Advances in Applied Mathematics 应用数学进展, 2023, 12(12), 5254-5266
Published Online December 2023 in Hans. https://www.hanspub.org/journal/aam
https://doi.org/10.12677/aam.2023.1212516

Hans汉斯

融合集成算法与宽度学习的商品需求量预测

曾诗懿[1*]，苏理云[1,2]，何青霞[1]，张　宇[1]，赵　锋[1]，张　彤[1]

[1]重庆理工大学理学院，重庆
[2]重庆理工大学时空大数据研究中心，重庆

收稿日期：2023年11月25日；录用日期：2023年12月19日；发布日期：2023年12月29日

摘　要

由于电商订单销售数据中全品类商品的种类繁多、分类多层化，且数据集还存在时序长度分布不均、地区差异、定性变量及非线性特征变量的处理等问题，导致需求量的预测任务较困难。为了解决上述问题，本研究提出一种宽度学习集成框架，将机器学习中的Random Forest、GBDT、XGBoost和LightGBM与宽度学习模型进行随机融合，并分别进行验证，对比模型效果。经实证分析结果表明：LightGBM-BLS模型具有最优的预测性能和计算性能，它在保持LightGBM模型计算优势的同时，大幅度地提升了模型本身的预测精度，使拟合优度达到0.99，评价指标RMSE、MSE降低90%以上，MAE降低85%以上。

关键词

需求量预测，特征工程，宽度学习(BLS)，XGBoost，LightGBM

Fusing Integrated Algorithms with Broad Learning System for Commodity Demand Forecasting

Shiyi Zeng[1*], Liyun Su[1,2], Qingxia He[1], Yu Zhang[1], Feng Zhao[1], Tong Zhang[1]

[1]College of Science, Chongqing University of Technology, Chongqing
[2]Research Center for Spatiotemporal Big Data, Chongqing University of Technology, Chongqing

Received: Nov. 25th, 2023; accepted: Dec. 19th, 2023; published: Dec. 29th, 2023

Abstract

Due to the wide variety of full-category commodities in e-commerce order sales data, multi-layered

*通讯作者。

文章引用: 曾诗懿, 苏理云, 何青霞, 张宇, 赵锋, 张彤. 融合集成算法与宽度学习的商品需求量预测[J]. 应用数学进展, 2023, 12(12): 5254-5266. DOI: 10.12677/aam.2023.1212516

categorization, and the dataset also has the problems of uneven distribution of time-series lengths, regional differences, and the treatment of qualitative variables and non-linear feature variables, which leads to a more difficult task of demand prediction. To solve the above problems, this study proposes a breadth learning integration framework, which stochastically fuses Random Forest, GBDT, XGBoost and LightGBM in machine learning with the breadth learning model, and validates and compares the model effects respectively. Empirical analysis results show that the LightGBM-BLS model has optimal prediction performance and computational performance, which maintains the computational advantages of the LightGBM model while substantially improving the prediction accuracy of the model itself, so that the goodness of fit reaches 0.99, and the evaluation indexes of RMSE and MSE are reduced by more than 90%, and MAE is reduced by more than 85%.

Keywords

Demand Forecasting, Feature Engineering, Broad Learning System, XGBoost, LightGBM

1. 引言

随着大数据信息技术的高速发展，电商行业的销售模式也随之而转变，会应个人喜好、生活习惯、价格区间等各类因素的综合参考，来决定某件商品的销售策略、方式及价格。因此，各大销售平台开始注重商品销售量及需求量的准确预测问题，可利用人工智能的方法[1]快速准确地实现商品销量、需求量的预测，高效提升整个商品生产供应链的绩效，大幅度提升市场竞争力，维持商品销量的稳定发展。然而，在过往的研究中，大部分商品的销量预测或需求量预测问题，均采用了时间序列预测方法[2] [3]或机器学习的预测方法进行处理。如：胡小芳(2022) [4]和 Yasaman (2022) [5]等采用了神经网络、传统时间序列模型等方法对零售店的商品销售情况进行了预测。但时间序列预测方法对原始数据集的时序长度要求较高，需在保证有较长的时序长度信息的条件下，才能更丰富地提取时序信息与目标变量间的关系，实现高精度的预测任务。考量本文采用的数据集，其时序信息既不连续也不充足，导致无法准确提取到丰富的时序信息。

同时，还有较多学者采用了机器学习或集成机器学习的方法实现商品销量的预测[6] [7] [8] [9]，如：刘辰阳(2022) [10]、吴霆辉(2023) [11]和谢勇(2019) [12]等分别采用了特征工程与机器学习模型结合的方法或是几种机器学习模型集成的方法来预测了电力负荷、住房月租金和商品销量等目标，其中吴霆辉就将 LightGBM 模型与 XGBoost 模型相融合，使模型的拟合优度提升到 0.98 左右。王晓辉(2020) [13]提出了基于遗传算法与随机森林的 XGBoost 改进方法的研究，使模型的拟合优度提高了 0.01%~2.1%左右。霍佳震(2023) [14]则在 GBDT 模型的基础上提出了基于 EEMD-Holt-Winters-GBDT 模型，在零售商品销量多步预测问题中表现出良好的预测性能。还有 Wang (2022) [15]和 Yuanyuan (2020) [16]等采用了 LightGBM 模型与 TCN 网络方法、K 近邻算法的结合。各类机器学习方法、特征工程处理方式的相融合的方法，能实现相对于单个机器学习模型预测精度的提升，但多种机器学习方法融合的方式存在一个计算时间缺陷。由于单个机器学习模型实现预测任务时，训练时间较长，融合多个模型的同时，虽然提高了一定的准确性，但计算速度较慢，使其模型若投入产品应用中并不能达到理想状态。

在该基础上，本文即思考和探索是否能在保证高效计算速度的同时，实现模型预测性能的提升，并将此作为文章的主要任务。近年来，陈俊龙教授(2018) [17]提出了一种宽度学习系统(BLS)，可通过增强节点的引入来增加网络中的非线性因素，以此更大信息化地提取输入数据集的特征信息。该方法在各类预测任

务中也得到了广泛的应用，如：苏理云(2023) [18]等提出了一种宽度学习方法与多头注意力机制相结合的框架，可捕获关键的时空特征信息以实现更高的预测性能；还有褚菲(2020) [19]则从理论上提出了基于lasso 和 elastic net 的宽度学习系统(BLS)网络结构稀疏方法；杨光雨(2022) [20]提出了一种基于最大信息挖掘宽度学习系统多核最小二乘支持向量机进行了短期电力负荷预测，得到了较高的预测精度。基于已有学者将宽度学习融合应用的方法，发现宽度学习模型优秀的运算速度、可自动捕捉数据集特征及非线性特征的优点，均能恰好弥补机器学习中树模型的缺陷，于是本研究即考虑通过实证分析的方式来验证是否宽度学习模型能帮助机器学习模型，提升在订单需求量预测任务上的预测性能，并增强其泛化能力。

2. 研究设计

数据来源

本研究的实证数据来源于第十一届泰迪杯 B 题中提供的企业面向经销商的出货数据，主要涵盖近 60 万条商品销售数据，其中数据集包含 5 个不同地区、8 种商品大类、12 种商品细类，近一千多种商品类别的销售情况数据。数据集的样本区间为 2015 年 9 月 1 日至 2018 年 12 月 20 日。

数据来源网站参考如下：https://www.tipdm.org/。

3. 研究方法

3.1. 特征工程

特征工程(Feature Engineering)是选择、操作和将原始数据转换为可用于监督学习变量特征的过程，以便于将所提取的变量特征应用到构建的预测模型中，达到提升模型预测对未知数据的准确度。简而言之，即通过特征工程的处理方式提取出自变量 X 中存在的显著影响特征和信息。

本研究中即应用特征工程提取了商品销售数据的节假日信息、促销日信息、月末月初等时序信息，继而还对数据进行了去重处理和商品价格信息的分箱处理，以平均价格($\bar{X}_{price}$)压缩了价格特征，即：

$$\bar{X}_{price}=\frac{\sum_{i}^{n}X_i^{T,code}}{n} \tag{1}$$

其中 T 为日期；$code$ 为同个商品的编码。

由于商品的大类和细类数据属于离散型数据，为了使机器学习模型能更好的识别到该类变量的特征信息，还采用了特殊的编码处理方式。最后，考虑到商品的历史订单需求量数据与当期需求量数据间的相关性，进而提取了该变量的滞后特征和趋势特征。

3.2. 特殊编码处理

1) 均值编码：

当数据集中存在定性特征时，由于定性特征表示某个数据属于一个特定的类别，其数据均表现为类别的离散型数据，为了充分提取分类变量的信息量，均值编码通常的处理方式即把概率替换成均值，y 为目标变量，x 为定性特征变量：

$$p(y,x_i)=(1-\lambda(n_i))\frac{\sum_{x=x_i}y}{n_i}+\lambda(n_i)\frac{\sum y}{N} \tag{2}$$

其中：$\frac{\sum_{x=x_i}y}{n_i}$ 表示 $x=x_i$ 对应的 y 均值，$\frac{\sum y}{N}$ 是整个训练集上 y 的均值。

2) 独热编码：

独热编码又称哑变量，是将离散特征的取值扩展到欧式空间上，且离散特征的某个取值分别对应欧式空间上的某个点，根据特征之间的距离进行编码，且编码后的特征，每一个维度的特征都可以看成连续的特征，能实现从离散到连续的转换。且每列变量被拆开为连续性的均标注为 0 或 1。

3.3. 宽度学习模型

宽度学习(Broad Learning System)系统是一种有效且高效的增量学习系统，与深度学习不同，它是一种不依赖于深度结构的神经网络结构。实质是一种随机向量函数链接神经网络，但与 CNN 不同，该网络并不通过反向传递改变特征提取器的核，而是通过求伪逆计算每个特征节点和增加节点的权重。

首先，需要构建输入数据到特征节点的映射，其映射的特征节点即实现了高效特征提取的能力。在特征学习阶段，原始输入数据通过特征映射节点随机转换为特征，然后将其连接到增强节点作为输入。

假定存在 m 组特征节点，且每个映射节点具有 q 个特征，假设输入数据为 $X \in R^{n\times d}$，其中 $X=\left[x_1,x_2,\cdots,x_n\right]^{\mathrm{T}}$ 为输入样本数，也即嵌入维数；则第一映射特征节点的映射特征公式如下：

$$D_i=\phi\left(XW_{e_i}+\beta_{e_i}\right),i=1,2,\cdots,m \tag{3}$$

其中：W_{e_i} 表示随机生成的权重，β_{e_i} 为第 i 组映射节点；$\phi(\cdot)$ 表示输入数据的激活函数。

级联的映射特征 $D=\left[D_1,D_2,\cdots,D_m\right]$ 和 D 会被进一步连接到增强节点。假设存在增强节点，则第 j 组的增强信息可通过如下的公式得到：

$$E_j=\xi\left(DW_{h_j}+\beta_{h_j}\right),j=1,2,\cdots,d \tag{4}$$

其中，W_{h_j} 和 β_{h_j} 分别表示随机产生的权重和偏差，且 $\xi(\cdot)$ 是激活函数。

则最终得到增强特征的输入即：

$$H=\left[D_1,D_2,\cdots,D_m\mid P_1,P_2,\cdots,P_d\right]=\left[D\mid P\right] \tag{5}$$

3.4. 基础模型介绍

3.4.1. 随机森林模型

随机森林(Random Forest, RF)属于 Bagging 算法之一，可通过组合多个弱分类器，最终结果以投票或取均值的方式，使整体模型的结果具有较高的精确度和泛化性能。且采用的 CART 决策树是基于基尼系数来选择特征的。

对未知样本 x 的预测可通过对所有单个回归树的预测取平均来实现：

$$\hat{f}=\frac{1}{B}\sum_{b=1}^{B}f_b\left(x'\right) \tag{6}$$

3.4.2. 梯度提升决策树模型(GBDT)和 XGBoost 模型

GBDT 和 XGBoost 模型均是属于集成学习中 Boosting 提升算法，其中 GBDT 主要是借助梯度下降的优化方法，且使用损失函数的负梯度，没有加入正则化项。而 XGBoost 则是基于二阶泰勒展开优化损失函数，在 GBDT 的基础上引入了正则化项 $\Omega\left(f_k\right)$，提高模型的计算精度，并将其目标函数变为：

$$L(\phi)=\sum_i l\left(y_i,\hat{y}_i\right)+\sum_k\Omega\left(f_k\right) \tag{7}$$

其中 $L(\phi)$ 是线性空间上的表达；k 表示第 k 棵树，$\hat{y}_i$ 即第 i 个样本 x_i 的预测值。

3.4.3. LightGBM 模型

LightGBM (Light Gradient Boosting Machine)是一种开源的分布式高性能梯度提升，原理和 GBDT 类

似，但具有更高效的训练，更低的内存使用以及更高的结果准确性等特性。相比较于 GBDT 和 XGboost 算法，LightGBM 使用的是直方图算法，占用的内存更低，且数据分割的复杂度也更低，能够在不损害准确率的条件下，加快 GBDT 模型的训练速度。其思想主要将连续的浮点特征离散成 K 个离散值，并构造宽度为 K 的 Histogram，通过遍历训练数据，统计每个离散值在直方图中的累计统计量。

综合对比以上梯度提升树模型各自的技术特点，发现随机森林(RF)模型具有对数据质量较低，且基学习器出错不会对整体结果造成较大影响的优势；GBDT 则将所有决策树的结果进行求和得到最终结果，在数据质量较好的情况下具有更好的精度；XGBoost 模型引入正则化项后，在 GBDT 模型的基础上缓解了过拟合的情况，并减少了运行时间；LightGBM 与 XGBoost 模型类似，保留了原本 XGBoost 的优势，并增进了可直接处理连续或离散特征的特点，在保证精度的同时再次提升了模型的计算速度。

3.5. 集成宽度学习算法

由于机器学习模型多从数据指标自身提取特征信息，来实现模型的预测，未充分考虑到特征变量间的非线性特征信息，继而导致降低了模型的预测精度。然而宽度学习算法(BLS)恰好可以解决这类问题，可充分提取特征变量间存在的非线性关系。于是本研究提出了集成宽度学习算法的模型框架，即分别将随机森林、GBDT、XGBoost 和 LightGBM 模型与宽度学习(BLS)算法进行集成融合，其集成算法的基本实现步骤如下：

Step1：经特征工程处理后得到的指标变量 $X_i\left(i=1,2,\cdots,n\right)$ 和目标变量 Y_{dem}，首先将指标变量 $X_i\left(i=1,2,\cdots,n\right)$ 分别输入 Random Forest(随机森林)、GBDT、XGBoost 和 LightGBM 模型中，经模型训练后得到不同区域商品订单需求量的预测值 $ord\_Y_{pre}^{m}$ (m 表示选取的预测模型)。

Step2：其次，将 Step1 中得到的四个模型预测值，根据区域的不同依次纳入 Step1 中的指标变量 $X_i\left(i=1,2,\cdots,n\right)$ 中，重新进行特征融合后，得到新的指标数据集：$X_j^m\left(j=1,2,\cdots,n+1\right)$，即分别添加一维不同区域对应的 $ord\_Y_{pre}^{m}$ 后得到的 $X_j^m\left(j=1,2,\cdots,n+1\right)$，其中 Step1 所提到的四种预测模型，将这些新的 $X_j^m\left(j=1,2,\cdots,n+1\right)$ 依次作为新的输入特征。

Step3：最后，采用宽度学习(BLS)的方法，分别对新的特征数据集 $X_j^m\left(j=1,2,\cdots,n+1\right)$ 实现商品需求量的预测任务，得到由 RF-BLS、GBDT-BLS、XGB-BLS、LGBM-BLS 这四类集成算法得出的最终预测结果 $\hat{Y}_i\left(i=1,2,\cdots,n-1,n\right)$。具体流程可参考如下图 1 和图 2，其中图 2 为图 1 中宽度学习算法(BLS)的基本框架图，作为图 1 的补充。

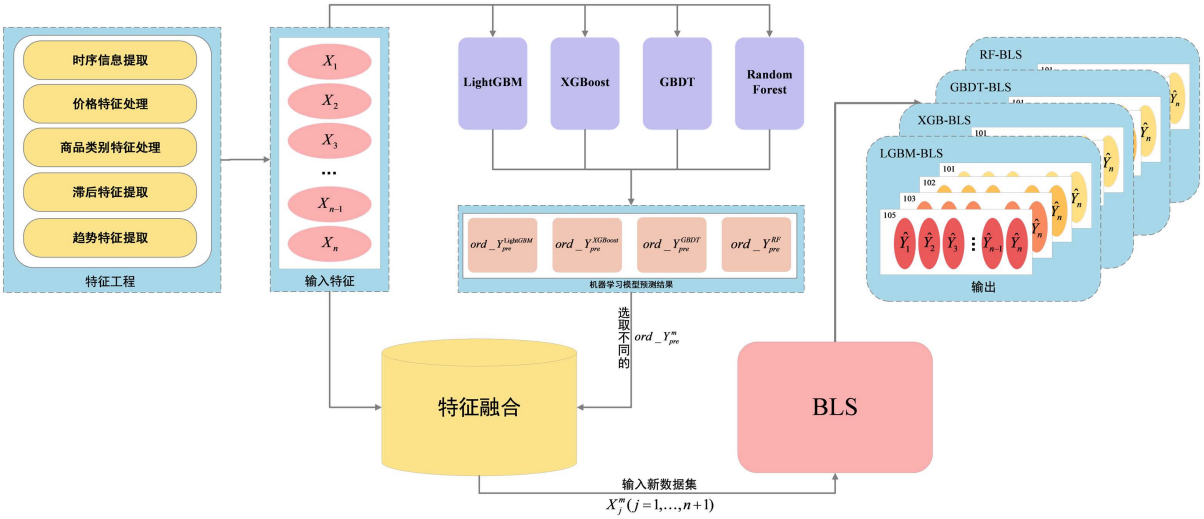


Figure 1. Integrated width learning algorithm framework diagram
图 1. 集成宽度学习算法框架图

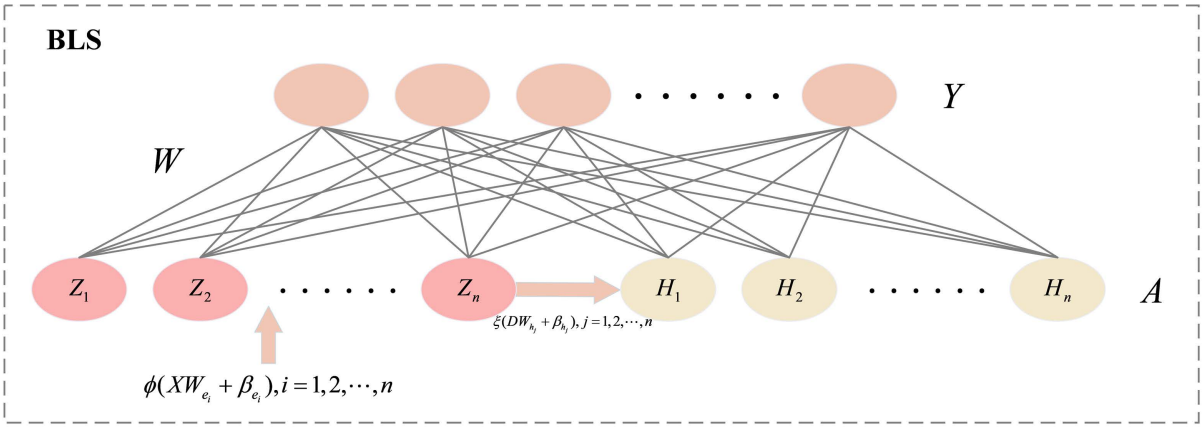


Figure 2. Width learning algorithm framework diagram
图 2. 宽度学习算法框架图

4. 实证结果与分析

4.1. 特征工程及其分析结果

为了提升模型的准确度，本研究构建了特征工程，来提取原始数据中对商品需求量的直接影响变量和潜在影响变量。原始实证数据涵盖的变量信息有：订单交易时间、商品区域代码、商品编码、商品大类编码、商品细类编码、销售方式、产品单价及订单需求量。目标预测变量即为订单需求量(ord_qty)，考虑到原始数据集本身存在时间信息，特征工程的第一步即提取了原始数据集中潜在的时间信息，如：是否节假日、是否工作日、是否月中月末、是否促销日等信息，其中促销日和节假日信息对商品的需求量存在显著的影响。

其次，观测整个数据各类商品的数据分布情况，发现共计有 1294 种商品，但每种商品对应的订单数据量并不相同，且每次订单数据产生的时序信息也并不均匀、时间段并不连续。某类商品还存在同一天具有多条订单信息的情况，为了充分提取商品的价格信息，本研究考虑将该类同天存在多条订单信息的价格数据进行压缩处理，采用均值处理日期相同、且商品相同的价格数据，使同天内，同类商品的价格数据信息仅保留其对应日期的均价；然而，由于数据集中各类商品的价格信息波动范围较大，其原因可能是由于商品种类的不同，而导致价格波动范围也存在较大差异，为了更好的提取因类别造成价格差异的特征信息，添加了平均价格分箱特征，来获取商品间的差异性。除此外，还根据商品线上线下的销售方式，添加了不同商品的线上线下销售比这一特征信息。其中，各类商品的时序分布情况如下图 3 所示。

根据不同商品时序图的分布情况，发现各类商品对应的时序长度存在较明显的差异，数量较多的时序长度基本分布在[0, 50]天，而超过 100 天的时序长度较少，且分布不均匀，同类商品中还存在时序信息不连续的情况。故针对该类时序数据分布不均，且长度较短的数据集，若采用时间序列模型预测商品需求量并不合适，因为数据本身体现的时序信息不能充分的识别到商品需求量信息随时间变化的规律，反而可能造成错误的信息干扰，从而影响模型的预测准确度。据此因素，且结合考虑数据集中存在定性特征，本研究则选择了随机森林、GBDT、XGBoost 类型的树模型及宽度学习模型来识别指标变量特征，预测订单需求量。

最后，因考虑到商品的历史需求量数据对当期订单需求量数据的影响，添加了需求量变量的 1, 2, 3, ···, 48, 60 阶的滞后项特征，获取历史需求量的波动信息；还添加了商品需求量的移动平均特征，获取相邻

日期间需求量数据间存在的影响；需求量数据的变化趋势特征也同样属于目标变量的直接影响变量，故特征工程中也添加了该变量的趋势特征。由于商品大类和细类等分类数据，属于定性数据、均是离散分布的，于是针对类别数据分别采用了均值编码和独热编码的特殊编码处理方式。

最终，经过特征工程处理后，得到 59 维输入特征变量。

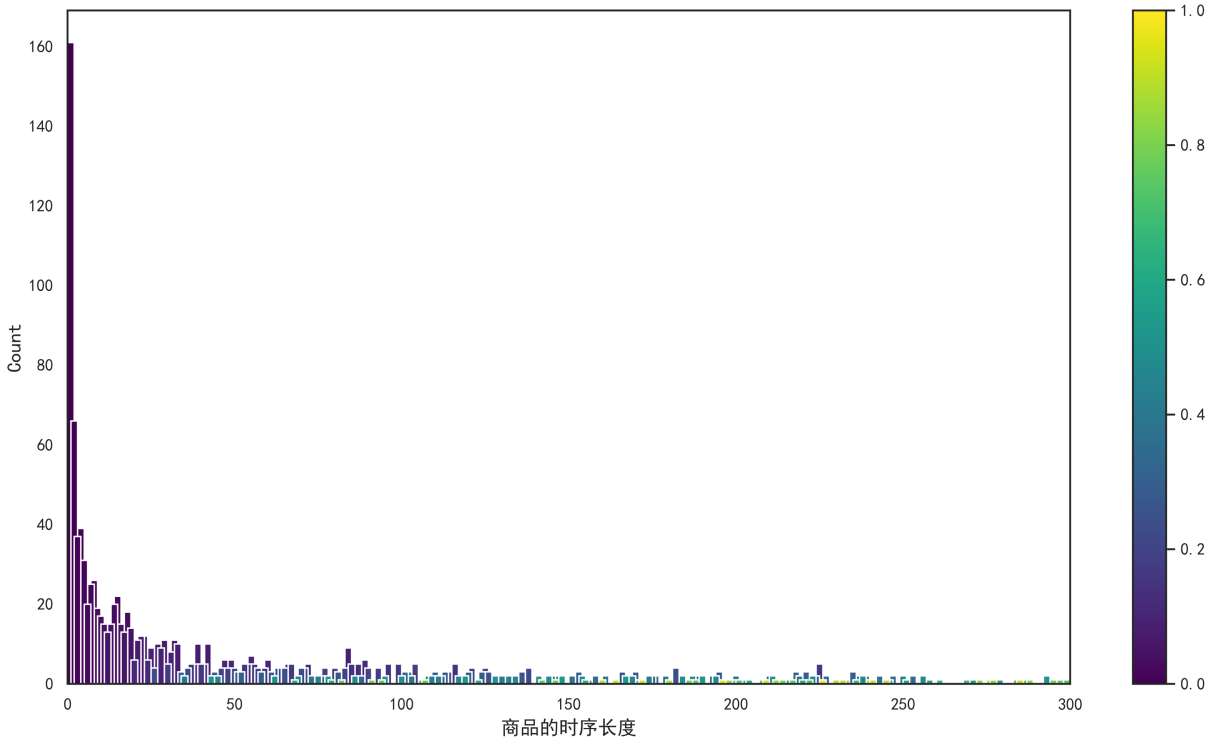


Figure 3. Statistical chart of chronological lengths for all categories of goods
图 3. 全类商品的时序长度统计图

4.2. 模型效果评价指标

为了衡量模型的预测性能，本研究选择了均方误差(Mean Square Error, MSE)、平均绝对误差(Mean Absolute Error, MAE)、均方根误差(Root Mean Squared Error, RMSE)、平均绝对百分比误差(Mean Absolute Percentage Error, MAPE)作为评价指标，并结合模型计算耗时(Time)来综合评定模型效果的优劣。各模型的评价指标计算公式如下：

$$\mathrm{MSE}=\frac{1}{n}\sum_{i=1}^{n}\left(Y_i-\hat{Y}_i\right)^2 \tag{8}$$

$$\mathrm{MAE}=\frac{1}{n}\sum_{i=1}^{n}\left|y_i-\hat{y}_i\right| \tag{9}$$

$$\mathrm{RMSE}=\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y_i-\hat{y}_i\right)^2} \tag{10}$$

$$\mathrm{MAPE}=\frac{100\%}{n}\sum_{i=1}^{n}\frac{\left|y_i-\hat{y}_i\right|}{y_i} \tag{11}$$

4.3. 集成模型与机器学习模型的对比分析

4.3.1. 基于机器学习模型的区域全商品需求量预测

综合文献分析，发现电商订单需求量预测任务中，较多学者采用随机森林(Random Forest)、梯度提升决策树(GBDT)、XGBoost 和 LightGBM 这四种机器学习模型[2] [3] [4]，而为了在此基础上，探索一种精度更高且计算速度更优的模型框架来帮助更好地解决电商订单预测问题，即需首先探究这四种模型在本研究中数据集上的表现能力，并便于根据不同模型的预测能力分析其优缺点。

因原始数据集中，商品存在区域差异，为了更准确的预测出全品类商品的订单需求量，将训练集数据按区域拆开，分别对各区域(101、102、103、105 区域)的订单需求量进行预测。其中，训练集数据的时间维度为 2015 年 9 月 1 日~2018 年 12 月 12 日，共 113,332 条商品订单数据，测试集数据时间维度为：2018 年 12 月 13 日~2018 年 12 月 20 日，共计 12,593 条不同商品的订单数据。经实证结果比较发现，四种机器学习模型中 Random Forest 和 XGBoost 模型是效果最好的，且各模型在 103 地区均表现最优，则此处仅列举了 103 地区各模型的评估结果：

Table 1. Area 103—evaluation results of machine learning models
表 1. 103 地区——机器学习模型的评估结果

区域	模型	MSE	MAE	RMSE	MAPE	Time(s)
103	**Random Forest**	**4.0595**	**0.4949**	**2.0148**	**0.0302**	184.1345
	GBDT	40.4232	2.3192	6.3579	0.2605	40.0599
	XGBoost	12.2152	0.6361	3.4950	0.0412	74.4189
	LightGBM	348.4430	8.9274	18.6666	0.8562	**1.4590**

综合比较表 1 中四种模型的结果发现，机器学习模型预测精度的排名依次为：Random Forest > XGBoost > GBDT > LightGBM，计算速度的排名依次为：LightGBM > GBDT > XGBoost > Random Forest。很明显 LightGBM 模型在本研究数据集上表现出了较差的预测精度，仅在计算速度占据了较强的优势，但通常情况下 LightGBM 较 XGBoost 模型相比具有更优越的准确性和计算速度，而在本研究的数据集上，Random Forest 和 XGBoost 却表现出了更优越的预测性能，其原因可能是由于数据特性造成的，机器学习模型的预测效果均较依赖于特征工程的处理，同时 XGBoost 模型中提供了较多的正则化选项，有助于控制模型的复杂性。即 GBDT、XGBoost 和 LightGBM 模型对数据质量的要求较高，更适用于数据特征信息丰富且存在显著关系的数据集，而 Random Forest 对数据质量的要求相对较低。

于是考虑到电商订单预测问题的实用性，需实现产品订单需求量的高效预测。本研究考虑探寻既能保持 LightGBM 模型计算速度的优势，又能高效提升 LightGBM 模型预测精度的方法。因宽度学习模型具有可自动从原始数据中学习有用的特征，不依赖于特征工程的处理；且模型中添加了 L1 和 L2 正则化，有助于限制模型的参数大小，控制其复杂性和防止过拟合等优点。且这些优点均是 LightGBM 模型所欠缺的，于是本研究就宽度学习和 LightGBM 的互补性，提出了一种融合宽度学习的集成算法框架，并通过实证分析的方式探究了机器学习模型和宽度学习模型的融合方式，是否能帮助模型提升其预测精度和计算速度，达到更好的预测性能，尤其是 LightGBM 模型是否能达到预期目标的效果，以下 3.3.2 小节对此展开的具体分析。

4.3.2. 集成宽度学习模型实证对比

根据 2.5 小节提出了集成宽度学习算法结果，分别采用 RF-BLS、GBDT-BLS、XGB-BLS、LGBM-BLS 这四种集成算法对同样的数据集进行实证分析，经模型评估结果的对比发现该四种模型在 103 地区的表现最优，105 地区表现最差，于是为对比集成算法模型和机器学习算法的预测能力，仅需在预测能力表现最差和最优的地区均呈现出集成模型更优的预测性能即可，以下即为集成宽度学习算法和机器学习算法分别在 103 和 105 地区的模型评估结果：

Table 2. Evaluation results of integrated algorithmic models and machine learning models
表 2. 集成算法模型及机器学习模型的评估结果

区域	模型	MSE	MAE	RMSE	MAPE	Time(s)
103	Random Forest	4.0595	0.4949	2.0148	0.0302	184.1345
	RF-BLS	21.3300	1.7109	4.6184	0.0909	184.5864
	GBDT	40.4232	2.3192	6.3579	0.2605	40.0599
	GBDT-BLS	27.9073	4.0104	5.2827	0.1601	40.5242
	XGBoost	12.2152	0.6361	3.4950	0.0412	74.4189
	XGB-BLS	6.8904	1.9452	2.6250	0.0949	74.8711
	LightGBM	348.4430	8.9274	18.6666	0.8562	1.4590
	LGBM-BLS	**0.3522↓**	**0.4893↓**	**0.5935↓**	**0.0319↓**	**1.9157 (+0.4567)**
105	Random Forest	405.2788	0.9463	20.1315	0.0355	269.2298
	RF-BLS	4.1200	0.7779	2.0298	0.0480	269.8824
	GBDT	474.2444	2.7645	21.7772	0.2458	60.0523
	GBDT-BLS	21.5067	2.3980	4.6375	0.1009	60.7133
	XGBoost	217.4326	1.0921	14.7456	0.0477	101.1275
	XGB-BLS	7.3558	1.5575	2.7122	0.0875	101.7807
	LightGBM	3248.2117	8.5242	56.9931	0.5298	5.0106
	LGBM-BLS	**1.2046↓**	**1.0929↓**	**1.0976↓**	**0.0738↓**	**5.6607 (+0.6501)**

*注：RF 为 Random Forest 缩写；GBDT 为梯度提升决策树的缩写；XGB 为 XGBoost 缩写；LGBM 为 LightGBM 缩写；BLS 为宽度学习(Broad Learning System)的缩写；(+0.4567)和(+0.6501)分别为 LGBM-BLS 模型与 LightGBM 模型的计算速度差。

根据表 2 中模型的评估结果可观测到，融合宽度学习的集成算法模型与原始机器学习的模型预测精度相比，基本均有所提升，使改进后的集成算法模型均在预测精度上得到了较大幅度的提升，且计算速

度基本保持不变，其时间误差控制在了 1 秒以内。此外，交叉比较四种传统机器学习模型和四种集成算法模型的效果，容易发现 LGBM-BLS 模型的预测性能最优，它不仅表现了模型最优的预测准确性，同时也兼顾了优越的计算速度，很好地融合了 BLS 和 LigntGBM 模型本身的优点，与 LigntGBM 模型相比其计算时间仅相差 0.5 秒左右，与其他三种集成算法相比，具备更强的预测能力，达到了订单需求量预测任务的预期目标。且表明文章所提出的该四类模型适用于样本量及特征量丰富、存在分类特征信息的数据集，例如：电商数据分析、股票数据分析以及能源相关数据分析等。

为了便于更直观地观测到集成宽度学习算法预测精度的大幅提升，在表 2 的基础上选取了 RMSE 指标，绘制了传统机器学习模型与集成宽度学习算法间 RMSE 指标对比图，并计算了两类模型对应 RMSE 指标的降幅(见图 4)：

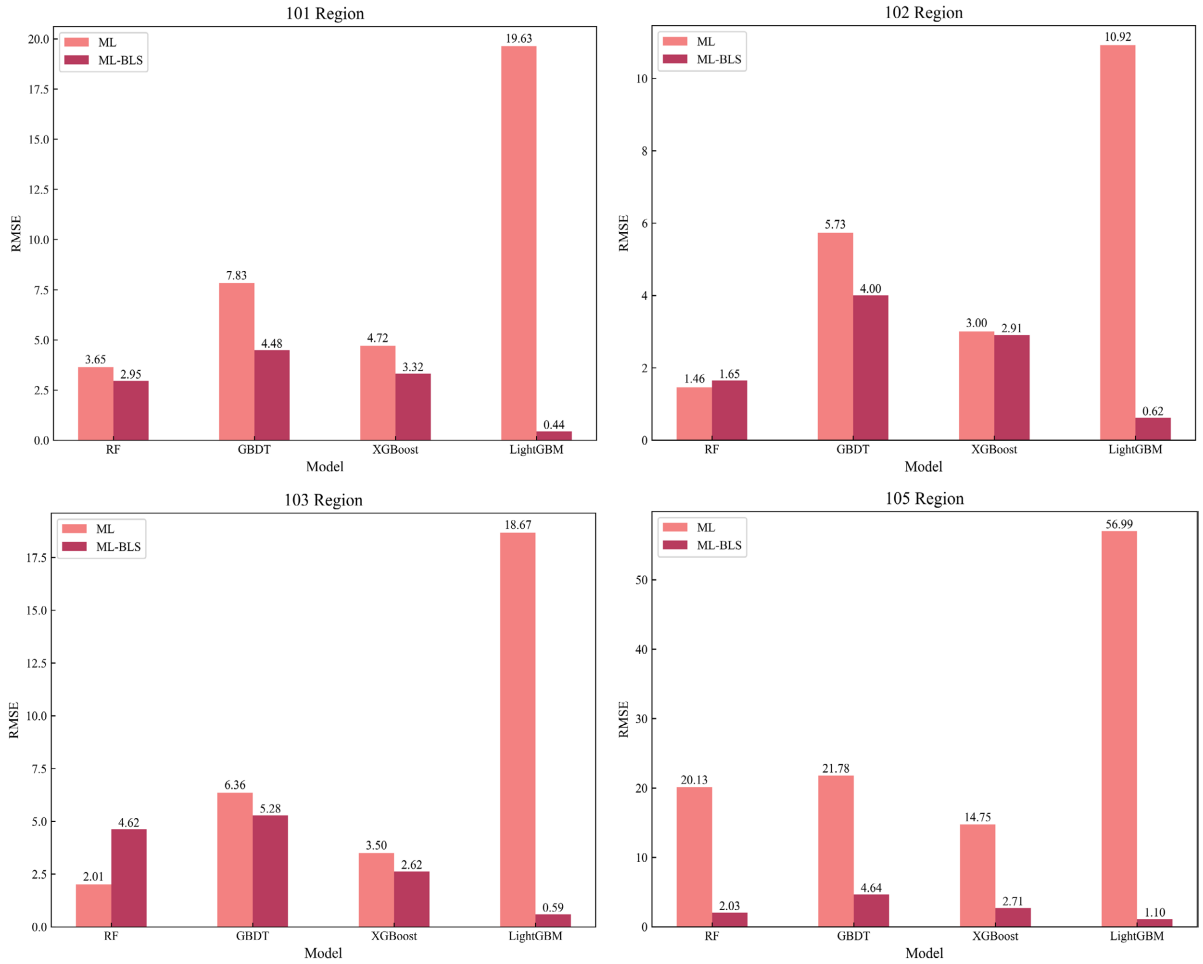


Figure 4. Comparison chart of RMSE metrics between models
图 4. 模型间 RMSE 指标对比图

图 4 中 ML 表示机器学习模型(Machine Learning)，ML-BLS 表示与宽度学习融合的集成算法；各柱子上方的数值标签即为集成算法与传统机器学习方法对应的 RMSE 误差值。对比 ML 和 ML-BLS，显然 LGBM-BLS 是降幅最大的，在各地区该模型的均方根误差(RMSE)均下降了 90%以上，说明 LGBM-BLS 的预测精度最优。进而，本研究继续将 LGBM-BLS 模型在测试集上的预测效果进行了可视化，得出了预测值与真实值间高度吻合的结论(见图 5)：

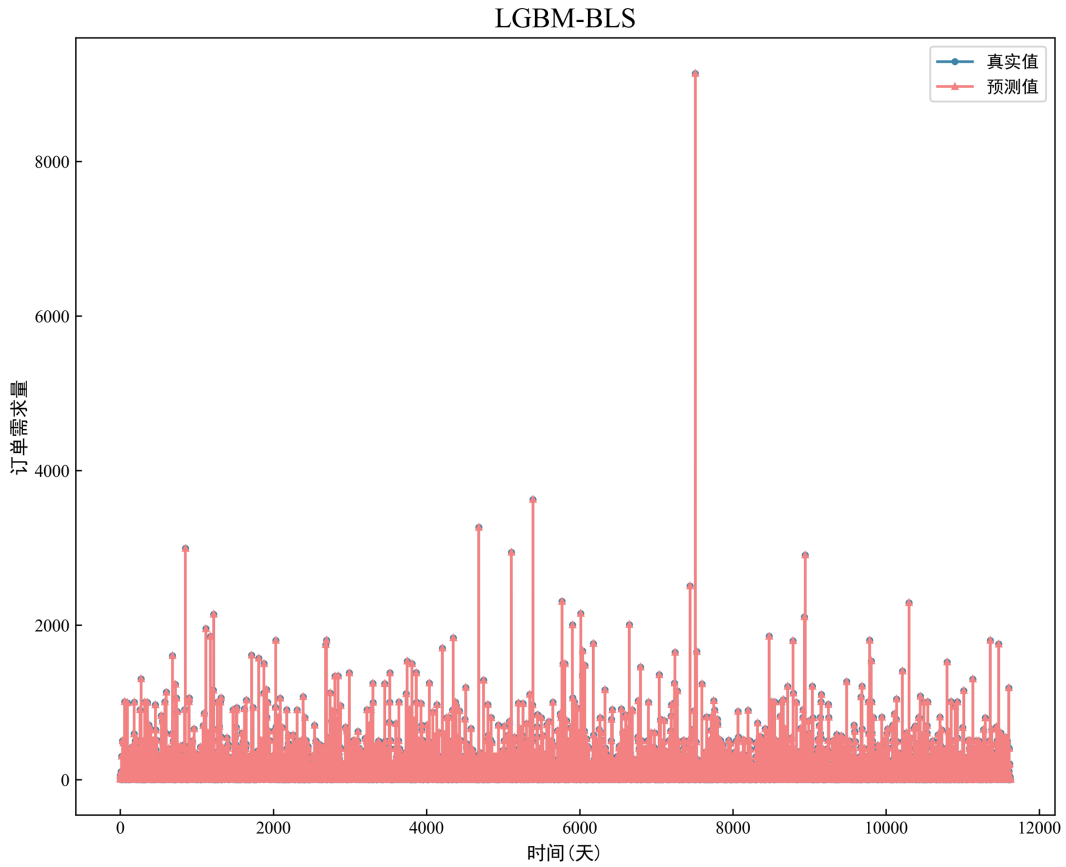
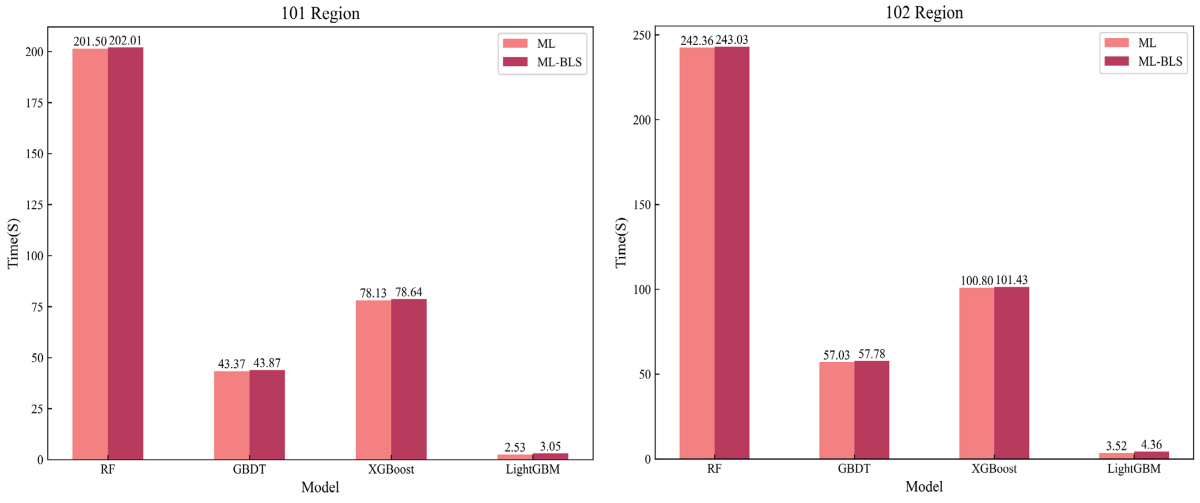


Figure 5. Test-set prediction effects of LGBM-BLS modeling
图 5. LGBM-BLS 模型的测试集预测效果图

最后，对集成宽度学习算法和机器学习模型的计算性能进行了可视化，以下为 ML 和 ML-BLS 分别在 101、102、103 和 105 地区测试集上的计算时间对比图。

图 6 中各柱子上的数值标签为集成宽度学习算法与对应的传统机器学习模型在测试集上的计算时长，从该数值的大小能直观观测到，集成算法均继续保持了 ML 模型原有的计算速度，并未因为算法的融合而出现更高的时间复杂度；同时，更容易发现 LGBM-BLS 模型的计算优势远超于其他模型。



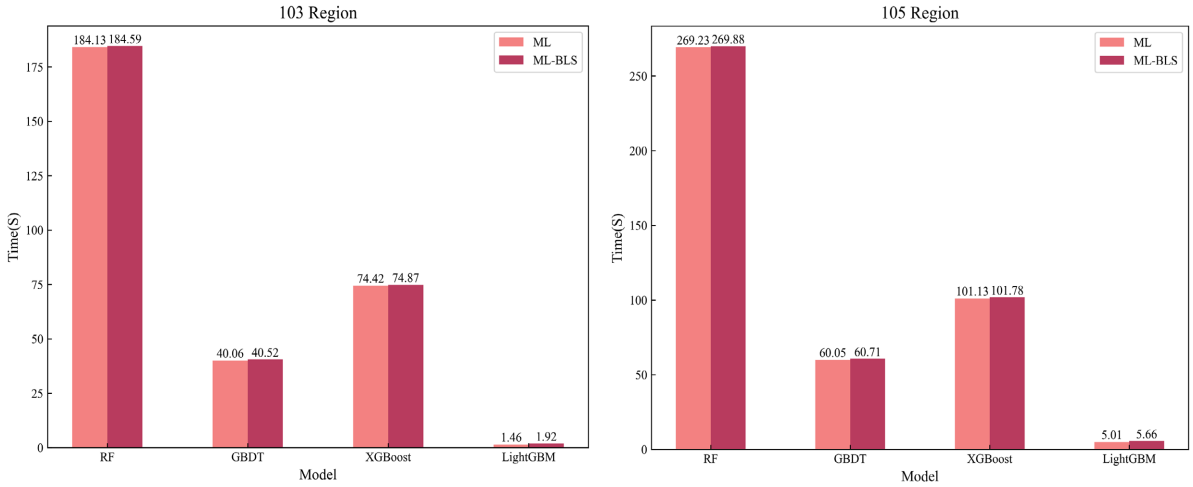


Figure 6. Comparison chart of computation time for each model (test set)
图 6. 各模型计算时长对比图(测试集)

5. 结论与建议

不同的模型均存在各自的优缺点，随机森林(Random Forest)、GBDT、XGBoost 和 LightGBM 模型中，LightGBM 具有较明显的计算速度优势；但是，若要同时兼顾预测模型的精度和计算速度，显然仅靠 LightGBM 模型的效果是不够的，且当采用 LightGBM 处理大规模数据集时还容易产生过拟合的情况。然而 LightGBM-BLS 的融合却很好地解决了该问题。由于正则化项的存在，有效地防止了模型的过拟合，且能避免模型精度容易受到特征工程效果的影响，因为宽度学习模型能自动从原始数据中学习有用的特征，降低了模型对特征工程的依赖度，使模型能最大信息化利用原始数据集的丰富特征，实现预测任务。

因此，本研究结合宽度学习模型的优势，提出了宽度学习模型与机器学习模型集成的算法框架，均在一定程度上提升了 RF、GBDT、XGBoost 和 LightGBM 模型的预测精度。尤其是 LightGBM-BLS 模型很好的达到了预期为实现高效计算且高精度预测的目标，为销售行业的订单分析及预测问题，提供了较好的思路方法，具有丰富的参考价值和实际意义。

基金项目

2023 年重庆市教育委员会人文社会科学研究重点项目(23SKGH251)和 2022 年重庆理工大学研究生教育高质量发展行动计划资助成果(gzlcx20223313 和 gzlcx20223314)。

参考文献

[1] Yin, X.T. and Tao, X.S. (2021) Prediction of Merchandise Sales on E-Commerce Platforms Based on Data Mining and Deep Learning. *Scientific Programming*, **6**, 1-9. https://doi.org/10.1155/2021/2179692

[2] Hu, W. and Zhang, X.C. (2020) Commodity Sales Forecast Based on ARIMA Model Residual Optimization. 2020 *IEEE 5th International Conference on Communication, Image and Signal Processing* (*CCISP*), Chengdu, 13-15 November 2020, 229-233. https://doi.org/10.1109/CCISP51026.2020.9273506

[3] Singh, B., Kumar, P., Sharma, N., *et al*. (2020) Sales Forecast for Amazon Sales with Time Series Modeling. 2020 *IEEE* 1*st International Conference on Power, Control and Computing Technologies* (*ICPC2T*), Raipur, 3-5 January 2020, 38-43. https://doi.org/10.1109/ICPC2T48082.2020.9071463

[4] 胡小芳. 基于电商数据的商品供需平衡研究[D]: [硕士学位论文]. 广州: 暨南大学, 2022.

[5] Ensafi, Y., Amin, S.H., Zhang, G.Q., *et al*. (2022) Time-Series Forecasting of Seasonal Items Sales Using Machine Learning—A Comparative Analysis. *International Journal of Information Management Data Insights*, **2**, Article ID: 100058. https://doi.org/10.1016/j.jjimei.2022.100058

[6] 黄国兴, 曹先怀, 钱晓飞. 一种基于随机森林的备件预测模型研究[J]. 运筹与管理, 2021, 30(10): 165-168.

[7] 荆浩, 刘垭, 唐金环. 基于多变量支持向量机的供应链需求预测分析[J]. 系统工程, 2018, 36(11): 121-126.

[8] 贺毅岳, 韩进博, 高妮. 基于 EEMD-SVR 的沪深 300 指数预测建模[J]. 统计与决策, 2020, 36(17): 152-156.

[9] 李杰, 王玉霞, 赵旭东. 电商企业商品销量的预测方法[J]. 统计与决策, 2018, 34(22): 176-179. https://doi.org/10.13546/j.cnki.tjyjc.2018.22.042

[10] 刘辰阳. J社区团购平台基于XGBOOST的快消品销量预测方法[D]: [硕士学位论文]. 大连: 大连理工大学, 2022.

[11] 吴霆辉. 基于遗传算法优化 LightGBM-XGBoost 模型的电力负荷预测[J]. 科学技术创新, 2023(3): 71-75.

[12] 谢勇, 项薇, 季孟忠, 彭俊, 黄益槐. 基于 Xgboost 和 LightGBM 算法预测住房月租金的应用分析[J]. 计算机应用与软件, 2019, 36(9): 151-155+191.

[13] 王晓晖, 张亮, 李俊清, 等. 基于遗传算法与随机森林的 XGBoost 改进方法研究[J]. 计算机科学, 2020, 47(S2): 454-458+463.

[14] 霍佳震, 徐骏, 陈铭洲. 基于 EEMD-Holt-Winters-GBDT 模型的零售商品销量多步预测[J]. 工业工程与管理, 2023: 1-14.

[15] Wang, D.-N., Li, L. and Zhao, D. (2022) Corporate Finance Risk Prediction Based on LightGBM. *Information Sciences*, **602**, 259-268. https://doi.org/10.1016/j.ins.2022.04.058

[16] Wang, Y.Y., Chen, J., Chen, X.Q., *et al*. (2021) Short-Term Load Forecasting for Industrial Customers Based on TCN-LightGBM. *IEEE Transactions on Power Systems*, **36**, 1984-1997. https://doi.org/10.1109/TPWRS.2020.3028133

[17] Chen, C.L. and Liu, Z.L. (2018) Broad Learning System: An Effective and Efficient Incremental Learning System without the Need for Deep Architecture. *IEEE Transactions on Neural Networks and Learning Systems*, **29**, 10-24. https://doi.org/10.1109/TNNLS.2017.2716952

[18] Su, L.Y., Xiong, L. and Yang, J.L. (2023) Multi-Attn BLS: Multi-Head Attention Mechanism with Broad Learning System for Chaotic Time Series Prediction. *Applied Soft Computing*, **132**, Article ID: 109831. https://doi.org/10.1016/j.asoc.2022.109831

[19] 褚菲, 苏嘉铭, 梁涛, 等. 基于 lasso 和 elastic net 的宽度学习系统网络结构稀疏方法[J]. 控制理论与应用, 2020, 37(12): 2543-2550.

[20] 杨光雨, 李晓航. 基于最大信息挖掘宽度学习系统短期电力负荷预测研究[J]. 电测与仪表, 2022, 59(3): 38-45.