.46元

(a) ARPU聚类

聚类	聚类-1	聚类-3	聚类-2
标签			
描述			
大小	 57.6% (286 905个)	 28.2% (140 598个)	 14.1% (70 447个)
输入	流量使用量 116.73 MB	流量使用量 1 317.32 MB	流量使用量 4 814.18 MB

(b) ARPU聚类

图 16 ARPU 聚类 和 流量 聚类

都保留在1	ARPU值中高，移到1	ARPU值高，移到1
	ARPU值低，保留在2	ARPU值中低，移到2
直接迁移到1	ARPU值高，移到1	ARPU值高，移到1
	ARPU值中，流量中高，移到2	ARPU值中，流量中高，移到2
	ARPU低，流量高，移到2	ARPU值中，流量低，移到5
	其余，保留在5	ARPU低，流量高，移到2
		其余，保留在6
直接迁移到1	ARPU值高，移到1	ARPU值高，移到1
	ARPU值中，流量中高，移到2	ARPU值中，流量高，移到2
	ARPU低，流量高，移到2	ARPU值中，流量中，移到5
	其余，移到5	ARPU低，流量高，移到5
		其余，移到6

图 17 迁转路径



倾向识别模型预测出 2016 年 9 月和 10 月具有换机倾向的用户,再结合终端推荐模型给出的这些用户的现用终端九宫格标签、用户价值标签,待营销的目标终端九宫格标签和终端迁转路径向其推荐合适的终端。

参考文献:

- [1] 邓逸斌,朱克隽. 大数据挖掘助力电信运营商终端营销[J]. 中国新通信, 2013, 43(23): 43-44.
DENG Y B, ZHU K J. Terminal marketing promotion based on big data [J]. China New Telecommunications, 2013, 43 (23): 43-44.
- [2] 张勇. 基于大数据挖掘的客户换机倾向评估模型研究 [J]. 数字通信世界, 2016, 144(7).
ZHANG Y. Terminal replaced inclination evaluated based on big data[J]. Digital Communication World, 2016, 144(7).
- [3] 赵一平. 运用数据挖掘技术控制手机客户离网浅析 [J]. 统计科学与实践, 2013(10): 46-47.
ZHAO Y P. Analysis of customer churn based on big data[J]. Zhejiang Statistics, 2013(10): 46-47.
- [4] 谷红勋, 杨珂. 基于大数据的移动用户行为分析系统与应用案例[J]. 电信科学, 2016, 32(3): 139-146.
GU H X, YANG K. Mobile user behavior analysis system and applications based on big data [J]. Telecommunications Science, 2016, 32(3): 139-146.
- [5] 沈超, 黄卫东. 数据挖掘在垃圾短信过滤中的应用 [J]. 电子科技大学学报, 2009, 38(s1): 21-24.

SHEN C, HUANG W D. Application of data mining in short message spam filtering [J]. Journal of University of Electronic Science and Technology of China, 2009, 38(s1): 21-24.

[作者简介]



王雪琼(1987-),女,中国电信股份有限公司上海研究院助理工程师,主要研究方向为数据分析、数据挖掘与建模。



熊珺洁(1983-),女,中国电信股份有限公司上海研究院工程师,主要研究方向为大数据分析、建模、无线网络的可靠性。



姚晓辉(1979-),男,中国电信股份有限公司上海研究院工程师,大数据领域首席技术支撑,主要研究方向为数据规划、数据挖掘、信息管理。