ISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 35,No. 5,Sep. 2020,pp. 842－849
DOI:10. 16337/j. 1004-9037. 2020. 05. 005

http://sjcj. nuaa. edu. cn
E-mail:sjcj@nuaa. edu. cn
Tel/Fax:＋86-025-84892742

基于特征工程的广告点击转化率预测模型

邓秀勤[1]，谢伟欢[2]，刘富春[3]，张翼飞[1]，樊　娟[1]

(1. 广东工业大学应用数学学院，广州，510520;2. 北京明略软件系统有限公司，广州，510300;3. 广东工业大学计算机学院，广州，510006)

摘　要：在大数据环境下，随着全球网络广告传播行业的快速发展，网络广告的计算也越来越受到人们的高度关注。计算广告旨在将广告投放到特定的受众人群，以广告环境和用户特征为基础进行数据分析计算，从候选广告库中选择出最佳匹配的广告。其核心问题是通过网络广告点击转化率预测的计算，将用户点击可能性最高的广告选择出来。广告点击转化率的精确预测与媒体、广告主和用户3方的利益密切相关。该研究基于TrackMaster平台提供的真实广告数据，以特征工程的视角，分别从用户信息特征、广告信息特征、上下文特征和统计特征4个角度进行特征分析，从而挖掘出对广告点击转化率影响较大的重要特征，构建广告点击转化率预测分层模型并训练，并且结合LightGBM算法模型得出广告点击转化率的重要特征排序。实验结果表明当特征选择阈值$\lambda=0.95$，特征选择数目为19，树的颗数为100时的受试者工作特征曲线下的面积(Area under receiver operating characteristic curve，AUC)值最大，模型的对数损失函数值约为0.136 8，此时模型具有最优的效果。预测模型和特征排序结果有助于企业制定最优的广告投放策略。

关键词：数据分析；点击转化率；计算广告；特征工程；特征分析

中图分类号：TP274　　**文献标志码**：A

A Prediction Model for Advertising Click Conversion Rate Based on Feature Engineering

DENG Xiuqin[1]，XIE Weihuan[2]，LIU Fuchun[3]，ZHANG Yifei[1]，FAN Juan[1]

(1. School of Applied Mathematics，Guangdong University of Technology，Guangzhou，510520，China;2. Mininglamp Technology，Guangzhou，510300，China;3. School of Computers，Guangdong University of Technology，Guangzhou，510006，China)

Abstract：Under the environment of big data，with the rapid expansion of the online advertising industry，the online advertising calculation has attracted more and more attention. Computational advertising aims at placing ads on a specific audience，performs data analysis and calculation based on the advertising environment and user characteristics，and selects the best matching ad from the candidate ad library. The core issue is the calculation of click conversion rate prediction for online advertising，which selects the ads with the highest probability of users clicking. The accurate prediction of advertisement click conversion rate is related to benefits of publishers，advertisers and users. Based on the advertising data provided by the TrackMaster platform，this study analyzes user information features，advertising information features，context features and statistical features from the perspective of feature engineering. The larger effects on the advertising click conversion characteristics are excavated out. Layered advertisement click conversion

基金项目：国家自然科学基金(61673122)资助项目；广东省自然科学基金(2019A1515010548)资助项目。
收稿日期：2020-02-16；**修订日期**：2020-04-27

rate prediction model is constructed and trained. The LightGBM algorithm model is adopted to obtain the important feature ranking of the ad click conversion rate. The experimental results indicate that when the feature selection threshold is 0.95, the number of feature choices is 19, and the number of trees is 100, the area under receiver operating characteristic (ROC) curve (AUC) value of the model is the maximum, and the logarithmic loss function value of the model is about 0.136 8. The model has the optimal effect. The prediction model and the result of feature ranking are helpful for the enterprise to make the optimal advertising strategy.

Key words: data analysis; click conversion rate; computational advertising; feature engineering; feature analysis

引　言

近年来,“大数据”成为热门,遍及全球,它指的是这个信息时代下的碎片化数据,而数据也逐渐成为各个企业不可或缺的战略性资源。企业创新和转型的主要驱动力逐渐转变成对数据的分析利用。目前,网络广告已经占据广告行业的一大部分,数据营销广告掘地而起,而传统广告停滞不前,其中一个原因是大众传媒已逐渐转变成针对用户兴趣爱好的小众传媒,因此对广告历史数据加以分析和利用是一种有效提高广告点击转化率的手段。

网络广告主要分为搜索广告和展示广告[1],搜索广告是指通过一些搜索引擎进行检索时所展示出来的广告,展示广告主要有文字链接、图片和视频等,它适用于一些应用界面或者视频媒体等。计算广告是一种网络广告定向技术[2],是指通过线上的渠道推送到准确的目标人群的网络广告,其主要方法是在广告库中选择出符合目标人群性别和年龄的广告。计算广告的宗旨是通过历史数据去了解用户,深入挖掘用户需求。Agarwal等[3]基于广告数据事先已有的不同粒度层次的概念,利用树状马尔可夫方法预测广告点击率。Wang等[4]则采用了贝叶斯模型对广告点击率进行评估。而Lee等[5]给出了4个多标准数学规划广告点击问题模型,通过分析用户点击或是浏览信息,解决了用户行为定向的问题。为了提高广告点击率预估的准确率,张志强等[6]提出通过深度学习的方法来学习特征间的高维关系。潘书敏等[7]提出了一种基于用户相似度和特征区分的混合模型,该方法利用用户相似度将用户划分为多个组,再对不同的组构建子模型并进行有效组合以探索相同特征对不同组的差异效应,从而准确地预测广告点击行为。

本文提出了结合特征选择的广告点击转化率预测混合模型,在混合模型的设计中加入特征选择LightGBM模型,其中利用XGBoost、逻辑回归模型来转换和训练特征,并利用LightGBM模型来选择重要特征,最终通过实验结果分析,得出模型的有效性和重要性特征排序。

1　数据分析及特征表示

1.1　实验数据集

本文的数据来源于某广告监测公司日常的车企数据日志,具体的时间范围为2019年10月27日至2019年11月30日,共计35 d,部分字段信息如表1所示。

TrackMaster系统每天收取海量的实时数据,存储到数据日志,实时数据并不可以直接使用,还需要进行数据预处理,这样数据才具有真实性、有效性。

表1　点击日志字段表

Table 1　Click on the log field table

字段名称	字段意义
Click	点击
Imp	曝光
Timestamp	点击时间戳
Device	设备回传
Gender	用户性别
Age	用户年龄
Edu	用户受教育程度

1.2 数据集预处理

该车企在2019年10月27日至2019年11月30日期间共有137 210次点击数据。经过数据过滤、数据清洗和异常数据的剔除之后，统计整理数据得到109 988次有效点击数据，广告位ID总计21个，媒体总计3个，以及计算了各维度下的广告点击转化率点击率(Click through rate，CTR)，如表2所示，为进行特征分析建立良好的数据基础。

表2 基础数据表

Table 2 Basic data table

媒体	曝光	点击	CTR/%
1	810 818	77 959	9.61
2	399 610	28 126	7.04
3	221 126	3 903	1.77

1.3 特征的分析与构建

特征工程是将原始数据转换成能被计算机算法所理解的特征体系的工程活动，为了提高模型的准确度和泛化能力，就要从原始数据中提取尽可能多的有用信息供算法使用。本文从用户信息特征、广告信息特征、上下文特征和统计特征中分析并提取影响广告点击转化率的重要特征。通过对数据集的深入分析，从用户信息特征中提取出性别、用户年龄段和用户兴趣标签；从广告信息特征中提取出广告主、广告位ID、广告图片ID、内容频道和内容URL；从下文特征中提取出媒体、广告投放时间、地理位置和操作系统，作为广告点击转化率预测模型的特征。

以上所描述的用户信息特征、广告信息特征和上下文特征属于类别特征，而统计特征是一种基于类别特征的统计特征构造框架[8]，图1给出了该框架。

构建的具体步骤如下：

(1)将类别特征分成两个特征组S_1和S_2，并设计统计指标组I，包括举止、标准差等。

(2)构造过程中，将S_1层中的一个特征$S_{1,i}$($1<i<m$)，和S_2层中的一个特征$S_{2,i}$($1<i<n$)进行两两组合，成为特征组合$\{S_{1,i},S_{2,i}\}$。

(3)在统计指标组I中选择一项I_i，在数据集中选择出符合$\{S_{1,i},S_{2,i}\}$的所有集合，并计算相关指标。

本文在类别特征的基础上，增加了用户ID，然后分为具有5个和9个特征的两个特征组，统计指标为I_1、I_2、I_3，因此可以扩展出$5\times9\times3=135$个统计特征，详见图2。

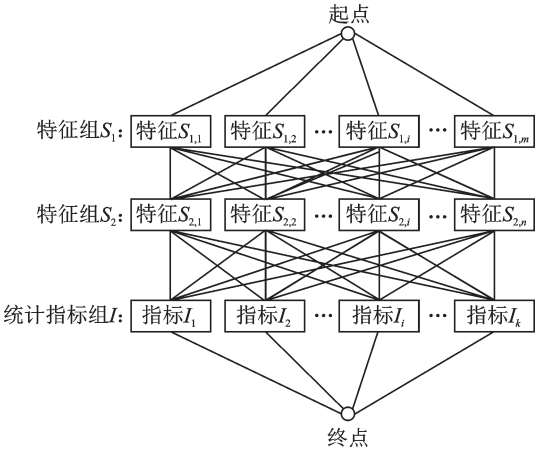

图1 统计特征构造框架

Fig.1 Statistical feature construction framework

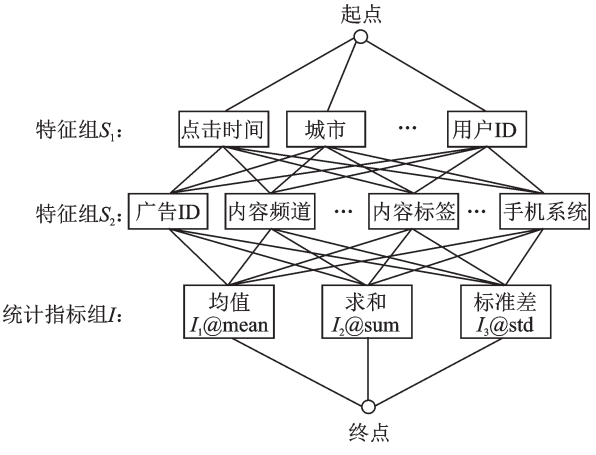

图2 统计特征构造具体示意图

Fig.2 Specific schematic diagram of statistical feature construction

1.4 特征表达

因为模型的训练过程需要输入的数据是数值型的,所以需要先将非数值型数据转化成数值型数据。在提取用户信息特征、广告信息特征和上下文特征之后,需要用到独热编码来表达这些类别特征,为了解决使用独热编码产生的共线性问题,本文主要使用虚拟编码,即用$n-1$个特征来代表具有n个可能取值的特征,这使得模型参数估计较为准确。将上述用户信息特征、广告信息特征和上下文特征3个类别特征进行虚拟编码,如对用户性别进行虚拟编码后得到2个特征向量,广告位ID进行虚拟编码后得到21个特征向量,地理区域进行虚拟编码后得到35个特征向量,关键词进行虚拟编码后得到8 969个特征向量,最后共计17 140个。最终形成的广告点击转化率预测模型的特征向量为17 275个,利用1.3节给出的方法,可以扩展出135个统计特征。

2 广告点击转化率预测模型

2.1 相关方法简介

2.1.1 XGboost算法

XGboost是华盛顿大学陈天奇博士于2016年开发的Boosting库,兼具线性规模求解器和树学习算法[9]。它是在Gradient Boosting的基础上进行改进得到的一种算法模型,可以说XGBoost是Gradient Boosting的高效实现。

2.1.2 LightGBM算法

LightGBM是微软2015年提出的新的boosting框架模型[10],类似XGBoost,LightGBM算法也是在Gradient Boosting的基础上进行改良所得,但两者在特征处理上有所不同。XGBoost在对特征进行选择时是通过预排序算法,而LightGBM则是利用HistoGram算法。预排序算法是严谨的对每个特征预先排好序,然后对每个特征进行选择;HistoGram算法则是将连续型特征离散化,然后按照离散后的数量形成相同数量的直方图,在选择特征时只需根据直方图中的离散值数量选出最优的分裂特征即可。这样虽然在严谨度上会比XGBoost低,但速度和内存开销会比XGBoost有所提升。XGBoost和LightGBM还有一个主要的不同之处在于它们决策树的生长策略不同,XGBoost的决策树生长策略是level-wise生长策略。这种生长策略的主要特点是对所有的叶子节点都一视同仁,对部分即使增益很小的树也会进行增长。这样做的优点是能够保证误差不会太大。LightGBM的决策树生长策略是leaf-wise生长策略,该策略是只对增益比较大的树进行增长,如果树分裂时没有增益或增益很小,则不会再对这些树进行生长。采用这种策略出现过拟合的风险会比较大,但是如果对树的生长加入深度限制,就能很好地解决过拟合的问题,这样既提升了算法的运算速率,也能尽量避免出现过拟合的风险。

总的来说,LightGBM是在保证了和XGBoost具有相近的准确度的同时,拥有比XGBoost更快的运行速率与更低的内存消耗,XGBoost追求比较完美的精确度,而LightGBM略微牺牲了精确度,大大提高了运行成本和速率。

2.2 广告点击转化率预测模型

本文模型通过逻辑回归模型[8]来学习XGBoost的叶子权重,计算权重之和sum,并将sum做sigmoid转换成0~1之间的值,作为最终预测值。在图3所示的模型结构,输入样本x进行集成树处理后,得到一个叶子节点标记为1,非叶子节点标记为0的节点序列,接着在线性分类器$\boldsymbol{\Sigma}$中通过逻辑回归训练就可得到逻辑回归模型。最后,通过LightGBM算法在模型训练结束后输出特征的相对重要性,得到模型下每一维特征的重要性排序。

先对该车企数据进行预处理形成L_1层的数据集T,然后进行特征转换和标准化处理后分别得到

$\boldsymbol{\Sigma}_{\text{base}}$和$\boldsymbol{\Sigma}_{\text{stat}}$,然后合并得到总训练矩阵$\boldsymbol{\Sigma}_T$,最后训练得到模型$M_{L2}$,具体步骤如算法1所示。

算法 1 广告点击转化率预测算法

输入:点击训练日志 L

输出:广告点击转化率预测模型

(1) 对L进行数据清洗、过滤,剔除异常数据后得到正常数据集;

(2) 对正常数据集进行计算整理得到点击数据集 T;

(3) 在 T 上提取用户信息特征、广告信息特征以及上下文特征并转换,得到基础特征训练矩阵$\boldsymbol{\Sigma}_{\text{base}}$;

(4) 从 T 中提取统计特征来进行标准化处理,得到统计特征训练矩阵$\boldsymbol{\Sigma}_{\text{stat}}$;

(5) 在$\boldsymbol{\Sigma}_{\text{stat}}$的基础上,采用LightGBM算法就可以得到特征选择后的统计特征训练矩阵$\boldsymbol{\Sigma}_{\text{select}}$;

(6) 将$\boldsymbol{\Sigma}_{\text{stat}}$与$\boldsymbol{\Sigma}_{\text{select}}$按行进行合并,得到总训练矩阵$\boldsymbol{\Sigma}_T$;

(7) 利用$\boldsymbol{\Sigma}_T$训练集成树分类器,得到集成树分类器$\boldsymbol{M}_{L1}$;

(8) 利用$\boldsymbol{M}_{L1}$在$\boldsymbol{\Sigma}_T$上处理得到$\boldsymbol{L}_1$的输出矩阵$\boldsymbol{\Sigma}_{L1}$;

(9) 利用$\boldsymbol{\Sigma}_{L1}$在$\boldsymbol{\Sigma}_T$上训练$\boldsymbol{L}_2$分类器,得到模型$\boldsymbol{M}_{L2}$;

(10) 将$\boldsymbol{M}_{L1}$和$\boldsymbol{M}_{L2}$进行结合,得到广告点击转化率预测模型。

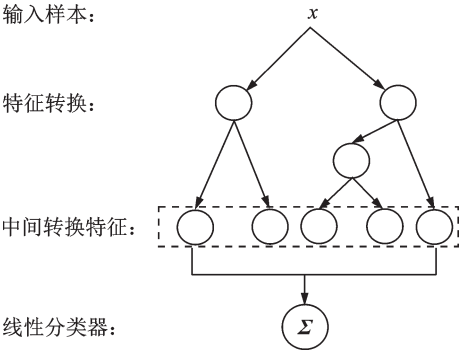

图3 模型结构

Fig.3 Model structure

3 模型评价指标

3.1 AUC

受试者工作特征(Receiver operating characteristic, ROC)[11]可以直观地反映模型在选取不同阈值时的敏感性和精确性。广告点击转化率预测指的是某一个广告被展示时,可能被点击的概率。因此可以采用ROC曲线来评估模型性能,其纵轴、横轴分别为

$$\text{TPR}=\frac{\text{TP}}{\text{TP}+\text{FN}} \tag{1}$$

$$\text{FPR}=\frac{\text{FP}}{\text{TN}+\text{FP}} \tag{2}$$

式中,TP表示实例,FP表示假阳例,TN表示真阴例,FN表示假反例。

ROC曲线下的面积(Area under ROC curve, AUC)值刻画ROC曲线下方的面积,是ROC曲线的一个直观反映,AUC值越大代表其正确性越高,选择不同的阈值,ROC曲线下方的面积也不同,即AUC值不同。一般0.5<AUC<1,本文的实验将AUC值作为评价指标。AUC的计算公式为

$$\text{AUC}=1-\frac{1}{m^+m^-}\sum_{x^+\in D^+}\sum_{x^-\in D^-}\left(I(f(x^+)<f(x^-))+\frac{1}{2}I(f(x^+)=f(x^-))\right) \tag{3}$$

式中D^+为所有正例组成的集合,x^+是其中的一个正例,D^-为所有反例组成的集合,x^-是其中一个反例,m^+表示正例的个数,m^-表示负例的个数。$f(x)$是模型对样本x的预测结果,$I(x)$在x为真时取1,否则取0。

3.2 Log-loss

AUC值更偏重于排序,当提升整体的预测概率时,Log-loss[12]值也会发生变化,而AUC值不变,因此可将Log-loss作为评价指标之一。对数似然损失,通过分类惩罚错误来保证分类器量化的精准度,即

对数损失函数$L(W)$越小,分类器的准确度越高。对数损失函数$L(W)$计算公式为

$$L(W)=-\frac{1}{N}\sum_{i}[\,y_i\log\hat{y}_i+(1-y_i)\log(1-\hat{y}_i)\,]+\lambda\|W\|^2 \tag{4}$$

式中,N为样本数,y_i的值为1或者0,分别表示有发生点击行为或没有发生点击行为,$\hat{y}_i$表示有发生点击行为的概率,$1-\hat{y}_i$则表示没有发生点击行为的概率。

4 实验结果与分析

本文以预处理之后的广告点击数据T作为实验数据集,其时间范围为2019年10月27日至2019年11月30日,共计35 d,由1.2节得知广告的有效点击次数为109 988次,在实验过程中随机选择数据集中的80%的数据量作为训练集,剩余的数据作为测试集,训练集的数据量为87 990条,测试集的数据量为21 998条。

本实验先将树的深度设定为3,然后逐步改变树的数量α,模型的AUC值会发生变化,如图4所示。

由图4可以看出,随着树的颗数α的递增,AUC值也逐渐增大,当递增到230颗时AUC值最大,此时模型最优。而当$\alpha=230$时,树的深度对模型的影响如图5所示。

由图5可以看出,树的深度为3时,AUC值最大,因此将树的深度固定为3。而转换后的特征长度也随着树的颗树的变化而变化,如图6所示。

由图6可以看出,每增加40颗树,转换特征长度就会增加200,即每增加1颗树,会多增加约5个叶子节点,呈稳定的增长趋势。 随着树的颗数的变化,分层模型的整体训练时间也会随着变化,如图7所示。

根据树的颗数对模型的影响,由图4和图7可以看出,当树的颗树为230时,AUC值最大,而分层模型训练时间适中。

综合上述分析,本文选择230为L_1的分类器数量。根据算法1,第一步要提取统计特征矩阵,进行

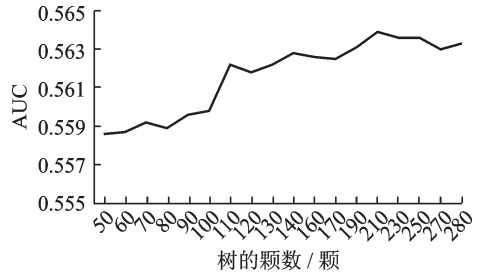

图4 树的颗数对模型的影响

Fig.4 Influence of tree number on the model

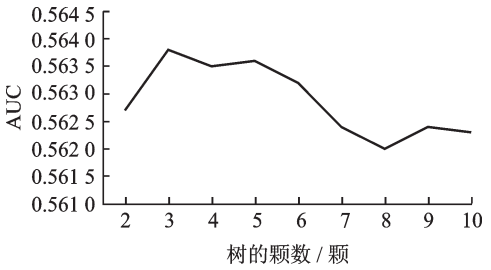

图5 树的深度对模型的影响

Fig.5 Influence of tree depth on the model

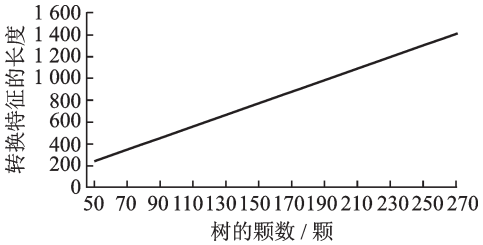

图6 转换特征长度的变化趋势

Fig.6 Change trend of the transformation feature length

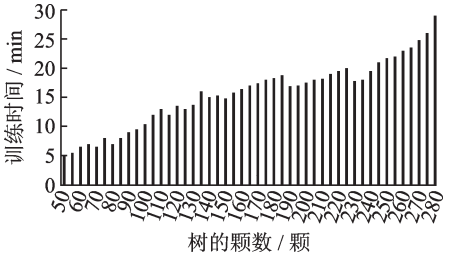

图7 分层模型整体训练时间的变化

Fig.7 Change of overall training time of the hierarchical model

LightGBM模型训练，提前设定特征选择阈值为$\lambda=0.95$，即允许选择器选择分数之和为λ的重要性特征。最终算法选择了靠前的19个特征，其分数之和占比约为95%。通过调整阈值，来控制特征个数的输出，最终输出30个重要性特征，如图8所示。

在$\lambda=0.95$的情况下，由图8可以得出，模型可以选择出19维特征，其他重要性得分非常低的特征可以忽略，这些特征的影响性较低。本文利用LightGBM模型对分层模型进行特征筛选，分析特征选择模型中树的颗数的变化对混合模型整体结果的影响，如图9所示。

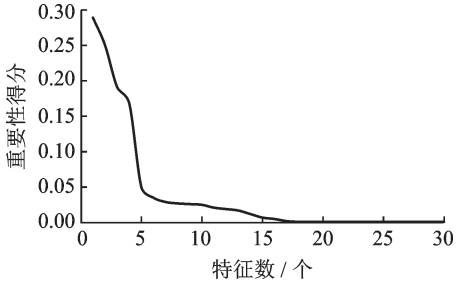

图8 特征重要性得分排序

Fig.8 Order of feature importance scores

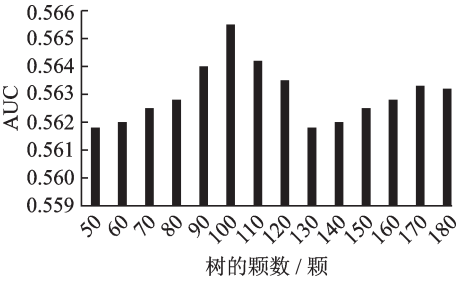

图9 特征模型树的颗数变化

Fig.9 Change of the number of feature model trees

由图9可以看出，当特征选择阈值$\lambda=0.95$，特征选择数目为19，树的颗数为100时的AUC值最大，此时模型具有最优的效果。

综上所述，以本文的汽车广告点击数据集为数据基础，为了使模型具有最优的效果，本文固定特征选择的树的颗数为100，$\lambda=0.95$，树的深度为3、颗数为230。而根据式(2)计算出此时混合模型的对数损失函数值约为0.136 8，显然对数损失函数值较小，模型表现良好。最后输出特征选择模型前10的重要特征排序，如图10所示。

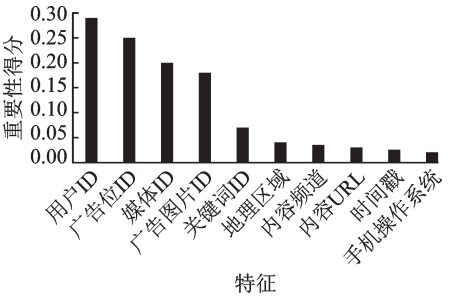

图10 前10名特征重要性排序

Fig.10 Top ten features in the importance ranking

5 结束语

广告点击转化率关系着广告主的切身利益，良好的点击转化率预测模型可以为广告主提高收益，而重要性特征无疑是广告主进行广告投放的重要参考。本文的主要工作是给出了一种统计特征的构建框架，并提出结合特征选择的广告点击转化率预测混合模型，利用XGBoost、逻辑回归模型来转换和训练特征，并利用LightGBM模型来选择重要特征并排序，广告主在投放商品广告时，可根据用户ID、广告位ID、媒体ID、广告图片ID、关键词ID和地理区域等重要性特征的排序，结合外在因素和以往的投放经验，制定较优的投放策略，来提升广告点击转化率，从而使广告主的利益最大化。

参 考 文 献：

[1] FAIN D C, PEDERSEN J O. Sponsored search: A brief history[J]. Bulletin of the American Society for Information Science & Technology, 2006, 32(2):12-13.

[2] 郭心语，刘鹏，周敏奇，等. 网络广告定向技术综述[J]. 华东师范大学学报:自然科学版，2013, 3: 93-105.
GUO Xinyu, LIU Peng, ZHOU Minqi, et al. Overview of online advertising targeting technology [J]. Journal of East China Normal University: Natural Science Edition, 2013, 3: 93-105.

[3] AGARWAL D, BRODER A Z, CHAKRABARTI D, et.al. Estimating rates of the rare events at multiple resolutions[C]//

Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Jose, USA: [s.n.], 2007: 16-25.

[4] WANG X, LI W, CUI Y, et al. Click-through rate estimation for rare events in online advertising[J]. Online Multimedia Advertising: Techniques and Technologies, 2011, 10: 1-12.

[5] LEE J, SHI Y, WANG F, et al. Advertise-ment clicking prediction by using multiple criteria mathematical programming[J]. World Wide Web, 2016, 19(4): 707-724.

[6] 张志强,周永,谢晓芹,等.基于特征学习的广告点击率预估技术研究[J].计算机学报,2016, 39(4): 780-794.
ZHANG Zhiqiang, ZHOU Yong, XIE Xiaoqin, et al . Research on advertising click rate prediction technology based on feature learning [J]. Chinese Journal of Computers, 2016, 39(4): 780-794.

[7] 潘书敏,颜娜,谢瑾奎.基于用户相似度和特征分化的广告点击转化率预测研究[J].计算机科学,2017, 44(2): 283-289.
PAN Shumin, YAN Na, XIE Jinkui. Prediction of click conversion rate of advertisements based on user similarity and feature differentiation [J]. Computer Science, 2017, 44(2): 283-289.

[8] 黄淦.基于大数据分析的广告点击转化率预测方法研究[D].广州:华南理工大学,2017.
HUANG Gan. Research on prediction method of advertising click conversion rate based on big data analysis[D]. Guangzhou: South China University of Technology, 2017.

[9] CHEN T, GUESTRIN C. XGBoost: A scalable tree boosting system[C]//Proceedings of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.[S.l.]: ACM, 2016: 785-794.

[10] KE G L, MENG Q, FINLEY T, et al. LightGBM: A highly efficient gradient boosting decision tree[C]//Proceedings of Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems. Long Beach, CA, USA: [s.n.], 2017: 4-9.

[11] FAWCETT T.ROC Graphs: Notes and practical considerations for data mining researchers[J].Pattern Recognition Letters, 2003, 31(8): 1-38.

[12] DAVIS J, GOADRICH M.The relationship between precision-recall and roc curves[C]//Proceedings of the 23rd International Conference on Machine Learning. NewYork, USA: ACM, 2006: 233-240.

作者简介:

邓秀勤(1966-),女,教授,硕士生导师,研究方向:机器学习、数据挖掘,E-mail:dxq706@gdut.edu.cn。

谢伟欢(1997-),男,学士,研究方向:机器学习。

刘富春(1971-),男,博士,教授,博士生导师,研究方向:算法分析与设计、控制理论与应用。

张翼飞(1996-),男,硕士研究生,研究方向:机器学习、聚类分析。

樊娟(1997-),女,硕士研究生,研究方向:数据挖掘、聚类分析。

(编辑:陈珺)