Statistics and Application 统计学与应用, 2023, 12(6), 1535-1546
Published Online December 2023 in Hans. https://www.hanspub.org/journal/sa
https://doi.org/10.12677/sa.2023.126157

Hans汉斯

基于LightGBM与深度兴趣网络Stacking融合模型的商品推荐算法

王　彤，熊　浪，彭俊杰

重庆理工大学理学院，重庆

收稿日期：2023年10月12日；录用日期：2023年12月4日；发布日期：2023年12月11日

摘　要

随着电子商务平台的迅速发展，如何提高用户对平台的忠诚度并稳定客流，进而调整平台运营方向以获得持续的收益，成为当前电子商务平台急需解决的关键问题。常见于电商平台的推荐系统利用用户的购买、收藏、浏览等数据，采用特定的算法向用户推荐商品。本研究提出了一种基于LightGBM与深度兴趣网络Stacking融合模型的商品推荐的新解决方案。该模型根据用户过去一年的交易记录提取相应的商品特征和用户特征，整合协同过滤的多路召回策略与这些特征，并将其作为模型的输入，以预测下单客户可能购买的产品并进行商品推荐。研究结果表明，在测试数据上，相对于其他常用推荐算法，本文提出的模型具有更高的准确性、更快的预测速度和更好的推荐效果。这些研究结果为电子商务企业提供了改进服务的契机，为相关研究和实践提供了有益的参考和借鉴，为商品推荐问题的解决提供了有价值的参考和帮助。

关键词

商品推荐，协同过滤，多路召回，LightGBM，深度兴趣网络

Commodity Recommendation Algorithm Based on the Fusion Model of LightGBM and Deep Interest Network Stacking

Tong Wang, Lang Xiong, Junjie Peng

College of Science, Chongqing University of Technology, Chongqing

Received: Oct. 12th, 2023; accepted: Dec. 4th, 2023; published: Dec. 11th, 2023

Abstract

With the rapid development of e-commerce platforms, how to improve user loyalty to the platform

文章引用: 王彤, 熊浪, 彭俊杰. 基于 LightGBM 与深度兴趣网络 Stacking 融合模型的商品推荐算法[J]. 统计学与应用, 2023, 12(6): 1535-1546. DOI: 10.12677/sa.2023.126157

and stabilize customer flow, and then adjust the direction of platform operation to obtain sustained revenue, has become a key issue that e-commerce platforms urgently need to solve. Recommendation systems commonly used in e-commerce platforms utilize user's purchase, collection, browsing and other data to recommend commodities to users using specific algorithms. In this study, we propose a new solution for commodity recommendation based on the fusion model of LightGBM and deep interest network Stacking. The model extracts the corresponding commodity features and user features based on the user's transaction records in the past year, integrates a collaborative filtering multiplexed recall strategy with these features, and uses them as inputs to the model in order to predict the commodities that the customers placing the order are likely to purchase and make commodity recommendations. The research results show that the model proposed in this paper has higher accuracy, faster prediction speed, and better recommendation effect than other commonly used recommendation algorithms on the test data. These findings provide e-commerce enterprises with opportunities to improve their services, provide useful references and lessons for related research and practice, and provide valuable references and assistance in solving the problem of commodity recommendation.

Keywords

Commodity Recommendation, Collaborative Filtering, Multiple Recall, LightGBM, Deep Interest Network

1. 引言

随着电商和在线服务的迅速发展，为用户提供个性化、精准化的商品推荐服务已成为各大企业和平台的竞争重点。因此，依据电商网站上的消费数据来推荐潜在的购买商品显得至关重要[1]。推荐系统作为一种信息过滤系统，旨在帮助用户选择和发现他们感兴趣的商品或服务，通过自动分析用户的历史行为、兴趣和偏好，提供个性化的推荐服务。这有助于用户在信息过载的环境中快速找到所需信息。推荐系统在电子商务领域得到广泛应用，其主要原理是通过分析用户的行为、兴趣和需求等信息，向用户推荐可能感兴趣的商品，从而提高购买转化率和满意度，增加商家的销售量和利润。推荐系统的核心是推荐算法[2]，目前主要包括：基于内容的推荐[3]、协同过滤算法[4] [5]和深度学习推荐算法[6] [7]等。本文研究的商品推荐算法主要是针对电子商务领域的商品推荐。

协同过滤算法是推荐系统领域中具有典型性的推荐算法，可以追溯至上世纪 90 年代，诸如矩阵因子分解等。协同过滤方法依赖用户与商品之间的交互信息为用户作出推荐，目前广泛应用于推荐系统[6]。协同过滤方法主要分为两种：基于用户的协同过滤和基于物品的协同过滤[8]。基于用户的协同过滤算法[8]首先计算用户之间的相似度，随后根据目标用户的历史行为数据找到与之相似的一组用户，最终将这些相似用户喜欢的商品推荐给目标用户。相似度计算通常采用余弦相似度、皮尔逊相关系数等方法。基于物品的协同过滤算法则先计算物品之间的相似度，再根据目标用户的历史行为数据找到与目标用户喜欢的物品相似的一组物品，最后将这些相似物品推荐给目标用户。多路召回策略[9]指采用不同策略、特征或简单模型召回一部分候选集，然后将这些候选集混合在一起供后续排序模型使用。多路召回的目的是在计算速度和召回率之间进行权衡，通过多路召回可以快速筛选候选集，确保召回率接近理想状态，从而不损害排序效果。本文综合了上述两种协同过滤方法进行多路召回。

LightGBM，全称轻量梯度提升机(Light Gradient Boosting Machine) [10]，类似于 XGBoost [11]，是梯度提升决策树(GBDT) [12]算法框架的一种工程实现，其原理与 GBDT 相似[13]。然而，LightGBM 是对 XGBoost 的一种优化方法，采用直方图方法、单边梯度抽样、互斥特征捆绑算法以及叶子生长策略对算法进行了优化，极大地提高了 GBDT 算法的性能。深度兴趣网络(Deep Interest Network, DIN) [14]是阿里妈妈精准定向广告团队于 2018 年发表的针对电商场景下深入理解用户兴趣的 CTR 模型。DIN 模型的核心在于将注意力机制与传统的嵌入和多层感知机(MLP)模型相结合。尽管注意力机制在计算机视觉和自然语言处理领域取得了巨大成功，但将其成功引入 CTR 预估领域得益于阿里工程师对电商领域的深刻理解。Stacking (又称为堆叠泛化)是一种模型融合技术[15]，它通过组合多个不同的基础模型来提高整体预测性能。在 Stacking 中，将多个基础模型的预测结果作为新特征，然后用一个元模型来训练和预测。相对于单个基础模型，Stacking 可以充分利用不同模型的优势，更好地适应数据并提高性能。

本研究采用数据集中提取的用户与商品行为特征，构建了基于 LightGBM 和 DIN 的堆叠融合模型，用于推荐和预测用户未来可能购买的商品。实验结果表明，我们的算法在商品推荐方面优于比较算法，各项评估指标均得到改善，验证了我们算法的有效性。

2. 相关算法

2.1. 协同过滤多路召回策略

在多路召回过程中，每个策略之间相互独立且可以并行运行。多路召回的示意图如图 1 所示。本研究综合了基于用户的协同过滤和基于商品的协同过滤两种协同过滤方法进行多路召回。

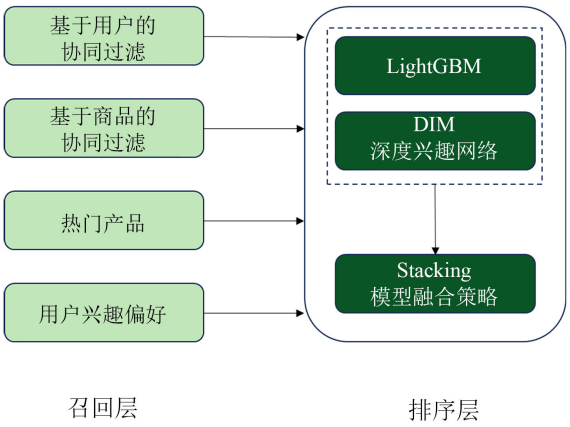

Figure 1. Schematic diagram of a multiplexed recall
图 1. 多路召回的示意图

两种不同方法的相似度计算采用的关联规则如下。

2.1.1. 基于物品的协同过滤的相似度计算关联规则

本文假设用户推荐商品与其最近购买商品具有一定的相关性，商品购买的时间越近，认为其相似度权重可能较大。因此，在相似度计算中引入时间权重，以使结果更具鲁棒性。时间权重的设定旨在考虑用户购买商品的时间因素，以优化推荐系统的准确性和用户满意度。其计算公式如下：

$$w_{ij}^{time} = \frac{|t_i - t_j|}{\log(N_U + 1)\sqrt{n_i n_j}}(N_U - L_U) \tag{1}$$

其中，w_{ij}表示用户 u 所购买的第 i 个商品与第 j 个商品的时间相似权重，t_i表示用户 u 购买第 i 个商品的时间，N_U为用户 u 所购买商品的个数，L_U为用户 u 所购买的当前商品所属的位置，n_i表示用户 u 所购买的第 i 个商品的个数。

考虑商品本身内容与属性的相似度权重，主要依据用户历史购买商品的属性特征，即 Description 特征。针对词袋模型[16]和 TF-IDF 模型[17]的文本，将其转化为词向量，再进一步转化为 Embedding 向量。随后，通过 Embedding 向量计算余弦相似度，以衡量商品间内容和属性的相似程度。需要注意的是，本文设定 Embedding 向量维度为 10。通过该方法，可以量化商品描述的相似程度，使得推荐系统能更好地考虑商品本身的内容和属性，进而提高推荐的精准度和效果。

$$w_{ij}^{content} = \cos\left(e_i, e_j\right) = \frac{e_i \cdot e_j}{\left\|e_i\right\| \cdot \left\|e_j\right\|} \tag{2}$$

其中，e_i为用户 u 购买第 i 个商品的 embedding 向量，e_j为用户 u 购买第 j 个商品的 embedding 向量。

最终，我们将上述权重进行整合，得到用户召回商品的相似度分数。结合用户购买行为的时间权重和商品本身内容与属性的相似度权重，以综合评估商品的相似度分数。这一综合评估的过程能够为推荐系统提供更为准确和个性化的商品推荐，满足用户的需求和偏好。

$$s_{ij} = w_{ij}^{time} \cdot w_{ij}^{content} \tag{3}$$

2.1.2. 基于用户的协同过滤的相似度计算关联规则

同物品的协同过滤相似度计算类似，用户召回的商品相似度计算公式：

$$s_{uv} = \frac{f_u + f_v}{\ln\left(N_I + 1\right)\sqrt{n_u n_v}} \cdot \left(N_I - L_I\right) \tag{4}$$

其中，S_{uv}为商品所交互的第 u 个用户和第 v 个用户的相似度分数，f_u为商品所交互的第 u 个用户的交易频数，f_v所为商品所交互的第 v 个用户的交易频数，N_I为商品所交互的用户个数，L_I为商品所交互的用户的位置，n_u为商品与第 u 个用户交互的次数，n_v为商品与第 v 个用户交互的次数。

2.2. LightGBM 模型

LightGBM 能够高效地处理类别特征并实现最优分割，这源于其能够充分利用类别特征的信息，并具备高度并行处理能力。这种优势使得 LightGBM 在提高计算速度的同时，能够提供良好的预测性能。值得注意的是，LightGBM 模型和梯度提升决策树模型的基本原理基本相同。梯度提升决策树通过迭代构建决策树并进行残差拟合，逐步改进模型的预测能力，从而实现对真实值的精准预测。LightGBM 在这一基本原理上进行了优化和改进，使其能够高效处理类别特征并并行化，进而提高了模型的训练和预测速度。提升树的数学表达式可以表示为：

$$f_M\left(x\right) = \sum_{m=1}^{M} T\left(x;\Theta_m\right) \tag{5}$$

其中，$T\left(x;\Theta_m\right)$为决策树表示的基模型，Θ_m为决策树参数，M 为决策树的棵数。初始提升树模型 $f_0(x) = 0$，第 m 步时的模型可以表示为：

$$f_m\left(x\right) = f_{m-1}\left(x\right) + T\left(x;\Theta_m\right) \tag{6}$$

通过使用经验风险最小化来确定下一棵决策树的参数：

$$\hat{\Theta}_m = \arg\min_{\Theta_m} \sum_{i=1}^{N} L\left(y_i, f_{m-1}\left(x_i\right) + T\left(x_i;\Theta_m\right)\right) \tag{7}$$

假设回归树的损失函数为平方损失为：

$$L\left(y, f\left(x\right)\right) = \left(y - f\left(x\right)\right)^2 \tag{8}$$

则对应到 GBDT 中，损失可推导为：

$$L\left(y, f_{m-1}\left(x\right) + T\left(x;\Theta_m\right)\right) = \left[y - f_{m-1}\left(x\right) - T\left(x;\Theta_m\right)\right]^2 \tag{9}$$

令：

$$r = y - f_{m-1}\left(x\right) \tag{10}$$

可得：

$$L\left(y, f_{m-1}\left(x\right) + T\left(x;\Theta_m\right)\right) = \left[r - T\left(x;\Theta_m\right)\right]^2 \tag{11}$$

通过损失函数对每次拟合进行梯度计算，利用负梯度作为当前模型值，用于估计回归提升树的残差近似值。通过不断迭代这一过程，从而求得参数值。即

$$r_{mi} = -\left[\frac{\partial L\left(y, f\left(x_i\right)\right)}{\partial f\left(x_i\right)}\right]_{f\left(x_i\right) = f_{m-1}\left(x\right)} \tag{12}$$

于是可以得到最终的梯度提升树为：

$$f\left(x\right) = f_M\left(x\right) = \sum_{m=1}^{M}\sum_{j=1}^{J} c_{mj} I\left(x \in R_{mj}\right) \tag{13}$$

$$c_{mj} = \arg\min_{c} \sum_{x \in R_{mj}} L\left(y_i, f_{m-1}\left(x\right) + c\right) \tag{14}$$

其中，R_{mj} 为对应的叶子区域，$j = 1, 2, \cdots, J$，J 为回归树 T 的叶子结点的个数。

基于上述原理，可以将输入数据中的非序列特征，如商品特征、顾客特征以及顾客与商品的交互特征，直接作为 LightGBM 模型的输入进行训练。这样能够快速得到一个具有较好分类能力的模型。这个模型可以作为与 DIN 深度兴趣网络模型融合的基础，进一步提高模型预测的精确度。

2.3. DIN 深度兴趣网络模型

DIN 深度兴趣网络的模型框架如图 4 右半部分所示。该模型的输入主要包括商品特征、用户特征、用户与商品的交互特征以及用户历史购买序列的特征。在这个模型中，最关键的部分是对商品序列和候选商品进行交互注意力计算。在交互注意力计算中，该模型利用候选商品与用户历史行为商品之间的相关关系来计算权重。这个权重反映了候选商品与历史行为商品之间的关注程度，代表了“注意力”的强弱程度。其计算公式如下：

$$V_u = f\left(V_a\right) = \sum_{i=1}^{N} w_i V_i = \sum_{i=1}^{N} g\left(V_i, V_a\right) V_i \tag{15}$$

其中，V_u 是用户的 Embedding 向量，V_a 是候选商品的 Embedding 向量，V_i 是用户 u 的第 i 次行为所交互的商品特征 Embedding 向量，这里的行为是指用户所购买的商品，其 Embedding 向量就是该商品的 Embedding 向量。由于输入向量具有高维的稀疏性，算法模型采用正则化对参数进行优化，即：

$$L_2(W)=\|W\|_2^2=\sum_{j=1}^{K}\|w_j\|_2^2=\sum_{(x,y)\in S}\sum_{j=1}^{K}\frac{I(x_j\neq 0)}{n_j}\|w_j\|_2^2 \tag{16}$$

其中，$w_j\in R^D$ 是第 j 个 Embedding 向量。$I(x_j\neq 0)$ 表示样本 x 是否有特征 j。

采用小批量梯度法进行参数迭代：

$$w_j^{t+1}=w_j^t-\eta\left[\frac{1}{|B_m|}\sum_{(x,y)\in S}\frac{\partial L(p(x),y)}{\partial w_j^t}+\lambda\frac{\alpha}{n_j}w_j^t\right] \tag{17}$$

激活函数采用 PReLU 函数[18]，公式如下：

$$f(s)=\begin{cases}s & ,s>0\\ \alpha s & ,s\leq 0\end{cases} \tag{18}$$

3. 数据集与预处理

3.1. 实验数据集

本文采用的数据集为 Online Retail II，该数据集包含了涵盖在线零售商销售数据的信息，由 UCI Machine Learning Repository 提供。在 Kaggle 平台上也有对该数据集的重新整理和分享。该数据集的时间跨度为 2010 年 12 月至 2011 年 11 月，提供了英国一家在线零售商用户 1 年的历史交易记录。总共包含 389,168 个样本，涵盖了 8 个特征，分别是产品订单号、商品代码、商品基本描述、商品交易数量、商品交易日期、顾客 ID、顾客所属国家。

3.2. 数据集预处理

3.2.1. 商品特征预处理

首先，对数据进行了预处理。根据特征价格理论[19]，将商品特征划分为畅销指标、盈利指标和退货指标。对原始数据进行了相应处理，得到了包括“净交易产品数”、“交易国家数”、“总销售额”、“净交易数”、“退货率”、“交易客户数”、“加权平均价格”和“退货数”在内的八个子特征变量。表 1 展示了商品商业价值特征变量的量化结果，图 2 则展示了商品相关特征的热力图。

Table 1. Quantification results of variables of characteristic variables of the commercial value of commodity
表 1. 商品商业价值特征变量的量化结果

特征分类	特征变量	变量量化
畅销指标	净交易产品数	每个产品成交总数 − 退货数量
	净交易数	每个产品成交交易数 − 退货交易数
	交易客户数	每个产品成功交易的客户数
	交易国家数	每个产品客户所在国家总数
盈利指标	总销售额	每个产品成交总价 − 退货总价
	加权平均价格	每个产品的加权平均价格
退货指标	退货数	每个产品退货数量
	退货率	每个产品退货数量/(产品成交总数 + 退货因子)

注：因部分产品没有成交只有退货，故退货因子取 0.5。

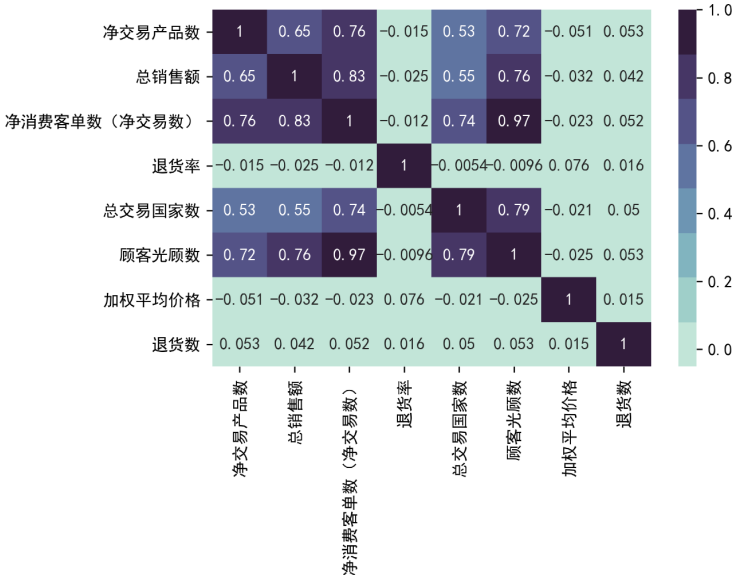

Figure 2. Heat map of commodity-related features
图 2. 商品相关特征热力图

本文采用熵权法[20]对三个特征的子特征变量分别进行内部熵权计算，确保每个指标的子特征熵权之和为 1。随后，使用两层 TOPSIS 法[21]进行得分计算。第一层 TOPSIS 法针对畅销指标、盈利指标和退货指标的子特征，分别得到相应的畅销得分、盈利得分和退货得分。第二层 TOPSIS 法通过熵权 TOPSIS 方法对畅销得分、盈利得分和退货得分进行熵权计算和加权距离计算，最终得到商品的商业评估得分。

3.2.2. 用户特征预处理

根据 RFM 模型[22]，首先计算每位顾客的购买总金额(商品单价乘以商品数量)作为 M 的指标值。接着利用交易日期和时间变量，确定每位顾客的最近购买天数作为 R 的指标值，并进一步确定每位顾客的购买次数作为 F 的指标值。通过熵权法计算这三个指标的权重，并将权重乘以对应指标的原始分值，得到基于熵权法 RFM 的指标得分：RS、FS、MS。电商客户价值综合得分的分布情况如图 3 所示，数值越高代表该客户的价值越大。计算公式为：

$$S = RS + FS + MS \tag{19}$$

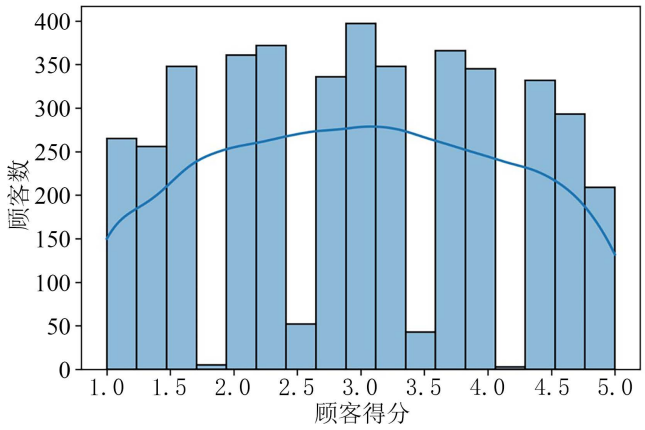

Figure 3. Customer score distribution map
图 3. 顾客得分分布图

3.2.3. 协同过滤多路召回

针对每位客户，其当前购买的商品很大程度上与其历史购买的商品具有较大的相关性。用户的历史购买行为可以很好地体现出用户的兴趣偏好。因此，可以根据兴趣偏好提取相应的特征，以为该用户推荐候选产品(见表 2)。最终，可以利用排序模型为用户推荐用户最感兴趣的前几个产品。

针对商品特征，本文主要选取了商品价值的几个关键指标，并根据这些指标对商品价值进行了打分，共计包括 3 个指标：总交易额、退货数、商品价值得分。而针对顾客特征，主要选取了几个重要的顾客价值指标，然后根据这些指标对顾客的价值进行了评分，共计包括 4 个指标：顾客的 R (活跃度)、F (购买频率)、M (购买金额)、S (顾客价值得分)。

商品特征：主要筛选了商品价值的几个指标以及最后对商品价值的打分指标，一共有 3 个指标：总交易额、退货数、商品价值得分。

顾客特征：主要筛选了顾客价值指标以及最后对顾客价值的打分指标，一共 4 个：顾客的 R (活跃度)、F (购买频率)、M (购买金额)，S (顾客价值得分)。顾客与商品的交互特征方面，主要考虑了用户召回的商品与其历史购买商品之间的相似度特征。这些相似度特征是基于商品的 Embedding 向量计算得出的，即 $sim_{ij} = Emb_i \cdot Emb_j$。$Emb_i$ 表示用户召回后的第 i 个商品的 Embedding 向量，Emb_j 表示用户历史购买的第 j 个商品的 Embedding 向量，以及最大最小相似性特征、相似度特征的均值等。至于商品 Embedding，由于原始数据中的 Description 特征是以字符形式存在的，本文利用词袋模型和 TF-IDF 模型将原始的 Description 转换成维度为 10 的向量。词袋模型和 TF-IDF 模型是基于文本词在文档中的频率构建的，用于展开原始 Description 特征为相应的向量。

Table 2. Feature classification of recommendation systems and their functionality
表 2. 推荐系统的特征分类和其功能

特征分类	特征组合	数据类型
用户特征	R (活跃度) F (购买频率) M (购买金额) S (顾客价值得分)	连续型
历史行为特征	用户购买的商品 ID	离散序列
用户与商品交互特征	相似度 总销售额 退货数 价值得分 交易次数 交易时间	连续型
商品特征	商品描述	字符型

最终，将协同过滤多路召回策略与特征工程的特征整合，得到每个用户历史购买商品的特征集合。这些特征集合将作为接下来排序模型的输入。为了为每个样本制作标签，采用了每个用户最后一次购买的商品作为预测目标。具体地，在最后一次购买之前的购买商品数据集将被作为训练集。如果最后一次购买的商品在训练集中出现，标签记为 1，否则标签记为 0。考虑到标签数据可能存在不平衡的情况，本文引入了随机噪声，以实现标签的平衡化。这样做可以使模型更具鲁棒性，适应不同类别标签的分布情况。

4. 实验分析与结果

4.1. 实验模型框架

在经过多路召回后，得到的特征包含商品信息、用户信息以及用户与商品的交互信息。通过排序模型，在众多候选商品中利用这些相关特征为每位顾客筛选出最优选的 k 个商品。传统的协同过滤模型在处理这种多结构化数据时已经显得较为困难。在这个基础上，本文采用了阿里巴巴提出的 DIN 深度兴趣网络模型来处理这种多结构化的数据，包括候选商品序列数据、商品和用户的离散类型 ID 以及连续性特征数据。DIN 模型通过注意力网络筛选出重要的特征，具有较强的实际业务解释能力。同时，为了更好地对商品进行排序，本文结合了目前在机器学习中具有较强排序和分类能力的模型 LightGBM。LightGBM 算法能够高效处理特征，因此在工程实践中得到了广泛的应用。本文的模型框架如下图 4。

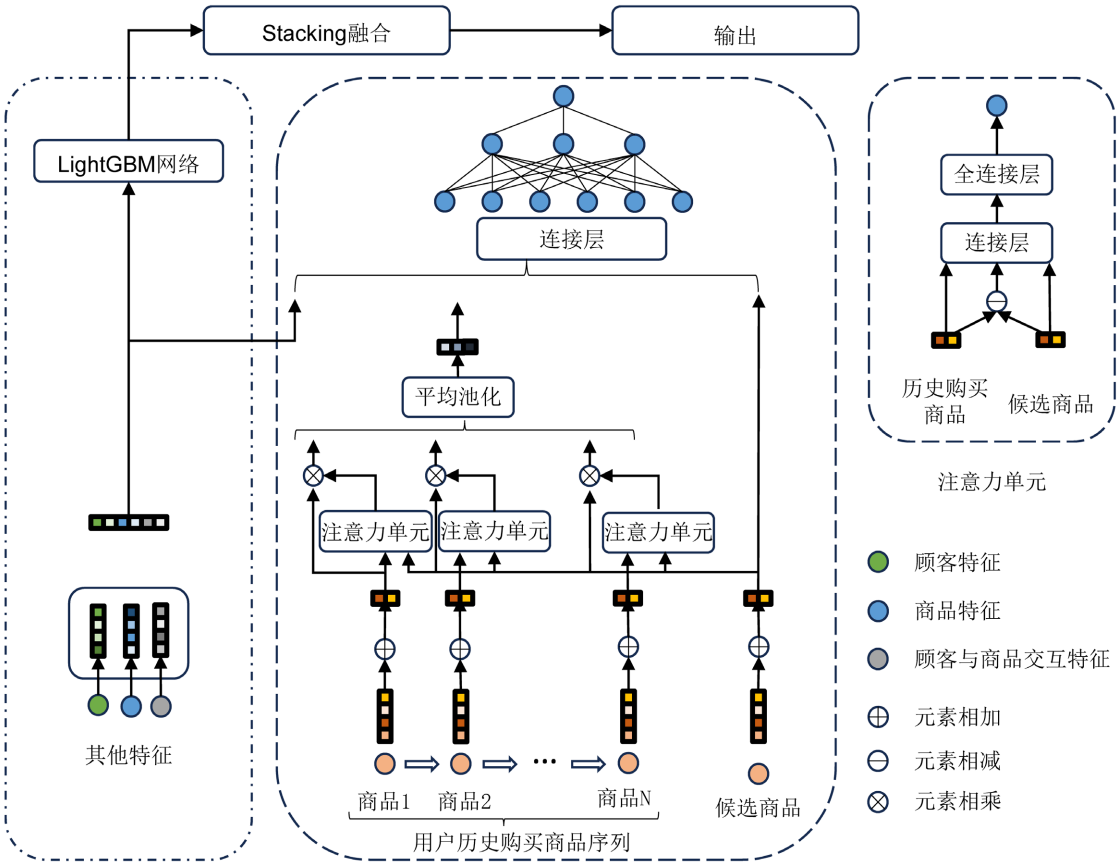

Figure 4. Fusion modeling framework of commodity recommendation based on DIN and LightGBM
图 4. 基于 DIN 与 LightGBM 的商品推荐融合模型框架

4.2. 模型结果与分析

4.2.1. LightGBM 模型结果

使用 LGBMClassifier 模型对训练数据进行训练，并对 AUC 值损失进行了可视化分析，具体结果如下图 5 所示。在测试数据上，LightGBM 模型达到的 AUC 值为 0.9107，展现了较好的预测效果。该模型的训练时间仅为 0.73 秒，呈现了较快的训练速度。

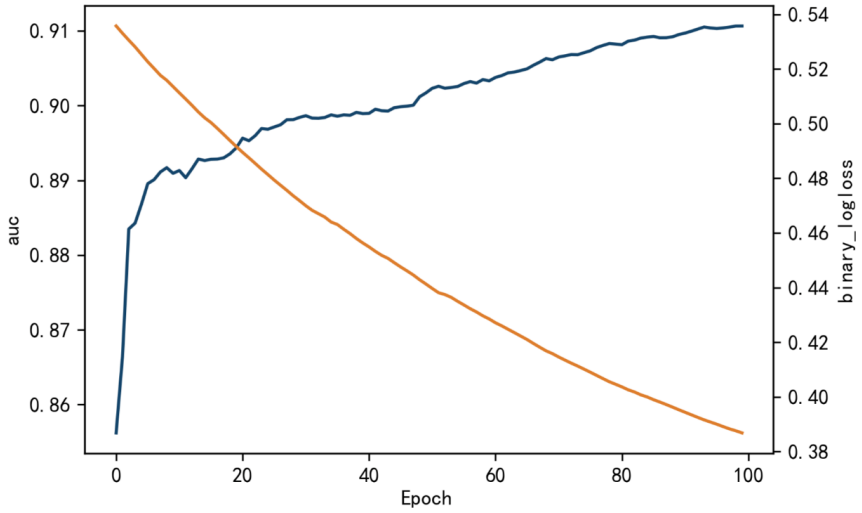

Figure 5. LightGBM model results
图 5. LightGBM 模型结果

4.2.2. DIN 模型结果

本文采用了 Adagrad 算法对 DIN 模型的参数进行迭代优化，设置了批量大小(batch size)为 50，进行了 20 个代(epoch)的迭代。训练过程总共用时 16.66 秒。最终，在测试集上，DIN 模型取得了 AUC 值为 0.9642 (见图 6)，展现了出色的预测效果。

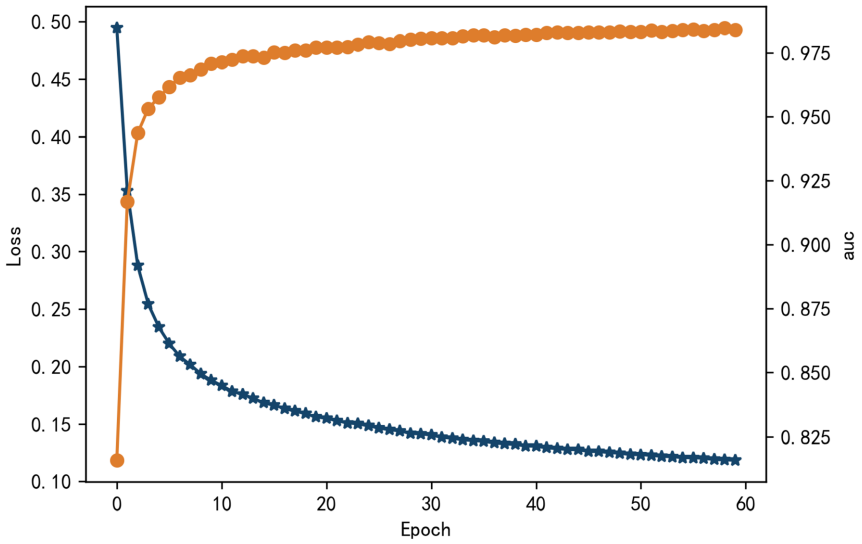

Figure 6. DIN model results
图 6. DIN 模型结果

4.2.3. Stacking 融合模型结果

本文采用了 5 折交叉验证方法，分别对 LightGBM 模型与 DIN 深度兴趣网络模型进行训练和验证。随后，将交叉验证的预测得分作为最后 Stacking 模型的特征。利用 Logistic 回归将融合后的特征作为训练数据的输入进行训练。最终，融合后的模型在测试集上达到了 AUC 值为 0.9685，表现出色。整个融合模型的训练时间为 13.2 秒，展现了较快的模型训练速度。

4.2.4. 模型结果对比与分析

表 3 展示了 3 个模型的结果，每个模型在实际训练过程中都经历了交叉验证，而交叉验证结果表明，本文提出的模型在商品推荐领域具有较好的泛化能力。通过 AUC 值的比较，可以看出融合模型，即 LightGBM 与 DIN 深度兴趣网络的 Stacking 融合网络模型，具有最优异的预测效果。相较于单独的 LightGBM 模型，融合模型的 AUC 值提高了 5%，相较于 DIN 深度兴趣网络模型则提高了 4%。在实际业务应用中，这种显著的提升效果通常会带来可观的收益。另一方面，就运行速度而言，LightGBM 模型的训练时间仅为 0.73 秒，明显快于 DIN 模型和融合模型，突显了 LightGBM 模型的优势。这也同时印证了融合模型的合理性。值得注意的是，本文的 LightGBM 与 DIN 深度兴趣网络的融合模型不仅在精度上高于了 DIN 模型，而且计算速度提升了 21%，这明确了融合模型的有效性。

Table 3. Comparison of model results
表 3. 模型结果对比

模型	AUC 值	训练时间/s
LightGBM	0.9107	0.73
DIN	0.9642	16.66
LightGBM + DIN	0.9685	13.2

5. 总结

本文采用了熵权法改进的 TOPSIS 方法从数据集中提取商品特征，并运用熵权法改进的 RFM 方法提取用户特征。在此基础上，利用提取出的用户与商品的行为特征，建立了基于 LightGBM 与 DIN 的 Stacking 融合模型，用于对用户未来可能购买的商品进行推荐和预测。我们将所提出的模型与单独的 LightGBM 模型以及 DIN 深度兴趣网络模型进行了比较。结果显示，相较于这两个单独的模型，我们所提出的模型的 AUC 值分别提高了 5%和 4%。同时，该模型的运行速度也比单独的 DIN 深度兴趣网络模型提高了 21%，验证了模型的有效性。这表明，我们提出的模型可以有效解决商品推荐的问题，为实际业务提供了有益的帮助。

参考文献

[1] 孙光福, 吴乐, 刘淇, 等. 基于时序行为的协同过滤推荐算法[J]. 软件学报, 2013, 24(11): 2721-2733.

[2] 包增辉, 宋余庆. 协同过滤算法的多样性研究[J]. 无线通信技术, 2013, 22(3): 5-9.

[3] Belkin, N.J. and Croft, W.B. (1992) Information Filtering and Information Retrieval: Two Sides of the Same Coin? *Communications of the ACM*, **35**, 29-38. https://doi.org/10.1145/138859.138861

[4] Adomavicius, G. (2012) Improving Aggregate Recommendation Diversity Using Ranking-Based Techniques. *IEEE Transactions on Knowledge & Data Engineering*, **24**, 896-911. https://doi.org/10.1109/TKDE.2011.15

[5] Linden, G., Smith, B. and York, J. (2003) Amazon.com Recommendations: Item-to-Item Collaborative Filtering. *IEEE Internet Computing*, **7**, 76-80. https://doi.org/10.1109/MIC.2003.1167344

[6] 黄立威, 江碧涛, 吕守业, 等. 基于深度学习的推荐系统研究综述[J]. 计算机学报, 2018, 41(7): 1619-1647.

[7] Hu, Z., Wang, J., Yan, Y., *et al*. (2021) Neural Graph Personalized Ranking for Top-N Recommendation. *Knowledge-Based Systems*, **213**, Article ID: 106426. https://doi.org/10.1016/j.knosys.2020.106426

[8] 邓灵斌, 申慧. 电子商务平台商品推荐信息特性对消费者购买意愿的影响实证研究[J]. 南华大学学报(社会科学版), 2019, 20(2): 60-65.

[9] Xing, L.J., Feng, X.W., Chen, H.M., Wang, Y. and Zhang, Y. (2020) Research on Fused Sorting Based on Logical Regression in News Recommendation System. *IOP Conference Series*: *Earth and Environmental Science*, **510**, Article ID: 062029. https://doi.org/10.1088/1755-1315/510/6/062029

[10] Wang, D., Zhang, Y. and Zhao, Y. (2017) LightGBM: An Effective miRNA Classification Method in Breast Cancer Patients. *Proceedings of the* 2017 *International Conference on Computational Biology and Bioinformatics*, Newark, 18-20 October 2017, 7-11. https://doi.org/10.1145/3155077.3155079

[11] 王天峥, 汤健, 夏恒, 等. 基于 XGBoost 串并联集成的数据驱动 MSWI 全流程模型[J/OL]. 计算机集成制造系统, 2023: 1-20. http://kns.cnki.net/kcms/detail/11.5946.TP.20230920.1143.014.html

[12] 王伟, 马乾伦, 白振华, 等. 基于梯度提升决策树的冷轧高强钢卷力学性能预测[J]. 中国机械工程, 2023, 34(18): 2222-2229.

[13] Friedman, J. (2001) Greedy Function Approximation: A Gradient Boosting Machine. *Annals of Statistics*, **29**, 1189-1232. https://doi.org/10.1214/aos/1013203451

[14] Zhou, G., Zhu, X., Song, C., *et al*. (2018) Deep Interest Network for Click-Through Rate Prediction. *Proceedings of the* 24*th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, London, 19-23 August 2018, 1059-1068. https://doi.org/10.1145/3219819.3219823

[15] 王飞, 黄涛, 杨晔. 基于 Stacking 多模型融合的 IGBT 器件寿命的机器学习预测算法研究[J]. 计算机科学, 2022, 49(z1): 784-789.

[16] 宋涛, 赵明富, 刘帅, 等. 基于有序视觉词袋模型的图像相似性衡量[J]. 华中科技大学学报(自然科学版), 2020, 48(8): 67-72, 78.

[17] 董伟, 董思遥, 王聪, 陶金虎. 基于 TF-IDF 算法和 DTM 模型的网络学习社区主题分析[J]. 现代教育技术, 2022, 32(2): 90-98.

[18] He, K., Zhang, X., Ren, S., *et al*. (2015) Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification. *Proceedings of the* 2015 *IEEE International Conference on Computer Vision* (*ICCV*), Santiago, 7-13 December 2015, 1026-1034. https://doi.org/10.1109/ICCV.2015.123

[19] Edmonds, R.G. (1984) A Theoretical Basis for Hedonic Regression: A Research Primer. *Real Estate Economics*, **12**, 72-85. https://doi.org/10.1111/1540-6229.00311

[20] 徐士伟, 苏业辉, 李慧文, 等. 基于熵权法的枢纽内公交站场布局评价研究[J]. 交通运输系统工程与信息, 2023, 23(5): 104-112.

[21] 邵垒, 彭阳, 张超, 等. 基于熵权改进 TOPSIS 理论的富氮气体最优分配方式研究[J/OL]. 航空动力学报, 2023: 1-9. https://doi.org/10.13224/j.cnki.jasp.20210486

[22] 师奥翔, 张洁. 基于改进 RFM 模型的电商用户价值分类的研究[J]. 计算机技术与发展, 2022, 32(12): 123-128.