钟佳玲, 赵芝泰, 何子安, 等. 基于 Blending 模型融合策略的产品订单需求预测[J]. 智能计算机与应用,2025,15(7):67-73. DOI:10.20169/j.issn.2095-2163.250710

基于 Blending 模型融合策略的产品订单需求预测

钟佳玲, 赵芝泰, 何子安, 陈国汉
(惠州学院 数学与统计学院, 广东 惠州 516000)

摘 要: 产品需求是企业供应链最重要的部分,稳定的产品需求预测需要更加的精准,本文提出基于 Blending 模型融合策略,实现了将随机森林(RF)、LightGBM、XGBoost 与 CatBoost 模型进行融合的机器学习算法。首先,结合产品订单数据自身具有多特征的特性,通过时间延迟等特征处理,分别建立随机森林、LightGBM、XGBoost 以及 CatBoost 机器学习模型进行预测,结果表明机器学习模型可用于产品需求预测,但单模型预测精度不高;其次,本文基于 Blending 模型融合策略,将建立的单模型进行融合,建立 RF-LGB-XGB-CB 融合模型。验证结果表明,采用 Blending 策略后的融合模型预测效果优于单模型,预测精度有所提升且泛化能力较好,可应用于未来产品需求预测。

关键词: 产品需求预测; 特征处理; Blending 模型融合; RF-LGB-XGB-CB 融合模型
中图分类号: F274;TP181 **文献标志码**: A **文章编号**: 2095-2163(2025)07-0067-07

Product order demand forecasting based on blending model fusion strategy

ZHONG Jialing, ZHAO Zhitai, HE Zian, CHEN Guohan
(School of Mathematics and Statistics, Huizhou University, Huizhou 516000, Guangdong, China)

Abstract: Product demand is the most important part of the enterprise supply chain, and stable product demand forecasting requires greater accuracy. This paper proposes a blending model fusion strategy that integrates machine learning algorithms such as Random Forest (RF), LightGBM, XGBoost, and CatBoost. First, by leveraging the multi-feature characteristics of product order data and processing features such as time delays, we establish machine learning models for Random Forest, LightGBM, XGBoost, and CatBoost for forecasting. The results indicate that machine learning models can be used for product demand forecasting, but the accuracy of single model predictions is not high. Secondly, based on the blending model fusion strategy, we fuse the established single models to create the RF-LGB-XGB-CB fusion model. Validation results show that the fused model using the blending strategy outperforms the single models, with improved prediction accuracy and better generalization ability, making it applicable for future product demand forecasting.

Key words: product demand forecasting; feature processing; Blending model fusion; RF-LGB-XGB-CB fusion model

0 引 言

需求预测是基于历史数据和未来的预判得出的有理论依据的结论,有利于公司管理层对未来的销售及运营计划、目标、资金预算做决策参考。需求预测有助于采购计划和生产计划的制定,减少受业务波动的影响。如果没有需求预测或者预测不准,公司内部很多关于销售、采购、财务预算等决策就都只能根据经验而来,会导致对市场预测不足,产生库存、资金的积压或不足等问题,增加企业库存成本。因此不断提高需求预测的精准度,是企业发展的必然要求。

目前,常见产品的需求预测方法有:基于统计模型预测方法。李成港等[1]利用自回归差分移动平均模型(AutoRegressive Integrated Moving Average Model,ARIMA)预测物流企业产品订单数据,该模型在短期预测上具有独特优势,预测结果精确度高,能为仓储优化提供辅助决策依据,解决企业仓储不合理等问题;牛凯等[2]基于 Prophet 模型预测电力物资需求,分析每类物资历史数据的规律,致力于完

作者简介: 钟佳玲(2000—),女,硕士研究生,主要研究方向:应用统计学; 赵芝泰(2002—),男,学士,主要研究方向:数据科学与大数据技术; 何子安(2002—),男,学士,主要研究方向:数学与应用数学。
通信作者: 陈国汉(1981—),男,博士,讲师,主要研究方向:统计学,计量经济学。Email:381588905@qq.com。

收稿日期: 2023-12-01

善电力物资的需求预测,预测精确度高、实用性好,科学地提高电力物资供应管理效益。基于机器学习模型预测方法。黄国兴等[3]将随机森林模型(Random Forest,RF)运用到舰船的备件预测,为舰船装备在海上任务期内备件配置问题提供参考价值;李婷婷等[4]使用轻量级梯度提升机模型(Light Gradient Boosting Machine,LightGBM)对离散型物料需求数据进行预测,效率和准确率更高,有利于提高离散制造企业的生产效率;李福等[5]使用极限梯度提升模型(eXtreme Gradient Boosting,XGBoost)结合天气和时间因素以及历史数据,实现对某区域每小时的共享单车用户借车需求量的有效预测。基于深度学习模型预测方法。李永锋[6]构建了基于混沌理论相空间重构算法的 BP(Back Propagation)神经网络预测模型,用于化工企业产品需求预测,平均预测精度大于 90%,优于其他类型的神经网络预测方法,对工厂及时调整生产具有一定的指导意义;郭继孚等[7]采用深度学习方法挖掘大量城市的出行分布规律,建立了通用的交通需求分布预测模型,既适用于现状需求分布估计,也适用于各类规划场景的需求分布预测,可节省调查和建模成本。基于组合模型预测方法。靖可等[8]为有效预测智能制造模式下的不确定性需求,提出自回归移动平均模型 ARIMA 和改进 BP 神经网络的组合模型,组合模型的预测精度较 ARIMA 模型有显著提高;吴庚奇等[9]提出了产品数据空间和一维卷积神经网络(One-Dimensional Convolutional Neural Networks,1D-CNN)-长短期记忆(Long Short-Term Memory,LSTM)神经网络的组合模型,用于某电气设备制造企业生产销售的环网柜产品需求预测,该模型的预测效果优于神经网络模型和单一的 LSTM 模型。深度学习组合模型更具优越性,预测效果好。

与组合模型预测方法不同,机器学习中利用集成学习思想进行的模型融合预测方法更加别具一格。赵茜茜等[10]采用集成学习的思想,将两个基础预测模型集成成一个较强的分类预测模型,用于预测一个新案例是否对指纹锁拉链产品有需求。然而,目前还较少人利用集成学习算法研究制造业的产品需求预测。

数据方面,制造企业产品订单数据往往包含着产品销售区域编码、类别以及销售方式等多特征,且不同产品之间需求量规模不同,上市或下架时间也不尽相同,产品数量有时也达到千乃至万级以上。

本文提出基于 Blending 模型融合策略的产品需求预测方法。利用机器学习模型具有较好泛化能力和提取特征能力,解决产品规模与冷启动问题;利用集成学习思想,将多个弱模型融合为较强模型,解决单一预测模型提取特征能力有限问题,提高预测精度;分别训练随机森林(RF)、LightGBM、XGBoost 和 CatBoost(Categorical Boosting)单模型,后将每个弱的单模型基于 Blending 模型融合策略进行融合,构建出拟合度高的 RF-LGB-XGB-CB 融合预测模型,提高产品需求预测精度;最后,基于产品订单需求的相关数据进行训练预测,并将构建出的各个模型的预测结果、精度进行比较,分析本文构建模型的有效性与可行性,并利用训练好的模型预测给定产品未来 3 个月的月需求量。

1 相关原理

1.1 机器学习模型

随机森林模型(RF)是基于 Bagging 和决策树的有监督学习模型,利用多棵决策树对样本进行训练并预测的一种学习器[11]。作为决策树的集成,随机森林很大程度上改善了决策树容易过拟合问题,不易受噪声和异常值干扰,对高维数据也能达到较好的拟合效果。

轻量级梯度提升机模型(LightGBM)是一种高效梯度提升决策树模型,是对梯度提升决策树(Gradient Boosting Decision Tree,GBDT)模型的改进,提高了准确率与效率,在不降低预测精度情况下,减少训练时间,大大降低了内存的占用[12]。

极限梯度提升模型(XGBoost)也是基于 GBDT 的一种模型,能够准确捕捉各种预测变量的非线性特征,在算法本身优化,除自身的损失,添加了正则化部分,有效防止过拟合,具有较强的泛化能力;在损失函数的误差部分做二阶泰勒展开,预测更加准确;训练引入并行运算,提升训练运行效率;数据上无需进行预处理,通过统计所有缺失值在当前节点分布规律来处理缺失值[13]。

CatBoost(Categorical Boosting)模型同样在 GBDT 模型的基础上进行改进,在模型精度方面,优于同属 GBDT 模型框架下的 XGBoost 模型和 LightGBM 模型。该模型主要的创新点在于采用排序提升的方法替换传统算法中梯度估计算法,解决了梯度偏差和预测偏移问题,使用对称决策树作为基模型,在提高模型分类正确能力的同时也兼顾泛化能力,有效防止模型过拟合[14]。

1.2　Blending 模型融合策略

集成学习是对多个个体学习器进行训练，通过一定的结合策略集成，最终形成一个强学习器。集成学习核心策略就是通过模型的集成，减少机器学习中的偏差和方差。集成学习方法有 Bagging、Boosting、Stacking 以及 Blending 方法，本文使用 Blending 集成学习方法，可以避免数据泄露问题，提高整体模型性能。

Blending 模型融合的策略：将训练数据划分为训练集、验证集、测试集。划分之后的训练集训练基模型（第一层模型），验证集经基模型预测后作为元模型（第二层模型）的训练集。测试数据同样经过基模型预测，形成新的测试数据。元模型对新的测试数据进行预测，得到最终结果，有效避免数据泄露问题。Blending 融合模型预测过程如图 1 所示。

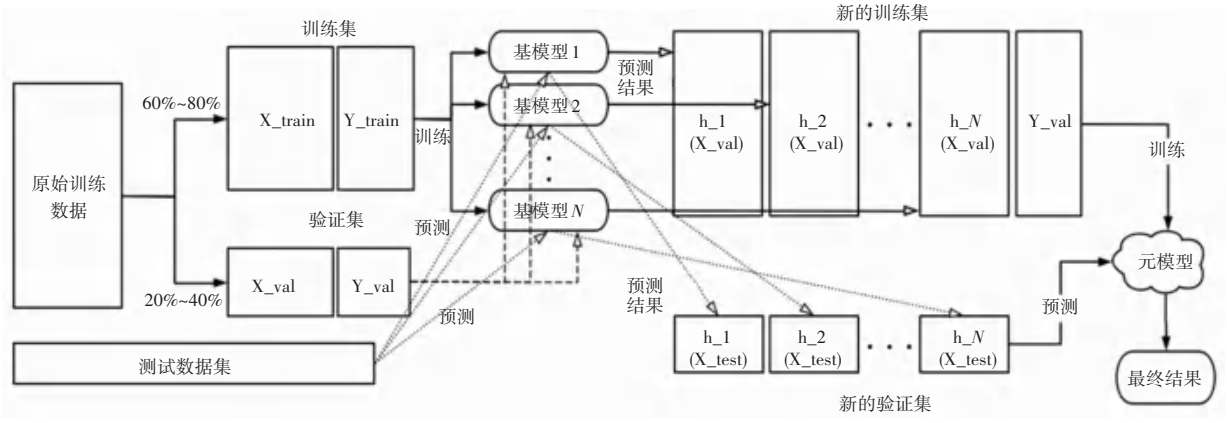

图 1　Blending 融合模型预测过程图

Fig. 1　Prediction process diagram of the blending model

实现融合模型具体流程：

（1）将原始训练数据划分为训练集和验证集；

（2）使用训练集训练 N 个不同的基模型；

（3）使用 N 个基模型对验证集进行预测，结果作为元模型的训练数据；

（4）利用新的训练数据，训练元模型；

（5）使用 N 个基模型，对测试数据集进行预测，结果作为元模型的测试数据；

（6）使用元模型对新的测试数据进行预测，得到最终结果。

2　基于 Blending 融合模型预测

2.1　数据预处理

以国内某大型制造企业 2015 年 9 月 1 日至 2018 年 12 月 20 日面向经销商的出货数据为训练数据，对企业给定的产品，预测未来 3 月（即 2019 年 1 月、2 月、3 月）的月需求量，即测试数据。对产品需求进行分析，训练集总共 597 694 条数据。产品订单部分数据集见表 1，待预测部分数据集见表 2。

表 1　产品订单部分数据集

Table 1　Product orders partial data set

订单日期	销售区域编码	产品编码	产品大类编码	产品细类编码	销售渠道名称	产品价格	订单需求量
2015/9/1	104	22069	307	403	offline	1 114	19
2015/9/1	104	20028	301	405	offline	1 012	12
2015/9/2	104	21183	307	403	online	428	109
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
2018/12/20	102	20215	302	408	offline	2 013	106
2018/12/20	102	20195	302	408	offline	2 120	187
2018/12/20	102	20321	302	408	offline	1 244	205

表 2 待预测部分数据集

Table 2 Some datasets to be forecasted

销售区域编码	产品编码	产品大类编码	产品细类编码
101	20002	303	406
101	20003	301	405
101	20006	307	403
⋮	⋮	⋮	⋮
105	22075	307	403
105	22083	303	401
105	22084	302	408

分析建模前需要对数据进行预处理,主要处理方式包括异常值、重复值、缺失值处理以及数据类型变换。

2.1.1 缺失值、重复值、异常值处理

筛选得出该数据集中没有缺失值,但 2018 年 12 月只有 20 天数据,后续 11 天需要进行预测填充;筛选剔除掉 312 条重复数据,利用 3δ 准则进行异常值剔除,最终有 582 796 条订单数据。

2.1.2 数据类型变换

将字符型的产品订单日期列转换为日期时间类型数据;将部分列的数据类型修改为占用内存较低的数据类型,如:将 int 64 类型转换为 int 32 类型。

将产品订单数据中的类别特征进行类别特征编码,产品销售方式进行数值化处理,完成数据预处理。

2.2 特征处理

模型建立前,需对数据进行特征处理,具体包括特征筛选、特征处理、特征融合。时间序列数据还需要进行时间特征分解、延迟操作,即利用历史信息预测未来。

处理数据中在最后半年没有需求量的产品,可将其视为“已下架”产品,处理预测数据中没有出现在训练数据中的“新品”。“已下架”的产品不用于训练,考虑模型的泛化能力,通过人工检验,决定是否将“新品”需求量填充为 0。对月数进行统计并添加至新列,共 40 个月,依次将其编码为 0~39,并与原数据进行合并。

特征筛选。去除掉无用特征列与“已下架”产品。由于销售方式列对需求量的影响不大,而价格列存在波动,较难确定,且待预测数据中不存在,因此去除掉销售方式列与价格列,保留最后 6 个月有需求量的产品,提取其月数编码与产品编号。

特征融合。对产品按月聚合需求量,并将筛选出来的产品与对应的需求量进行数据表连接,对待预测数据进行月份编码即 40、41、42,并与训练数据进行连接;添加具体的年份和月份列;添加更多需求量特征,包括产品月平均需求量、大类产品月平均需求量、小类产品月平均需求量、区域产品月平均需求量。

特征处理。将类别特征进行编码,利用字典、map 函数对大类、小类、区域进行编码。

延迟操作。利用历史信息对未来进行预测,采用延迟操作产生历史信息,如将第 0~33 个月的销量作为第 1~34 个月的历史特征,即延迟一个月,给予当月产品的上个月需求量历史信息特征。

本文利用前 1 个月、前 2 个月、前 3 个月、前半年、前一年的产品历史信息需求量特征,来预测未来月份的产品需求量。由于使用了 12 个月作为延迟特征,导致大量的数据为空值,因此将最开始 11 个月的原始特征删除,对于其他空值则把其填充为 0 后进行数据类型转换,降低模型训练时的压力。

对数据进行划分。由于时间数据连续,因而不采用随机分割,而是将数据根据月份进行分割,月数编码为 0~38 为训练集,编码 39 为验证集,40、41 和 42 为测试集。

2.3 构建模型

2.3.1 随机森林模型

建立随机森林模型后对参数进行调优,对决策树的数量、树的最大深度、拆分内部节点所需的最少样本数、叶节点处需要的最小样本数、寻找最佳分割时考虑的特征数量等参数进行随机搜索(RandomizedSearchCV),进一步进行网格搜索(GridSearchCV),得出随机森林最优参数见表 3。

表 3 随机森林最优参数

Table 3 Random forest optimal parameters

参数名	参数值
树的数量	100
树的最大深度	2
内部节点所需最少样本数	2
叶节点处需要最小样本数	8
特征数量	"sqrt"

2.3.2 LightGBM 模型

将训练集与验证集转换为 LightGBM 能识别的数据格式,建立 LightGBM 模型,设置参数字典,放入数据进行训练,利用模型内置 cv 函数对参数进行交叉验证调优。调优思路 1:对叶子个数、树深度这两个参数进行组合遍历,提高准确率;调优思路 2:对

特征随机采样比例、不进行重采样的情况下随机选择数据比例，给这两个参数设置一定范围，进行组合遍历，降低过拟合。最终确定 LightGBM 模型最优参数见表 4。

表 4　LightGBM 模型最优参数

Table 4　Optimal parameters of the LightGBM model

参数名	参数值
基学习器	GBDT(默认)
学习任务及相应学习目标	“regression”:L2 正则项回归模型
学习率	0.01
叶子个数	200
停止训练次数	50
样本比例	0.90
特征随机采样比例	0.30
决策树数量	1 000

2.3.3　XGBoost 模型

训练 XGBoost 模型，初始模型训练完成后，定义参数范围，利用网格搜索调整参数，继续训练模型，得出 XGBoost 模型最优时参数见表 5。

表 5　XGBoost 模型最优参数

Table 5　Optimal parameters of the XGBoost model

参数名	参数值
决策树数量	400
树的最大深度	10
叶子上最小样本数	1
列采样比例	0.7
随机采样比例	0.7
学习率	0.1

2.3.4　CatBoost 模型

用默认参数训练 CatBoost 模型，遍历参数范围进行调参，继续训练，得出 CatBoost 模型最优参数见表 6。

表 6　CatBoost 模型最优参数

Table 6　Optimal parameters of the CatBoost model

参数名	参数值
迭代次数	700
学习率	0.02
树的深度	8
采样权重	0.20
过拟合检测器的类型	‘Iter’
迭代后以最佳度量值继续训练的迭代次数	100

2.3.5　模型对比

本文选取可决系数（R^2）、均方根误差（*RMSE*）和平均绝对误差（*MAPE*）来分析模型拟合情况与预测值偏差情况，衡量模型的泛化能力以及预测效果与精度。*RMSE* 越小、*MAPE* 越小、R^2 越接近 1，表示模型精度越高，预测效果越好[15]，计算如下：

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{1}$$

$$RSME = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{2}$$

$$MAPE = \sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \times \frac{100}{n} \tag{3}$$

其中，y_i 为真实值；$\hat{y}_i$ 为预测值；$\bar{y}$ 表示平均值；n 为样本数。

不同单模型预测指标见表 7，可见训练集的 R^2 都在 0.5 左右，验证集都在 0.4 左右，表明机器学习模型可以应用于产品需求预测，提取特征能力较好，但单模型预测精度不高，需要进一步进行模型融合。

表 7　不同模型预测指标结果

Table 7　Different Single-Model Prediction Indicator results

模型	*RMSE*		*MAPE*		R^2	
	训练集	验证集	训练集	验证集	训练集	验证集
RF	1 171.70	729.20	1 459.71	1 917.16	0.53	0.36
LightGBM	1 139.40	662.42	1 739.73	2 328.48	0.55	0.47
XGBoost	1 118.57	569.03	316.35	458.43	0.57	0.61
CatBoost	1 227.84	651.90	1 772.36	2 370.75	0.48	0.49

2.3.6　基于 Blending 模型融合策略

基于 Blending 模型融合策略，选取 CatBoost 作为元学习器，将上述模型验证集的预测结果作为新训练集，测试集的预测结果作为新测试集，输入到模型中训练，训练测试，进行模型评价验证，并利用融合模型进行应用预测，模型融合过程如图 2 所示。

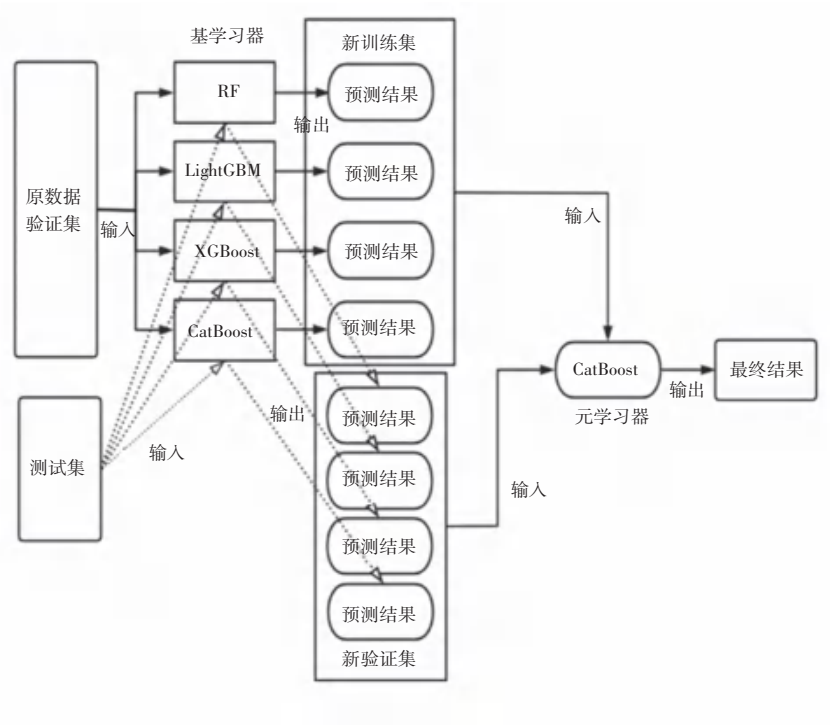

图 2 模型融合过程图

Fig. 2 Diagram of the model fusion process

数据训练后,进一步通过测试,得出该模型效果 *RMSE* 为 294.98, *MAPE* 为 335.41, R^2 为 0.894,可见融合后的模型,模型精度都有所提高。利用 RF-LGB-XGB-CB 融合预测测试集部分预测结果见表 8。

表 8 融合模型预测测试集部分预测结果

Table 8 Partial prediction results of the fusion model prediction test set

销售区域编码	产品编码	2019 年 1 月预测需求量	2019 年 2 月预测需求量	2019 年 3 月预测需求量
101	20002	82.55	67.37	55.06
101	20003	181.40	152.09	140.17
101	20006	71.06	50.05	66.04
101	20011	180.61	261.25	372.12
101	20014	78.53	83.34	63.38
⋮	⋮	⋮	⋮	⋮
105	22066	1 273.51	828.63	549.08
105	22072	380.45	225.94	200.52
105	22075	171.51	202.76	197.34
105	22083	693.87	435.91	304.13
105	22084	20.07	52.19	65.98

观察预测结果,发现某些产品月需求量存在大于 0 小于 1 和小于 0 的情况,与特征处理过程中存在“新品”或“已下架”产品猜测基本吻合,结合原数据将其需求量填充为 0。对预测结果进行人工检验,验证得出本文提出基于 Blending 融合策略的模型可以预测出“新品”的需求量,泛化能力较好,预测精度较高。

3 结束语

针对制造企业产品订单数据,首先对数据进行预处理,通过特征处理进一步完善前期数据处理,后分别建立 4 个机器学习模型:随机森林(RF)、LightGBM、XGBoost 以及 CatBoost 模型,结果表明机器学习模型可以较好提取特征,并进行产品需求预测,但单模型提取特征能力有限,预测精度不高;进而引入集成学习思想,本文提出利用基于 Blending 模型策略,对 4 个单模型进行融合,建立 RF-LGB-XGB-CB 融合预测模型。预测结果表明,融合后的模型有效,相较于单模型更优,预测精度进一步提升,具有较好的泛化能力,可以应用于产品未来需求预测,为公司管理层对未来的销售及运营计划、目标,资金预算做决策参考。

参考文献

[1] 李成港,李雨萌,黄芊芊. 基于 ARIMA 模型的产品需求预测研究[J]. 物流工程与管理,2018,40(4):77-78.

[2] 牛凯,洪芳华,费冬,等. 基于 Prophet 算法的电力物资需求预测方法研究[J]. 科学技术创新,2020(33):163-164.

[3] 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.
[4] 李婷婷,黄欣迪,曹萌萌,等. 基于LightGBM 模型的离散制造业产品物料需求智能预测[J]. 智能计算机与应用,2023,13(9):59-66.
[5] 李福,徐良杰,朱然博,等. 基于 XGBoost 算法的共享单车借车需求量预测[J]. 武汉理工大学学报(交通科学与工程版),2021,45(5):880-884.
[6] 李永锋. 改进的混沌理论和 BP 神经网络化工产品需求预测模型设计[J]. 粘接,2022,49(8):177-181.
[7] 郭继孚,李寻,白盛光,等. 基于深度学习的城市交通需求场景库[J]. 城市交通,2023,21(1):74-85.
[8] 靖可,唐亮,赵礼强,等. 智能制造模式下基于改进 BP-ARIMA 组合模型产品需求预测方法[J]. 数学的实践与认识,2017,47(4):15-24.
[9] 吴庚奇,牛东晓,耿世平,等. 多价值链视角下基于深度学习算法的制造企业产品需求预测[J]. 科学技术与工程,2021,21(31):13413-13420.
[10]赵茜茜,张洋溢,聂焱. 一种基于集成学习的指纹锁拉链产品需求预测方法[J]. 数字技术与应用,2019,37(7):50-51.
[11]刘智玉,陈南梁,汪军. 基于随机森林算法的小样本纱线质量预测[J]. 东华大学学报(自然科学版),2023,49(6):80-86.
[12]李华洋,曹志鹏,吴小龙,等. 基于 LightGBM 算法的地层破裂压力预测方法及应用[J]. 中国测试,2024,50(4):134-143.
[13]刘敏,周健,胡月明,等. 基于 XGBoost 算法的可恢复耕地宜耕性评价:以湘阴县为例[J]. 农业资源与环境学报,2024,41(1):49-60.
[14]梁宏涛,孔翎超,刘国柱,等. 融合数字孪生的风电机组故障检测 ASL-CatBoost 方法[J]. 系统仿真学报,2024,36(4):873-887.
[15]蒋其容,魏勇,高先松,等. LSTM-LightGBM 组合模型的短期电力负荷预测[J]. 中国设备工程,2023,552(8):78-81.