基于机器学习的终端换机预测模型

Prediction Model of Terminal Replacement Based on Machine Learning

欧阳秀平，万源沅，邹俊德（中国联通广东分公司，广东 广州 510627）
Ouyang Xiuping，Wan Yuanyuan，Zou Junde（China Unicom Guangdong Branch，Guangzhou 510627，China）

摘　要：
终端业务不仅可以提升用户收入，还会对用户维系产生影响，对运营商有重要战略意义。通过机器学习等大数据预测技术，可以精准甄别潜在终端换机用户和用户偏好的终端，为运营商终端精准营销提供依据。模型投产以来终端营销转化率由原先的3%提升至4.5%，效果显著。

关键词：
运营商；大数据；终端；机器学习
doi：10.12045/j.issn.1007-3043.2020.04.015
文章编号：1007-3043（2020）04-0075-05
中图分类号：TN919
文献标识码：A
开放科学（资源服务）标识码（OSID）：

Abstract：
Terminal services can not only increase user revenue，but also have an impact on user maintenance，which has important strategic significance for operators. Machine learning and other data prediction techniques can accurately identify potential terminal switching users and user preferences of terminals，and provide a basis for operators' terminal precise marketing. Since the model was put into production，the conversion rate of terminal marketing has increased from 3% to 4.5%，which has achieved remarkable results.

Keywords：
Operators；Big data；Terminal；Machine learning

引用格式：欧阳秀平，万源沅，邹俊德. 基于机器学习的终端换机预测模型[J]. 邮电设计技术，2020（4）：75-79.

1　概述

对于运营商，终端营销既可以为公司带来终端收入，提升用户价值[1]，又可以通过终端合约等维系用户在网，减少用户流失。预测用户换机行为可以帮助运营商向用户精准推荐相关终端活动，实现终端成本资源精准投放，完善全省自有终端运营体系，为5G到来储备终端运营能力。当前终端营销主要通过业务规则等方法筛选目标用户，存在营销成本高、成功率低、无法针对终端市场变化做出快速反应等问题。因此，如何精准预测用户换机行为成为一个亟待解决的问题。

收稿日期：2020-02-20

根据用户换机动机可以将用户的换机行为划分为品牌粉丝换机、常规性换机和偶发性换机3种，如图1所示。

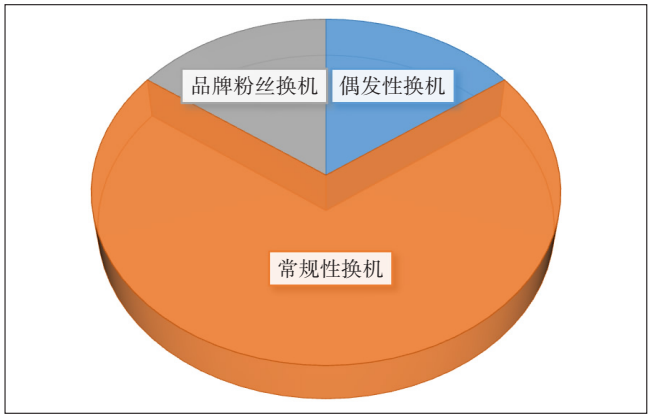


图1　终端换机用户划分

a）品牌粉丝换机是指在品牌新机发布时购买新机，或者因为当前同系列旧款机型降价促销等原因，购买该品牌旧款机型。通过分析用户历史终端购买行为，可识别品牌粉丝用户。针对此部分用户，在该品牌终端新品发布时向其推荐相应的机型，将会有较高的营销转化率。

b）常规性换机是指用户周期性的换机行为，而不是由于新机发布、终端活动或者终端损坏等突发性原因产生的换机行为。通过数据分析和机器学习算法甄别潜在换机用户和用户偏好的终端。本文主要研究分析这部分用户。

c）偶发性换机是指因为手机丢失、损坏，或者因为终端促销活动等外界突发因素影响而产生的换机行为。针对此部分换机用户，目前尚无成熟的逻辑进行预判，此部分用户暂不纳入终端预测模型当中。

为了对常规性换机用户进行精准预测，必须解决以下2个问题。

a）哪些用户需要换机。针对此问题，基于机器学习算法建立模型，预测即将有换机行为的用户，输出用户的换机概率，供业务部门综合考虑触点投放，根据用户换机概率进行营销策略匹配。

b）用户需要换什么机型。决定用户终端选择的最重要的因素是终端品牌和价格，在筛选出潜在换机用户的基础上，进一步预测用户的终端品牌和价格倾向，可以给用户推荐其偏好的终端，提高换机营销转化率，增加公司收入。

本文主要贡献如下：通过对用户终端持有情况和终端换机情况进行数据分析，发现用户终端选择及换机规律等；建立算法模型，预测潜在换机用户和用户倾向的终端，为终端精准营销提供数据基础。

2 终端数据分析

对用户终端持有概况和终端换机概况2个方面的数据进行分析，发现用户换机和终端选择的规律，为后文的模型建设提供数据分析基础。

2.1 终端持有概况分析

根据用户数占比，市场份额排名前3的终端品牌分别为：苹果（24.90%），华为（16.84%）和OPPO（16.82%）。其中，男性更偏好华为手机，女性在OPPO和VIVO手机中占比相对较高（见图2）。用户终端价格主要集中在1 000~2 000和2 000~3 000价格档位（见图3）。苹果手机受到各个年龄段的喜爱，其中18~35岁成年人是苹果手机的主力消费人群，未成年人更倾向于选择VIVO和OPPO手机，中老年人更偏爱华为手机（见图4）。据此，可在终端营销中向年轻女性推荐OPPO和VIVO手机，向中年男性推荐华为手机。

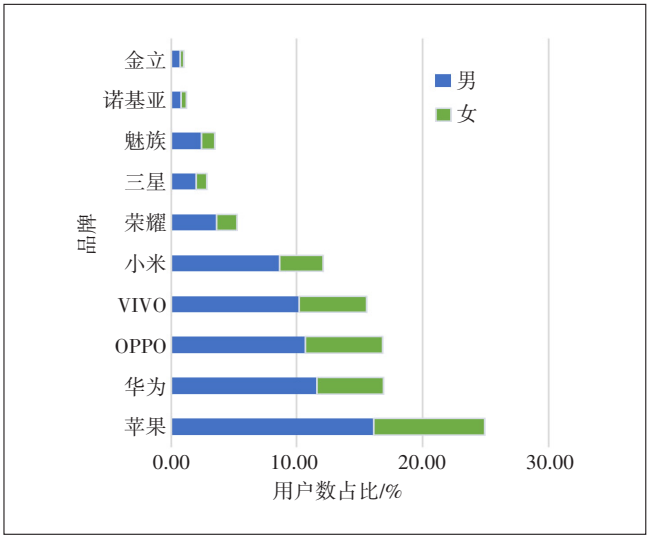


图2 用户终端品牌分布占比图

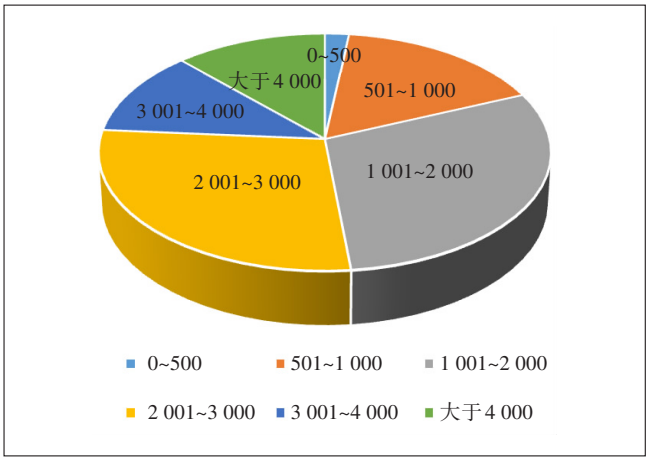


图3 用户终端价格分布占比图

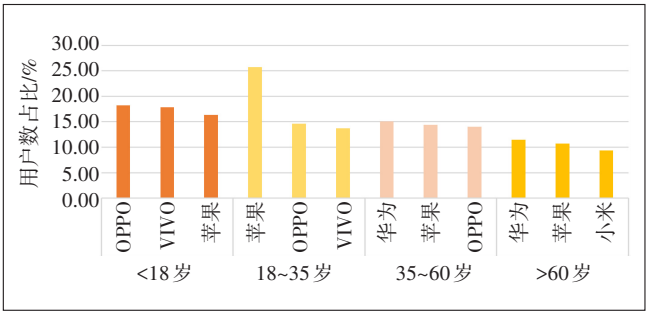


图4 分年龄段TOP3品牌

2.2 终端换机概况分析

用户平均换机时间基本上稳定在21.5个月左右，22岁以下的学生用户平均终端换机周期相对较长（见图5）。终端品牌选择方面，采用保有率、换出率和换入率3个指标刻画用户换机时的品牌选择，用户换机时倾向于选择原先使用的手机品牌，其中，苹果手机保有率最好，50%以上的苹果用户在换机后仍会选择苹果终端。所以对于苹果老用户，可在营销中直接向其推荐苹果手机。苹果、华为这2个品牌的用户存在10%~15%的流动性。OPPO、VIVO和华为的用户存在15%~20%的流动性（见图6）。终端价格方面，用户倾向于选择与原手机同价格档位的手机或者向更高价格档位迁移，很少有用户会选择超低档位的手机（见图7）。

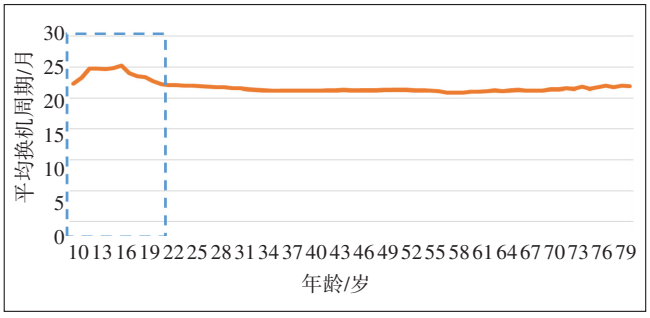
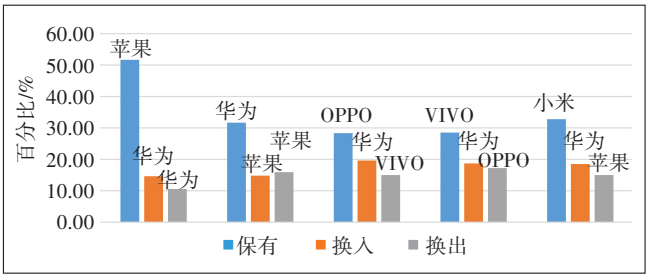


图5　不同年龄平均换机周期图



本图中仅展示该品牌换入和换出用户数最多的品牌。

图6　品牌保有、换入、换出率

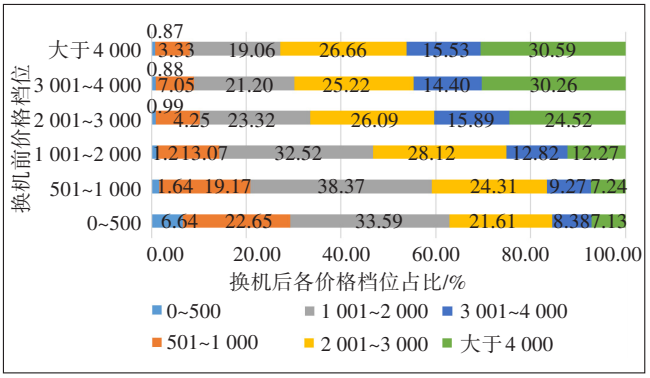


图7　换机前后终端价格变化

3 模型建立与应用

终端换机预测模型主要由数据准备与数据预处理、特征工程、模型训练与验证、模型应用4部分组成（见图8）。

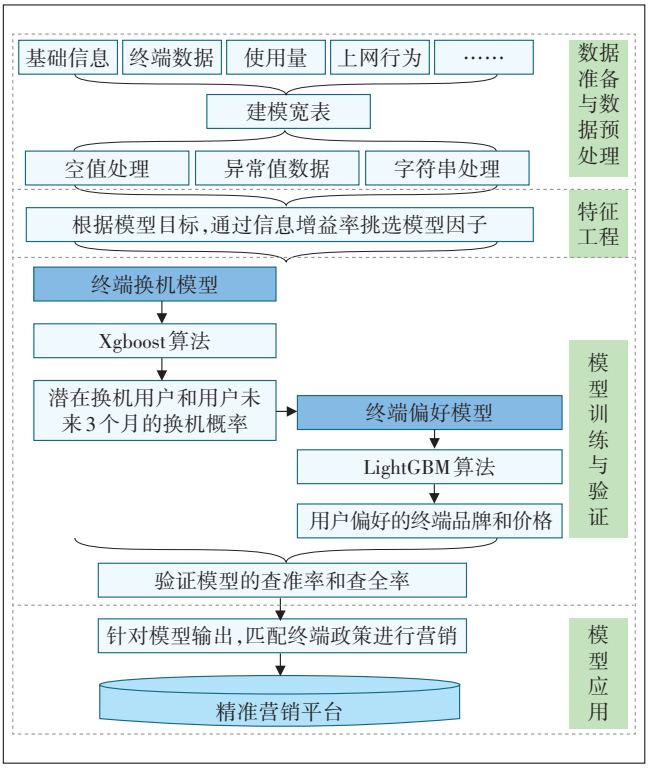


图8　终端预测模型框架图

3.1 数据准备与数据预处理

筛选出与用户终端选择相关的因子，并且参考前文中终端数据分析结论和相关终端换机论文[2]，选取用户基础信息（例如年龄、性别、入网渠道）、终端信息（例如用户上一次换机时间、当前终端厂商、价格、型号、屏幕尺寸等）、使用量数据（例如通话时长、使用流量等）、账务数据（例如用户出账收入等）、上网行为数据（例如购物类APP使用次数、游戏类APP使用次数等），加工成建模宽表。对数据进行以下预处理。

a）空值处理：对于空值达到50%以上的因子，认为该因子数据质量较差，放入模型中会影响模型的判断，进行剔除处理。对于通话次数、使用流量等数值为空的数据，经核查确认该用户没有话单或者流量详单，则将空值改为0。对于年龄等基础信息类因子为空的值，用中位数填充。对于终端型号无法解析或者部分终端参数缺失的信息用“其他”代替[3]。

b）异常值处理：对于不符合业务常识的数据，例如性别中除了男、女之外的其他记录，用“未知”进行填充，例如年龄大于100岁的用户，用“未知”进行填充；对于通话、流量等数值过大的数据，用均值+标准差替代[4]。

c）字符串处理：对字符型变量进行one-hot-encoding编码转化。

3.2 特征工程

文献[5]采用信息增益率挑选因子，信息增益率越大说明包含的可供分类决策的信息越多，信息增益率的计算过程如下。

步骤1：计算信息增益。

信息增益表示由于已知特征X的信息而致使Y的信息不确定性减少的程度。假定特征A对训练数据集D的信息增益为$g(D,A)$，根据定义其值为集合D的熵$H(D)$与特征A给定条件D下的条件熵$H(D|A)$之差。

$$g(D,A)=H(D)-H(D|A) \tag{1}$$

数据集D的熵$H(D)$的定义如下：

$$H(D)=-\sum_{m=1}^{M}\frac{|C_m|}{|D|}\log_2\frac{|C_{ml}|}{|D|} \tag{2}$$

计算特征A对于训练集D的条件熵$H(D|A)$，如式(3)所示。

$$H(D|A)=\sum_{i=1}^{n}\frac{|D_{il}|}{|D|}H(D_i)=-\sum_{i=1}^{n}\frac{|D_{il}|}{|D|}\sum_{m=1}^{M}\frac{|D_{im}|}{|D_i|}\log_2\frac{|D_{im}|}{|D_i|} \tag{3}$$

其中，$|D|$为样本大小，假设有M个类C_m，$m=1,2,\cdots,M$。$|C_m|$为属于类C_m的样本个数。设变量A的取值有n个，根据变量A的取值把集合D划分为n个子集$D_1,D_2,\cdots,D_n$。D_{im}为子集D_i中属于类C_m的集合。

步骤2：计算信息增益率。

信息增益率为特征A对训练数据集D的信息增益，如式(4)所示。

$$g_R(D,A)=\frac{g(D,A)}{H_A(D)} \tag{4}$$

其中，$H_A(D)=-\sum_{i=1}^{n}\frac{|D_i|}{|D|}\log_2\frac{|D_i|}{|D|}$，$n$是特征$A$的数量。

根据信息增益率筛选出平均换机时长、上一次换机时间、终端品牌、终端价格、电池容量、屏幕尺寸、出账收入等20个因子放入模型。

3.3 模型训练与验证

通过机器学习算法建立终端换机和终端偏好模型。终端换机模型用于预测用户在未来3个月的换机概率，终端偏好模型用于预测用户换机时品牌和价格档位的选择倾向。通过使用决策树、逻辑回归、神经网络等单模型算法和随机森林、lightGBM、Xgboost等集成学习算法进行训练，综合考虑模型的查全率和查准率，选取效果最优的算法。终端换机模型最终选择Xgboost算法进行训练，终端偏好模型最终选择lightGBM算法。为了保证训练的准确性和稳定性，防止过度拟合，对模型进行如下处理。

a）样本平衡：因为换机用户仅占训练样本的一小部分（约6%），如果不进行样本平衡处理，模型预测将倾向于在训练样本中占比较大的一类。针对此问题，对于负样本（非换机用户）进行欠采样处理[6-10]，并且调整正负样本比例（在1∶1~1∶4），根据模型结果选择最优的正负样本比例，本文最终采用正负样本比例为1∶2。

b）交叉验证：这里用10折交叉验证[11]，即将数据集分为10份，轮流将其中9份作为训练数据，1份作为测试数据，进行训练，综合之后，使模型具有较高的准确性和稳定性[12]。

c）列采样：对模型进行列采样，从M个因子中随机选取m个（$m<M$），避免模型在个别因子上产生过度拟合[13]。

d）剪枝：根据样本数和因子数设置剪枝规则，设置通过最末端叶子节点的最小样本数为20，这样既保证了模型的准确性，又避免出现过度拟合。

这里引入查准率和查全率来评估模型[14]，查准率用于衡量模型的准确性，查全率用于衡量模型的覆盖率。以换机模型为例，查准率和查全率的定义如下，终端偏好模型的查全率和查准率计算公式以此类推。

查准率=预测换机且实际换机的用户数/预测换机的用户数

查全率=预测换机且实际换机的用户数/实际换机的用户数

通过平移时间窗口的方法，计算测试时间内模型的查准率和查全率，据此评估模型的稳定性。最终统计出来终端换机模型查准率为50%左右，查全率为40%左右，每月预计输出数据量为80万左右；在换机模型的基础上，终端偏好模型的查全率查准率均为40%，模型的准确率和覆盖率均较好。

3.4 模型应用

针对模型输出的潜在换机用户，根据用户换机后

品牌和价位选择倾向，匹配相应的终端政策（见表1），将结果反馈给精准营销平台，对用户进行精准营销。模型应用之后，终端营销转化率由原先的3%提升至4.5%，模型应用效果显著。

表1 终端预测模型输出及营销策略匹配示例表

模型输出				策略匹配	
用户号码	换机概率	品牌偏好	价格偏好	匹配终端	终端政策
186××××××××	0.95	苹果	3 000~4 000	iPhone 7/ iPhone 6s plus	存费送机
166××××××××	0.57	华为	2 000~3 000	华为P20	购机送费

4 运营商终端大数据预测展望

技术方面，随着用户终端类数据的积累，可以使用协同过滤等推荐系统算法对用户偏好的终端型号和终端活动进行预测和精准推荐。此外，用户的换机行为会随着当前终端市场不断变化，因此模型需具备自迭代框架，对数据预处理、特征选择、算法选择、模型训练等流程进行自动化能力封装，这样才可以适应不断变化的市场，自动调优。

业务方面，5G即将来临，移动终端的形态正在经历变革，智能手环、手表等可穿戴设备正不断涌现[15]。泛终端的发展将是未来运营商终端营销的重要着力点，未来可深入挖掘泛终端用户特征，对泛终端潜在用户进行精准预测，抢占5G终端市场。

5 总结

终端业务既可以为公司带来终端收入，还会对用户维系产生影响，对运营商具有重要意义。本文通过数据分析寻找用户终端选择和换机规律，基于机器学习等大数据预测技术精准预测用户在未来3个月换机概率和终端选择倾向。模型投产后终端营销转化率由原先的3%提升至4.5%，应用效果显著。后续将进一步优化模型预测技术，开发模型自迭代框架，实现模型自动调优，以适应不断变化的终端市场，并且将终端分析及预测技术投入泛终端领域中，在5G时代，为泛终端营销提供决策依据。

参考文献：

[1] 刘力凯，王国胤，邓维斌. 优势关系粗糙集的移动用户换机预测方法[J]. 小型微型计算机系统，2015，36(8)：1789-1794.
[2] 张灵明. 品牌忠诚及其影响因素研究[D]. 厦门：厦门大学，2006.
[3] 曹林. 基于统计学习的数据预处理缺失值清洗方法研究[D]. 哈尔滨：哈尔滨工程大学，2011.
[4] 吴翌琳，房忠祥. 大数据探索性分析[M]. 北京：中国人民大学出版社，2016：16-24.
[5] 李虹利，蒙祖强. 运用信息增益和不一致度进行填补的属性约简算法[J]. 计算机科学，2018，45(10)：224-231.
[6] CHAWLA N V，BOWYER K W，HALL L O，et al. SMOTE：synthetic minority over-sampling technique [J]. Journal of Artificial Intelligence Research，2011，16(1)：321-357.
[7] BARUA S，WASLAM M M，YAO X，et al. Majority Weighted Minority Oversampling Technique for Imbalanced Data Set Learning [J]. IEEE Transactions on Knowledge & Data Engineering，2013，26(2)：405-425.
[8] KUBAT M，MATWIN S. Addressing the Curse of Imbalanced Training Sets：One-Sided Selection [C]// International Conference on Machine Learning. 2012.
[9] YEN S J，LEE Y S. Cluster-based under-sampling approaches for imbalanced data distributions [J]. Expert Systems with Applications，2009，36(3)：5718-5727.
[10] NIKULIN V，MCLACHLAN G J. Classification of Imbalanced Marketing Data with Balanced Random Sets [J]. Journal of Machine Learning Research，2009(7)：89-100.
[11] 吴喜之. 复杂数据统计方法[M]. 北京：中国人民大学出版社，2015：43-48.
[12] WU X，KUMAR V，QUINLAN J R，et al. Top 10 algorithms in data mining[J]. Knowledge and Information Systems，2008，14(1)：1-37.
[13] 范明，范宏建. 数据挖掘导论[M]. 北京：人民邮电出版社，2014：100-119.
[14] 周志华. 机器学习[M]. 北京：清华大学出版社，2016.
[15] 宋春涛，张帆，曹振. 电信运营商的终端大数据分析及应用[J]. 邮电设计技术，2016(8)：41-46.
[16] 王雪琼，熊珺洁，姚晓辉. 基于大数据挖掘的终端换机模型[J]. 电信科学，2016，32(12)：43-52.
[17] 杨盾. 基于手机用量分析的换机预测模型研究[D]. 南京：南京财经大学，2017.
[18] 张鹏，程乔，韦亮，等. 基于客户业务行为的潜在换机客户挖掘模型研究与应用[J]. 信息通信技术，2017，11(5)：51-57.
[19] 张志勇. 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.
[20] 战培志，关芳芳. 电信运营商大数据应用实践探析[J]. 江苏通信，2018(34)：92-94.
[21] 王洋，何阳. 电信运营商大数据价值转化与应用策略研究[J]. 信息通信技术与政策，2018(1).

作者简介：
欧阳秀平，高级工程师，硕士，中国联通广东分公司信息化部总经理，中国联通广州软件研究院院长；万源沅，工程师，硕士，主要从事用户行为方向大数据建模，机器学习算法研究工作；邹俊德，工程师，硕士，主要从事用户维系方向大数据建模，机器学习算法研究工作。