DOI:10.13718/j.cnki.xsxb.2019.03.017

基于 XGBoost 的新零售优惠券使用行为预测①

徐 宁, 喇 磊

对外经济贸易大学 信息学院, 北京 100029

摘要: 为实现新零售优惠券的定向投放, 提出了对用户优惠券使用行为预测的模型. 该文采用 XGBoost 算法, 突破了以 TAM 模型(技术接受模型)为基础解释个人优惠券使用意愿的传统方法, 并基于口碑网的真实交易数据进行了特征提取和用户使用行为建模. 在 K 折交叉验证之后通过变量重要性评分, 确定了对消费者使用决策贡献度较高的特征, 并与随机森林和 GBDT(梯度提升决策树)算法进行了 AUC(Area under curve)准确率的对比. 该研究证明了基于 XGBoost 的集成学习算法在优惠券使用行为预测中的有效性, 对新零售精准营销有重要的现实意义.

关 键 词: XGBoost; 优惠券使用预测; 新零售

中图分类号: F224-39 **文献标志码**: A **文章编号**: 1000-5471(2019)03-0101-05

新零售是通过运用互联网新技术与新思维, 以消费者体验为核心的数据驱动零售模式. 在移动互联网时代, 消费主体产生的数据规模呈爆炸式增长, 为精准挖掘用户需求和行为预测创造了客观条件[1].

以优惠券盘活老用户或者吸引新客户进店消费, 是新零售推动线上与线下业务跨界融合的重要方式[2], 然而优惠券的随机投放可能会对用户造成一定程度的干扰. 对商家来说, 优惠券的滥发也可能会降低品牌声誉, 同时显性增加营销成本[3]. 所以, 个性化投放是提高新零售优惠券核销率的重要技术, 需要针对用户的消费偏好进行预测, 直接的预测目标是用户在领取优惠券之后的一定期限内是否会进行消费. 所以, 优惠券使用预测是一个典型的二分类问题.

XGBoost[4]是近年来诞生的基于梯度提升决策树的集成学习算法, 在系统优化和机器学习原理方面都进行了深入的拓展. XGBoost 因其可并行化的特点、优良的学习效果以及高效的训练速度而获得了广泛关注.

现有的优惠券投放相关研究主要以 Im 等[5]基于 TAM 和 IDT(创新扩散理论)的理论为基础, 探究优惠券使用意图的影响因素. 汪明远等[6]通过使用 Smart PLS 建立了结构方程模型, 并对问卷调查数据进行了分析, 发现使用态度和从众行为均能提高消费者移动优惠券的使用意愿. 但以上研究均未利用消费者使用优惠券的真实数据. 王小平等[7]发现 Boosting 类算法能显著提高分类模型的训练速度. 任浩等[8]将 XGBoost 应用于文本情感识别, 取得了较高的分类准确度. 王重仁等[9]将 XGBoost 算法引入了互联网用户的行为预测, 表明 XGBoost 对用户行为预测有理想的效果.

本文使用真实的消费者线上线下交易数据, 并将 XGBoost 引入到优惠券使用预测中. 通过挖掘用户的行为信息, 建立分类预测模型, 从而精准预测用户是否会在期限内使用相应的优惠券. 结果表明, 在预测优惠券使用的问题上, 与传统的机器学习算法相比, XGBoost 具有准确度高、速度快等优势.

1 数据描述

本文所使用的数据来源于口碑网在 2016 年 1 月 1 日—2016 年 6 月 30 日之间的真实线上线下消费行为, 目的是预测用户在 2016 年 6 月领取优惠券后是否会在 7 月 1 日前核销. 消费券核销的度量方法是: 如

① 收稿日期: 2018-03-21
基金项目: 北京市社会科学基金项目(16GLC067).
作者简介: 徐 宁(1997-), 男, 本科生, 主要从事商务数据挖掘研究.

果 Date=null&Coupon_id!=null，表示领取优惠券但未使用；如果 Date!=null&Coupon_id=null，则表示普通消费；如果 Date!=null&Coupon_id!=null，则表示优惠券已核销.

数据集共有 477 355 条记录，字段描述如表 1.

表 1 字段描述

字段	描述
User_id	用户 ID
Merchant_id	商户 ID
Coupon_id	优惠券 ID
Discount_rate	优惠率
Distance	User 离该 merchant 的门店距离
Date_received	领取优惠券日期
Date	消费日期

1.1 数据预处理

缺失值的存在一般会严重影响模型质量和预测精度. 首先进行缺失值填充，将缺失值统一转化为 NULL(空)类型.

消费行为具有很强的时间相关性，在节假日前后会出现较大波动. 数据集中的日期记录为字符串格式，需要进行日期离散化. 将日期转换为星期几、是否节假日、是否节前、是否节后等离散变量的形式.

将距离转换成[0，10]之间的数字. x in(0，10)表示距离为 $500\times x$ 米，0 表示低于 500 米，10 表示大于 5 km.

数据集中的折扣形式有 2 种：x\in[0，1]代表折扣率；x：y 表示满 x 减 y. 需要将 2 种方式进行统一. 将满减转换成折扣率的形式更好操作，得到 discount_rate，discount_man 和 discount_jian 这 3 个变量.

1.2 特征选择

在过去的研究中，学者主要根据个体动机来研究优惠券的使用行为. 巫红霞[10]解构了电商消费者购买的行为模式. Im 等[5]基于 TAM 和 IDT 的理论研究，发现优惠券使用意图主要受消费者对商家的态度和个人信息披露影响. 龚艳萍等[11]所验证的影响因素包括优惠券属性、消费者属性、产品属性. 结合以往的研究与数据集特点，本研究划分了 4 个特征集(表 2).

表 2 特征集描述

特征集	描述	特征数量
优惠券特征	描述优惠券自身的特点及消费历史规律	11
商户特征	描述商户的受欢迎程度和商品消费规律	10
用户特征	描述用户自身特征和消费偏好	13
用户－商户交互特征	用户对商家的态度	9

2 分类建模

2.1 XGBoost 算法

从数学模型角度上来说，XGBoost 本质上是一个加强版的梯度提升树，可以用来做回归和分类. 和以往梯度提升树 (gradient boosted tree，GBT) 相比，XGBoost 的优势在于提高了泛化度（正则项、缩减率、列抽样)，提高了精确度（二阶导)，提高了速度（算法优化、系统优化).

1) 目标函数定义：其中 i 表示第 i 个样本，$l((\hat{y}_i, y_i))$ 表示第 i 个样本的预测误差，$\sum_k \Omega(f_k)$ 表示树的复杂度，T 表示叶节点的个数，w 表示叶节点的数值.

$$J(\varphi) = \sum_i l(\hat{y}_i, y_i) + \sum_k \Omega(f_k)$$

$$\text{其中 } \Omega(f) = \gamma T + \frac{1}{2}\lambda \| w \|^2 \tag{1}$$

2) 目标函数采用加法训练，不是直接优化整个目标函数，而是分步骤优化目标函数，首先优化第一棵树，之后优化第二课，直至优化完成第 K 棵树.

$$\hat{y}_i^{(0)} = 0$$

$$\hat{y}_i^{(1)} = f_1(x_i) = \hat{y}_i^{(0)} + f_1(x_i)$$

$$\hat{y}_i^{(2)} = f_1(x_i) + f_2(x_i) = \hat{y}_i^{(1)} + f_2(x_i)$$

……

$$\hat{y}_i^{(t)} = \sum_{k=1}^{t} f_k(x_i) = \hat{y}_i^{(t-1)} + f_t(x_i) \tag{2}$$

3）在第 t 步时，我们添加了一棵最优的CART树（分类回归树）f_t，这棵最优的CART树就是在现有 $t-1$ 棵树的基础上，使得目标函数最小的那棵 CART 树.

$$J^{(t)} = \sum_{i=1}^{n} l(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)) + \Omega(f_t) \tag{3}$$

4）对目标函数进行二阶泰勒展开.

$$J^{(t)} \simeq \sum_{i=1}^{n} [l(y_i, \hat{y}_i{}^{(t-1)}) + g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i)] + \Omega(f_t)$$

$$where\ g_i = \partial_{\hat{y}^{(t-1)}} l(y_i, \hat{y}^{(t-1)}),\ h_i = \partial^2_{\hat{y}^{(t-1)}} l(y_i, \hat{y}^{(t-1)}) \tag{4}$$

5）将 J 看作 f 的函数，因此 $l((y_i, \hat{y}_i{}^{(t-1)}))$ 可以看作常数项拿掉.

$$\tilde{J}^{(t)} = \sum_{i=1}^{n} \left[g_i f_t(x_i) + \frac{1}{2} h_i f_t{}^2(x_i) \right] + \Omega(f_t) \tag{5}$$

6）定义分裂的候选集合 $I_j = \{i \mid q(x_i) = j\}$ 为叶子 j 的集合，化简上式得

$$\tilde{J}^{(t)} = \sum_{j=1}^{T} \left[(\sum_{i \in I_j} g_i) w_j + \frac{1}{2} (\sum_{i \in I_j} h_i + \lambda) w_j^2 \right] + \gamma T \tag{6}$$

7）定义 $G_j = \sum_{i \in I_j} g_i$，$H_j = \sum_{i \in I_j} h_i$，并对 w_j 求导，得到目标函数最优解和最优权重.

$$\tilde{J}^{(t)}(q) = -\frac{1}{2} \sum_{j=1}^{T} \frac{G_j^2}{H_j + \lambda} + \gamma T \tag{7}$$

$$w_j^* = -\frac{G_i}{H_j + \lambda} \tag{8}$$

2.2　评价标准

分类准确率（Accuracy）是指所有正确的分类占全部标签的百分比. 分类准确率这一衡量分类器的标准比较容易理解，但是该标准不能反应出响应值的潜在分布，也不能输出分类器犯错的类型. 所以，本文主要以 *AUC* 作为主要评判标准.

AUC（Area under curve）是用于二分类模型的评价指标. *AUC* 表示随机选择一个正样本和一个负样本，分类器能够正确地给出正样本的 *score* 高于负样本的概率. 假设 M 为正样本的数量，N 为负样本的数量. 首先对 *score* 进行排序，之后令最大 *score* 对应的 *sample* 的 *rank* 为 n，次大的 *score* 对应 *sample* 的 *rank* 为 $n-1$，以此类推. 然后把所有正样本的 *rank* 相加，再减去 2 个正样本相组合的情况，得到的结果表示所有样本中有多少对正样本的 *score* 大于负样本的 *score*，然后再除以 $M \times N$，即 ：

$$AUC = \frac{\sum_{i \in positive\ class} rank_i - \frac{M(M+1)}{2}}{M \times N} \tag{9}$$

2.3　*K* 折交叉验证法（K-CV）

类别数据分布不均衡是分类任务中一个常见的问题，所以在分类模型构建之前，需要对分类不均衡性的问题进行处理. 本数据集中标签为 0 的样本数远大于标签为 1 的样本数.

本文采用 K 折交叉验证法来解决这一问题[12]. K 折交叉验证实质上是将实验重复地进行 K 次，首先将原始数据随机分成 K 个数据集，每次实验都从这 K 个数据集中选择一个不同的集合作为测试集，剩余的 $K-1$ 个作为训练集进行实验，最后将得到的 K 个实验结果取平均值.

K 折交叉验证能够避免欠拟合和过拟合，更好地说明模型结果. 在 XGBoost 库中，通过 xgb.cv 函数进行交叉验证.

3　实验结果与分析

3.1　变量重要性分析

通过 XGBoost 可以判断每个特征变量对模型的贡献程度，从中可知哪些特征变量对于用户使用优惠

券行为的影响更为显著. 分析结果如图1所示.

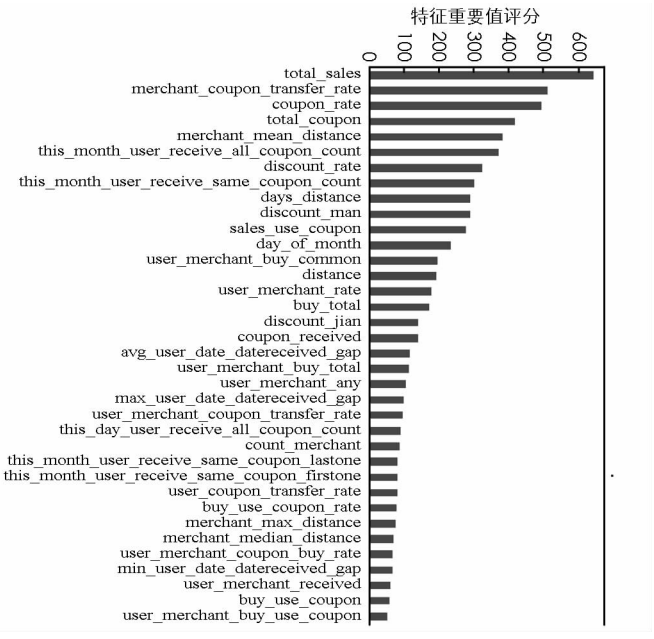

图1 特征重要性评分

由图1可知，商户特征集对结果的贡献度最大. 商户被消费次数(total_sales)，商户优惠券转化率(merchant_coupon_Transfer_rate)，优惠券消费占总消费比例(coupon_rate)，商户发放优惠券次数(total_coupon)，使用优惠券消费的用户与商户的平均距离(merchant_mean_distance)，这5个商家特征拥有最高的重要性评分，很大程度上说明商家想要提高优惠券核销率，需要为消费者提供更大的感知便利性. 具体方法可以从两方面着手：①增加优惠券发放次数，优化发放渠道；②缩短线下活动与目标用户群之间的距离，提高线下网点覆盖率.

优惠券折扣率(discout_rate)，消费日期(day_of_month)，用户从领取优惠券到消费的平均时间间隔(avg_user_date_datereceived_gap)，用户与商户交互的比例(user_merchant_rate)，这些特征也说明省钱体验、客户消费习惯和客户对商家的忠诚度也是消费者使用的重要驱动力. 而且，在节假日之前发放优惠券，也可以提高优惠券的核销率.

3.2 其他分类算法

为了比较模型的性能，本文也采用了另外2种常用的分类算法——随机森林和梯度提升树. 随机森林(Random Forest)，是利用多棵CART树对样本进行训练并预测的一种分类器. 随机森林能处理连续、离散变量，适用于多分类问题，在一定程度上可以防止过拟合，模型稳定性强，对噪声不敏感，能并行分布式处理.

GBDT(Gradient Boosting Decision Tree)又叫MART(Multiple Additive Regression Tree). GBDT是一种迭代的决策树算法，该算法由多棵决策树组成，将所有树的结果累加起来得出最终答案. GBDT的优点是对离散和连续型变量的处理很灵活，而且不需要进行特别复杂的特征工程. GBDT的缺点也很明显，计算复杂度高，对高维稀疏特征的应用性能较差.

3.3 算法结果比较

分别采用GBDT，RandomForest和XGBoost对该口碑网爬取的优惠券使用行为数据集进行分类预测，训练集与测试集的比例均为7∶3，进行多次调参优化后取最好结果如图2所示. 由图2中显示的AUC score和Accuracy可知，无论是对准确率和召回率的平衡能力，还是对0－1标签正确分类的能力，XGBoost对该不平衡数据集的分类效果显著强于RandomForest与GBDT.

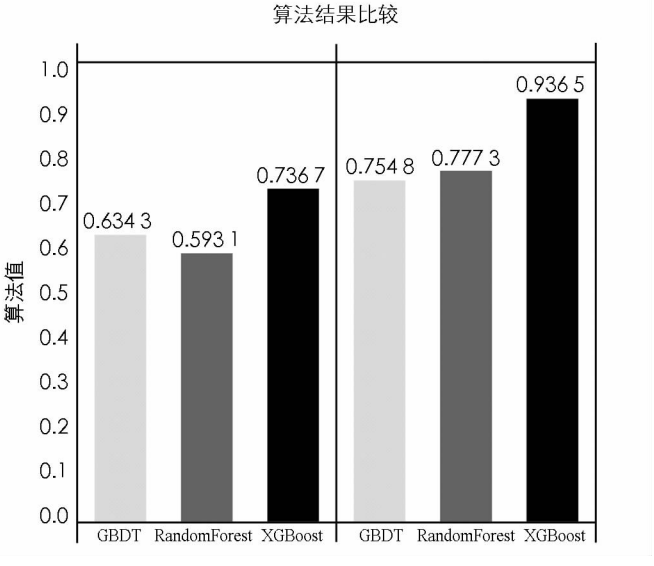

图2 各算法准确率对比

4 总 结

优惠券营销作为服务商连接新零售线上线下交易的桥梁，是一种非常重要的营销新趋势. 所以，预测用户使用情况并改进投放方法是商家面临的重要问题.

本文应用口碑网真实的消费者线上线下交易数据，突破了大部分研究以 TAM 模型为理论基础构建个人优惠券使用意愿的传统方法，并将商户属性、优惠券属性和用户－商户交互属性引入影响因素．通过基于集成学习的 XGBoost 进行消费券使用行为预测，并与随机森林、GBDT 算法进行了比较，结果表明 XGBoost 显著提升了消费券使用的预测准确率．

本文还通过对变量重要性的分析，确定了对消费者使用决策贡献度较高的变量．该研究有助于理解优惠券核销率的影响因素，以及用户收到优惠券之后的购买行为决策，对优惠券的投放与精准营销有重要的现实意义．

参考文献：

[1] WANG S, ZHANG Y. The New Retail Economy of Shanghai [J]. Growth & Change, 2005, 36(1): 41-73.
[2] 赵树梅，徐晓红．“新零售”的含义、模式及发展路径 [J]．中国流通经济，2017，31(5)：12-20.
[3] LUO Jia-ning. An Offline Transferable and Divisible Mobile Coupon Based on NFC [C]//Proceedings of International Conference on Computing, Mechanical and Electronics Engineering. Sin gapore: International Institute Engineers, 2015.
[4] CHEN T, GUESTRIN C. XGBoost: A Scalable Tree Boosting System [C]//ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Francisco: ACM, 2016.
[5] IM H, HA Y. Determinants of Mobile Coupon Service Adoption: Assessment of Gender Difference [J]. International Journal of Retail & Distribution Management, 2014, 42(5): 441-459.
[6] 汪明远，赵学锋．消费者调节定向和从众行为对移动优惠券使用意愿的影响研究 [J]．管理学报，2015，12(7)：1045-1050.
[7] 王小平，李柳柏．基于 AdaBoost 算法的图像自动标注 [J]．西南大学学报(自然科学版)，2015，37(7)：174-180.
[8] 任　浩，叶　亮，李　月，等．基于多级 SVM 分类的语音情感识别算法 [J]．计算机应用研究，2017，34(6)：1682-1684.
[9] 王重仁，韩冬梅．基于社交网络分析和 XGBoost 算法的互联网客户流失预测研究 [J]．微型机与应用，2017，36(23)：58-61.
[10] 巫红霞．基于改进 Shapley 权力指数的特征选择算法 [J]．西南师范大学学报(自然科学版)，2017，46(11)：62-71.
[11] 龚艳萍，许志忠．优惠券促销有效性的影响因素研究 [J]．全国商情(经济理论研究)，2008(2)：50-51.
[12] 胡局新，张功杰．基于 K 折交叉验证的选择性集成分类算法 [J]．科技通报，2013，29(12)：115-117.

On XGBoost-Based Prediction of New Retail Coupon Usage Behavior

XU　Ning,　LA　Lei

College of Information Technology & Management, University of International Business and Economics, Beijing 100029, China

Abstract: To achieve targeted delivery of new retail coupons, a model for predicting coupon usage behavior has been proposed. The XGBoost algorithm is adopted, which breaks through the traditional method of interpreting personal willingness to use coupons based on the TAM model. Feature extraction and user behavior modeling are formulated based on real transaction data. After K-fold cross-validation, the variable importance score was used to determine characteristics that have a significant contribution to consumer decision-making. AUC accuracy comparison with random forest and GBDT algorithm is also performed. This research proves the effectiveness of the XGBoost-based ensemble algorithm in predicting the use behavior of coupons and has important practical significance for precise new retail marketing.

Key words: XGBoost; coupon usage prediction; new retail

责任编辑　夏　娟