基于GBDT算法的潜在5G用户预测研究与实现

Study and Implementation of Potential 5G User Prediction Based on GBDT Algorithm

陈　锋，李张铮，庄毅莹（中国联通福建省分公司，福建 福州 350000）
Chen Feng，Li Zhangzheng，Zhuang Yiying（China Unicom Fujian Branch，Fuzhou 350000，China）

摘　要：
5G用户规模发展是新时代新基建大背景下运营商5G网络建设的终极目标。传统的通过人工方式进行营销发展存在诸多不足，浪费大量人力物力财力。针对这些缺点，基于运营商O域和B域数据引入梯度提升决策树（GBDT）分类算法，通过学习存量5G用户正负样本在历史网络上产生的出账数据和网络数据建立5G用户分类预测模型，做到精准挖掘5G潜在用户，提升市场营销的命中率。研究结果表明，基于GBDT算法的潜在5G用户预测模型能有效预测5G目标用户，提高5G用户转化率，对5G用户发展起到积极推动作用。

关键词：
潜在5G用户预测；机器学习；GBDT算法；5G用户画像；5G用户营销
doi：10.12045/j.issn.1007-3043.2021.04.010
文章编号：1007-3043（2021）04-0045-05
中图分类号：TN929.5
文献标识码：A
开放科学（资源服务）标识码（OSID）：

Abstract：
The development of 5G user scale is the ultimate goal of operator 5G network construction in the new era of new infrastructure. There are many shortcomings in the traditional marketing development through artificial way，which wastes a lot of human and material resources. In view of these shortcomings，based on the operator O-domain and B-domain data，it introduces gradient boosting decision tree（GBDT） classification algorithm，by learning the stock of 5G user positive and negative samples on the historical network of accounting data and network data，it establishes 5G user classification prediction model，so as to accurately mine 5G potential users，and improve the marketing hit rate. The results show that the potential 5G user prediction model based on GBDT algorithm can effectively predict 5G target users，improve the conversion rate of 5G users，and actively promote the development of 5G users.

Keywords：
Potential 5G user prediction；Machine learning；GBDT algorithm；5G user portrait；5G user marketing

引用格式：陈锋，李张铮，庄毅莹. 基于GBDT算法的潜在5G用户预测研究与实现[J]. 邮电设计技术，2021（4）：45-49.

0 引言

随着国家5G新基建时代的来临，5G移动用户规模发展带来的高流量高收益成为当下及今后运营商收入的主要来源。运营商移动网络5G用户传统营销方式较为粗放，主要体现在5G用户营销策略和定位不够清晰；5G用户目标缺乏针对性；营销成功与否和5G营销人员的营销水平相关；事前没对用户进行有效的筛选，营销成功率低；已有的传统网络用户迁转到5G过程中形成的历史数据没有得到利用。如何规避上述问题，精准有效地推动传统移动网络用户向5G转化成为业界研究的热点方向。作为电信运营商的优势之一，多年的包含日常运营过程中形成的B域和O域的大数据集可以用来对5G用户进行画像，通过大数据手段充分挖掘这些数据中包含的用户基础信息、用户消费信息、用户上网行为偏好和用户网络感知等能够为5G用户智能营销开辟新的方向的信息。

作为人工智能的重要组成部分，机器学习技术是

收稿日期：2021-02-23

国家发展战略重点扶持的目标[1]，也是当下各行业关注的焦点。为了推动传统5G用户营销方式的数字化，提升网优专业5G市场支撑智能化水平，有必要对基于机器学习算法的潜在5G用户预测进行研究。

1 移动网络5G用户传统营销方式的痛点

移动网络传统用户营销方法存在诸多短板，比如营销策略模糊、目标用户存在盲目性、营销成效与人员水平相关等。

1.1 5G用户营销策略和定位不够清晰，没有形成差异化营销

受到长期传统标准化大生产经验的影响，运营商在制定5G用户营销策略时往往是一刀切，对所有用户采用统一的口径和指标做营销宣传，没有考虑用户个体差异性；但实际上5G敏感用户始终比不敏感用户容易发展，对2类用户不加区分地采用相同营销手段容易造成参差不齐的营销结果。

1.2 5G用户目标不精准，营销目标对象与实际结果存在偏差

由于5G用户营销数据的局限性和分析方法不当，运营商在发展5G用户时没能形成5G用户特征评估体系，未能对5G用户进行精准画像，导致常规方法评估出来的目标用户与实际营销结果偏差较大，浪费不必要的人力物力。

1.3 5G用户营销成效与营销人员主观水平相关，降低了营销效率

在现场营销或代理商营销场景中，营销人员只能通过个人主观判断该用户是否是潜在5G用户，缺乏客观的评估手段，不同营销水平的人员营销结果千差万别，判断能力不强的人员消耗了不必要的时间在5G不敏感用户上，降低了营销效率。

2 基于GBDT算法预测潜在5G用户

随着5G网络规模的不断扩大，运营商越来越需要进行精准的5G用户营销来拉动收入。影响传统移动网络用户转化为5G用户的因素很多，其中用户基本属性、用户消费信息、用户上网行为偏好和用户网络感知是影响用户转化为5G用户的最核心因素，充分挖掘这些数据有利于指导5G用户营销。

本文通过利用GBDT机器学习算法学习5G用户正负样本历史上的B域出账数据和O域网络数据，建立5G用户分类预测模型预测出传统移动网络用户是否是潜在5G用户。该模型可在5G用户营销支撑、5G网络感知保障等网优日常工作中起到积极作用。

2.1 GBDT分类算法原理概述

GBDT分类算法属于集成学习中的Boosting方法。Boosting方法使用多个弱基分类器，训练基分类器时采用串行的方式，每个基分类器之间有依赖，它的基本思路是将基分类器一个个叠加，每个基分类器在训练的时候，对前一个基分类器分错的样本，给予更高的权重。测试时，根据各个分类器的结果加权得到最终结果。GBDT的原理就是所有弱分类器的结果相加等于预测值，然后下一个弱分类器去拟合误差函数对预测值的残差（残差就是预测值与真实值之间的误差），其中弱分类器的表现形式就是各棵决策树。该算法具体原理如下[2]：

假设输入训练集样本$D=\{(x_1,y_1),(x_2,y_2),\cdots,(x_m,y_m)\}$，最大迭代次数$T$，损失函数$L(y,f(x))=\log(1+exp(-yf(x)))$，其中$y\in\{-1,+1\}$。输出是强学习器$f(x)$。

a）初始化弱学习器：$f_0(x)=\arg\min_c\sum_{i=1}^{m}L(y_i,c)$。

b）对迭代次数$t=1,2,\cdots,T$，有：

（a）对样本$i=1,2,\cdots,m$计算负梯度误差：

$$r_{ti}=-\left[\frac{\partial L(y,f(x_i))}{\partial f(x_i)}\right]_{f(x)=f_{t-1}(x)}=\frac{y_i}{1+exp(y_i f(x_i))}$$

（b）利用$(x_i,r_{ti})(i=1,2,\cdots,m)$，拟合一棵CART回归树，得到第$t$棵回归树，其对应的叶子节点区域为$R_{tj}$（$j=1,2,\cdots,J$），其中$J$为回归树$t$的叶子节点个数。

（c）对叶子区域$j=1,2,\cdots,J$，计算最佳负梯度拟合值：

$$c_{tj}=\arg\min_c\sum_{x_i\in R_{tj}}\log(1+exp(-y_i(f_{t-1}(x_i)+c)))\approx\frac{\sum_{x_i\in R_{tj}}r_{ti}}{\sum_{x_i\in R_{tj}}|r_{ti}|(1-|r_{ti}|)}$$

（d）更新强学习器：$f_t(x)=f_{t-1}(x)+\sum_{j=1}^{J}c_{tj}I(x\in R_{tj})$

c）得到强学习器f(x)的表达式：

$$f(x)=f_T(x)=f_0(x)+\sum_{t=1}^{T}\sum_{j=1}^{J}c_{tj}I(x\in R_{tj})$$

2.2 训练集和测试集样本生成

2.2.1 样本的采集

提取某省联通2020年3月份5G用户46 170个和等量的非5G用户生成正负样本标签，5G用户作为正样本标记为1，非5G用户作为负样本标记为0。样本字段都是用户在传统网络（3G/4G）用户时的历史数据，这些原始字段包含B域的用户基础信息和用户消费信息、O域的用户上网行为和用户网络感知KQI指标（见表1）。

表1　5G用户正负样本原始字段

5G用户样本原始字段	用户基础信息								
	用户号码	账期	年龄	性别	归属城市	套餐名称	入网时间	终端厂家	单双卡终端
	用户消费信息								
	出账收入	语音通话时长	数据流量	语音通话时长2G	语音通话时长3G	流量2G	流量3G	流量4G	
	用户上网行为								
	最大使用APP协议大类			最大使用APP流量占比			前5名的APP流量占比		
	用户网络感知KQI指标								
	页面响应成功率			视频流媒体初始播放成功率			视频流媒体有效下载速率		

这些原始字段中，用户基础信息使用2019年8月份的当月数据（2019年8月份开始5G放号）；用户消费信息使用当月及前3个月的数据；用户上网行为使用当月数据，其中最大使用APP指的是当月用户产生最大流量的APP；用户网络感知KQI指标是用户当月每天流量最高的10个小区的KQI指标值汇总，形成每天的KQI指标字段。

2.2.2 样本划分为训练集和测试集

机器学习一般将样本划分为训练集和测试集，训练集用于模型训练，测试集用于测试模型性能。本文利用scikit-learn的train_test_split()函数将样本划分为训练集和测试集，其中参数测试集比例test_size取0.2，即训练集和测试集比例为8∶2。

2.3 数据预处理

数据预处理主要是检查每个特征是否有缺失值或非法字符，对不合理的值进行校正替换，对类别值过多的高基数类别特征进行降基处理，类别特征不平衡字段需重新归并。检查样本数据发现，数值型特征的用户消费信息存在缺失值，比如语音通话时长、流量字段；类别型特征的性别、终端厂家等字段存在缺失值，对这些列调用scikit-learn的SimpleImputer对象进行均值填充；有609个类别特征套餐名称值和204个终端厂家值存在高基数问题，需要降基处理，这里根据特征的分布情况使用pandas的分箱操作cut()方法对高基数特征进行分段编码[3]；归属地（市）、最大APP协议大类存在特征取值不均衡问题，对比例较低的类别值重新归并。

2.4 特征工程

特征工程是机器学习过程的重要环节，样本特征的好坏决定了机器学习性能的上限，而模型只是逼近这个上限而已。特征工程的主要内容包括特征构造、特征抽取和特征选择[4]。本文的原始特征包括B域的用户基础信息和用户消费信息、O域的用户上网行为和用户网络感知KQI指标共100多个维度。为了满足特征选择的需要，在此先进行特征构造和特征抽取，最后进行特征选择，避免过高的特征维数导致模型过拟合。

2.4.1 特征构造

原始字段中的入网时间是Object类别特征，无法进行数值计算提取有效信息。本文通过设置一个标杆时间2020年12月来构造用户从入网到标杆时间的在网月数特征。

2.4.2 特征抽取

用户网络感知KQI共一个月（30天）的数据，每天有页面响应成功率、视频流媒体初始播放成功率、视频流媒体有效下载速率3个指标，总计有90个维度的特征。数据特征维度太高，首先会导致计算很麻烦，其次增加了问题的复杂程度，分析起来也不方便。但盲目减少数据的特征会损失数据包含的关键信息，容易导致模型预测性能下降。主成分分析（PCA——Principal Component Analysis）降维方法，既减少了需要分析的指标，又尽可能多地保持了原来数据的信息。本文使用scikit-learn的PCA估计器对KQI数据进行降维，由于不确定具体变换的合适维数，就取PCA的n_components参数为0.95，即变换后的结果保留95%的原始信息，计算后维数降至67。将67维的PCA分量与目标列做相关性分析，最相关的是第1个分量kqi_data_pca_0相关系数0.14，后续只采纳该分量进行训练。

2.4.3 特征/目标相关性分析

特征选择不仅具有减少特征数量（降维）、减少过拟合、提高模型泛化能力等优点，而且还可以使模型获得更好的解释性，增强对特征和特征、特征和目标之间关系的理解，加快模型的训练速度获得更好的预

测性能。此处采用pandas的相关系数计算函数corr()来分析特征和目标间的相关性(见表2)。

表2　部分特征和目标间的相关系数值

目标	特征	特征与目标间的相关系数
是否5G用户	归属地(市)_福州	−0.658 375
	倒数第2个月4G网络流量	0.081 093
	倒数第2个月数据流量	0.082 157
	倒数第2个月3G网络语音通话时长	0.082 159
	倒数第2个月语音通话时长	0.083 254
	倒数第3个月3G网络语音通话时长	0.083 726
	倒数第3个月4G网络流量	0.084 020
	倒数第3个月语音通话时长	0.084 547
	3G网络语音通话时长	0.084 618
	倒数第3个月数据流量	0.085 101
	上月3G网络语音通话时长	0.085 993
	语音通话时长	0.087 054
	上月语音通话时长	0.087 534
	性别_AE557AA113ADF6F9	0.096 392
	最大使用APP协议大类_Streaming	0.118 620
	倒数第3个月出账收入	0.120 133
	倒数第2个月出账收入	0.121 612
	单双卡终端_双卡	0.123 278
	出账收入	0.126 348
	上月出账收入	0.129 829
	kqi_data_pca_0	0.139 687
	归属地市_泉州	0.419 942
	归属地市_厦门	0.428 678

由于部分特征间的相关性过高，将造成特征间的多重共线性，影响模型效果，这里剔除相关系数大于0.8的特征，保留与目标相关性最大的特征。

2.5 模型训练

2.5.1 基于交叉验证的分类预测模型选择

机器学习中常用的分类预测模型有逻辑回归、KNN、朴素贝叶斯、随机森林、GBDT和XGBoost等。这里分别使用这些模型进行5折交叉验证打分，评估标准为正确率accuracy，选出最好的模型。实验结果表明，最佳模型为GBDT，平均cross_val_score得分最高为0.814(见图1)。后续就使用GBDT模型进行建模训练。

2.5.2 基于随机搜索的GBDT模型超参数优化

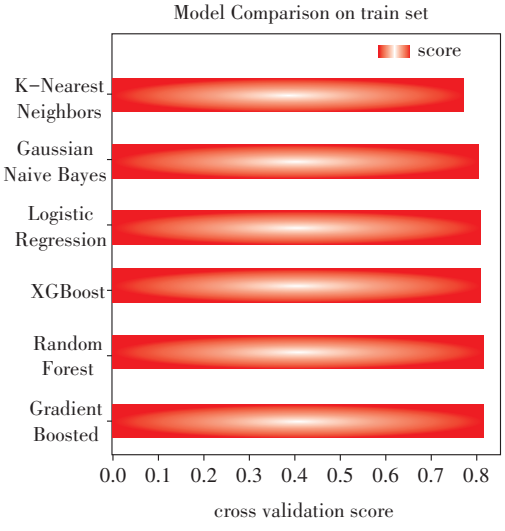
GBDT模型的超参数分2类：第1类是Boosting框架的重要参数，调节模型中boosting的操作，主要包括n_estimators、learning_rate和subsample，第2类是弱学习器即CART回归树的重要参数，调节模型中每个决策树的性质，主要包括max_depth、min_samples_split、min_samples_leaf和max_features等[5]。

本文利用scikit-learn库自带的RandomizedSearchCV随机搜索算法来调整GBDT算法超参数，候选超参数值集合如下：

learning_rate =[0.005,0.01,0.05,0.1]

n_estimators =[100,400,800,1000]

subsample =[0.5,0.6,0.7,0.8]




```
def evaluate_model_performance(df_features,df_target,model_estimators_lst):
    model_scores_lst=[]
    for estimator in model_estimators_lst:
        model_score=cross_val_score(estimator,df_features,df_target,cv=5,scoring='accuracy')
        print('current model score is:{0}'.format(model_score))
        model_score_lst.append(model_score)
# Dataframe to hold the results
    model_comparison=pd.DataFrame('model':['Logistic Regression',
                                           'Random Forest',
                                           'Gradient Boosted',
                                           'Gaussian Naive Bayes',
                                           'K-Nearest Neighbors',
                                           'XGBoost'],
                                  'score':[model_scores_lst[0],model_scores_lst[1].mean(),
                                           model_scores_lst[2],model_scores_lst[3].mean(),
                                           model_scores_lst[4],model_scores_lst[5].mean(),])
 # Horizontal bar chart of test mae
    model_comparison.sort_values('score'ascending=False).plot(x='model',y='score',
                                           kind='barh', color='red',edgecoler='black')
# Plot formatting
plt.ylabel('');plt.yticks(size=14);plt.xlabel('cross validation score');plt.xticks(size=14)
plt.title('model Comparison on train set',size=20);
```

图1　基于交叉验证的分类模型选择

min_samples_split =[500,700,900,1100]
min_samples_leaf =[100,200,300,400]
max_depth =[5,10,15,20]
max_features=[13,20,27,34]

最终搜索得到的最佳超参数组合是：{'subsample'：0.6，'n_estimators'：400，'min_samples_split'：1100，'min_samples_leaf'：300，'max_features'：13，'max_depth':5,'learning_rate':0.01}。在测试集上进行评估，分类正确率acurracy为0.808，召回率0.632。

2.5.3 基于GBDT分类模型的潜在5G用户预测

运营商可根据5G用户GBDT分类模型特征字段采集数据，构成样本输入模型对潜在5G用户进行预测。实验结果表明，现网5G用户预测命中率为71%，即真实5G用户中有71%被模型预测出来。

3 5G用户预测模型在现网中的应用

从2020年4月份开始收集某市联通全网3G/4G用户的B域和O域数据进行5G用户预测，将预测出的5G用户清单交市场部进行5G精准营销。市场部反馈营销结果及建议给项目组，项目组人员根据实际结果修正训练数据的特征，重新进行样本建模学习，整个流程不断闭环迭代开发，提高预测的命中率（见图2）。

2020年4月前按每月营销目标人数6万计算，平均每月营销成功的5G用户数约为3 335人，占营销用户总数的5.56%，即营销成功率为5.56%；在开始使用5G用户预测模型后，平均每月营销成功的5G用户数约为14 659人，营销成功率提升至24.43%，每月多发展5G用户11 324人（见图3）。按每用户月平均ARPU值50元计算，2020年4月份、5月份、6月份3个月共增加收入339万元。

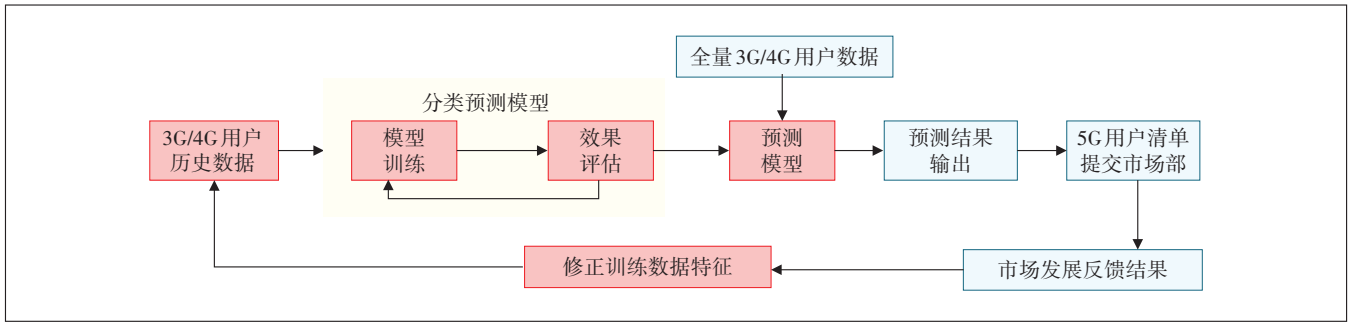


图2　5G用户预测项目运行环节流程

4 总结

5G用户传统营销方式存在诸多痛点，人工标准化营销费时费力。通过引入机器学习算法学习5G用户正负样本历史出账数据和网络数据，建立分类预测模型，可精准预测全网潜在的5G用户，解决了5G时代用户规模发展的困境，极大程度地提高了5G用户营销的成功率。

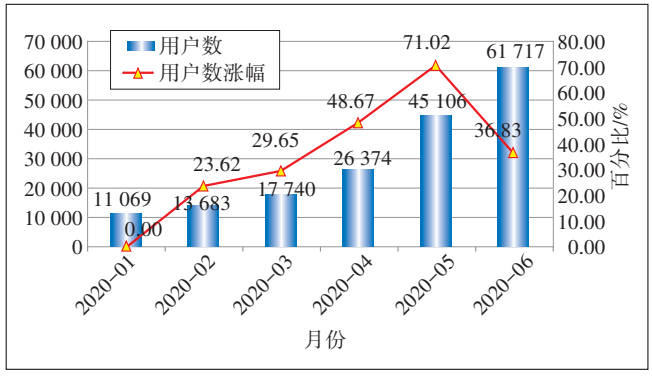


图3　使用5G用户预测模型前后用户数增长情况

参考文献：

[1] 工业和信息化部．促进新一代人工智能产业发展三年行动计划（2018-2020）[EB/OL]．[2020-12-22]．https://www.sohu.com/a/210606521_99964548.
[2] 刘建平．梯度提升树（GBDT）原理小结[EB/OL]．[2020-12-07]．https://www.cnblogs.com/pinard/p/6140514.html.
[3] Wes McKinney．利用Python进行数据分析[M]．2版．北京：机械工业出版社，2018：199-201.
[4] 周志华．机器学习[M]．北京：清华大学出版社，2016：225-266.
[5] Ensemble methods[EB/OL]．[2020-12-07]．https://scikit-learn.org/stable/modules/ensemble.html#gradient-tree-boosting.

作者简介：
陈锋，毕业于福州农林大学，高级工程师，主要从事无线网络优化工作；李张铮，毕业于大连理工大学，工程师，主要从事无线网络优化工作；庄毅莹，毕业于福州大，工程师，主要从事数据支撑工作。