技术应用

基于机器学习组合模型的用户消费行为预测

张　峰,张丽娜,李静静

(河北工程大学,河北　邯郸　056038)

摘　要:大数据背景下,如何利用大量的用户消费行为数据精准识别高质量的用户和渠道,进而预测用户的未来需求,已成为企业面临的难题,对用户消费行为进行深入挖掘与分析具有重要的现实意义。本文首先对用户消费行为数据进行可视化分析;其次,结合随机森林和 Logistic 模型,提出了一种基于机器学习组合模型的用户消费行为预测方法。结果表明:该方法的 AUC 值为 94.85%,明显优于其他模型。该方法可为企业分析用户消费行为、制定科学的营销策略提供借鉴。

关键词:用户消费行为;组合模型;随机森林;Logistic

【中图分类号】TP181;F274　　【文献标识码】A　　【文章编号】2096-6059(2022)02-019-09

Prediction of User Consumption Behavior Based on Machine Learning Combination Model

ZHANG Feng,ZHANG Li´na,LI Jingjing

(Hebei University of Engineering,Handan　056038,China)

Abstract: In the context of big data,it has become a difficult problem for enterprises to accurately identify high-quality users and channels by using a large number of user consumption behavior data,so as to predict the future needs of users. Therefore,in-depth mining and analysis of user consumption behavior has important practical significance. Therefore,this paper first carries on the visualization analysis to the user consumption behavior data; Secondly,combined with random forest and Logistic,a predictive method of user consumption behavior based on machine learning combination model is proposed. The results show that the AUC value of this method is 94.85%,which is obviously better than other models. It can provide a good reference for enterprises to analyze consumer behavior and formulate reasonable strategies.

Key words: user consumption behavior; combination model; random forest; logistic

0　引言

在企业的日常运营中,无论是线上还是线下都产生了大量的用户消费行为数据。这些数据为企业带来了新的发展机遇但也使企业面临巨大的挑战,如何判别高质量的用户和渠道、优化营销成本成为各领域企业的痛点。对企业而言,传统的营销渠道已经无法有效地满足用户的个性化和多样化需求,互联网和数据挖掘技术的发展,为公司拓宽了获客渠道。因此,以用户消费行为分析为核心,依托互联网技术和数据挖掘技术的精准营销日渐成为各大企业关注的焦点。2020 年,根据中研普华产业研究

收稿日期:2022-03-16

作者简介:张　峰(1996-),男,安徽滁州人,硕士研究生,主要研究方向:模糊数学、机器学习;
张丽娜(1996-),女,安徽阜阳人,硕士研究生,主要研究方向:统计学习;
李静静(1996-),女,河北邢台人,硕士研究生,主要研究方向:时间序列分析。

基金项目:河北省自然科学基金项目(A2021402008,F2021402010);河北省高等学校科学研究项目(ZD2020185,QN2020188)

院调研数据显示，有超过 60%的企业将大数据应用于营销分析[1]，用户行为数据对企业的帮助作用日益突出，各企业也逐渐开始重视将大数据加入营销的各个环节。因此，如何充分挖掘和分析用户消费行为数据，并依据分析结果制定营销策略，已成为各行业企业亟需解决的重要问题。

为提高对用户消费行为的精准预测，针对用户消费行为预测方法，国内外学者做了一些研究。Schmittlein 等[2]针对用户消费行为预测问题提出了经典的概率预测模型，即 Pareto/NBD 模型。李美其和齐佳音[3]基于大众点评网站的用户数据，使用 Pareto/NBD 模型对用户购买行为进行预测，实验表明该方法的精度得到了提升。随着机器学习方法的发展，不少学者开始将机器学习方法应用到用户消费行为预测问题上。白婷等[4]利用网站上的用户消费行为数据，提取有效特征，使用加权 GBDT（Gradient Boosting Decision Tree）模型对用户购买商品进行了预测。葛绍林等[5]提出深度森林模型，对用户消费行为进行预测分析，结果表明该方法具有较好的预测效果。

因为用户消费行为预测问题的复杂性，单一模型常常会产生过拟合现象，因此也有不少学者利用组合模型对用户消费行为数据进行挖掘和预测分析。张韶[6]基于京东大数据平台上的真实数据，经过数据处理和特征选择，然后选取了 LightGBM、CatBoost 和 XGBoost 模型进行单项训练，通过加权投票和 Stacking 融合策略构建组合模型，并进行对比实验，结果表明基于加权投票的组合模型的预测效果要优于其余单项模型。张建彬和霍佳震[7]基于已有的销售数据，提出了一种基于机器学习和 Stacking 集成的综合预测模型，结果表明该融合模型的预测效果优于单一模型，准确率达 85%。

综上所述，对用户消费行为预测问题的研究仍处于不断发展阶段，国内外学者从最初的统计学方法发展到现在的机器学习方法，通过模型构建方式对用户消费行为预测进行了深入研究。然而，在具体的实际问题中，当前方法的预测性能还不是十分理想。因此，本文将针对某平台上的用户消费行为数据，分析用户消费行为与商品之间的潜在关系，结合处理效率较高的随机森林和 Logistic 模型，提出一种基于组合模型的用户消费行为预测方法，以提升用户的购买转换率，增强预测模型对实际问题的适用性。

1 数据来源及清洗

1.1 数据来源

数据来源于 2021 年全国大学生数据统计与分析竞赛（https://m.saikr.com/dsa/2021），原始数据集包括用户信息表（user_info）、用户登录情况表（login_day）、用户访问统计表（visit_info）、用户下单表（result）4 部分，各部分的特征字段和样本情况，如表 1 所示。

1.2 数据清洗

由于原始数据中存在大量缺失、异常以及重复等情况，为了对用户的消费行为进行可视化和预测分析，本文需要对初始数据进行清洗，进一步提高数据集的质量。

首先，对缺失值进行删除。缺失数据是指数据集中存在空白或未知数据的情况。针对用户信息表中“城市”字段存在缺失（共计 28209 条）问题，进行删除处理。

其次，对异常值进行清除。异常值是指在数据记录中存在不符合实际情况的数据，比如在用户登录情况表和用户访问统计表中，用户没有领券访问次数的记录却存在已经领券的情况、平台开课数为 0 但用户学习课节数和完成课节数不为 0 的情况、

表 1 用户消费行为数据情况

Table 1 Data of user consumption behavior

数据	特征字段	字段数	样本数
用户信息（user_info）	ID、age、city 等	8	135968
用户登录情况（login_day）	ID、登录天数、领券数量等	16	135617
用户访问统计（visit_info）	ID、首页访问数、是否领券访问数等	26	135617
用户下单表（result）	ID、是否购买	2	4639

用户登录时长为 0 但用户的登录天数和最后登录距期末天数的值却不为 0 等多种不切实际的情况，约占整体数据的 18.66%。将这些异常值进行删除，剩余有效数据共计 110306 条。

再次，对重复值进行处理。重复数据是指同一数据多次出现的情况，比如在用户下单表中，用户 ID 为 “2000002390697240”、“2000002516432100”和“2000002480841520”等均重复出现多次，在用户信息表中用户 ID 为 “2000002352923140”、“200000235 2922980”的用户均重复出现多次。因此，本文对用户信息表中的 9979 条重复值、用户登录情况表与用户访问统计表中的 4 条重复值、用户下单表中的 13 条重复值进行删除。

通过上述步骤对 4 个部分的数据进行清洗处理后，以用户 ID 进行匹配合并，得到新的样本数据共计 86776 条。

2　数据的可视化分析

为找出其中的行为规律以及挖掘数据中更为丰富的潜在价值，本文根据数据清洗得到的用户消费行为数据进行可视化分析。这里主要对数据集中的用户城市分布情况、用户登录情况（包括登录天数、登录间隔、最后登录距期末天数和登录时长）两个方面进行可视化分析。

2.1　用户城市分布情况

对数据中的城市字段（city_num），首先按照各城市所属的省（市、自治区）进行统计划分，然后统计各省市中总用户数量和购买用户数量，最后借助 ArcGIS 软件，利用自然间断法将用户数量分成 5 个等级（city_rank），可视化结果如图 1 所示。

在图 1 中，左图为总用户数量地区分布情况，右图为购买用户数量地区分布情况。由于不同地区的用户数量不同，在图中呈现出的颜色存在较大差异，颜色越深表示该省市用户数量越大，反之用户数量越小。从图中可以看出，总用户数量和购买用户数量在空间分布上不均匀，呈现“东高西低、南高北低”的空间分布格局，其中购买用户数量在空间分布上的这种格局表现尤为显著。总用户数量较高的地区主要集中在重庆、广东、四川、山西、山东；对应的下单购买用户数量较高的地区主要集中在东部沿海地区和经济发达地区。而青海、西藏等省市由于人口基数小、互联网普及率相对较低等原因，用户数量较少。

2.2　用户登录情况

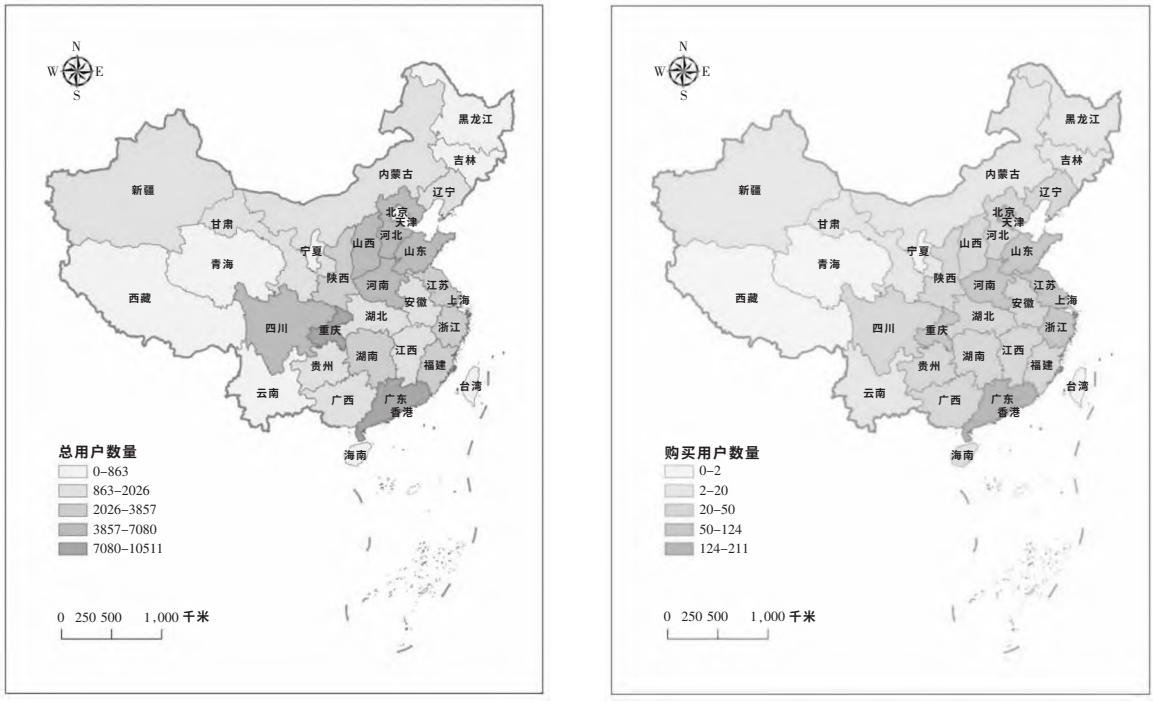

图 1　总用户和购买用户所在地区空间分布图

Figure 1　Spatial distribution of users and purchasing users

从登录天数（login_time）、登录间隔（login_diff_time）、最后登录距期末天数（distance_day）和登录时长（login_time）四个方面对用户的登录情况进行分析，由于字段中的数据均为离散型数据，因此先对数据进行分段处理，统计该区间内用户数量并绘制图表，如图 2、图 3、图 4 及图 5 所示，其中折线表示总用户数量，条形图表示购买的用户数量。

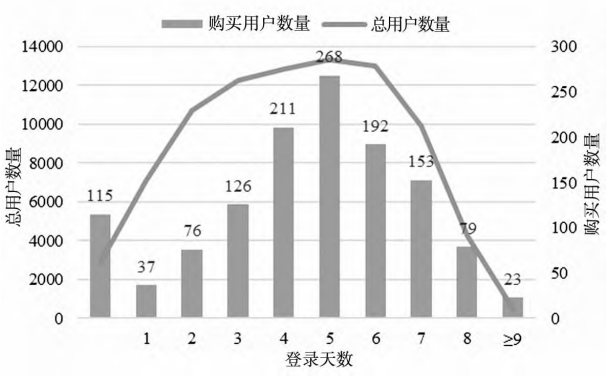

图 2　用户登录天数（login_day）情况

Figure 2　Number of login days

如图 2 所示，横轴表示用户登录的天数，纵轴表示总用户数量和购买用户数量。可以看出，随着用户登录天数的增加，总用户数量和购买用户数量变化趋势基本相同，都呈现出先上升后下降的趋势，但是购买用户数量的下降趋势与上升趋势相比较为平缓，总用户数量的上升趋势与下降趋势相对较为平缓。当登录天数为 5 时，总用户数量和购买用户数量同时达到峰值，此时总用户数量为 13307，约占总体的 15.33%，其中购买用户数量为 268。

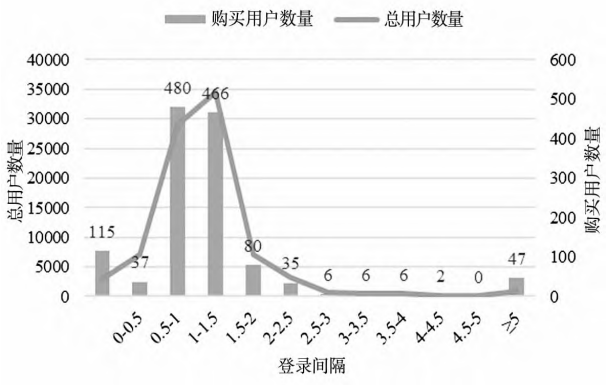

图 3　用户登录间隔（login_diff_time）情况

Figure 3　User login interval（login_diff_time）

如图 3 所示，横轴表示用户登录间隔，纵轴表示总用户数量和购买用户数量。可以看出，无论是购买用户还是未购买用户，其登录时间间隔都较为集中，主要分布在 0.5~1 和 1~1.5 两个时间间隔内。在该间隔内的购买用户总数达 946 人，占总体比例的 73.9%；用户数达 63608 人，占总用户数量的 73.3%。这说明选择购买的用户一般登录的时间间隔都比较短，会及时地进行登录并产生消费行为。时间间隔为 0.5~1 的总用户数量少于时间间隔为 1~1.5 的用户数量，但是时间间隔为 0.5~1 的购买用户数量却多于时间间隔为 1~1.5 的购买用户。同时，当登录间隔超过 1.5 时，随着登录间隔的增加，总用户数量和购买用户数量逐渐趋近于 0。

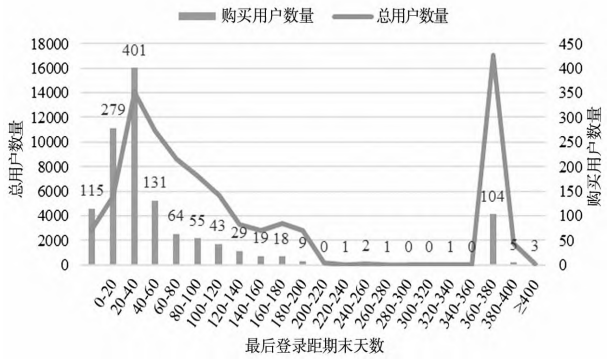

图 4　用户最后登录距期末天数（distance_day）情况

Figure 4　The number of days between the user's last login and the end of the term（distance_day）

如图 4 所示，横轴表示用户最后登录距期末的天数，纵轴表示总用户数量和购买用户数量。可以看出，总用户数量和购买用户数量都随着最后登录距期末天数的增加呈现先增加后减小的走势，但是在最后登录距期末天数为 360~380 范围内的总用户数量和下单购买的用户数量陡然上升，且总用户数量达到最高。这说明存在大量的用户在近一年的时间内都未曾消费该企业的产品，其中包含 104 个下单购买过的用户，表明该企业存在用户大量流失的情况。其次，购买过的用户和其他用户一般最后登录距期末天数集中于 0~60 这个范围内；其中处于20~40 范围内的人数最多，占购买用户数的比例为 31.33%，占总用户数的比例为 16.15%，说明一般用户的登录周期可能在 20~40 之间。

如图 5 所示，横轴表示用户登录时长，纵轴表示总用户数量和购买用户数量。可以看出，随着登

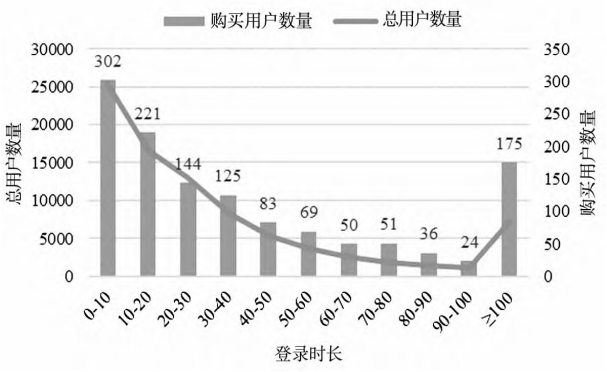

图 5　用户登录时长(login_time)情况

Figure 5　User login duration (login_time)

录时间越长，总用户数量和购买用户数量越来越少，并逐渐趋近于 0。在登录时长为 0~10 的范围内，登录的总用户数量最多，此时购买的用户也高达 302 人，占总购买用户的 23.57%，说明用户在登录前已经具有明确的消费目标。随着登录时长的增加，购买的人数逐渐减少，原因是用户不存在明确的消费目标，只是随机浏览并进行一些非理性的消费。

3　数据的特征选择

通过对数据的清洗，共收集有效数据 86776 条，包含 49 个字段。特征个数过多，会增加模型的训练时间成本、模型的复杂度，易发生过拟合问题，因此需要进行特征选择。特征选择的目的在于根据统计学方法或机器学习模型的特征需要找到最优的特征子集。为构建更准确的用户消费行为预测模型，本文将分两个步骤对数据集进行特征选择，即剔除无关变量与 Lasso 特征选择。

首先对数据集中的无关变量字段进行剔除，包括 user_id、app_num、model_num、platform_num、age_month、first_order_time 和 first_order_price 等，从而降低数据量，提高运算速度。处理后数据集的具体变量描述如表 2 所示。

表 2　用户消费行为数据变量及其描述

Table 2　Data variables of user consumption behavior and their description

符号	变量名	变量含义	符号	变量名	变量含义
x_1	main_home	首页访问数	x_{22}	baby_info	宝宝访问数
x_2	main_home2	首页访问数	x_{23}	click_notunlocked	课程未购买弹窗访问数
x_3	mainpage	课程计划访问数	x_{24}	share	点击分享访问数
x_4	schoolreportpage	课程访问数	x_{25}	click_dialog	首页广告弹窗点击访问数
x_5	main_mime	我的访问数	x_{26}	login_day	登录天数
x_6	lightcoursetab	轻课访问数	x_{27}	login_diff_time	登录间隔
x_7	main_learnpark	学习乐园访问数	x_{28}	distance_day	最后登录距期末天数
x_8	pgbarrierspage	小屋首页访问数	x_{29}	login_time	登录时长
x_9	evaulationcenter	测评中心访问数	x_{30}	launch_time	隔天再次访问落地页数
x_{10}	coupon_visit	是否领券访问数	x_{31}	subscribe1_num	关注公众号 1
x_{11}	click_buy	购买按钮点击访问数	x_{32}	subscribe2_num	关注公众号 2
x_{12}	progress_bar	拖动进条访问数	x_{33}	add_friend	添加销售好友
x_{13}	ppt	ppt 下一步访问数	x_{34}	add_group	进群
x_{14}	task	任务结束页访问数	x_{35}	camp_num	开课数
x_{15}	video_play	视频跟读访问数	x_{36}	learn_num	学习课节数
x_{16}	video_read	界面继续访问数	x_{37}	finish_num	完成课节数
x_{17}	next_nize	识字访问数	x_{38}	study_num	课程重复学习
x_{18}	answer_task	答案解析访问数	x_{39}	coupon	领券数量
x_{19}	chapter_module	点击模块访问数	x_{40}	course_order_num	有年课未完成订单
x_{20}	course_tab	今日课程访问数	x_{41}	city_rank	城市等级
x_{21}	slide_subscribe	上课页访问数	y	result	是否下单购买

注：除 city_rank 与 result 是字符型变量外，其余变量皆为数值型。

其次，进行 Lasso 特征选择。用户消费行为指标应具有较强的解释意义，并且要符合一定逻辑。然而该数据集中的指标变量包含着大量的冗余信息，这将直接影响用户消费行为预测模型的性能，甚至会出现较大的偏差。因此，还需对上述指标变量进行二次选择，选择出更具重要性的指标。此外，考虑到变量间多重共线性对模型的影响，尤其是对 Logistic 模型的解释性会产生极大影响，所以选用 Lasso 方法进行变量选择，以有效克服上述问题[8]。

经过上述两个步骤，最终选择出 20 个有效变量 $x_1,x_2,x_3,x_{10},x_{12},x_{13},x_{15},x_{16},x_{18},x_{19},x_{20},x_{24},x_{26},x_{27},x_{28},x_{29},x_{36},x_{39},x_{40},x_{41}$。最后重新组成新的样本数据，有效数据样本总计 84104 条，其中下单购买的用户样本有 1209 个，未购买的用户样本有 82895 个。

4 基于随机森林和 Logistic 回归的用户消费行为预测

用户消费行为预测是一个典型的机器学习分类任务。因此，选取了处理效率较高的随机森林（Random Forest，RF）和 Logistic 回归对用户消费行为数据进行学习。

4.1 随机森林

随机森林[9]是一种集成多棵决策树的集成学习算法，可用于解决分类及回归问题。随机森林的“随机”体现在两个方面：

（1）随机抽取样本。针对分类问题，RF 的训练集通过有放回的自助法随机产生，每一轮训练所使用的训练集均以同样方式生成，以保证所有样本都有机会参与训练。

（2）随机属性选择。首先从该节点的全部属性集合中随机抽取若干个属性组成子集；其次从属性子集中找到最优分裂属性进而划分。每一棵决策树在其生成中都会随机生成不一样的分裂属性子集，随机属性选择增强了树之间的独立性，也增加了算法的随机性。

经过模型内部处理，在每个训练集上构建一种决策树，N 棵树就会有 N 种分类结果，根据投票原则，将投票最多的类别指定为模型的最终输出。而正因为随机森林的“随机”，使模型不易过拟合。此外，该模型在处理高维度数据中具有明显的优势，在预测准确度上也有较好的效果。

4.2 Logistic 回归

Logistic 回归[10,11]是将多元线性回归的思想拓展成一种用于解决分类问题的模型。该模型对数据分布没有严格的条件，并且具有结构简单、参数易解释、节约算力、稳健性较好等优点。假设 y 表示用户是否下单购买，即“0”表示未下单购买，“1”表示下单购买。若模型的预测结果是 $y=1$ 的概率，其表达式可以表示为：

$$P(y=1)=\frac{\exp(\omega\cdot x+b)}{1+\exp(\omega\cdot x+b)} \tag{1}$$

其中，$x\in R^n$ 是输入，$y\in\{0,1\}$ 是输出，$\omega\in R^n$ 和 $b\in R$ 是参数，ω 称为权值向量，b 为偏置，$\omega\cdot x$ 为 ω 和 x 的内积。模型的输出结果可通过与阈值 0.5 比较，若大于 0.5，则表示下单购买，否则表示未下单购买。

4.3 基于随机森林和 Logistic 回归的组合模型

经过数据清洗和特征选择后，新的用户消费行为数据共计 84104 条，其中下单购买的用户样本有 1209 个，未购买的用户样本有 82895 个，存在着严重的类别不平衡问题。因此，本文采用欠采样技术[12]进行数据层面上的处理，以平衡正负类样本数量。首先，从未购买用户样本中随机抽取 1209 个样本，与已购买用户的 1209 个样本组成第一平衡训练集。其次，从未购买用户样本与已购买用户样本中分别随机抽取 800 个样本，组成第二平衡数据集，并按 8:2 对其划分数据集。

为进一步提高用户消费行为预测模型精度，将 RF 与 Logistic 模型进行串行组合，其构建原理如图 6 所示。RF 与 Logistic 的组合模型具体构建思路[13]：首先，用第一平衡训练集对 RF 进行训练。其次，将训练好的 RF 对第二平衡数据集进行预测，将得到的输出结果作为一个新的输入变量添加到 Logistic 模型中，而 Logistic 模型中其他的输入变量保持不变，得到组合模型。

最后，本文将第二平衡数据集的训练集部分用朴素贝叶斯（Naïve Bayes，NB）、支持向量机（Support Vector Machine，SVM）等其他单一模型进行训练，并在测试集上作对比，以保证各自模型最终所得出的预测准确率在比较分析中更具有说服力。

5 对比实验

5.1 模型性能评估指标

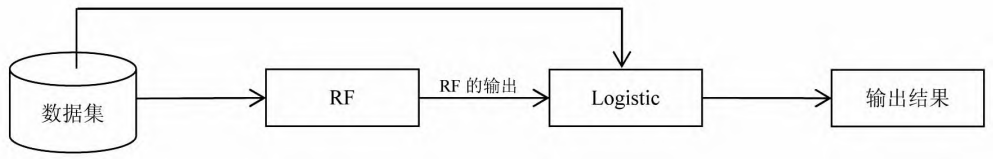

图 6　RF-Logistic 组合模型的构建原理

Figure 6　Construction principle of RF-Logistic combination model

根据用户消费行为预测用户是否购买产品，是一个典型的二分类任务。本文使用二分类问题中常用的评估指标，包括准确率 A（Accuracy）、精确率 P 以及 F_1 分数来评估模型性能[14]。

$$A=\frac{TP+TN}{TP+TN+FP+FN} \tag{2}$$

$$P=\frac{TP}{TP+FP} \tag{3}$$

$$F_1=\frac{2TP}{2TP+FP+FN} \tag{4}$$

其中，FN 表示正类样本（即标签“1”）预测为负类（即标签“0”）的样本数，TP 表示正类样本预测为正类的样本数，FP 表示负类样本预测为正类的样本数，TN 表示负类样本预测为负类的样本数。

此外，本文采用 ROC（Receiver Operating Characteristic）曲线和 AUC（Area Under the Curve of ROC）值来验证模型的判别能力和预测精度。ROC 曲线[15]一般应用于二分类模型的评估，其绘制方法基于两个重要的指标，即灵敏度（True Positive Rate，TPR）和特异度（False Positive Rate，FPR）。灵敏度表示预测为正类的样本数占所有正类样本数的比例；特异度是指当前被误分到下单购买用户中真实的没有下单购买的用户占所有用户数的比例。其具体计算公式如下：

$$\mathrm{TPR}=\frac{TP}{TP+FN} \tag{5}$$

$$\mathrm{FPR}=\frac{FP}{TN+FP} \tag{6}$$

根据以上原理，对样本按预测结果排序，再对每个样本分别以 TPR 和 FPR 为坐标点绘制 ROC 曲线。若坐标点离左上角越近，则表示分类器的预测准确率越高；若坐标点离右下角越近，则表示其预测准确率越低。此外，若画出的曲线足够平滑，基本可以判断没有太过拟合。ROC 曲线线下面积即 AUC 值，也是分类任务中的常用评估指标。若AUC 值越大，表明模型的分类准确率就越高。

5.2　结果分析

本文将构建的组合模型和常用的单一模型在同一测试集上进行预测，其分类效果如表 3 所示。

表 3　常用分类模型和组合模型的测试评估指标结果

Table 3　Test evaluation index results of common classification models and combination models

Model	A	P	F_1	AUC
NB	0.6688	0.7778	0.5431	0.7198
SVM	0.6781	0.7087	0.6589	0.7214
Logistic	0.7156	0.7586	0.6592	0.7481
RF	0.7406	0.7833	0.7367	0.7857
RF-Logistic 组合模型	0.9250	0.8671	0.9259	0.9485

由表 3 可知，NB 与 SVM 模型的预测准确率 A 值分别为 66.88%和 67.81%，其分类效果在其余评价指标上也表现得非常不理想；Logistic 和 RF 模型的预测准确率分别为 71.56%和 74.06%，均显著优于 NB 和 SVM 模型；RF-Logistic 组合模型的预测准确率高达 92.50%，与 Logistic 和 RF 模型相比，在预测准确率上分别提高了 20.94%和 18.44%。综合来看，RF-Logistic 组合模型的 F_1 分数高达 92.59%，与 NB、SVM、Logistic 和 RF 单一模型相比，组合模型的分类效果得到了大幅度提高。

此外，本文绘制了 ROC 曲线来进一步直观地反映组合模型与其他单一模型在下单购买行为预测（即标签“1”）上的分类效果，如图 7 所示。

从图 7 可知，NB 模型、SVM 模型、Logistic 模型、RF 模型以及 RF-Logistic 组合模型的 ROC 中 AUC 值分别为 0.7198，0.7214，0.7481，0.7857 和 0.9485。其中 RF-Logistic 组合模型的 AUC 值最高，说明组合模型对判别用户是否购买的分类效果较好。根据以上评估结果及分析，验证了 RF-

Logistic 组合模型可作为最终的用户消费行为预测模型。

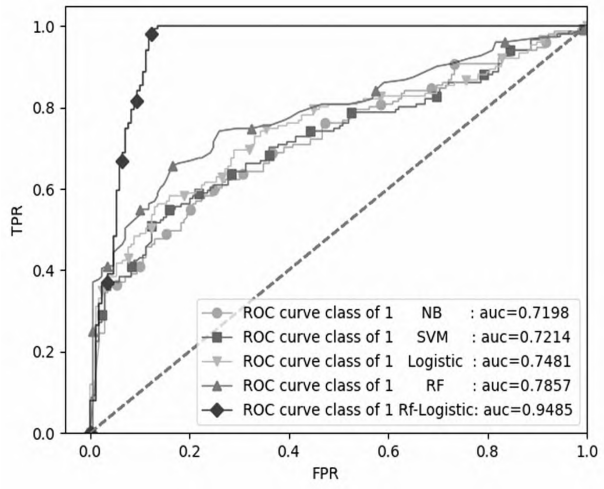

图 7　组合模型与常用分类模型的 ROC 曲线

Figure 7　ROC curves of the combined model and the commonly used classification model

6　结论

本文通过对用户消费行为数据进行清洗、可视化分析和特征选择，建立了基于随机森林和 Logistic 回归的用户消费行为预测模型，较大地提高了用户消费行为数据的预测准确率。这为帮助企业分析用户的消费行为规律、判别高质量用户，进而制定合理的营销策略提供了有益的借鉴。结合模型的分析结果，本文提出了如下的营销建议。

6.1　地区差异性营销

在对用户城市分布情况进行可视化分析后，发现总用户数量和购买用户数量存在同增同减关系，且在空间分布上呈现“东高西低、南高北低”的不均匀空间分布格局。根据这一现象，相关企业可以按区域进行营销，若营销推广所在区域位于东南地区，可以采取促销满减、开通会员打折等形式吸引用户注册购买，增加用户数量和购买用户数量。若营销推广所在区域位于西北地区，应该先注重增加用户数量，因为位于这些地区的用户往往比较看重价格，因而可采取降价、打折等形式进行促销。

6.2　登录签到有奖

在对用户登录情况进行可视化分析后发现，用户登录的时间间隔越短，下单购买的用户越多。根据这一情况，企业可以采取一定的措施来减少用户登录的时间间隔，即增加用户登录的次数，来促进用户下单购买，比如对登录时间间隔较长的用户设置签到有奖的活动，签到的次数越多，获得的奖励就越多越丰厚。此外，还可以设置用户在有限时间内登录平台达到一定次数时发放优惠券，来调动用户登录的积极性。

6.3　优惠券精准投放

根据用户消费行为数据分析，其中领券访问数仅占 7%，而发生领券购买行为的用户数仅占 4%，说明优惠券并未达到预期的营销效果。这说明企业在投放优惠券的时机选择和人群选择上具有较大的盲目性，定位不够准确，并且部分用户在浏览平台产品时收到不感兴趣的优惠券推送消息时，会产生反感情绪而导致用户不断流失。因此，建议企业在发放优惠券时，按照场景进行设定：获取新用户、提高活跃度、提高转化率和自传播。针对从未下单购买的用户，可以通过注册激活发券、下单有礼等方式获取新用户；通过发放优惠券的方式将已注册激活的用户唤醒召回；通过满减券或者折扣券来实现用户从低价值向高价值的转化。通过场景设定，将优惠券发放给最有可能使用的人，以达到精准投放的目的。同样，也可考虑设定优惠券的具体面值、有效期和使用范围。

6.4　社交分享激励

根据数据分析，点击分享访问的用户占比 61.43%，说明用户乐于与好友互动，将产品分享给好友。相关企业可以通过增加分享、关注、进群、做任务、添加好友等社交互动方式，鼓励用户和亲朋好友一起参与，促进用户的增长，提升用户的触达范围和转化效果：通过登陆和访问页面的推送，激励用户点击分享内容，提高用户活跃度，同时促进产品的宣传和推广。

参考文献：

[1] CIRN. 2020-2025 年中国大数据应用行业全景调研与发展战略研究咨询报告[R]. 深圳：中研普华产业研究院，2020.

[2] Schmittlein D C，Morrison D G，Colombo R. Counting your customers: Who-are they and what will they do next?[J]. Management Science，1987，33(1): 1-24.

[3] 李美其，齐佳音. 基于购买行为及评论行为的用户购买预测研究 [J]. 北京邮电大学学报（社会科学版），2016，18(4): 18-25.

[4] 白婷，文继荣，赵鑫，等. 基于迭代回归树模型的跨平台长尾商品购买行为预测[J]. 中文信息学报，2017，31(5)：185-193.
[5] 葛绍林，叶剑，何明祥. 基于深度森林的用户购买行为预测模型[J]. 计算机科学，2019，46(9)：190-194.
[6] 张韶. 基于机器学习的用户购买行为预测研究 [D]. 西安：长安大学，2020.
[7] 张建彬，霍佳震. 基于 Stacking 模型融合的用户购买行为预测研究[J]. 上海管理科学，2021，43(1)：12-19.
[8] 白玥，田茂再. 几种高维变量选择方法的比较及应用[J]. 统计与决策，2017(22)：11-16.
[9] Breiman L. Random Forests [J]. Machine Learning，2001，45(1)：5-32.
[10] 金海月. 逻辑斯蒂回归模型在电信领域中的应用[J]. 沈阳理工大学学报，2018，37(2)：34-38.
[11] 马文苑，冯仲科，成竺欣，等. 山西省林火驱动因子及分布格局研究 [J]. 中南林业科技大学学报，2020，40(9)：57-69.
[12] Devi D，Biswas S K，Purkayastha B. Learning in presence of class imbalance and class overlapping by using one-class SVM and undersampling technique [J]. Connection Science，2019，31(2)：105-142.
[13] 冉霞. 基于机器学习组合模型的乳腺癌生存预测[D]. 济南：山东大学，2020.
[14] 李航. 统计学习方法[M]. 北京：清华大学出版社，2019.
[15] Mamitsuka H. Selecting features in microarray classification using ROC curves[J]. Pattern Recognition，2006，39(12)：2393-2404.

(上接第 13 页)

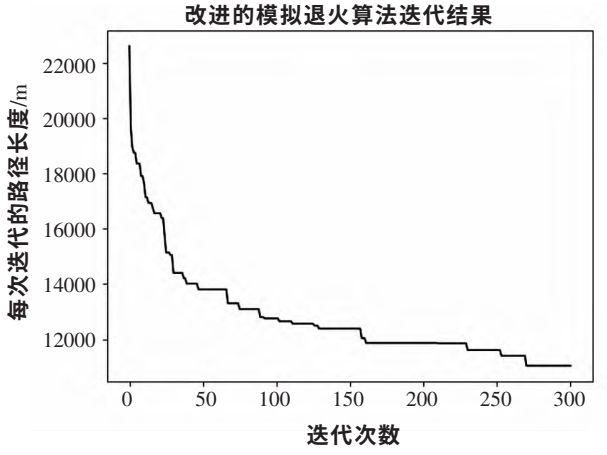

图 6　改进模拟退火算法的迭代过程图

Figure 6　Iterative process diagram of the improved simulated annealing algorithm

4　结语

本文将无线传感器网络中单车充电路径规划问题抽象为 NP-hard 经典旅行商问题(TSP)，建立数学模型，并提出基于动态概率的改进模拟退火算法。改进算法通过动态概率和交换法、移位法、倒序法调节模拟退火算法的参数来优化局部寻优能力和全局寻优能力。仿真实验结果表明，与路径最短的贪婪算法和传统模拟退火算法求解结果相比，本文的改进模拟退火算法在求解无线传感器网络中单车充电路径规划问题时具有收敛精度高、全局寻优能力强的优点。

参考文献：

[1] 陈浩，王光林，郝询.移动机器人的路径规划综述与发展趋势[J].物流技术与应用，2020，25(10):158-160.
[2] 梁晓辉，慕永辉，吴北华，江宇.关于路径规划的相关算法综述[J].价值工程，2020，39(03):295-299.
[3] 姚明海，王娜，赵连朋.改进的模拟退火和遗传算法求解 TSP 问题[J].计算机工程与应用，2013，49(14):60-65.
[4] 谢燕丽，许青林，姜文超.一种基于交叉和变异算子改进的遗传算法研究 [J]. 计算机技术与发展，2014，24(04):80-83.
[5] 于莹莹，陈燕，李桃迎.改进的遗传算法求解旅行商问题[J].控制与决策，2014，29(08):1483-1488.
[6] 孙文彬，王江.一种基于遗传算法的 TSP 问题多策略优化求解方法[J].地理与地理信息科学，2016，32(04):1-4.
[7] 徐练淞，潘大志.一种求解 TSP 问题的改进遗传蚁群算法[J].智能计算机与应用，2017，7(03):34-36+40.
[8] 何庆，吴意乐，徐同伟.改进遗传模拟退火算法在 TSP 优化中的应用[J].控制与决策，2018，33(02):219-225.
[9] 包强. 一种求解旅行商问题的混合遗传模拟退火算法[J].中国储运，2021(11):204-205.
[10] 许焕，王博源，庄泽琰.基于模拟退火算法的充电路径规划[J].数学建模及其应用，2021，10(02):65-76.
[11] 深圳市尚龙数学技术与交叉学科产业化研发中心. 2020 年“深圳杯”数学建模挑战赛 C 题[EB/OL].[2020-07-15].http://www.m2ct.org/viewpage.jsp?editId=12&uri=0D00233&gobackUrl=modular-list.jsp & pageType=smxly&menuType= flowUp.
[12] 柳毅，毛峰，李艺.Python 数据分析与实践[M].北京：清华大学出版社，2019.