计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2021,30(10):264−270 [doi: 10.15888/j.cnki.csa.008074]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于时序特征和集成算法的用户购买预测①

盛钟松, 朱海景, 余　谅

(四川大学 计算机学院, 成都 610041)
通讯作者: 余　谅, E-mail: yuliang@scu.edu.cn

摘　要: 大数据时代, 电商平台积累了大量用户在平台上的行为数据, 比如浏览、点击、下单和加入购物车等等. 如何使用机器学习算法去探索大数据背后的用户消费喜好和习惯成为了一个新的研究热点. 本文主要在特征工程和模型搭建两个方面对用户购买预测的效果做出提高. 通过深入理解电商业务知识, 利用统计学知识, 分别从用户、商品和评论数据等多个方面的数据构建了 115 个特征; 在模型搭建方面, 主要设计了一个两层融合模型, 第一层采用了 XGBoost、CatBoost 和逻辑回归作为基分类器, 从不同的角度考虑用户购买预测, 第二层采用加权平均的方法对基类模型的预测结果进行融合, 其权重由线性分类器学习生成. 实验结果表明该融合模型的 *F*1 评分要高于个体分类器, 并且多次实验证明, 融合模型的稳定性也要比个体分类器好.

关键词: 用户购买; 融合模型; 特征工程; 电商平台; 线性分类

引用格式: 盛钟松,朱海景,余谅.基于时序特征和集成算法的用户购买预测.计算机系统应用,2021,30(10):264–270. http://www.c-s-a.org.cn/1003-3254/8074.html

User Purchase Prediction Based on Timing Features and Ensemble Algorithms

SHENG Zhong-Song, ZHU Hai-Jing, YU Liang

(College of Computer Science, Sichuan University, Chengdu 610041, China)

Abstract: In the era of big data, e-commerce platforms have accumulated a large number of user behavior data, such as browsing, clicking, placing orders and adding commodities to shopping carts. How to use machine learning algorithms to explore the consumer preferences and habits behind big data has become a new research hotspot. This study mainly improves the user purchase prediction from two aspects: feature engineering and model building. After the deep understanding of e-commerce knowledge, we have constructed 115 features with statistical knowledge and data from many aspects such as users, commodities and comments. Moreover, a two-layer fusion model is designed. The first layer uses XGBoost, CatBoost, and logistic regression as the base classifiers which predict user purchase behaviors from different perspectives. The second layer employs a weighted average method to fuse the prediction results of the base class model, and its weight is generated by linear classifier learning. The experimental results show that the *F*1 score of the fusion model is higher than that of the individual classifier, and many times of experiments prove that the fusion model has high stability compared with the individual classifier.

Key words: user purchase; fusion model; feature engineering; e-commerce; linear classifier

近年来, 随着互联网技术的高速发展, 网上购物已然成为大部分消费者购物的第一选择, 国内电商平台的规模也越做越大, 电子商务已经成为中国国民经济的重要贡献者. 过去 10 年, 中国网络零售额快速增长, 同比增长 27.3%, 高于世界平均增长速度. 2019 年, 中国网络零售额占比达到新高, 网上零售额占总零售额

① 收稿时间: 2020-12-03; 修改时间: 2021-01-04; 采用时间: 2021-01-20

的 20% 以上[1]. 如此庞大的消费群体, 使得各大电商平台积累了海量的原始数据, 如何从海量的消费者数据中发现消费者购买行为背后的规律开始成为一个新的研究热点. 这对预测消费者未来的购买行为、帮助电商平台实现高效营销以及有效的客户服务具有重要意义.

1 相关工作

目前, 随着数据挖掘技术和机器学习算法的逐渐成熟, 各种预测算法纷纷应用在电商用户购买预测的研究中. Liu 等人[2] 通过建立一个包含 1000 多个特征的预测模型, 分别从用户、品牌和品类等方面对数据进行特征描述, 该模型证实了在预测用户“双十一”之后是否会再次购买商家商品的有效性, 取得了不错的效果. Lee 等人[3] 通过研究多个不同的电子商务网站, 发现不同类型的电子商务网站中不同用户的行为轨迹, 从中发现用户在购买前的一些行为习惯. Liu 等人[4] 通过使用 SVM 的方法对用户网上购买做出预测, 提高了电子商务的产品推荐准确率和转化率. 祝歆等人[5] 通过构建 Logistic 回归-支持向量机融合算法模型对网络购买行为的预测做出研究, 证明混合模型的预测效果要优于个体模型. Dong 等人[6] 分别基于日常场景和促销场景下, 构建基于时间演变的特征, 研究用户品牌的购买预测, 发现用户在促销场景下的购买更多是冲动性的, 而日常场景下的购买则受用户历史行为活动的影响. 随着神经网络和深度学习的发展, 胡晓丽等人[7] 提出了一种基于 CNN-LSTM 的用户购买行为预测模型, 不再人工构建大量的特征, 通过神经网络的方法实现用户和商品特征之间的交互, 模型最终的 $F1$ 值要比基准模型平均提高 7%–11%.

用户购买预测问题可以被描述为一个典型的分类问题, 模型训练和其它分类任务差不多, 特征工程才是机器学习项目成功的关键, 是数据科学不可分割的一部分. 特征工程的工作往往比较困难, 因为它属于特定的领域, 而机器学习算法在很大程度上是通用的. 特征工程中存在很多的尝试和试错, 机器学习项目的大部分工作通常都是在这方面进行的. 虽然各大研究团队已经提出了很多的分类算法, 但在电子商务中预测任务的特征工程方面的文献并不多. 本文在特征工程方面花费了大量的时间, 分别从用户、商品、品类和品牌多个角度构建了大量特征. 我们将描述如何从用户行为数据中生成各种类型的特征, 进行特征选择[8,9], 并通过实验验证了这些特征的重要性. 数据挖掘技术通常是在某一数据集上训练出一个学习器, 再对测试数据做出预测, 得到一个准确的结果. 研究表明, 个体学习器的训练效果往往不尽人意. 一般而言, 为了是模型有更好的预测效果, 研究人员通常会通过多次训练来逐步拟合目标值. 在数据挖掘技术中, 集成学习是一种提高模型预测准确率的有效策略, 集成学习通过训练多个弱学习器, 根据一定的规则对结果做出预测, 从而提高整个模型的泛化能力. 本文在模型搭建方面, 主要选择了集成学习模型中比较有代表性的 XGBoost 模型[10] 和 CatBoost 模型[11], 以及逻辑回归模型作为基分类器, 再以基分类器的输出作为融合模型的输入特征, 从而实现用户对商品的购买预测. 选择一个合适的融合方法可以提升由弱分类器组成的融合模型的鲁棒性和泛化能力. 也有部分学者在融合方法领域做出研究, Tumer 等人[12] 提出了一种自适应投票聚类集成算法, 实验结果证明该方法不仅适用于无噪声环境, 在有噪声环境中也非常有效. Peng 等人[13] 使用投票法对训练的基分类器融合, 得出最后的预测结果, 准确率要比个体分类器高出 12%.

2 用户基于目标品类下商品的购买预测模型

2.1 问题场景描述

用户购买预测是电子商务推荐系统的一个主要分支[14], 目的是预测用户在未来的某段时间内是否会购买某种商品. 电商平台往往拥有海量的用户历史消费数据和商品数据, 通过数据挖掘的方法从海量数据中发现用户潜在的兴趣爱好, 预测出用户未来的购买意向, 将很大的提高平台的交易成功率. 图 1 表示的是京东商城 3 月份不同评论数商品的用户购买数统计, 图中可以看出大部分用户更加倾向于购买评论数多的商品, 用户可以从多条评论中了解商品的好坏, 说明用户的购买意向可能和商品评论数有一定的正相关性. 同时也可以从用户的一些历史操作行为 (如关注、加入购物车等) 去分析出用户的潜在消费习惯. 因此, 通过以上分析, 可以从已有的用户数据和商品数据中构建出相关特征, 构建训练集和测试集, 再对用户在未来 5 天内基于目标品类下的商品做出购买预测.

2.2 特征工程

原始数据通常具有很大的噪声, 存在各种缺失值和异常值, 并且用户信息和商品信息相对比较独立, 数

据交互比较分散, 能用于预测的特征比较少. 因此, 首先必须要对原始数据做一定的去噪处理和数据探索性分析, 发现数据的分布规律以及数据属性之间相似性, 再通过特征工程从原始特征中提取出一组有对任务预测有促进作用的特征, 对机器学习预测模型往往能起到决定性的作用.

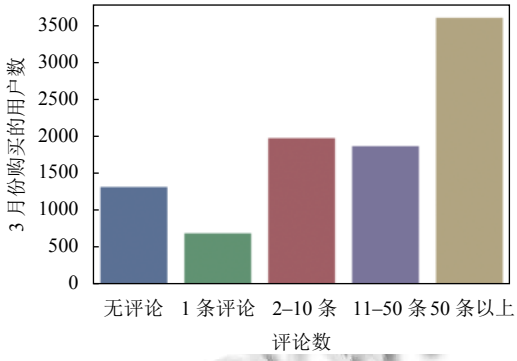

图 1 3 月份用户购买数和商品评论数的关系

本文基于对统计分析知识和电商业务的掌握和了解, 主要从 5 个方面来构建新的特征, 用于模型的训练. 图 2 是基于原始数据的特征交互图.

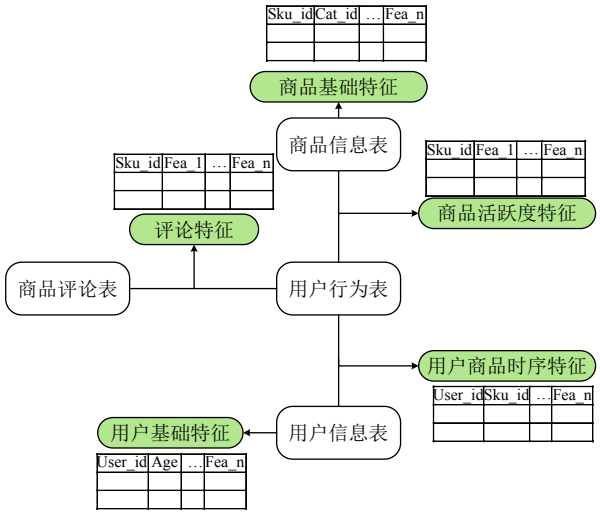

图 2 特征交互图

构建的主要 5 类特征如下:

1) 用户基础特征

描述用户个人信息, 主要包含用户年龄、用户等级 (会员等级)、用户性别等; 分析发现用户的年龄和性别可能会影响用户对目标商品的购买. 例如, 某种商品的用户年龄段在 30–40 岁左右, 20 岁的用户的购买意愿可能较低. 除此之外, 用户的会员等级也反映着用户的一些购买习惯, 高等级会员可能更偏向于消费一些奢侈用品.

2) 商品基础特征

用于描述商品的基本信息, 包含商品的品牌、品类等特征. 同一个商品可能属于不同的品牌, 有些用户由于自己的兴趣爱好可能只会选择其中的一个品牌, 或者从不同的角度去考虑这多个品牌, 再做出选择.

3) 用户-商品的时序行为特征

表示用户在某段时间内对商品和品类的行为特征统计, 一定程度上反映了用户近期是否会对商品进行购买. 时间窗口可以选择距离用户是否购买商品的预测日的前 7 天、15 天和 1 个月这 3 个时间段, 对每个用户, 统计出其在某个时间窗口内对商品或者品类的行为 (点击, 加入购物车, 购买和关注) 次数, 再计算出用户基于特定行为的购买转化率, 作为预测模型的特征. 假设用户 u_i 在时间窗口 T 内的点击次数为N_i, 购买次数为M_i, 则用户 u_i 基于时间窗口 T 的点击购买转化率为N_i/M_i.

4) 商品 (品类或者品牌) 活跃度特征

描述商品、品类或品牌在某段时间内的受欢迎程度, 商品越受欢迎, 表示用户越可能购买该产品或品类下的商品. 基于 3) 中设定的时间窗口, 对每一个商品 (品类或品牌), 统计出其在某个时间窗口内关注 (购买、加入购物车) 过该商品 (品类或品牌) 的不同用户的数量, 该值越大, 表明商品的客户群大, 被大量不同用户喜欢, 用户购买它的概率也越大. 此外统计购买商品的不同年龄段用户的总数, 用以区分商品在不同年龄段的用户群体的受欢迎程度. 同时, 我们发现一些品牌具有特定的目标用户, 这将影响不同用户的购买, 本文通过计算购买过某品牌的用户的平均年龄和平均性别, 以代表在此期间访问过该品牌的所有用户的这些特征.

5) 评论特征

描述用户对购买商品的评价, 商品的评价是用户在购买商品前的一个参照依据. 包含商品在某个时间点的累积评论数、商品的差评率、以及有无差评等特征. 对商品差评率进行分箱处理, 把连续型特征转化为类别特征.

表 1 是特征工程构建的用户购买行为预测特征.

表 1 构建的特征表示和特征表述

特征类型	特征表示	特征描述
用户基础特征	x_1–x_3	用户编号、性别、会员等级
	x_4–x_{10}	所属年龄段, one-hot编码处理
	x_{11}–x_{14}	用户浏览数、关注数、加入购物车数、点击数
商品基础特征	x_{15}–x_{17}	商品编号、品类编号、品牌编号
	x_{18}–x_{20}	商品脱敏参数p_1, p_2, p_3
用户-商品时序行为特征	x_{21}–x_{24}	距预测日7天内, 用户基于商品的行为购买转化率, 行为包括点击、浏览、加购、关注
	x_{25}–x_{28}	距预测日15天内, 用户基于商品的行为购买转化率, 行为包括点击、浏览、加购、关注
	x_{29}–x_{32}	距预测日一个月内, 用户基于商品的行为购买转化率, 行为包括点击、浏览、加购、关注
用户-品类时序行为特征	x_{33}–x_{36}	距预测日7天内, 用户基于品类的行为购买转化率, 行为包括点击、浏览、加购、关注
	x_{37}–x_{40}	距预测日15天内, 用户基于品类的行为购买转化率, 行为包括点击、浏览、加购、关注
	x_{41}–x_{44}	距预测日一个月内, 用户基于品类的行为购买转化率, 行为包括点击、浏览、加购、关注
用户-品牌时序行为特征	x_{45}–x_{48}	距预测日7天内, 用户基于品牌的行为购买转化率, 行为包括点击、浏览、加购、关注
	x_{49}–x_{52}	距预测日15天内, 用户基于品牌的行为购买转化率, 行为包括点击、浏览、加购、关注
	x_{53}–x_{56}	距预测日1个月内, 用户基于品牌的行为购买转化率, 行为包括点击、浏览、加购、关注
商品活跃度特征	x_{57}–x_{60}	距预测日7天内, 购买 (点击、加购、浏览、关注) 过商品 (品类) 的用户数量
	x_{61}–x_{64}	距预测日15天内, 购买 (点击、加购、浏览、关注) 过商品 (品类) 的用户数量
	x_{65}–x_{68}	距预测日一个月内, 购买 (点击、加购、浏览、关注) 过商品 (品类) 的用户数量
	x_{69}–x_{84}	购买过商品 (品牌) 的用户的平均年龄段、平均性别
	x_{85}–x_{98}	购买过商品 (品牌) 的不同年龄段 (共7个年龄段) 的用户数量
	x_{99}–x_{112}	购买过商品 (品牌) 的不同年龄段的用户和不同年龄段有过购买行为人数的比值
评论特征	x_{113}–x_{115}	商品累计评论数、差评率、是否有差评

2.3 集成学习算法模型

图 3 所示是本文预测模型的整体结构图, 表 1 中基于原始数据生成的新特征作为基预测模型的输入, 基预测模型采用 XGBoost、CatBoost 和逻辑回归, 对于 XGBoost 模型, 选取表 1 中特征的两个不同的特征子集分别训练 XGBoost 模型, 得到两个不同的 XGBoost 模型, 基预测模型的训练采用 10 折分层交叉验证的方法, 把数据分成 5 份, 即 train1–train10, 单个模型的每次都把 9 份分好的数据集作为训练, 1 份用于评估模型的性能, 重复 10 次以上操作, 确保每份数据都预测一遍, 对测试数据 test 而言, 10 个模型分别对其做出预测, 最终对结果取均值. 个体模型的训练过程采用网格搜索的方法选取最佳参数. 为进一步提高性能, 本文使用集成技术对上述个体预测模型的结果做一定的处理, 即采用加权平均的方法得出最终的预测结果, 融合模型定义如下:

$$p(u,i)=\sum_{j=1}^{k}\omega_j p_j(u,i) \tag{1}$$

其中, u表示用户, i表示商品, $p(u,i)$表示用户u在未来 5 天内购买商品i的最终概率, j表示基预测模型的个数, p_j表示第j个基预测模型的预测结果, ω_j指的是分配给第j个基预测模型的权重. 对于权重ω, 采用构建线性分类器的方法学习权重, 给每一对用户-商品对(u,i)生成一个k维特征向量, 第j维表示第j个基预测模型生成的概率$p_j(u,i)$, 该k维特征向量作为线性模型的输入, 学习出最终融合模型的权重.

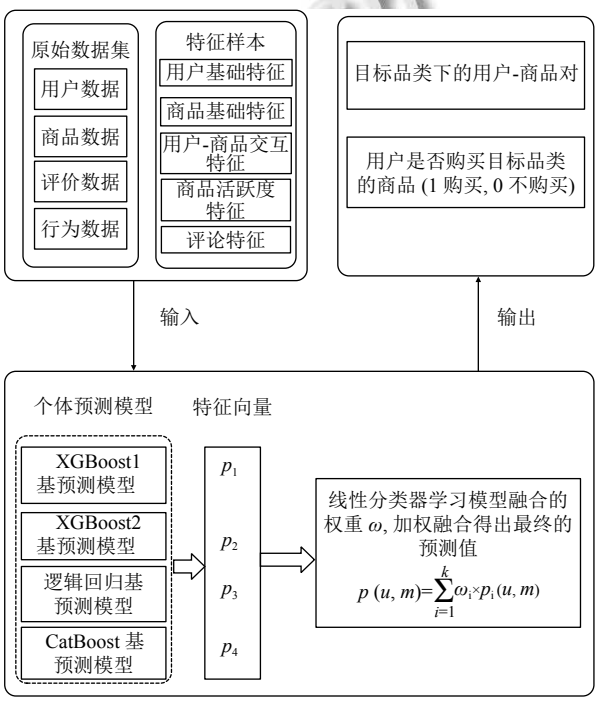

图 3 融合模型整体框架

模型融合采用线性分类器方法学习融合模型的权重, 其目的是为了进一步降低基预测模型的预测误差, 提高模型整体的预测准确率. 融合模型并没有采用基于树模型的分类器或者其它复杂分类器, 主要考虑最终预测模型的泛化能力. 相比于简单的线性分类而言, 采用复杂模型来实现模型融合的方法更加容易造成过拟合的现象, 并且考虑到模型性能的问题, 复杂模型的训练时间也相对要长. 在融合模型权重取值方面, 通过人工赋值的方法具有很大的随机性, 可能多次尝试也不一定会有一个良好的结果, 随着基预测模型的数量增多, 融合模型权重的选择会更加多样和复杂; 而通过采用线性分类器的方法来学习权重, 把不同类型的基预测模型的输出作为线性分类器的输入, 任务的真实值作为输出, 通过梯度下降的方法, 最终学习出融合模型的权重, 从理论上讲, 更加科学和有效.

算法的详细流程如算法 1.

算法 1. 模型融合算法

输入: 生成特征组成的数据集$D=\{(x_1,y_1),(x_2,y_2),\cdots,(x_n,y_n)\}$;
输出: 用户未来 5 天是否购买商品 (1 表示购买, 0 表示不购买);

1. 选择基分类器: XGBoost, CatBoost, Logistic Regression;
2. 选取两个不同的特征子集, 生成数据集 $D1$, $D2$ ($D1$, $D2$ 用于后续 XGBoost 模型的训练);
3. XGBoost1 个体模型把数据集 $D1$ 作为输入数据, 进行 10 折交叉验证得到预测结果为$P^1_{\text{train}}=(p_1,p_2,p_3,\cdots,p_{10})^{\text{T}}$;
4. XGBoost2 个体模型把数据集 $D2$ 作为输入数据, 进行 10 折交叉验证得到预测结果为$P^2_{\text{train}}=(p_1,p_2,p_3,\cdots,p_{10})^{\text{T}}$;
5. CatBoost 和 Logistic Regression 个体模型把数据集 D 作为输入数据, 类似以上两个 XGBoost 模型, 得出预测结果$P^3_{\text{train}}=(p_1,p_2,p_3,\cdots,p_{10})^{\text{T}}$和$P^4_{\text{train}}=(p_1,p_2,p_3,\cdots,p_{10})^{\text{T}}$;
6. 构建 4 维特征向量$P=(P^1_{\text{train}},P^2_{\text{train}},P^3_{\text{train}},P^4_{\text{train}})^{\text{T}}$, 作为线性分类器 Linear Classifier 的输入, 学习出融合模型式 (1) 中的权重ω;
7. 所有模型训练完成, 对 test 数据集做出预测, 每个基分类器会生成 10 个结果, 对其取均值, 则 4 个基分类器得到的最终预测结果为$P^1_{\text{test}}=\left(\bar{p}_1\right)^{\text{T}}$, $P^2_{\text{test}}=\left(\bar{p}_2\right)^{\text{T}}$, $P^3_{\text{test}}=\left(\bar{p}_3\right)^{\text{T}}$, $P^4_{\text{test}}=\left(\bar{p}_4\right)^{\text{T}}$, 其中, $\bar{p}_i(i=1, 2, 3, 4)$表示均值;
8. 构建 4 维特征向量$P=(P^1_{\text{test}},P^2_{\text{test}},P^3_{\text{test}},P^4_{\text{test}})^{\text{T}}$, 进行模型融合, 得出最后的预测结果$p(u,i)$.

3 实验与分析

3.1 实验数据

本次实验的数据集来自京东平台举办的算法大赛“高潜用户购买意向预测”, 包含用户信息表、商品信息表、商品评价数据表和用户行为数据表, 表 2–表 5 描述了各表格的字段信息. 数据集总共包含 105 231 个用户, 28 710 种商品以及 442 种品牌. 用户行为数据包含 2016-02-01 到 2016-04-15 这段时间内用户对商品的各种行为动作, 预测任务是用户在未来 5 天内对目标品类 cate=8 下的商品的购买意向预测. 划分数据集: 用 2016-02-01 到 2016-02-29 的数据来预测 2016-03-01 到 2016-03-05 的购买意向, 用 2016-02-15 到 2016-03-14 的数据来预测 2016-03-15 到 2016-03-19 的购买意向, 将这两部分作为训练集; 用 2016-03-13 到 2016-04-10 的数据预测 2016-04-11 到 2016-04-15 的购买意向, 该部分数据作为验证集.

表 2 用户表

字段名	字段说明	备注
user_id	用户编号	脱敏
sex	性别	0男, 1女, 2保密
age	年龄段	–1表示未知
user_lv_cd	用户等级	顺序枚举
user_reg_time	用户注册日期	粒度到天

表 3 商品表

字段名	字段说明	备注
sku_id	商品编号	脱敏
a1	属性1	枚举, –1表示未知
a2	属性2	枚举, –1表示未知
a3	属性3	枚举, –1表示未知
cate	品类编号	脱敏
brand	品牌编号	脱敏

表 4 评论数据表

字段名	字段说明	备注
dt	截止到时间	粒度到天
comment_num	累计评论数分段	枚举
has_bad_comment	是否有差评	0无, 1有
bad_comment_rate	差评率	差评数占总评论数据的比重
sku_id	商品编号	脱敏

表 5 行为数据表

字段名	字段说明	备注
user_id	用户编号	脱敏
sku_id	商品编号	脱敏
time	行为发生的时间	—
model_id	点击模块编号, 只有行为是点击才有	脱敏
type	1浏览, 2加购, 3购物车删除, 4下单, 5关注, 6点击	—
cate	品类编号	脱敏
brand	品牌编号	脱敏

3.2 评估指标

本文采用京东算法比赛给出的 $F1$ 值作为预测模型的评估指标, $F1$ 的定义由式 (2) 所示, P表示准确率,

R表示召回率, 准确率表示预测正确的样本数占总数的比例, 召回率表示预测正确的样本数占总的正确样本数的比例, 该值能综合考虑准确率和召回率来评估分类模型的性能.

$$F1 = \frac{5RP}{2R+3P} \tag{2}$$

3.3 实验结果分析

为了比较融合模型和基分类器的预测效果, 本文另外训练了 LightGBM[15] 和随机森林两种算法模型. 获取每个基分类器和融合模型的预测结果, 每种算法运行 5 次, 取 5 次的平均值作为最终比较的参考标准. 对比结果如图 4 所示, 图中展示了用于融合的 4 个基分类模型、融合模型以及另外训练的 LightGBM 模型和随机森林的 $F1$ 评分对比, 可以看出个体模型中, CatBoost 的效果最佳, $F1$ 评分为 0.735, 要优于其它个体预测模型, 表现最差的个体模型为随机森林. 与个体模型相比, 本文采用的融合模型的 $F1$ 评分为 0.757, 要明显优于其它个体模型, 证明了本文设计的融合模型的有效性.

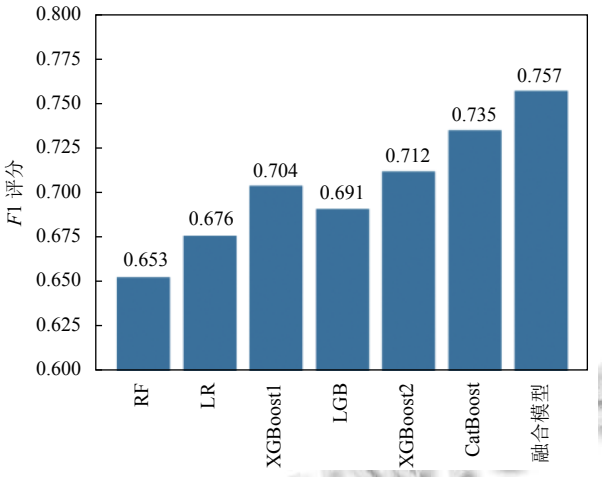

图 4 各模型 $F1$ 评分对比

基预测模型中, CatBoost 和 XGBoost 模型的预测表现最佳, 考虑到两种模型都是基于 boosting 的集成算法, 由多个回归树组合而成. 其中, 每一个回归树的训练过程都是去拟合上一次迭代的负梯度方向的值, 从而最小化损失函数. 同时模型能很好的处理输入特征中的类别特征, 并且在训练过程中能自动生成组合特征并且进行特征选择, 因此选择 CatBoost 和 XGBoost 模型作为融合模型的基预测模型, 可以很好的实现用户购买预测任务. 同时, 融合模型的好坏往往取决于基预测模型之间的差异性, 所以本文把逻辑回归模型作为一个基预测模型, 主要考虑该模型从不同于树模型的角度去训练数据, 逻辑回归可以看成对不同的特征赋予不一样的特征权重, 最终学习出一个函数来预测购买的概率. 因此本文设计的融合模型可以从不同的角度考虑, 得出用户的购买概率, 并且还具有更强的泛化能力和鲁棒性.

图 5 是基预测模型 XGBoost2 输出的特征重要性排名 (提取排名前 20 的特征), 其中对模型训练最有利的前 3 个特征分别为 7 天内用户对品类的加入购物车行为的购买转化率、7 天内用户对品类的关注购买转化率和品类在 15 天内被关注的用户数, 说明了用户在距离购买预测日的时间间隔越近做出的行为, 对用户最后做出购买选择的影响越大, 其中加入购物车和关注行为最能够反映用户对该品类的喜爱程度, 并且品类在 15 天内被关注的用户数在一定程度上反映了该品类近期的受欢迎程度, 该特征重要性排名第三也符合实际生活中, 用户购买商品前会考虑商品品类的受欢迎度再做出选择.

图 6 可以看出, 融合模型的预测效果总体上要优于其它模型, 且 5 次实验的结果相对比较稳定. 基于个体模型而言, CatBoost 和 XGBoost 这两个模型的 $F1$ 评分相对较高, 但是缺乏稳定性, 随着实验次数的增加, 模型预测效果波动较大. 模型融合的出发点就是最大化个体模型预测的优点, 忽略个体模型中的缺点. 通过在同一测试集上的实验结果对比, 发现本文设计的融合模型, 可以很好地提高任务的预测准确性, 同时也解决了个体模型预测不稳定的问题.

4 总结与展望

本文主要通过生成时间演变特征和设计融合模型进行用户对商品的购买预测. 在特征工程方面, 生成了大量的特征来捕捉用户的偏好和行为, 包括商品、品类和品牌的特征以及它们之间的交互, 实验表明在用户和商品的基本属性稳定的情况下, 用户的时序行为特征能够更好的反映用户的购买意图, 说明了特征工程工作一定程度上提升了模型预测的精度; 在模型的构建方面, 本文设计一个基于 XGBoost、CatBoost 和逻辑回归的两层融合模型, 第一层通过选取不同的特征子集对个体学习器进行训练, 第二层融合模型通过线性分类器的方法学习出融合模型的权重, 再做出最后的预测. 从结果看出, 本文设计的融合模型的效果要优于其

它常用的方法. 在未来的学习中, 希望能够从原始数据中挖掘出更多用户潜在的行为特征, 来提高模型的预测效果, 同时希望能够在基预测模型中引入深度学习领域相关算法, 通过自动学习特征, 对用户购买做出预测.

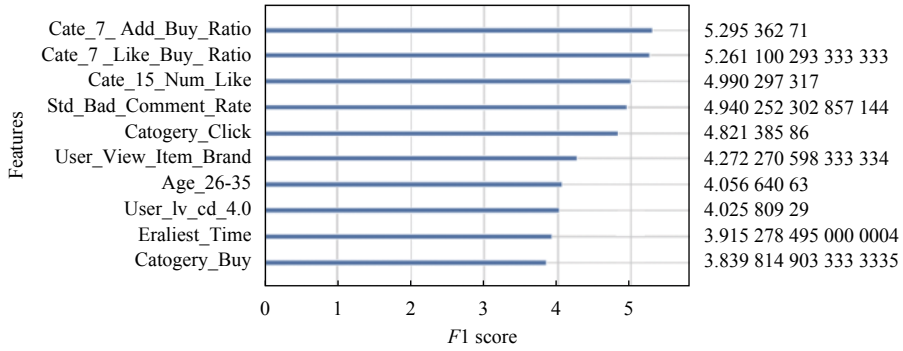

图 5 XGBoost2 特征重要性排名

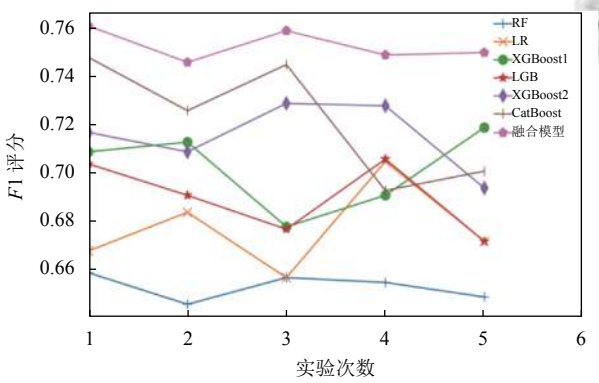

图 6 各模型多次实验的结果比较

参考文献

1 E-commerce in China. https: //www.statista.com/topics/1007/e-commerce-in-china/. [2020-09-17].

2 Liu GM, Nguyen TT, Zhao G, *et al*. Repeat buyer prediction for e-commerce. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Francisco: ACM, 2016. 155–164.

3 Lee M, Ha T, Han J, *et al*. Online footsteps to purchase: Exploring consumer behaviors on online shopping sites. Proceedings of 2015 ACM Web Science Conference. Oxford: ACM, 2015. 15.

4 Liu XM, Li J. Using support vector machine for online purchase predication. Proceedings of 2016 International Conference on Logistics, Informatics and Service Sciences. Sydney: IEEE, 2016. 1–6.

5 祝歆, 刘潇蔓, 陈树广, 等. 基于机器学习融合算法的网络购买行为预测研究. 统计与信息论坛, 2017, 32(12): 94–100. [doi: 10.3969/j.issn.1007-3116.2017.12.014]

6 Dong YQ, Jiang WJ. Brand purchase prediction based on time-evolving user behaviors in e-commerce. Concurrency and Computation: Practice and Experience, 2019, 31(1): e4882. [doi: 10.1002/cpe.4882]

7 胡晓丽, 张会兵, 董俊超, 等. 基于 CNN-LSTM 的用户购买行为预测模型. 计算机应用与软件, 2020, 37(6): 59–64. [doi: 10.3969/j.issn.1000-386x.2020.06.012]

8 董敏, 曹丹, 刘皓熙, 等. 基于动态规划和 K-means 聚类的特征选择算法: 中国, 106022385A. 2016-10-12.

9 Yang JB, Ong CJ. An effective feature selection method via mutual information estimation. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), 2012, 42(6): 1550–1559. [doi: 10.1109/TSMCB.2012.2195000]

10 Chen TQ, Guestrin C. XGBoost: A scalable tree boosting system. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Francisco: ACM, 2016. 785–794.

11 Prokhorenkova L, Gusev G, Vorobev A, *et al*. CatBoost: Unbiased boosting with categorical features. Proceedings of the 32nd International Conference on Neural Information Processing Systems. Montréal: ACM, 2017. 6639–6649.

12 Tumer K, Agogino AK. Ensemble clustering with voting active clusters. Pattern Recognition Letters, 2008, 29(14): 1947–1953. [doi: 10.1016/j.patrec.2008.06.011]

13 Shi LJ, Mao XC, Peng ZL. Method for classification of remote sensing images based on multiple classifiers combination. Applied Mechanics and Materials, 2012, 263–266: 2561–2565. [doi: 10.4028/www.scientific.net/AMM.263-266.2561]

14 KE GL, MENG Q, FINLEY T, *et al*. LightGBM: A highly efficient gradient boosting decision tree. Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach: ACM, 2017. 3149–3157.

15 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. 计算机科学, 2019, 46(9): 190–194. [doi: 10.11896/j.issn.1002-137X.2019.09.027]