第 19 卷　第 18 期　2019 年 6 月
1671—1815(2019)018-0234-05

科　学　技　术　与　工　程
Science Technology and Engineering

Vol. 19　No. 18　Jun. 2019
© 2019　Sci. Tech. Engrg.

引用格式:陆　平, 陈笑天. 基于梯度提升树模型的网络优惠券使用预测[J]. 科学技术与工程, 2019, 19(18): 234-238
Lu Ping, Chen Xiaotian. Prediction of internet coupon usage based on gradient boosting decision tree model[J]. Science Technology and Engineering, 2019, 19(18): 234-238

基于梯度提升树模型的网络优惠券使用预测

陆　平　陈笑天*
(中国电子信息产业发展研究院,北京 100846)

摘　要　互联网与实体经济融合发展背景下,网络优惠券往往承担了提升用户体验、促进再次消费的重要功能。构建梯度提升树、随机森林等模型,预测网络优惠券使用行为;并对影响因素的重要性进行排序。结果表明:梯度提升树算法的五折交叉验证平均测试精度、曲线下面积值分别为 0.804 与 0.886,高于随机森林与单棵决策树算法。优惠券折扣率对于用户使用优惠券行为起着决定性影响,用户经常活动的地点离该商户最近门店的距离、领取优惠券时间等特征对用户使用优惠券行为具有重要影响。

关键词　网络优惠券　梯度提升树　随机森林　预测
中图法分类号　TP391.75　N94;　**文献标志码**　A

互联网与实体经济融合发展背景下,越来越多的企业已开始应用线上到线下(online to offline)营销模式拓展自身业务;即互联网平台提供商业信息、服务预订和优惠信息等,将主体商户的线下实体业务信息如优惠、新品等信息推送给互联网线上用户,随后通过优惠券或折扣等手段引导线上互联网用户到实体店进行线下消费。在该模式下,网络优惠券承担了提升用户体验、促进再次消费的重要功能。随着我国经济进入高质量发展阶段,以更为精准高效的网络优惠券替代传统的粗放式营销模式,不仅能够大幅减少用户接收到的垃圾信息,进而优化客户消费体验,还有利于商户提升商品销量并减少广告成本。这对于提升经济运行和信息流动的有效性、有效挖掘潜在消费能力具有重要意义。

一些学者针对优惠券使用的相关问题进行了研究。优惠券功效方面。Clark 等[1]指出优惠券有利于建立客户忠诚或维护客户关系,可以激发顾客再次购买行为。Blattberg 等[2]研究发现,持有优惠券的用户倾向于增加他们所购买商品的数量,或者加速购买商品的时间;且优惠券具有广告效用,能够通过优惠券吸引到新的顾客。影响因素方面。Hsu 等[3]发现优惠券持有人的主观态度、对优惠券使用规范的理解、使用优惠券效用感知等因素对于用户的优惠券使用意愿存在一定程度影响。Traver[4]的研究显示那些移动电子优惠券的优惠程度也就是打折率是决定用户使用这种优惠券效率的决定性因素。Jayasingh 等[5]认为用户使用优惠券的可感知性、可兼容性和可获得性决定了用户如何使用优惠券。Chakraborty[6]指出用户认为高面值的优惠券比低面值提供了更多的信息和优惠,因此面值较高的优惠券使用率更高。国内相关研究起步较晚,主要探讨优惠券的投放策略、积分系统和顾客使用意愿。王玉平[7]依据期望确认理论,构筑了一个研究优惠券的行为模型;并且使用该模型分析用户使用欲望与感知度等要素的关联。管轶楠[8]研究发现历史优惠使用行为以及优惠券发放前访问行为特征、历史交易总金额、历史优惠使用行为以及优惠券有效期内访问行为特征与优惠券赎回行为存在相关关系。张建同等[9]基于实验研究确定了线上优惠券发放对用户消费的正向影响。黄正等[10]使用聚类分析优惠券用户类型,并尝试找出分发逻辑。吕丽辉等[11]对用户使用优惠券的行为方式进行了研究,提出可能与用户的期望确认有关。刘芬等[12]对用户使用优惠券的动机及行为特征进行了研究。

根据上述文献,以商户特征、优惠券自身折扣力度、用户因素为综合框架的网络优惠券使用行为的实证研究仍较为欠缺。现基于阿里天池的公开数据,结合商户、优惠券、用户行为三个维度,研究影响用户使用优惠券行为的关键因素,为优惠券的精准投放提供借鉴和参考。

2019 年 1 月 25 日收到　工信部规划司研究项目资助
第一作者简介:陆　平(1988—),男,汉族,江西南昌人,博士,副研究员。E-mail:luping101010@126.com。
* 通信作者简介:陈笑天(1992—),男,汉族,北京人,硕士,助理研究员。E-mail:chenxiaotian@ccidthinktank.com。

1　样本数据

数据来源于阿里天池平台。原始样本是已经过消除隐私处理后的数据。数据集的时间区间从2016 年 1 月初到同年 6 月底。其中,在线交易记录有 11 429 826 条元组,包括用户编号、产品编号、用户在线行为、优惠券编号、折扣率、优惠券获得日期、线上事务发生日期这七个特征。线下事务记录有 1 754 884条元组,包括用户编号、商品编号、优惠券编号、优惠券折扣率、用户离商户的距离、优惠券获得的日期、线下事务发生日期七个特征。如果在线下商户交易的日期不为空值;并且优惠券获得日期不为空值,该情形意味着用户在消费中核销了优惠券,将其看作为正样本。如果线下商户交易的日期为空值,但优惠券领取的日期不为空,该情形意味着用户虽获得优惠券但并没有对它进行核销,将其看作为负样本。

根据线下事务记录数据,正样本数量约占负样本数量的十分之一左右,存在比较严重的类不平衡问题。这可能导致模型更多地关注占多数的类别,会对训练出来的模型性能产生不利影响。解决该问题的做法是通过过采样、欠采样或两者结合的方式来对样本集进行处理,从而使得正样本和负样本的数量比例大致相同。在数据处理阶段,对样本空间中的负样本进行稀疏采样,比例设置为 10%。

根据线上事务记录和线下事务记录的基本特征,可生成优惠券、商户、用户三类特征,并派生出用户与商户的交叉特征,见表 1。

表 1　特征选择

Table 1　Feature selection

类别	特征
优惠券特征	优惠券折扣率、满减优惠券中的满元额度、满减优惠券中的减免额度、是否满减优惠等
商户特征	商户被消费次数、商户优惠券发放次数、商户被消费次数中核销优惠券比率等
用户线下特征	用户领取优惠券次数、用户核销优惠券次数、用户经常活动的地点离该商户最近门店的距离等
用户线上特征	用户点击率、用户购买率、领取优惠券时间等

2　模型构建与评估方法

2.1　模型构建

随机森林模型中的每一棵决策树都随机选择一些样本和一些特征,因此可以在一定程度上避免过拟合问题,且具有良好的抗噪能力。算法步骤如下:使用自举法随机从样本集合 S 中选择 s 个样本;随机从特征集合 F 中选择 f 个特征,在该样本集上使用选出的 f 个特征建立决策树;重复上述两个步骤 n 次,能够生成 n 棵决策树,所有的决策树汇总成为森林;对于测试数据,森林中的每棵树分别进行预测,将每棵树的结果汇总并计算众数(可以看作一种投票机制),该结果即为随机森林算法的预测结果。

梯度提升树算法以 CART 树作为弱分类器,经过多次迭代,最终组合成一个具有较强预测能力的集成模型,体现了 boosting 集成建模思想。

对于二元梯度提升树,采用对数似然损失函数:

$$L[y,f(x)] = \lg\{1 + \exp[-yf(x)]\} \tag{1}$$

式(1)中:$y \in \{-1, +1\}$。负梯度误差为

$$r_{ti} = -\left\{\frac{\partial L[y,f(x_i)]}{\partial f(x_i)}\right\}_{f(x)=f_{t-1}(x)} = \frac{y_i}{1 + \exp[y_i f(x_i)]} \tag{2}$$

各子节点的最佳负梯度拟合值为

$$c_{tj} = \underbrace{\arg\min}_{c} \sum_{x_i \in R_{tj}} \lg(1 + \exp\{-y_i[f_{t-1}(x_i) + c]\}) \tag{3}$$

特征 j 的重要性通过其在单棵决策树中重要度的平均值来衡量:

$$\hat{J}_j^2 = \frac{1}{M}\sum_{m=1}^{M} \hat{J}_j^2(T_m) \tag{4}$$

式(4)中:M 是树的数量。特征 j 的重要度计算方式如下:

$$\hat{J}_j^2 = \sum_{t=1}^{L-1} \hat{i}_t^2 1(v_t = j) \tag{5}$$

式(5)中:v_t 是与节点 t 有关联的特征集合,i_t^2 是节点 t 分裂之后平方损失的减少值。

遍历各节点的所有可能的分割情形,以最小 *Gini* 作为分割标准。

$$Gini(t) = 1 - \sum_{j=1}^{k} [p(j \mid t)]^2 \tag{6}$$

式(6)中:$p(j \mid t)$ 是 t 节点属于类别 j 的概率。

2.2　评估方法

为了比较不同模型之间的性能好坏,采用交叉验证方法。步骤如下:把总训练样本采用分层采样的方式随机分成相等容量的 5 份子集 S_1、S_2、S_3、S_4 和 S_5;把其中一个子集合用作测试集 $TEST_i$,其余四个合并作为训练数据集 $TRAIN_i$,从而形成五组训练-测试数据集($TRAIN_i$, $TEST_i$)($i = 1,2,\cdots,5$);基于每组训练-测试集,分别训练单棵决策树、随机森林、梯度提升树三类模型,并评估各个模型在五组训练-测试集上的预测性能,进而计算平均性能。

模型评估方面,采用曲线下面积(area under curve, AUC)作为衡量模型预测性能好坏的主要标准。该指标是指受试者工作特征曲线(receiver op-

erating characteristic curve,ROC)所覆盖的面积。通过依次遍历真正率(TPR)和假正率(FPR)这两个指标阈值来绘制 ROC 曲线。可基于混淆矩阵(表 2)计算真正率和假正率,公式见式(7)和式(8)。

$$TPR = \frac{TP}{TP + FN} \tag{7}$$

$$FPR = \frac{FP}{FP + TN} \tag{8}$$

曲线下面积的取值范围处于 0.5 ~ 1。曲线下面积越大,则意味着模型性能越高。

表 2 混淆矩阵

Table 2 Confusion matrix

		实际结果	
		1	0
预测结果	1	*TP*	*FP*
	0	*FN*	*TN*

3 训练与测试

3.1 超参数设置

超参数设置往往对模型预测性能存在重要影响,以下采用敏感性分析方法,对单棵决策树、随机森林、梯度提升树进行超参数设置。

对于单棵决策树模型,树最大深度取值在 2 ~ 16 变化过程中,模型在交叉验证测试集上的曲线下面积平均值成“倒 U”形。根据参数敏感性测试,树最大深度取值在 10 的时候,单棵决策树模型的曲线下面积平均值达到峰值。基于上述分析,把 10 作为单棵决策树模型的最大深度参数值。

对于随机森林模型,树最大深度、弱分类器数是对结果影响较大的超参数。树最大深度在 4 ~ 32 内变化过程中,模型在交叉验证测试集上的曲线下面积平均值成“倒 U”型变化。从敏感性测试结果看,树最大深度取值为 16 的情形下,模型曲线下面积平均值最高。弱分类器个数取值 10、100、1 000 情形下,模型曲线下面积平均值分别为 0.876 9、0.883 1 和 0.883 2,可见弱分类器数的增加通常能够改善模型预测能力,但提升空间随着该参数的取值变大而缩小。基于上述分析,把随机森林模型的树最大深度、弱分类器个数参数分别设置为 16 和 100。

对于梯度提升树模型,弱分类器数、树最大深度、学习率等超参数对模型预测性能好坏起到关键作用。学习率参数通常设置为 0.1。弱分类器个数取值 10、100、200 与 300 情形下,模型曲线下面积平均值分别为 0.864 2、0.885 3、0.889 1 和 0.885 9,由此将弱分类器的数量设置为 200。不断调整梯度提升树模型的树最大深度参数值,敏感性测试结果表明,树最大深度取值为 6 的情形下,模型曲线下面积平均值达到最大。该取值要远小于随机森林模型的树最大深度取值。原因在于:梯度提升树作为 boosting 集成算法的代表,更加关注偏差的减少,该模型能够基于泛化性能相对较弱的分类器构建出很强的集成学习模型;而随机森林模型作为 bagging 集成算法的代表,它更加关注方差的减少,因此它在不剪枝的决策树上效果更明显。基于此,把梯度提升树模型的树最大深度、弱分类器个数分别设置为 6 和 200。

3.2 模型评估

为了评估选出最为合适的模型,对比不同模型的五折交叉验证中的预测精度和曲线下面积,结果见表 3。随机森林、梯度提升树模型的测试精度和曲线下面积大于单个决策树模型,这意味着对于网络优惠券使用行为预测问题,集成模型往往性能更好。

梯度提升树模型的测试精度和曲线下面积大于随机森林,可见梯度提升树模型的性能比随机森林模型稍好一些。随机森林模型采取有放回的均匀取样,而梯度提升树模型根据每次训练集之中每个样本的分类是否正确,重新确定样本权重,并把更新后的数据集作为接下来层次的训练样本,这是梯度提升树模型的分类精度高于随机森林的原因。

3.3 影响因素分析

影响优惠券使用的主要因素来自三个方面,即优惠券自身因素、用户行为因素和商户因素,重要性排名见表 4。根据梯度提升树的结果,优惠券自身因素和用户行为因素重要性靠前,在重要性得分前十位中分别占据第 1、5、6、7 位和第 2、3、4 位,商户因素的重要性靠后,在重要性得分前十位中占 8、9、10 位。

优惠券方面。优惠券的折扣率、满减优惠券中的满元额度、满减优惠券中的减免额度、是否满减优惠四个特征对用户是否核销优惠券起到关键作用。其中,最重要的因素为优惠券折扣率,其重要性得分 0.089 8,约超出第二名 0.03。

表 3 五折交叉验证测试精度与曲线下面积

Table 3 Five-fold cross-validation and AUC value

五折交叉验证子集	单棵决策树		随机森林		梯度提升树	
	测试精度	曲线下面积	测试精度	曲线下面积	测试精度	曲线下面积
S_1	0.774	0.853	0.798	0.884	0.805	0.887
S_2	0.780	0.854	0.802	0.885	0.804	0.887
S_3	0.779	0.855	0.798	0.882	0.806	0.885
S_4	0.781	0.852	0.800	0.883	0.803	0.887
S_5	0.781	0.853	0.797	0.881	0.800	0.886
平均值	0.779	0.854	0.799	0.883	0.804	0.886

用户行为因素方面。用户经常活动的地点离该商户最近门店的距离、领取优惠券是月中的第几天、优惠券领取日期与预测日期间隔天数对用户优惠券使用行为起到重要影响。其中,用户经常活动的地点离该商户最近门店的距离是较为重要的影响因素。

商户因素方面。商户被消费次数、商户被消费次数中核销优惠券比率、商户发放优惠券次数等特征,对用户优惠券使用行为的影响力较高。其中,商户被消费次数反映了商户本身的热门程度。商户发放优惠券次数体现了商户使用网络优惠券这种促销方式的倾向性。

表 4　梯度提升树的重要性排名前十特征

Table 4　Top ten characteristics of gradient boosting decision tree

排名	特征	重要性
1	优惠券折扣率	0.089 8
2	用户经常活动的地点离该商户最近门店的距离	0.063 3
3	领取优惠券是月中的第几天	0.062 8
4	优惠券领取日期与预测日期间隔天数	0.060 6
5	满减优惠券中的满元额度	0.057 5
6	满减优惠券中的减免额度	0.048 8
7	是否满减优惠	0.046 7
8	商户被消费次数	0.045 9
9	商户被消费次数中核销优惠券比率	0.034 6
10	商户发放优惠券次数	0.030 3

4　结论

基于梯度提升树与随机森林等模型,对网络优惠券使用行为进行预测,并对影响因素进行排序分析,得到以下结论。

(1)模型选择方面,单棵决策树模型、随机森林模型、梯度提升树模型的曲线下面积分别为 0.854、0.884 与 0.886。对于网络优惠券行为预测问题而言,采用梯度提升树将取得更为精准的预测效果。

(2)对于用户而言,优惠券折扣率、满减优惠券中满减额度、是否满减优惠四个因素对用户是否核销优惠券具有关键影响,其中,优惠券折扣率是影响用户是否核销的关键。用户经常活动的地点离该商户最近门店的距离、领取优惠券的月中时间等因素对用户使用优惠券行为也存在重要影响。

(3)对于商户而言,为提高网络优惠券投放效率,提升自身的优惠力度起到决定性作用,其次是根据用户行为特征进行布局和规划,如在潜在客户密集区布营业网点,根据用户消费习惯在特定时间进行促销等活动。

参　考　文　献

1　Clark R A, Zboja J J, Goldsmith R E. Antecedents of coupon proneness: a key mediator of coupon redemption [J]. Journal of Promotion Management, 2013, 19(2): 188-210

2　Blattberg R, Eppen G D, Lieberman J. A theoretical and empirical evaluation of price deals for consumer nondurables [J]. Journal of Marketing, 1981, 45: 116-129

3　Hsu T, Wang Y, Wen S. Using the decomposed theory of planned behavior to analyse consumer behavioral intention towards mobile text message coupons[J]. Journal of Targeting, Measurement and Analysis for Marketing, 2006, 14(4): 309-324

4　Traver R P A. Factors affecting coupon redemption rates[J]. Journal of Marketing, 1982, 46(4):102-113

5　Jayasingh S, Eze U C. An extended model for analyzing adoption behavior of mobile coupon[C]//Proceedings of the 7 th International Conference on Advances in Mobile Computing and Multimedia. Malaysia: ACM Digital Library, 2009: 282-287

6　Chakraborty G, Cole C. Coupon characteristics and brand choice [J]. Psychology & Marketing, 1991, 8(3): 145-159

7　王玉平. 移动优惠券用户持续使用意愿影响研究——以移动旅游优惠券为例[D]. 杭州: 杭州电子科技大学, 2018
Wang Yuping. Research on the influence of users' continuing use of mobile coupons: a case study of mobile travel coupons[D]. Hangzhou: Hangzhou Dianzi University, 2018

8　管轶楠. 面向电商企业的电子优惠券投放决策研究[D]. 南京: 南京理工大学, 2018
Guan Yinan. Research on the decision making of electronic coupon marketing for e-commerce enterprises[D]. Nanjing: Nanjing University of Science & Technology, 2018

9　张建同, 方陈承. 顾客历史行为和优惠券对其购买决定的影响——基于一项实验研究[J]. 软科学, 2017, 31(2): 109-112
Zhang Jiantong, Fang Chencheng. The effects of historic behaviors and e-coupon promotion on consumer purchase decision: based on an experimental study[J]. Soft Science, 2017, 31(2): 109-112

10　黄　正, 孙景仰, 刘丹阳, 等. 基于聚类分析在 O2O 平台上优惠券发放的研究[J]. 现代商业, 2018(12): 32-34
Huang Zheng, Sun Jingyang, Liu Danyang, et al. Research on coupon issuance on O2O platform based on clustering analysis [J]. Modern Business,2018(12): 32-34

11　吕丽辉, 王玉平. 移动旅游优惠券用户持续使用意愿研究[J]. 东岳论丛, 2017, 38(5): 147-152
Lu Lihui, Wang Yuping. Research on the continuous use willingness of mobile travel coupon users[J]. Dongyue Tribune, 2017, 38(5): 147-152

12　刘　芬, 赵学锋, 张金隆, 等. 移动优惠券的消费者使用意愿研究:基于个人特征和动机的视角[J]. 管理评论, 2016, 2(2): 94-102
Liu Fen, Zhao Xuefeng, Zhang Jinlong, et al. Study on consumer usage intention of mobile coupons: from the perspective of personal characteristics and motivations [J]. Management Review, 2016, 2(2): 94-102

Prediction of Internet Coupon Usage Based on Gradient Boosting Decision Tree Model

LU Ping, CHEN Xiao-tian*
(China Center for Information Industry Development, Beijing 100846, China)

[**Abstract**] The online business and the offline business of physical stores are being more closely integrated. Internet online coupons can play a role in improving the user experience and promoting re-consumption. Gradient boosting decision tree and random forest model were built to predict the usage of internet coupons and rank the importance of influencing factors. The results show that the average test accuracy and area under curve value of gradient boosting decision tree algorithm are 0.804 and 0.886 respectively, which are higher than those of random forest and decision tree algorithm. The discount rate of the coupon plays the most important role in use of coupons. The distance between the place where the user often moves and the nearest store of the business, the day on which the coupon is received have an important influence on the use of coupons.

[**Key words**] internet coupons gradient boosting decision tree random forest prediction