王记红, 邱奕桐, 黎永贵, 等. 基于 LGBM 方法的商品销量预测模型研究[J]. 智能计算机与应用,2025,15(4):191-197.
DOI:10.20169/j.issn.2095-2163.250427

基于 LGBM 方法的商品销量预测模型研究

王记红, 邱奕桐, 黎永贵, 林伊泳, 彭俊丰, 徐 俊, 周如旗
(广东第二师范学院 计算机学院, 广州 510000)

摘 要: 为了实现更好的销量预测模型,本文以某大型企业的出货数据为对象,通过深入研究特征工程和 LGBM 集成学习方法,实现特征标准化,建立基于 LGBM 方法的商品销量预测模型,并将 LGBM 模型预测结果与 RF、LSTM 等模型进行统一对比分析。通过实验结果可以看出,本文所建立的基于 LGBM 商品销量预测模型相较其它预测模型测试 *RMSE* 更小,表明预测效果更好,并且又对单独区域建模展开了探索研究,从而进一步减小了 *RMSE*。综合来看,基于 LGBM 方法进行商品销量预测是一个比较有效的方法,可广泛应用在商品销售等领域。

关键词: LGBM; 特征编码; 销量预测; *RMSE*
中图分类号: F713.3; TP181 **文献标志码**: A **文章编号**: 2095-2163(2025)04-0191-07

Research on commodity sales forecast model based on LGBM method

WANG Jihong, QIU Yitong, LI Yonggui, LIN Yiyong, PENG Junfeng, XU Jun, ZHOU Ruqi
(School of Computer Science, Guangdong University of Education, Guangzhou 510000, China)

Abstract: To achieve a superior sales forecasting model, shipment data from a large enterprise are selected to serves as the subject of investigation. By delving deeply into feature engineering and the LGBM ensemble learning method, feature normalization is accomplished, and a sales forecasting model based on the LGBM method is established. The prediction results of the LGBM model are compared with those of the RF, LSTM, and other models. Experimental results demonstrate that the LGBM-based sales forecasting model had a smaller *RMSE* compared to other models, indicating superior prediction performance. Additionally, *RMSE* is further decreased upon modeling individual regions. Overall, sales forecasting using the LGBM method proves effectiveness and holds broad applicability in the field of merchandise sales.

Key words: LGBM; feature encoding; sales forecasting; *RMSE*

0 引 言

近年来,电商经济进入精细化经营阶段,跨境电商也在蓬勃发展,使得预测商品销量成为企业经营的关键环节。分析商品销量数据能帮助企业洞察客户需求,制定有效的营销策略,优化生产计划和库存管理,降低成本,提高效率,从而获取更大的市场份额和收益。基于历史订单数据进行未来需求预测对企业经营至关重要,能帮助企业合理规划生产、销售、供应链管理,降低资金成本和滞销风险,实现生产资源的高效配置和利用。

当前主流商品销量预测模型通常包含时间序列分析、回归分析和机器学习等方法,取得了一定的成效。但随着数据规模和特征维度的扩大,传统方法面临着模型复杂度不足、特征处理及筛选困难等问题,需要继续研究能适应大规模、高维度学习的方法,能够自动提取特征,提高模型预测精度和泛化能力。

LGBM(Light Gradient Boosting Machine)[1]是一种基于决策树算法的梯度提升框架,以其高准确率、高性能而著称。该方法的特点包括支持类别特征、缺失值处理、高效处理大规模数据、并行学习和 GPU 加速等。LGBM 采用基于直方图的学习方法和带深度限制的 Leaf-wise 策略,使其在保持较高准确

基金项目: 广东省教育厅 2022 高等教育专项(2022GXJK287); 广东省普通高校重点领域(乡村振兴)(2020ZDZX1023); 2021 年度广东省普通高校科研平台和科研项目(2021ZDZX3016)。

作者简介: 王记红(1978—),男,博士,讲师,主要研究方向:大数据,深度学习,推荐系统。Email:redblue04@163.com。

收稿日期: 2023-10-07

率的同时,具有较低的内存消耗和更快的训练速度。LGBM 适用于各种场景,尤其是在大规模数据集上表现优异。现已广泛应用于各类机器学习竞赛和实际应用问题中,如点击率预测[2]、欺诈检测[3]、用户行为分析[4]等。由于其高效和灵活性,LGBM 成为了数据科学家和研究人员在各种复杂任务中的首选方法,取得了良好的应用效果。

本文采用 LGBM 方法对商品销量进行预测研究,基于商品销量数据的特点,设计特征工程,对模型参数进行优化,并综合对比随机森林(RF)[5-6]、长短记忆网络(LSTM)[7-8]等主流模型,验证 LGBM 模型的相关性能指标,评估其可行性。

1 数据集描述及处理

1.1 数据集描述

本文采用数据来自国内某大型企业在 2015 年 9 月 1 日至 2018 年 12 月 20 日的出货数据(见表 1)[9]。数据规格为 597 694 rows×8 columns,包括:订单日期(order_date)、销售区域编码(sales_region_code)、产品编码(item_code)、产品大类编码(first_cate_code)、产品细类编码(second_cate_code)、销售渠道名称(sales_chan_name)、产品价格(item_price)和订单需求量(ord_qty)。

表 1 出货数据

Table 1 Shipping data

序号	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
0	2015/9/1	104	22069	307	403	offline	1114	19
1	2015/9/1	104	20028	301	405	offline	1012	12
2	2015/9/2	104	21183	307	403	online	428	109
3	2015/9/2	104	20448	308	404	online	962	3
…	…	…	…	…	…	…	…	…
597691	2018/12/20	102	20215	302	408	offline	2013	106
597692	2018/12/20	102	20195	302	408	offline	2120	187
597693	2018/12/20	102	20321	302	408	offline	1244	205

在样本数据中,“订单日期”代表了每个订单需求量的日期,每个“产品大类编码”可以对应多个“产品细类编码”,“销售渠道名称”分为 2 个类别,即“online”(线上)和“offline”(线下)。样本时间跨度较大,属性包括日期型、离散型、连续型等多种类型。由于数据质量决定模型上限,需要对样本进行必要的预处理,从而保证数据质量,提高模型拟合效能。

1.2 数据预处理

销量预测本质上是基于历史数据拟合趋势,而历史数据通常由于业务流程的复杂性和商品的多样性导致实际商品订单数据可能存在偏差或遗漏,这些数据会对模型预测质量带来一定的干扰。因此,在进行销量预测建模前,首先要对所选样本进行重复值、缺失值、异常值等预处理,从而得到尽可能干净的数据。数据基本预处理流程如图 1 所示。

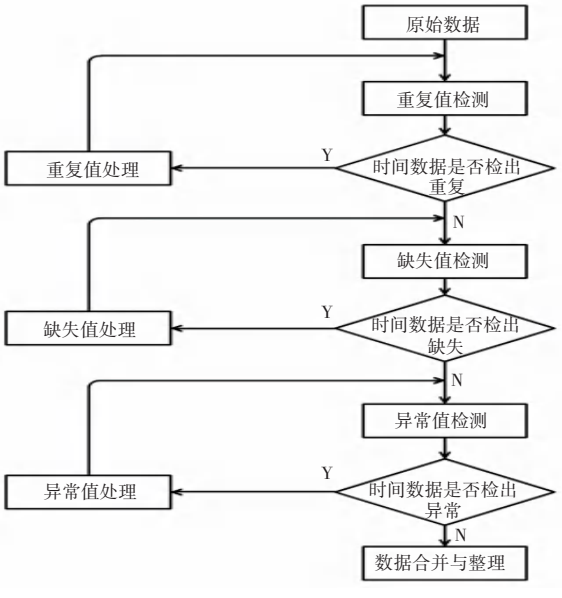

图 1 数据预处理流程

Fig. 1 Data preprocessing process

数据样本经过预处理后，可以进行数据编码和特征工程工作。根据对促销日的特性分析，考虑不同日期属性对商品销量的影响，将日期类型分为 2 类，分别为：促销日、非促销日；根据对销售渠道的特性分析，将销售方式类型分为 2 类，分别为：线上销售、线下销售；以上均采用 0、1 编码。根据价格区间高低，考虑不同价格区间对商品销量的影响，将商品价格类型分为 4 类，分别为低价、中价、较高价、高价。通过对产品价格分布分析，采用四分位策略，以其小于价格的 25%为低价格，25%～50%为中价格，50%～75%为较高价，75%～100%为高价格。根据日期序列特征，按照日期归属当月几号，划分为月初、月中、月末，依次用 0、1、2 进行编码。本数据集提供的是天粒度的统计数据，可以将日期属性通过 Python 内置日期函数进行解析，提取年、季、月、周、日信息。经过上述编码后的样例数据见表 2。

表 2　特征编码

Table 2　Feature encoding

Year	Month	day	week	grade	旬	是否节假日	销售渠道
2015	9	1	36	0	0	0	0
2015	9	1	36	0	0	0	0
2015	9	2	36	1	0	0	0
2015	9	2	36	1	0	0	1
2015	9	2	36	2	0	0	1
…	…	…	…	…	…	…	…

2　基于 LGBM 的商品销量预测模型

2.1　LGBM 算法描述

GBDT（Gradient Boosting Decision Tree）是一种在机器学习领域广受欢迎的模型[10-12]。通过迭代弱分类器（决策树）来不断优化模型。其主要优点在于强大的训练效果和出色的抗过拟合能力。

LGBM（Light Gradient Boosting Machine）是基于 GBDT 算法的高效实现框架。与其他 GBDT 实现相比，LGBM 具有以下显著优势：

（1）更快的训练速度。LGBM 采用了基于 Histogram 的决策树算法，大大加速了训练过程。

（2）高效的数据采样。LGBM 引入了单边梯度采样（GOSS）技术，只考虑高梯度的数据，从而在计算信息增益时减少了计算量。

（3）更低的内存消耗。通过互斥特征捆绑（EFB）技术，LGBM 能够将多个互斥特征捆绑为一个特征，实现降维。

（4）优化的树生长策略。与传统的按层生长策略不同，LGBM 采用了带深度限制的 Leaf-wise 生长策略，更加高效且准确。

（5）并行处理及 Cache 优化。LGBM 支持高效的并行训练，特别适合处理大数据，LGBM 进一步优化了 Cache 命中率，提高了数据处理速度。

LGBM Histogram 算法的代码描述具体如下。

算法 1　Histogram-based Algorithm

输入 I: training data, d: max depth

输入 m: feature dimension

nodeSet←{0}　▷ tree nodes in current level

rowSet←{{0,1,2,…}}　▷data indices in tree nodes

for i = 1 to d do

　for *node* in *nodeSet* do

　　usedRows←*rowSet*[*node*]

　　for k = 1 to m do

　　H←new Histogram( )

　　▷Build histogram

　　for j in *usedRows* do

　　bin←$I.f[k][j]$. bin

　　H[bin]. y←H[bin]. $y + I.y[j]$

　　H[bin]. n←H[bin]. n + 1

　　Find the best split on histogram H.

　　…

Update *rowSet* and *nodeSet* according to the best split points.

…

LGBM Histogram 算法的基本思想是先对特征值进行装箱处理，实现特征值的离散化，遍历数据时，根据离散化后的值作为索引在直方图中累积统计量。当遍历一次数据后，直方图累积了需要的统计量，然后根据直方图的离散值，遍历寻找最优的分割点。使用离散化后的特征值，可以非常快速地计算出每个区间的梯度和二阶导数的和，这 2 个和可用于快速评估分割一个区间的增益。当一个数据点的特征值落入一个区间时，可以直接更新该区间的统计量，而不需对所有数据点进行遍历。由于特征值被转换为区间索引，因此在存储时可以使用较小的数据类型，极大减少内存的使用。LGBM 直方图方法可以支持数据并行和特征并行，从而大大加速计算。

在 GBDT 中，梯度可以作为衡量样本训练误差的指标。LGBM 引入了 Gradient-based One-Side Sampling（GOSS）。算法流程如下。

算法 2 Gradient-based One-Side Sampling

输入 I: training data, d: iterations

输入 a: sampling ratio of large gradient data

输入 b: sampling ratio of small gradient data

输入 *loss*: loss function, L: weak learner

Models←{ }, fact← $\frac{1-a}{b}$

top N←$a \times len(I)$, $randN$←$b \times len(I)$

for i = 1 to d do

preds←*models. predict*(I)

g←*loss*(I,*preds*), w←{1,1,…}

sorted←*GetSortedIndices* (abs (g))

topSet sorted[1:*topN*]

randSet←*RandomPick*(*sorted*[*topN*: *len*(I)], *randN*)

usedSet←*topSet* + *randSet*

w[*randSet*] ×= fact▷Assign weigh fact to the small gradient data

newModel←L(I[*usedSet*], − g[*usedSet*], w[*usedSet*])

models. append(*newModel*)

算法的核心思想是保留那些梯度较大的样本,因为这些更可能是未被充分训练的样本,而对梯度较小的样本进行随机抽样。GOSS 的步骤如下:

(1)根据梯度的绝对值对所有样本进行降序排序。

(2)选择前 $a \times 100\%$ 的样本,记为集合 A。

(3)从剩余的 $(1-a) \times 100\%$ 的样本中随机选择 $b \times 100\%$,记为集合 B。

(4)在计算信息增益时,将集合 B 中的样本梯度放大 $(1-a)/b$ 倍。

通过 GOSS,LGBM 能够更加关注那些未被充分训练的样本,同时只会轻微地改变原始数据的分布,从而在减少计算量的同时,仍然保持了模型的准确性。

2.2 建立商品销量预测模型

2.2.1 模型参数

LGBM 模型参数的选择多数情况下依赖于数据的性质、任务类型和特定的业务需求。通常使用默认参数作为起点,设置 *learning_rate* 为一个较小的值,以确保模型可以充分学习。LGBM 可以自动处理类别特征,但确保能被正确标记为类别类型,可以尝试创建交互特征、衍生特征等,一般对模型有帮助。若模型过拟合,考虑增加 *lambda_l*1(*L*1 正则化)和 *lambda_l*2(*L*2 正则化)。考虑树的复杂性,需要设置 *num_leaves*(每棵树的最大叶子数)、*max_depth*(限制树的深度可以帮助防止过拟合)。针对样本和特征的子采样,需要设置 *bagging_fraction*、*bagging_freq*、*feature_fraction* 的取值。可以使用小的 *learning_rate* 与大的 *n_estimators* 组合,或者使用大的 *learning_rate* 与小的 *n_estimators* 组合。使用早停(*early stopping*)来自动选择 *n_estimators* 的值。针对本论文数据集,模型主要参数设置见表 3。

表 3 LGBM 主要参数

Table 3 Main parameters of LGBM

参数	值	含义
boosting_type	gbdt	梯度提升决策树
objective	regression	学习任务目标函数
metric	mse	评估指标
max_depth	15	树的最大深度
min_child_samples	5	分裂一个内部节点所需的最小样本数
subsample	0.8	样本采样比例
colsample_bytree	0.8	每棵树使用的特征的比例
reg_alpha	0.2	*L*1 正则化系数
reg_lambda	0.2	*L*2 正则化系数
n_estimators	200	树的数量

2.2.2 模型比较

(1)随机森林(Random Forest,RF)。RF 是一种广泛应用的机器学习方法,属于集成学习方法范畴,通过构建多个决策树并进行投票或平均来进行预测。RF 具有较好的准确性、能处理大量特征且不易过拟合等优点,通常作为基线模型进行比较。RF 的重要参数包括:*n_estimators*(决定森林中树的数量,通常越多越好,但会增加计算量),*max_depth*(控制树的最大深度,以防止过拟合),*min_samples_split*(决定节点分裂的最小样本数),*min_samples_leaf*(控制叶子节点最少的样本数),*max_features*(限制每次分裂考虑的特征数量,通常设置为总特征数的平方根),*bootstrap*(是否进行样本的有放回抽样)。模型主要参数设置见表 4。

表 4 模型训练参数

Table 4 Model training parameters

参数	值	含义
n_estimators	80	森林中树的数量
max_features	*sqrt*	分裂最大特征数
max_depth	60	树的最大深度
min_samples_leaf	15	一个叶节点所需的最小样本数
min_samples_split	5	分裂一个内部节点所需的最小样本数

(2)深层 LSTM(D-LSTM)。长短时记忆网络

(LSTM)是一种特殊的循环神经网络(RNN)，适合于处理和预测时间序列中间隔和延迟非常长的重要事件。LSTM 通过引入门结构和细胞状态来解决传统 RNN 的长期依赖问题，使其能够学习长序列数据的依赖关系。LSTM 的重要参数包括：*hidden_size*(隐藏层的大小，决定模型的容量)，*num_layers*(网络的层数，增加层数可以增加模型的复杂性)，*dropout*(用于减少过拟合)，*learning_rate*(学习率，控制权重更新的幅度)，*sequence_length*(输入序列的长度)，以及 *batch_size*(每个批次的样本数量)等。本论文研究数据集属于长时间序列预测范畴，D-LSTM 可能是一个合适的选择，该模型包含多个 LSTM 层、Dropout 层、Relu 层、Batch_Norma 层和 Dense 层。该模型具有处理复杂的序列数据、防止过拟合、适应性强特点。模型网络结构及参数见表 5。总共包含 6 个 LSTM 层，每层单元数分别为 64、128、256、256、128 和 64，采用了 LeakyReLU、Dropout、BatchNormalization 等优化技术，有助于学习复杂模式，防止过拟合，并加速训练，最后接 2 个 Dense 层，输出预测值。

表 5　D-LSTM 网络结构

Table 5　D-LSTM network structure

Layer (type)	Output Shape	*Param*
lstm (LSTM)	(1, 30, 64)	20 480
leaky_re_lu (LeakyReLU)	(1, 30, 64)	0
dropout (Dropout)	(1, 30, 64)	0
batch_normalization (BatchNo)	(1, 30, 64)	256
lstm_1 (LSTM)	(1, 30, 128)	98 816
leaky_re_lu_1 (LeakyReLU)	(1, 30, 128)	0
dropout_1 (Dropout)	(1, 30, 128)	0
batch_normalization_1(BatchNo)	(1, 30, 128)	512
lstm_2 (LSTM)	(1, 30, 256)	394 240
leaky_re_lu_2 (LeakyReLU)	(1, 30, 256)	0
dropout_2 (Dropout)	(1, 30, 256)	0
batch_normalization_2(BatchNo)	(1, 30, 256)	1 024
lstm_3 (LSTM)	(1, 30, 256)	525 312
leaky_re_lu_3 (LeakyReLU)	(1, 30, 256)	0
dropout_3 (Dropout)	(1, 30, 256)	0
batch_normalization_3(BatchNo)	(1, 30, 256)	1 024
lstm_4 (LSTM)	(1, 30, 128)	197 120
leaky_re_lu_4 (LeakyReLU)	(1, 30, 128)	0
dropout_4 (Dropout)	(1, 30, 128)	0
batch_normalization_4(BatchNo)	(1, 30, 128)	512
lstm_5 (LSTM)	(1, 64)	49 408
leaky_re_lu_5 (LeakyReLU)	(1, 64)	0
dropout_5 (Dropout)	(1, 64)	0
batch_normalization_5(BatchNo)	(1, 64)	256
dense (Dense)	(1, 32)	2 080
leaky_re_lu_6 (LeakyReLU)	(1, 32)	0
dense_1 (Dense)	(1, 1)	33

3　实验结果及分析

3.1　实验结果

对 3 个模型进行初步训练和验证，得到了 3 个模型的验证拟合图，如图 2~图 4 所示。

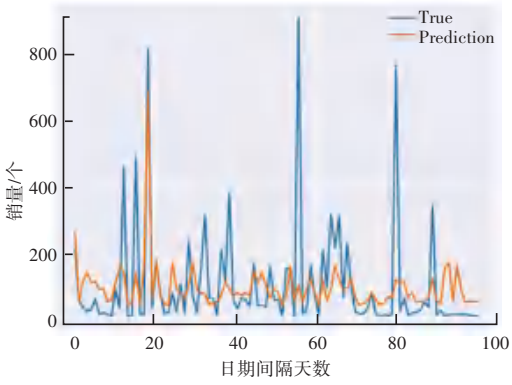

图 2　LGBM 预测

Fig. 2　LGBM prediction

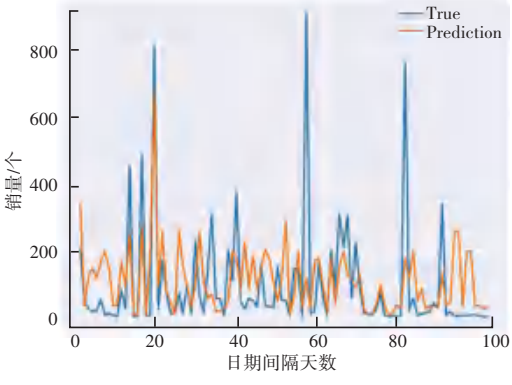

图 3　RF 预测

Fig. 3　RF prediction

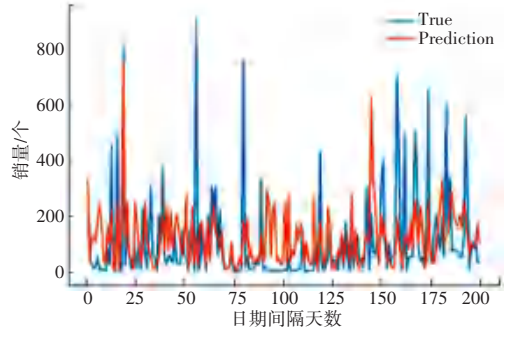

图 4　D-LSTM 预测

Fig. 4　D-LSTM prediction

在 3 种模型的初步训练和评估中，以均方根误差(*RMSE*)作为评估指标。LGBM 方法获得了最小的 *RMSE*，体现了较强的拟合能力。模型评价结果见表 6。

表 6 模型评价指标

Table 6 Model evaluation indicators

预测模型	*RMSE*
D-LSTM	641.0
RF	100.9
LGBM	93.3

通过模型初步对比,选择最优的 LGBM 模型进行优化与验证,首先使用网格搜索自动调参,再进行手动微调。网格搜索是一种自动调参的方法,通过对一系列超参数的组合进行评估和比较来寻找最佳的模型超参数组合。通过网格搜索调参后,LGBM 在验证集上的 *RMSE* 由原来的 93.3 降为 71.7, *MAPE* 也由原来的 5.0%降为 1.1%。

由于各销售区域间的销量差异显著,使用单一模型预测所有产品的销量可能无法得到满意的结果。因此,选择对主要销售区域进行独立训练。通过区域分析,最后对 101、102、103 和 105 四个区域进行单独建模和验证。各销售区域的 *RMSE* 见表 7。从表 7 中可以看出,LGBM 模型在 105 区域的表现最佳,其次是 102 区域。

表 7 区域模型预测效果对比

Table 7 Comparison of regional model prediction effects

区域编码	*RMSE*	*MAPE* /%
101	97.5	1.3
102	92.6	1.2
103	93.2	1.3
105	88.8	1.4

通过对 105 区域的预测模型进行了拟合验证,如图 5 所示。可以看出拟合效果较好,说明了针对独立区域建模能够获得一定的性能提升。

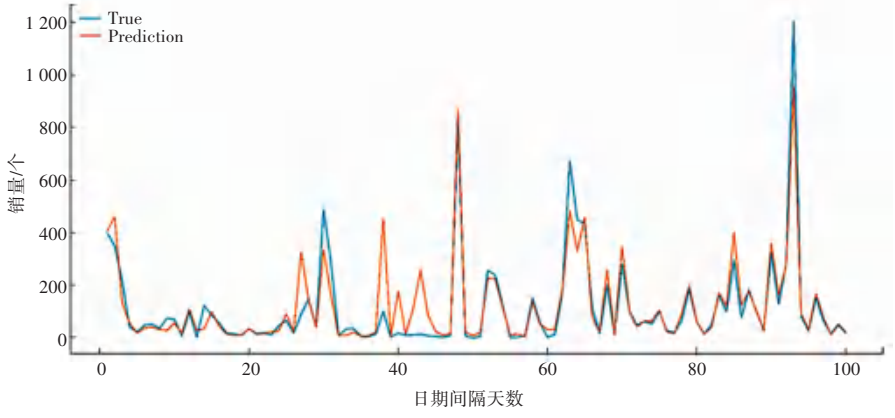

图 5 105 区域模型验证

Fig. 5 105 regional model validation

3.2 结果分析

为了分析文中预测模型的有效性以及预测效果,本文将其与 RF 模型、LSTM 模型进行了对比实验,通过 *RMSE* 评估其拟合能力,值越小说明拟合得越好,因此本文提出的基于 LGBM 的销量预测模型具有较好的性能。

通过预测值与真实值对比曲线,可以较为直观的发现分析预测值与真实值的趋势走向以及拟合程度。预测趋势与实际值的趋势比较吻合,但是在拐点处波动稍大。

考虑到区域间的差异性,通过区域分布分析,选择主要区域进行单独建模,可以得到更好的拟合效果。不过,从实际应用看,建立多个模型会使得模型维护代价增加,实际应用中需要平衡模型研发的成本和收益问题。适量选择少数大的区域进行独立建模,而其它区域则选择统一模型比较合适,使得收益最大化。

4 结束语

数据的质量决定了模型的上限,在本文中,首先对数据进行清洗、填补缺失、处理异常值,根据建模需要进行特征编码。通过数据分析,发现产品的销量存在着季节性规律、节假日规律、活动促销等高度相关性,符合实际规律。

本文基于 LGBM 模型进行销量预测,对天粒度的数据进行拟合,通过对比 RF、D-LSTM 模型,得到了最优的结果,证明其有较好的学习和泛化能力。从测试集趋势图看,模型在拐点处预测差异稍大。分析可能的原因:一是由于竞争、促销等、环境等动态因素,导致销量的变化具有不确定性。二是模型本身不适合做较大波动的时间序列趋势预测。

尽管商品销量预测模型研究取得了一定的成

果，但仍然面临着一些问题和挑战。首先，销量预测模型如何满足不同的市场和环境下的更高要求的预测需求；此外，实时性、长周期的销量预测如何保证性能。销量预测需要进一步融合大数据和人工智能等先进技术，从数据、特征和算法等方面深入挖掘，从而实现更加高效的模型，提升公司在市场的竞争力。本项目得到恒巨人工智能产学研联合实验室的技术支持。

参考文献

[1] KE Guolin, MENG Qi, FINLEY T, et al. Lightgbm: A highly efficient gradient boosting decision tree[C]//Proceedings of the 31st Conference on Neural Information Processing System. Long Beach, USA: NIPS Foundation, 2017: 1-9.

[2] WANG Z, GLYNN P W, YE Y. Likelihood robust optimization for data - driven problems [J]. Computational Management Science, 2016, 13(2): 241-261.

[3] MINASTIREANU E A, MESNITA G. Light GBM machine learning algorithm to online click fraud detection[J]. Journal of Information Assurance & Cybersecurity, 2019, 2019: 263928.

[4] 王予涵. 基于 LightGBM 的用户购买行为预测研究[D]. 兰州：兰州大学，2021.

[5] HU Lan, CHUN Y, GRIFFITH D A. Incorporating spatial autocorrelation into house sale price prediction using random forest model[J]. Transactions in GIS, 2022, 26(5): 2123-2144.

[6] SALES M H R, BRUIN D S, SOUZA C, et al. Land use and land cover area estimates from class membership probability of a random forest classification [J]. IEEE Transactions on Geoscience and Remote Sensing, 2021, 60: 4402711.

[7] HE Qiqiao, WU Cuiyu, SI Y W. LSTM with particle Swam optimization for sales forecasting [J]. Electronic Commerce Research and Applications, 2022, 51: 101118.

[8] 马超群，王晓峰. 基于 LSTM 网络模型的菜品销量预测[J]. 现代计算机（专业版），2018(23)：26-30.

[9] 广东泰迪智能科技股份有限公司. 2023 年(第 11 届)“泰迪杯”数据挖掘挑战赛[EB/OL]. [2023-02]. https://www.tipdm.org:10010/#/competition/1620719578957127680/question.

[10] PANARESE A, SETTANNI G, VITTI V, et al. Developing and preliminary testing of a machine learning-based platform for sales forecasting using a gradient boosting approach [J]. Applied Sciences, 2022, 12(21): 11054.

[11] ISLAM S, AMIN S H. Prediction of probable backorder scenarios in the supply chain using Distributed Random Forest and Gradient Boosting Machine learning techniques[J]. Journal of Big Data, 2020, 7: 65.

[12] 徐欢. 需求预测与库存决策的集成研究[J]. 上海管理科学，2018, 40(4)：102-106.