Software Engineering and Applications 软件工程与应用, 2022, 11(1), 50-59
Published Online February 2022 in Hans. http://www.hanspub.org/journal/sea
https://doi.org/10.12677/sea.2022.111007

Hans汉斯

基于ResNet和DF融合的用户购买预测算法研究

张嵌嵌，何利力

浙江理工大学，信息学院，浙江 杭州

收稿日期：2021年12月3日；录用日期：2022年2月1日；发布日期：2022年2月8日

摘　要

各大电商平台在前期为客户提供服务的同时已经积累了大量用户及商品数据，如何充分利用这些数据为企业增加收入、为用户提供个性化服务已成为研究热点。基于电商平台的环境及数据情况，针对电商平台用户及商品种类数量众多，但平台方无法准确预测用户是否购买这一问题，本文提出了一种基于残差神经网络(ResNet)和深度森林(Deep Forest)融合的用户购买行为预测算法。对某在线商城的大量数据处理为150维的用户特征数据和120维的商品特征数据。首先利用残差神经网络对用户购买行为进行预测，后通过深度森林进行预测，最后通过线性叠加的方式将两种模型融合。通过对残差神经网络进行调参，对深度森林中的随机森林深度进行调整进一步提高预测精度。实验结果表明，该融合模型相比传统算法具有更好的预测效果。

关键词

残差网络，深度森林，用户购买行为预测，组合预测

Research on User Purchase Prediction Algorithm Based on the Fusion of ResNet and DF

Qianqian Zhang, Lili He

College of Information, Zhejiang Sci-Tech University, Hangzhou Zhejiang

Received: Dec. 3rd, 2021; accepted: Feb. 1st, 2022; published: Feb. 8th, 2022

Abstract

Major e-commerce platforms have accumulated a large amount of user and product data while providing services to customers in the early stage. How to make full use of these data to increase

文章引用: 张嵌嵌, 何利力. 基于 ResNet 和 DF 融合的用户购买预测算法研究[J]. 软件工程与应用, 2022, 11(1): 50-59.
DOI: 10.12677/sea.2022.111007

revenue for enterprises and provide personalized services for users has become a research hotspot. Based on the environment and data of the e-commerce platform, in view of the large number of e-commerce platform users and product types, the platform which cannot accurately predict whether the user will buy or not, this paper proposes a user purchase behavior prediction algorithm combined the residual neural network (ResNet) with deep forest (Deep Forest). A large amount of data of an online shopping mall is processed into 150-dimensional user characteristic data and 120-dimensional commodity characteristic data. First, the residual neural network is used to predict the user's purchase behavior, and then the deep forest is used to predict, and finally the two models are merged by linear superposition. By adjusting the parameters of the residual neural network, the depth of the random forest in the deep forest is adjusted to further improve the prediction accuracy. Experimental results show that the fusion model has a better prediction effect than traditional algorithms.

Keywords

Residual Network, Deep Forest, User Buying Behavior Prediction, Combined Forecast

1. 引言

随着计算机技术的快速发展，电商行业的发展也十分迅速。据报道，在 2021 年的双十一活动中，京东电商平台的总交易额达到 3491 亿元，淘宝及天猫平台更是达到了 5403 亿元的总交易额，每天都有数以百万笔订单生成。因此，在大数据时代，如果可以合理预测用户的购买行为，将有效降低各大商家的备货等损耗，极大提升商家收益率。

用户购买行为受众多因素影响，如用户偏好因素，商品价格及品牌等因素。这些因素都将在很大程度上增加用户购买行为的不确定性，从而对预测效果的准确性提出挑战。用户的购买预测可以归纳为典型的二分类问题，即购买与不购买。目前有关用户购买预测的研究主要聚焦在两个方面：① 特征工程的构建；② 相关模型的选择。在特征工程的构建方面：刘潇蔓将[1]影响用户购买的数据分为基础特征、组合特征和衍生特征，从这三个方面构建数据的特征工程；卞天宇[2]则根据数据集构建了包含基础统计指标特征、排序特征、标签特征和用户平均加权选择倾向特征四个特征的特征工程；吴非[3]则主要从时间序列的角度完成对用户行为及基本特征的特征构建。在相关模型选择方面的研究大致可以分为两类，一类是单一算法模型的应用，如盛钟松[4]提出用 CatBoost 模型对用户购买行为进行预测；葛绍林[5]则用深度森林模型进行预测研究；Anindita A. Khade [6]则基于统计分类器 C4.5 决策树算法构建客户数据可视化平台对用户行为进行研究；X. Liu [7]使用支持向量机(SVM)的方法对用户购买进行预测；另一类则是组合算法模型的应用，吴非[3]在构建特征工程后，通过对 GBDT 模型与 lightGBM 模型的融合模型中的单模型进行网格调参实现对用户线上购买行为的预测；C. Okan Sakar [8]等人则提出融合具有权重回溯的弹性反向传播特性的多层感知器(MLP)和长短期记忆网络(LSTM)两种算法的组合模型，构建一个实时在线购物者行为分析系统，并根据预测结果来提高网站的购买转化率。Bruno J. D. Jacobs [9]等人采用将潜在狄利克雷分配(LDA)和狄利克雷多项式(MDM)混合的新方法应用在大型产品分类中，从而进行用户购买行为的预测，实验表明，其预测准确性优于协同过滤法和离散选择模型。曾宪宇[10]等人则提出用基于潜在因子的方式建立对用户购买商品和最佳替代商品的一种选择模型，称之为 LF-CM (latent factor based

choice model)，随后为了提升预测的精度，又提出了一种针对购买周期中所有商品的排序学习模型LFS-CM (latent factor and sequence based choice model)。

通过对他人研究的学习，本文一方面将某在线商城的大量数据处理为 150 维的用户特征数据和 120 维的商品特征数据；另一方面由于随机森林较稳定，即数据集中出现了一个新的数据点，整个算法不会受到过多影响，它只会影响到一颗决策树，很难对所有决策树产生影响，所以基于随机森林的稳定性和残差网络更好地拟合分类函数及层数较深时训练的优化性，提出一种融合 ResNet 和 DF 的用户购买预测算法。

2. 理论基础

2.1. 残差网络

随着神经网络的发展，更多层的神经网络被用在各种研究中。然而，神经网络的梯度在反向传播的过程中要不断地被传播，这也就导致了当神经网络的层数在加深时，梯度在传播过程中可能会出现逐渐消失的现象，梯度消失可能会导致无法对前面网络层的权重进行有效的调整。ResNet 的提出有效地解决了在加深网络层数时导致的梯度消失的问题。如图 1，该图是残差网络的基本结构：

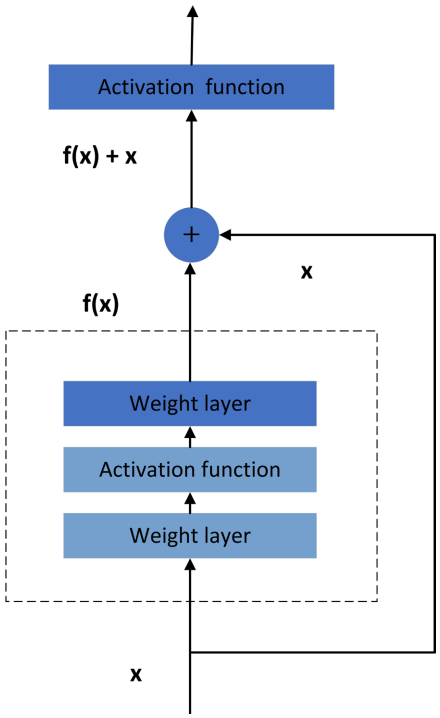

Figure 1. The basic structure of the residual network
图 1. 残差网络的基本结构

残差网络借鉴了高速网络(Highway Network)的跨层连接思想，在此思想的基础上加以改进，原先的残差项是带有权值的，在 ResNet 中，将残差项用恒等映射的方式进行替代。

若某段神经网络的输入为 x，期望输出是 $F(x)$，则在残差网络中，直接将输入 x 作为输出的初始结果，则输出结果为 $F(x)=f(x)+x$，若 $f(x)=0$，则 $F(x)=x$，即此时为恒等映射。所以，在 ResNet 中，不再是通过训练及学习得到一个完整的输出，而是 $f(x)=F(x)-x$，即残差。这也使得当网路层数很深时，当残差结果接近于 0 时，梯度不会消失，模型的准确率也不会下降。

2.2. 随机森林(Random Forest)

随机森林是 Leo Breiman 在 20 世纪 90 年代提出的一种方案，它利用一组生长在随机选择的数据在空间中的决策树来构建预测器集合[11]。Breiman 在[12] [13] [14]证明了通过树的系综，可以在分类和回归精度方面获得实质性的提高，其中系综中的每棵树是根据随机参数生长的，最终的预测是通过集合的聚合得到的。由于系综的基本成分是树结构的预测因子，并且由于这些树中的每一个都是使用随机性注入构建的，因此这些过程被称之为“随机森林”。

信息、熵和信息增益是决策树的根本，对于决策树而言，如果带分类的事物集合可以划分为多个类别当中，则某个类(x_i)的信息可以定义为：

$$I\left(X=x_i\right)=-\log_2 p\left(x_i\right) \tag{1}$$

其中，$I\left(X\right)$用来表示随机变量的信息，$p\left(x_i\right)$指的是当x_i发生时的概率。

在随机森林中，熵是用来衡量不确定性的，当熵越大，$X=x_i$的不确定性越大，反之越小，可以计作：

$$H\left(X\right)=\sum_{i=1}^{n} p\left(x_i\right)I\left(x_i\right)=-\sum_{i=1}^{n} p\left(x_i\right)\log_b p\left(x_i\right) \tag{2}$$

在决策树算法中，大家通常会选择信息增益作为用来选择特征的指标，其中，如果信息增益越大，通常代表这个特征的选择性好，在概率中定义为：待分类的集合的熵和选定某个特征的条件熵(X 在给定条件下 Y 的条件干率分布的熵对于 X 的数学期望)之差，计作：

$$IG\left(Y\mid X\right)=H\left(Y\right)-H\left(Y\mid X\right) \tag{3}$$

其中，$H\left(Y\mid X\right)$表示条件熵，且$H\left(Y\mid X\right)=\sum_x H\left(Y\mid X=x\right)$。

在随机森林模型中，最后的分类结果通常依靠于决策树的投票结果来确定。

3. 基于 ResNet 和 DF 融合的用户购买预测模型

用户购买预测模型是用来预测某个用户是否会购买某个商品的一类模型，用户和商品会通过特征提取表示为能被计算的向量形式。本文中使用 U 表示用户特征向量，I 表示商品特征向量，$P\left(U,I\right)$ 表示预测模型，其输出表示该用户购买该商品的概率，如果预测的概率超过预设的购买阈值 p_t，则认为该用户会购买该商品。预测模型可由多种预测模型线性叠加 $P=\sum_i w_i P_i\left(U,I\right)$，本文通过基于 ResNet 的深度神经网络模型和深度森林模型的线性叠加，提升了预测的精准度。

3.1. 基于 RestNet 的用户购买预测

图 2 所示的是基于 ResNet 的神经网络的购买预测模型结构，首先将用户特征 U 输入一个全连接的神经网络 Net1，将用户特征 U 映射为维数为 k^2 的一维向量，然后将该一维向量变形为 $k\times k$ 维的二维向量并输入 ResNet1 中，输出和商品特征 I 维数相同的权重向量 α，该向量表示该用户特征所代表的一类用户对商品不同特征的重视程度。商品集合 $S=\left\{I_1,\cdots,I_n\right\}$ 包含 n 个商品的所有特征向量，然后将权重向量 α 依次和商品集合中的每一个商品的特征 $I_k,k\in n$ 做点乘运算，得到一个长度为 n 的一维向量 S，再将该向量 S 输入全连接的神经网络 Net2，输出维数为 k^2 的一维向量，然后将该一维向量变形为 $k\times k$ 维的二维向量并输入 ResNet2 中，输出长度为 n 的一维向量 y，向量 y 中的每一位数代表每个商品会被该用户购买的预测概率。购买的标签 $\hat{y}$ 由用户的历史购买行为生成，$\hat{y}$ 是长度为 n 的一维向量，如果商品 $k,k\in n$ 被该用户购买过，则 $\hat{y}$ 中对应位置上为 1，否则为 0。损失函数使用的是二分类交叉熵(Binary Cross Entropy)，如式(4)所示，通过最小化该损失函数更新模型参数。

$$Loss = -\left(\hat{y}\cdot\log(y)+(1-\hat{y})\cdot\log(1-y)\right) \tag{4}$$

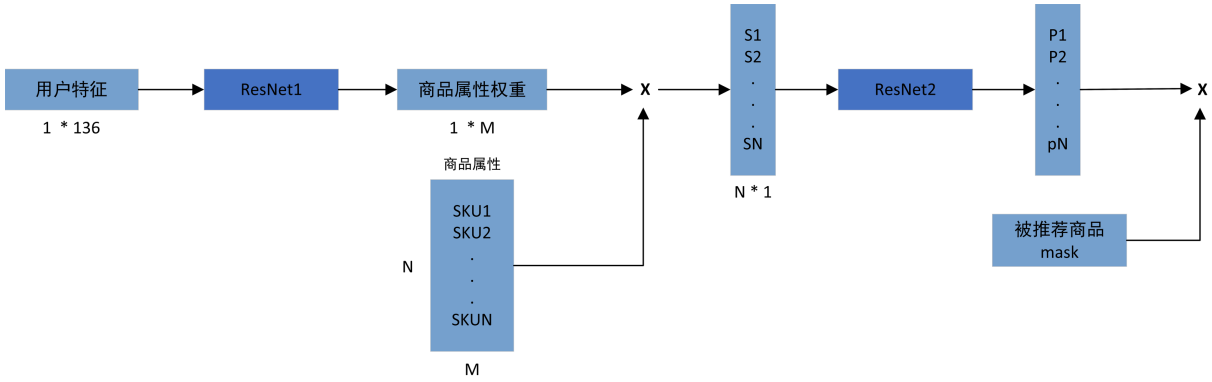

Figure 2. Purchase prediction model based on ResNet
图 2. 基于 ResNet 的购买预测模型

本文模型采用的残差块细节如图 3 所示：

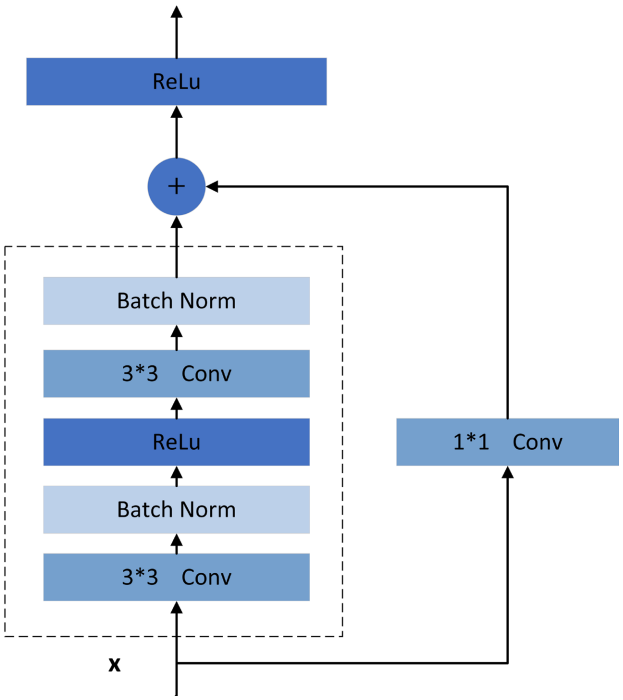

Figure 3. Detail map of residual block
图 3. 残差块细节图

3.2. 基于深度森林的购买预测

深度森林是周志华[15]提出的一种新的决策树集成方法：即生成一个具有级联结构的深度森林集成。由于本文数据集较大，深度神经网络训练起来又需要调大量参数，DeepForest 训练起来过程效率高且可扩展且前人[16]的实验效果教好，所以本文采取该方法进行训练预测。

其主要包括两个阶段，如图 4 所示，一是借助卷积神经网络思想，采用滑动窗口进行特征提取，称之为多粒度扫描阶段。该模型将用户特征向量 U 和商品特征向量 I 拼接作为输入，设置滑动窗口维度为

100，步长为 1，生成扫描子样本，子样本经过随机森林 A 和随机森林 B 进行训练，生成概率特征向量；二是级联森林阶段，级联森林由多级随机森林组成，级联级别的数量可以根据模型的复杂性进行自适应确定。在本文中，将多粒度扫描阶段获得的概率特征向量作为输入，经 4 个不同的随机森林分类生成 4 个 2 维增强特征向量，随后将增强特征向量与原始概率向量组合生成新的特征向量作为下一级森林的输入向量。重复此过程，最后将最终输出的平均值中的最大值作为最终预测结果。

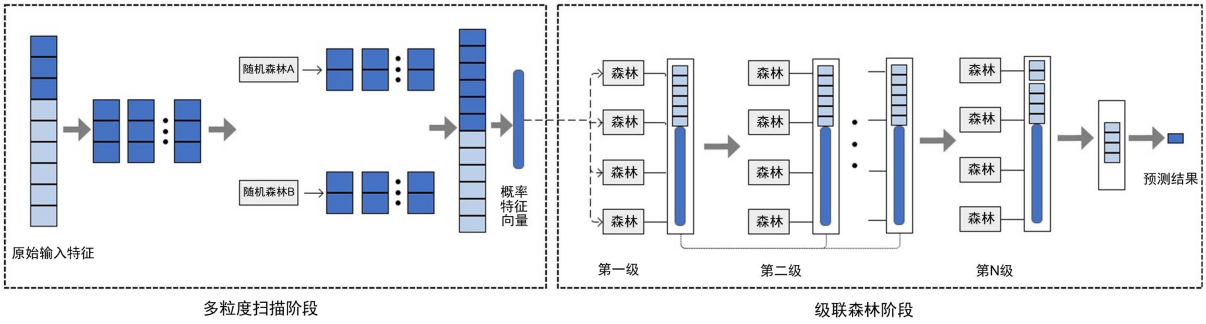

Figure 4. Part of the model of the deep forest
图 4. 深度森林部分模型图

3.3. 组合预测模型

最后基于 ResNet 的神经网络购买预测模型 $P_r(U,I)$ 和基于深度森林的购买预测模型 $P_f(U,I)$ 通过线性叠加构成最终的用户购买预测模型 $P(U,I)=w_r\cdot P_r(U,I)+w_f\cdot P_f(U,I)$，其中 $w_r+w_f=1$。

4. 实验

本实验数据来源于某电商平台提供的公开数据源，该数据主要包括两部分：第一部分为该电商平台的用户相关数据，经过特征提取，表示为长度为 150 的用户特征向量，用户的特征主要有年龄、性别、所属地区等；第二部分为该电商平台上有关商品数据，经过特征提取，表示长度为 120 的商品特征向量，商品特征主要有商品品牌、颜色、价格、产地等。表 1 为基于 ResNet 的神经网络购买预测模型中超参数的设定。

Table 1. Improved model structure parameters
表 1. 改进模型结构参数

参数名	参数值
Batchsize	512
残差单元数量	2
卷积核尺寸	3 * 3
卷积核步长	1
输入层维数	Net1: 150, Net2: 30,000, Resnet1: 64 * 64, Resnet2: 128 * 128
隐藏层层数	50
输出层维度	Net1: 64 * 64, Net2: 30,000, Resnet1: 120, Resnet2: 30,000
激活函数	RELU, Resnet2 的最后一层为 sigmoid
学习率	0.001

4.1. 数据

该数据源自某在线商城 2016 年 2 月 1 日至 2016 年 4 月 15 日的历史数据，该数据经过脱敏处理。该数据源可将数据分为三类：某类商品的基本信息，如表 2 所示，用户基本信息，如表 3 所示，用户和商品的交互数据，如表 4 所示。经过特征工程处理，将用户特征和商品特征处理成 one-hot 形式的向量。

Table 2. Product basic information data sheet
表 2. 商品基本信息数据表

字段名	含义
sku_id	商品 id
brand_name	品牌名
avg_price	均价
sale_qtty	销量
para_1	产品参数 1
para_2	产品参数 2
para_3	产品参数 3
para_4	产品参数 4
para_5	产品参数 5
para_6	产品参数 6
comment_num	评论数
good_comment_rate	好评率

Table 3. Basic table of user interaction data
表 3. 用户交互数据基本表

字段名	含义
user_id	用户 id
sku_id	商品 id
rank	商品的排序编号
is_purchase	是否购买该商品
a_type	行为(浏览，加购物车等)
a_date	行为日期

Table 4. User basic information data sheet
表 4. 用户基本信息数据表

字段名	含义
user_id	用户 id
age	年龄
sex	性别
user_lv_cd	用户等级
…	…

4.2. 评价指标

预测用户的购买与否行为是典型的二分类问题，所以在本文中我们采用二分类问题中常用的评价指标，即本文实验结果采用预测精度(Precision)和召回率(Recall)和 F1 综合评价三个指标，计算公式如下：

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{5}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{6}$$

$$\text{F1} = \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \tag{7}$$

其中，TP 为预测正确的购买正样本数，FP 则表示预测错误的负样本数，即预测购买但实际没有购买的样本数，FN 表示预测错误的正样本数，即用户实际购买了但是在预测中没有购买的样本数。

4.3. 实验结果

图 5 所示式基于 ResNet 的购买预测模型训练过程中损失的变化曲线，可以观察到，在前 10,000 步左右的时候，损失降低的比较快，从 0.9 左右快速降到 0.5 左右，随后 loss 降低的速度开始放缓，最终训练到 70,000 步左右时，损失降低到 0.3 左右。

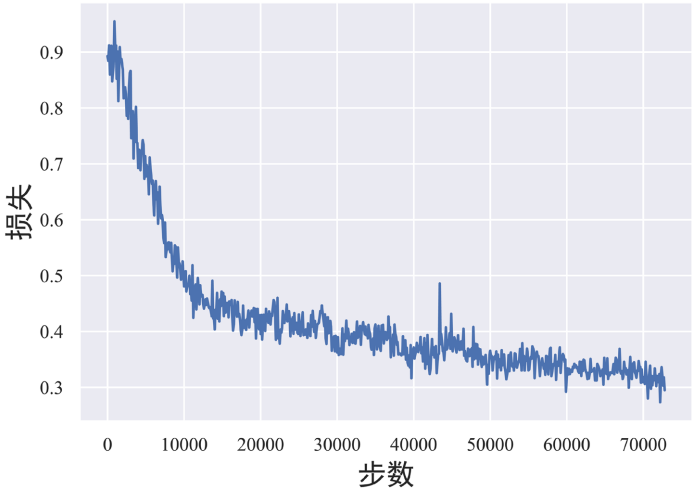

Figure 5. ResNet-based purchase prediction model training process
图 5. 基于 ResNet 的购买预测模型训练过程

图 6 所示是深度森林中随机森林的最大深度对最终结果 F1 指标的影响，可以看到在最大深度为 40 的时候，F1 的指标会达到最高，所以最终采用的随机森林的最大深度为 40。

为了对比基于深度残差网络和随机森林的组合模型对于用户购买行为的预测效果。在相同的训练集上，本文采用了其他三种预测方法训练了模型。在集成学习算法模型中，基于相同的用户特征、商品特征和用户行为数据训练了 XGBoost 模型和 LightGBM 模型；在线性分类模型中，训练了 SVM 模型；在机器学习算法中，训练了随机森林算法模型。最后，将在 5 个训练模型中训练的预测效果进行对比比较。如图 7 所示，展示了多种模型在该任务下的表现，本文提出的 ResNet + DF 的融合模型方法在 F1 和 Precision 指标的比较中均高于其他方法，仅在 Recall 指标中略低于 XGBoost 模型，表 5 中展示各指标的具体数值。

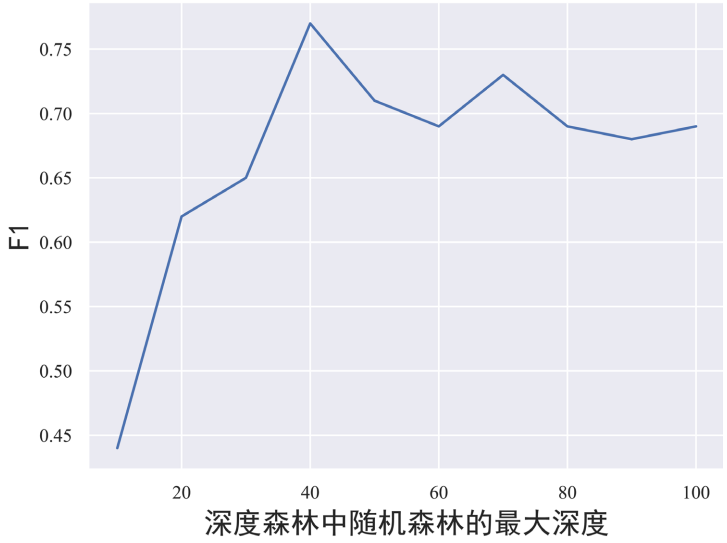

Figure 6. The impact of the maximum depth of different random forests in the deep forest on the final F1
图 6. 深度森林中不同随机森林的最大深度对最终 F1 的影响

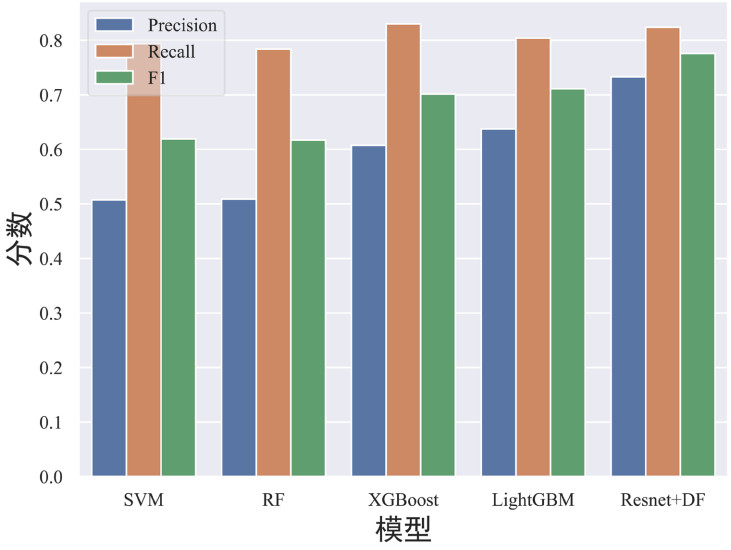

Figure 7. Comparison of results of multiple models
图 7. 多种模型结果比较

Table 5. Comparison of the performance of various algorithm models on the test set
表 5. 各种算法模型在测试集上的表现比较

模型	Precision	Recall	F1
SVM	0.5076	0.794	0.6193
RF	0.5090	0.784	0.6173
XGBoost	0.6075	0.830	0.7015
LightGBM	0.6375	0.804	0.7111
ResNet + DF	0.7330	0.824	0.7758

5. 总结

本文主要介绍了应用深度残差网络和随机森林两种模型的组合模型解决电商用户购买行为预测的问题。首先提出了基于深度残差网络和随机森林两种模型的组合模型来预测用户购买与不购买这一典型的二分类问题，后使用真实的电商数据来验证评估模型。根据实验结果可以看出，该组合模型比其他单一模型具有更高的 F1 值，其 F1 值为 0.7758。但本文的模型未对用户购买时间进行预测，即用户在浏览或收藏某一产品后产生购买行为的时间进行预测，基于时间的预测将是未来研究的重要方向。预测用户的购买行为及购买时间这种方法可以为企业的库存决策和精准营销提供有力的支持。

基金项目

课题编号：2020C01157；

课题名称：全流程供应链协同企业服务平台开发及应用——全流程供应链协同企业服务平台开发及应用；

计划类别：浙江省省级重点研发计划；

构建快消品及装备两个制造行业的全流程供应链协同模式，开发一个全流程供应链协同企业服务平台，实现市场感知、预警、响应与主动服务，具有市场预测、计划投放、补货决策、物流服务等功能，快速响应市场业务需求。

参考文献

[1] 刘潇蔓. 基于特征选择和模型融合的网络购买行为预测研究[D]: [硕士学位论文]. 北京: 北京交通大学, 2017.

[2] 卞天宇. 基于隐式反馈数据的用户行为分析及购买预测[D]: [硕士学位论文]. 南京: 南京邮电大学, 2020. https://doi.org/10.27251/d.cnki.gnjdc.2020.000253

[3] 吴非. 基于特征工程的用户购买预测模型研究[D]: [硕士学位论文]. 西安: 长安大学, 2019.

[4] 盛钟松. 基于 CatBoost 集成算法的用户购买预测研究[J]. 现代计算机, 2021(9): 15-18.

[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] Khade, A.A. (2016) Performing Customer Behavior Analysis Using Big Data Analytics. *Procedia Computer Science*, **79**, 986-992. https://doi.org/10.1016/j.procs.2016.03.125

[7] Liu, X. and Jing, L. (2017) Using Support Vector Machine for Online Purchase Predication. 2016 *International Conference on Logistics, Informatics and Service Sciences* (*LISS*), Sydney, 24-27 July 2016, 1-6. https://doi.org/10.1109/LISS.2016.7854334

[8] Sakar, C.O., Polat, S.O., Katircioglu, M., *et al.* (2018) Real-Time Prediction of Online Shoppers' Purchasing Intention Using Multilayer Perceptron and LSTM Recurrent Neural Networks. *Neural Computing and Applications*, **31**, 6893-6908. https://doi.org/10.1007/s00521-018-3523-0

[9] Jacobs, B., Donkers, B. and Fok, D. (2016) Model-Based Purchase Predictions for Large Assortments. Social Science Electronic Publishing, Rochester.

[10] 曾宪宇, 刘淇, 赵洪科, 等. 用户在线购买预测: 一种基于用户操作序列和选择模型的方法[J]. 计算机研究与发展, 2016, 53(8): 1673-1683.

[11] Biau, G. (2012) Analysis of a Random Forests Model. *Journal of Machine Learning Research*, **13**, 1063-1095.

[12] Breiman, L. (1996) Bagging Predictors. *Machine Learning*, **24**, 123-140. https://doi.org/10.1007/BF00058655

[13] Breiman, L. (2000) Some Infinite Theory for Predictor Ensembles. University of California, Berkeley.

[14] Breiman, L., Breiman, L. and Cutler, R.A. (2001) Random Forests Machine Learning. *Journal of Clinical Microbiology*, **2**, 199-228.

[15] Zhou, Z.H. and Feng, J. (2017) Deep Forest: Towards An Alternative to Deep Neural Networks. *Proceedings of the* 26*th International Joint Conference on Artificial Intelligence Main Track*, Melbourne, 19-25 August 2017, 3553-3559. https://doi.org/10.24963/ijcai.2017/497

[16] 张宾, 付玥, 周晶, 王帅, 李晓明. 基于深度森林的电商平台用户行为预测方法[J]. 信息技术, 2021(6): 96-101. https://doi.org/10.13274/j.cnki.hdzj.2021.06.018