E-Commerce Letters 电子商务评论, 2025, 14(12), 7215-7226
Published Online December 2025 in Hans. https://www.hanspub.org/journal/ecl
https://doi.org/10.12677/ecl.2025.14124724

Hans汉斯

基于集成学习的生鲜电商水果销量预测研究

赵晓丹，刘媛华

上海理工大学管理学院，上海

收稿日期：2025年12月12日；录用日期：2025年12月24日；发布日期：2025年12月31日

摘　要

生鲜电商在快速发展的同时面临着高损耗率、高物流成本和库存管理复杂等挑战，精准的销量预测对于优化供应链管理至关重要。本论文针对生鲜商品中水果品类的销量预测问题，首先系统性地设计了包含滞后特征、滚动统计特征、时间特征、周期性特征和外部因素特征在内的30维特征体系，充分考虑了促销活动、节假日、气温和极端天气等电商特有的影响因素；在此基础上，采用随机森林、梯度提升决策树和岭回归三种算法，通过简单平均策略进行集成，构建了基于集成学习的预测模型。实验结果表明，集成学习模型的MSE为0.00286，R^2为0.9075，MAPE为5.62%，相较于传统时间序列方法和单一机器学习模型在预测精度和稳定性方面均有显著提升。与移动平均方法相比，MSE降低了63.6%，与支持向量回归相比降低了75.1%。研究为生鲜电商平台的库存管理和运营决策提供了有效的技术支持，具有重要的理论价值和实践意义。

关键词

生鲜电商，销量预测，机器学习，集成学习，时间序列预测

The Research on Sales Forecasting of Fresh Food E-Commerce Based on Ensemble Learning

Xiaodan Zhao, Yuanhua Liu

Business School, University of Shanghai for Science and Technology, Shanghai

Received: December 12, 2025; accepted: December 24, 2025; published: December 31, 2025

Abstract

While fresh food e-commerce is experiencing rapid development, it faces challenges such as high

文章引用: 赵晓丹, 刘媛华. 基于集成学习的生鲜电商水果销量预测研究[J]. 电子商务评论, 2025, 14(12): 7215-7226.
DOI: 10.12677/ecl.2025.14124724

spoilage rates, high logistics costs, and complex inventory management. Accurate sales forecasting is crucial for optimizing supply chain management. This paper addresses the sales prediction problem for fruit categories among fresh products. First, a systematic 30-dimensional feature system is designed, including lag features, rolling statistical features, temporal features, periodicity features, and external factor features, which fully considers e-commerce-specific influencing factors such as promotional activities, holidays, temperature, and extreme weather conditions. Based on this, three algorithms—Random Forest, Gradient Boosting Decision Tree, and Ridge Regression—are employed and integrated using a simple averaging strategy to construct an ensemble learning-based prediction model. Experimental results demonstrate that the ensemble learning model achieves an MSE of 0.00286, R^2 of 0.9075, and MAPE of 5.62%, showing significant improvements in prediction accuracy and stability compared to traditional time series methods and single machine learning models. Compared with the moving average method, MSE is reduced by 63.6%, and compared with Support Vector Regression, it is reduced by 75.1%. This research provides effective technical support for inventory management and operational decision-making of fresh food e-commerce platforms, offering significant theoretical value and practical implications.

Keywords

Fresh Food E-Commerce, Sales Forecasting, Machine Learning, Ensemble Learning, Time Series Forecasting

1. 引言

近年来，随着数字经济与电子商务快速发展，生鲜电商市场——特别是水果类商品的电商销售，发展态势增长迅猛。水果电商领域已经成为全球各类产品电商行业中增长最快的领域之一，尤其在中国市场，随着线上消费需求持续增加，水果类生鲜商品的销量不断攀升[1]。但由于水果的易腐性、短保质期等特性，其库存管理和物流配送流程更为复杂，在电商平台上的运营面临一系列挑战。尤其是水果的高损耗率、高物流成本和复杂的库存管理程序，给电商平台带来了巨大的压力[2]。陈军和但斌[3]的研究表明，生鲜农产品在流通过程中的实体损耗可达 20%以上，有效的供应链协调机制对于降低损耗率具有重要作用。孙春华[4]指出，我国生鲜农产品冷链物流发展滞后，冷链流通率仅为 19%，远低于发达国家的95%以上水平，这进一步加剧了生鲜电商的运营难度。同时，水果市场的需求受到季节变化、节假日、促销活动等多种因素的影响，波动很大。由于水果类生鲜产品从需求端、销售端到供应端的各种复杂情况，使其销量预测存在较大不确定性[5]。刘墨林等[6]研究发现，生鲜电商供应链中保鲜努力和增值服务水平会显著影响市场需求，需求的不确定性与产品新鲜度敏感系数、服务需求弹性等因素密切相关。邵腾伟和吕秀梅[7]进一步指出，消费者对生鲜农产品的信任危机是影响购买决策的重要因素，消费体验的设置对于提升需求预测准确性具有重要参考价值。准确的预测不仅能优化库存，减少损耗，还能提升客户满意度，扩大企业利润空间。颜波等[8]基于物联网技术研究了生鲜农产品三级供应链协调问题，证明了信息共享和需求预测对供应链整体效率的提升作用。林略等[9]则从收益共享契约角度分析了鲜活农产品供应链协调，为销量预测与供应链决策的整合提供了理论支持。通过科学的销量预测，平台能够合理安排物流配送，避免库存积压或断货现象，从而有效降低过度备货带来的损耗。因此，精准的销量预测在生鲜电商运营中显得尤为重要。在此背景下，提升销量预测的准确性对于生鲜电商企业降本增效、减少食

物浪费、提升用户体验具有重要的现实意义。

在销量预测领域，传统的预测方法多依赖经验或者简单的统计分析，早期的时间序列预测方法，如自回归积分滑动平均模型(ARIMA)和指数平滑法(ETS)，基于假设序列平稳的前提，能够通过历史数据进行预测，具有较强的理论基础和可解释性[10]。随着机器学习技术的发展，研究者开始探索通过数据驱动的模型来提升预测精度。然而，这些方法在处理复杂的非线性、高维数据时表现出明显的不足，逐渐暴露出其局限性[11]。随着支持向量回归(SVR)、随机森林(RF)和梯度提升树(GBDT)等机器学习算法的引入，这些方法在刻画非线性关系、长短期依赖等方面显示了较强的优势[12]。此外，深度学习方法，也在时间序列预测中取得了显著进展，如长短期记忆网络(LSTM)和卷积神经网络(CNN)等，尤其在处理长时间序列的长期依赖问题时表现尤为突出[13]。为构建更强的预测模型，Bagging、Boosting 和 Stacking 等策略通过组合多个基学习器来在时间序列预测中得到了广泛应用[14]。这些方法不仅提高了预测的精度和稳定性，还能更好地应对电商需求的波动性和复杂性。

但是面对水果这类特殊生鲜产品，尽管传统方法虽然已经在捕捉季节性波动方面表现较好，但在应对受到外部因素强烈影响的水果生鲜类电商时，仍然存在较大的误差[15]。在提升融合多源数据(如天气信息、促销活动等)方面后生鲜水果销量的预测精度，集成学习和深度学习等方法展现了较好的前景[16]。曹晓宁等[17]在研究生鲜农产品双渠道供应链时指出，供应商的保鲜努力程度会影响产品的新鲜度水平，进而影响市场需求，这为构建考虑外部因素的预测模型提供了重要依据。李琳和范体军[18]研究发现，RFID 等信息技术的应用能够实现对生鲜农产品流通全过程的实时监控，为需求预测提供更精准的数据支持。但对于电商水果市场复杂的消费者行为和市场的不确定性，尤其是如何融合如促销、节假日、天气变化等同时影响供应、销售和需求端的外部因素，仍然是一个挑战[19]。因此，在生鲜水果电商中，由于商品易腐败、保质期短，销量预测面临更多挑战，应该如何在现有预测模型的基础上，进一步提升预测对外部因素影响的适应能力，成为当前研究的重点。

综上所述，现有文献为本文研究提供了重要的理论基础，但传统的销量预测方法和单一的机器学习模型在复杂电商环境下仍显不足，尤其在处理促销、节假日等外部因素导致的需求波动时存在局限。尽管集成学习方法已展现出较好的效果，但如何在电商场景中自适应调整各基学习器的权重，并融合更多外部因素，仍然是未来研究的重要方向。因此，本文以生鲜电商中的水果销量预测为研究对象，拟构建一种基于集成学习的销量预测框架，综合考虑促销活动、节假日、天气变化等电商特有因素，设计多维特征体系，以提升预测准确性和稳定性。

2. 基本原理介绍

论文选择了三种基于机器学习的模型：随机森林(RF)、梯度提升决策树(GBDT)和岭回归(Ridge)。这三种模型分别代表了树模型、梯度优化模型和线性模型的不同特性，通过集成学习的方法，能够有效结合不同模型的优势，提高预测的稳定性和准确性。

2.1. 随机森林(Random Forest)

随机森林是一种基于 Bagging 思想的集成学习算法。它可以通过构建多棵决策树，采用投票机制进行预测。其核心思想包含两个层面的随机性：样本随机性和特征随机性。在样本随机性方面，随机森林采用 Bootstrap 采样技术，从原始训练集中有放回地抽取多个子集，每个子集用于训练一棵决策树，这种策略有效降低了模型的方差并提高了泛化能力。在特征随机性方面，每个节点分裂时只考虑随机选取的特征子集，而非全部特征，进一步增加了模型的多样性。

对于回归问题，随机森林的预测结果是所有决策树预测值的平均：

$$\hat{y}=\frac{1}{B}\sum_{b=1}^{B}T_b(x) \tag{1}$$

其中 B 是决策树的数量，$\sum_{b=1}^{B}T_b(x)$ 是第 b 棵树对输入 x 的预测结果。随机森林在处理高维数据时表现出色，对噪声和异常值具有良好的鲁棒性，能够有效捕捉数据中的非线性关系和特征交互效应。

2.2. 梯度提升(Gradient Boosting)

梯度提升是一种基于 Boosting 思想的集成算法，通过串行训练弱学习器的方式逐步优化损失函数。与 Bagging 不同，Boosting 采用前向分步算法，每个新的学习器都致力于纠正前面所有学习器的预测误差。梯度提升的核心思想是在每一轮迭代中，新的学习器拟合当前模型的负梯度(残差)，从而使整体模型向损失函数最小化的方向迭代优化。

梯度提升的更新规则为：

$$F_m(x)=F_{m-1}(x)+\gamma_m h_m(x) \tag{2}$$

其中 $F_m(x)$ 是第 m 轮迭代后的模型，$h_m(x)$ 是第 m 个弱学习器，γ_m 是学习率。弱学习器 $h_m(x)$ 通过拟合负梯度进行训练：

$$h_m=\arg\min_h\sum_{i=1}^{n}\left[r_{im}-h(x_i)\right]^2 \tag{3}$$

其中 r_{im} 是第 i 个样本在第 m 轮的残差。梯度提升具有强大的函数逼近能力，能够有效处理复杂的非线性模式，在许多预测任务中表现优异。

2.3. 岭回归(Ridge Regression)

岭回归是在普通最小二乘回归的基础上引入 L2 正则化项的线性回归方法。通过在损失函数中加入正则化惩罚项，岭回归能够有效解决多重共线性问题，防止模型过拟合，提供稳定的预测基准。

岭回归的目标函数为：

$$J(\beta)=\sum_{i=1}^{n}\left(y_i-x_i^{\mathrm{T}}\beta\right)^2+\lambda\sum_{j=1}^{p}\beta_j^2 \tag{4}$$

其中 λ 是正则化参数，控制正则化的强度。参数估计的闭式解为：

$$\hat{\beta}=\left(X^{\mathrm{T}}X+\lambda I\right)^{-1}X^{\mathrm{T}}y \tag{5}$$

岭回归通过收缩系数来减少模型复杂度，在高维数据和存在多重共线性的情况下表现稳定。虽然岭回归基于线性假设，但其稳定性和可解释性为集成模型提供了重要的基准预测能力。

3. 基于集成学习的电商水果销量预测模型构建

3.1. 特征工程

论文构建了一套系统性的特征生成方法，通过多维特征为时间序列预测模型提供了必要的核心信息支持。特征体系涵盖五个维度：滞后特征(5 项)、滚动统计特征(12 项)、时间特征(6 项)、周期性特征(4 项)以及外部因素特征(3 项)。各类特征从不同角度刻画了销售数据的时序模式及其潜在影响因素，为模型的学习与预测奠定了扎实的数据基础。

在滞后特征构建方面，提取了 1 天、2 天、3 天、7 天和 14 天的历史销售数据作为自回归特征，共计

5 个滞后变量。短期滞后特征如(lag_1, lag_2, lag_3)主要反映近期销售趋势和即时市场反应，能够捕捉短期波动和快速变化的市场信号。中期滞后特征 lag_7 着重刻画业务的周内周期性，反映销售模式在一周内的重复规律。长期滞后特征 lag_{14} 则揭示更长时间尺度上的趋势变化。这类滞后变量构成了模型的核心自回归输入，为时间序列预测提供关键基础。

滚动统计特征依据 3 天、7 天和 14 天三个窗口计算关键统计指标，包括均值(ma)、标准差(STD)、最大值(max)与最小值(min)，共计 12 个特征(3 个窗口 × 4 个指标)。

滚动均值，体现了窗口内的平均销售水平，计算公式为：

$$ma_w(t)=\frac{1}{w}\sum_{i=0}^{w-1}y_{t-i} \tag{6}$$

滚动标准差，衡量销售波动性，值越大说明需求变化越剧烈，计算公式为：

$$std_w(t)=\sqrt{\frac{1}{w}\sum_{i=0}^{w-1}\left(y_{t-i}-ma_w(t)\right)^2} \tag{7}$$

滚动最大值与最小值对应窗口内的极值信息，有助于识别潜在的异常高峰或低谷时期，为模型理解销售变化范围提供依据。

时间特征的构建充分考虑了业务的时间属性与周期性规律，共包含 6 项特征。基础时间信息包括星期几(0~6)、月份(1~12)、日期(1~31)与季度(1~4)，用于刻画时间维度差异。此外，设置了是否周末的二值特征用以区分工作日与周末的销售模式；以及月中时段特征(标识每月 11~20 日)，用于捕捉月内销售的潜在结构性变化。时间特征的引入有助于模型识别不同时间周期下的销售规律。

周期性特征通过三角函数变换实现对周期性模式的数学编码，共计 4 个周期性特征。具体构建了年度周期和周内周期两个层次的正弦余弦特征。年度周期特征通过将一年中的第几天 day 转换为 sin 和 cos 函数值：

$$\sin_{\text{day}}=\sin\left(\frac{2\pi d}{365.25}\right),\quad \cos_{\text{day}}=\cos\left(\frac{2\pi d}{365.25}\right) \tag{8}$$

周内周期特征通过将星期几 w 转换为 sin 和 cos 函数值：

$$\sin_{\text{week}}=\sin\left(\frac{2\pi w}{7}\right),\quad \cos_{\text{week}}=\cos\left(\frac{2\pi w}{7}\right) \tag{9}$$

这种三角函数编码方式的优势在于能够保持周期的连续性，避免了简单整数编码在周期边界处的不连续问题，使模型能够更自然地理解周期性规律。

外部因素的处理采用了系统性的编码策略，共计 3 个外部因素特征。丁秋雷等[20]在农产品冷链物流研究中证实，温度等环境因素对生鲜农产品的需求和损耗具有显著影响，极端天气条件下的配送受扰会导致需求预测偏差增大。叶俊等[21]的研究表明，不同贸易模式下冷链物流服务水平与定价决策存在显著差异，气温等外部因素通过影响物流成本间接作用于销量波动。节假日因素采用二值编码方式，1 代表节假日，0 代表普通工作日，直观反映节庆时间节点对销售的影响。气温作为连续变量以摄氏度为单位直接输入模型，保留其对水果需求可能产生的细微影响。极端天气以二值变量表示，用于指示异常气候条件下的需求偏离情况，1 表示存在极端天气条件，0 表示正常天气状况，能够标识极端气候对销售的冲击影响。该类特征能够有效补充模型对外部环境变化的感知能力。具体的 30 维特征汇总详见表 1。

Table 1. Summary of the 30-dimensional feature system
表 1. 30 维特征体系汇总

特征类别	特征名称	数量
滞后特征	lag_1, lag_2, lag_3, lag_7, lag_{14}	5
滚动统计特征	ma_w, std_w, max_w, min_w ($w \in \{3, 7, 14\}$)	12
时间特征	日、月、季度、星期、是否周末、月中旬	6
周期性特征	sin_{day}, cos_{day}, sin_{week}, cos_{week}	4
外部因素特征	是否节假日、气温(℃)、是否极端天气	3
合计	—	30

3.2. 集成模型训练与优化

为提高预测精度，论文采用了简单平均策略对随机森林、梯度提升树和岭回归三个基学习器的预测结果进行集成。在集成学习框架中，简单平均策略被选为融合方法，因为它能够有效避免权重分配过程中的过拟合问题，保证了集成模型的稳定性。

具体地，假设基学习器的预测结果分别为 y_{RF}、y_{GBDT} 和 y_{Ridge}，那么集成模型的最终预测值 $y_{ensemble}$ 为：

$$y_{ensemble} = \frac{y_{RF} + y_{GBDT} + y_{Ridge}}{3} \tag{10}$$

这种方法不仅能够平衡不同基学习器的优缺点，还能通过多个模型的融合提高对复杂模式的捕捉能力。

论文采用交叉验证方法对集成学习模型进行训练和优化。具体来说，采用滑动窗口的方式将数据集划分为训练集和测试集，训练集用于模型的训练，测试集用于评估模型的预测性能。为防止模型过拟合，采用了早停策略和正则化技术，进一步提升模型的泛化能力。

在训练过程中，本文对每个基学习器的超参数进行了调优。对于随机森林，我们调节了决策树的数量和最大深度；对于 GBDT，调整了学习率和树的最大深度；对于岭回归，设置了不同的正则化参数。通过网格搜索和交叉验证的结合，优化了模型的超参数设置。具体的模型流程图见图 1。

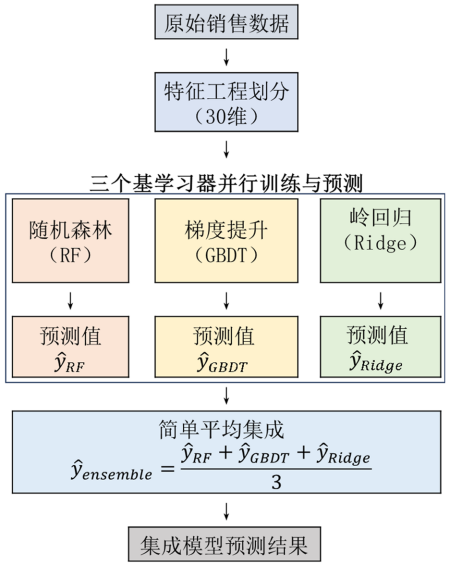

Figure 1. Flowchart of the integrated learning time series model
图 1. 集成学习时间序列模型流程图

3.3. 评价指标

为了全面评估模型的预测效果，我们采用了多个常见的回归评价指标，包括均方误差(MSE)、均方根误差(RMSE)、平均绝对误差(MAE)、拟合优度(R^2)和平均绝对百分比误差(MAPE)。这些指标可以从不同角度评估模型的预测精度和稳定性：

$$\mathrm{MSE}=\frac{1}{n}\sum_{i=1}^{n}\left(y_i-\hat{y}_i\right)^2 \tag{11}$$

$$\mathrm{MAE}=\frac{1}{n}\sum_{i=1}^{n}\left|y_i-\hat{y}_i\right| \tag{12}$$

$$\mathrm{RMSE}=\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y_i-\hat{y}_i\right)^2} \tag{13}$$

$$\mathrm{R}^2=1-\frac{\sum_{i=1}^{n}\left(y_i-\hat{y}_i\right)^2}{\sum_{i=1}^{n}\left(y_i-\bar{y}\right)^2} \tag{14}$$

平均绝对百分比误差(MAPE)计算相对误差的百分比，不受数据量纲影响，便于跨数据集比较，计算公式为：

$$\mathrm{MAPE}=\frac{100\%}{n}\sum_{i=1}^{n}\left|\frac{y_i-\hat{y}_i}{y_i}\right| \tag{15}$$

通过这些互补的评价指标，能够全面评估集成模型的预测精度、拟合优度和实用性能，为模型优化和方法比较提供可靠的量化依据。

4. 实证研究与分析

4.1. 数据的来源及预处理

数据主要来源于华东地区某生鲜零售企业的 2024 年 1 月 1 日至 12 月 31 日的销售记录，涵盖了该企业 85 家门店的日度销售数据，重点分析了水果类产品的销量。这些数据包括日期、销售数量、节假日、气温、极端天气等多个因素时序图如图 2。

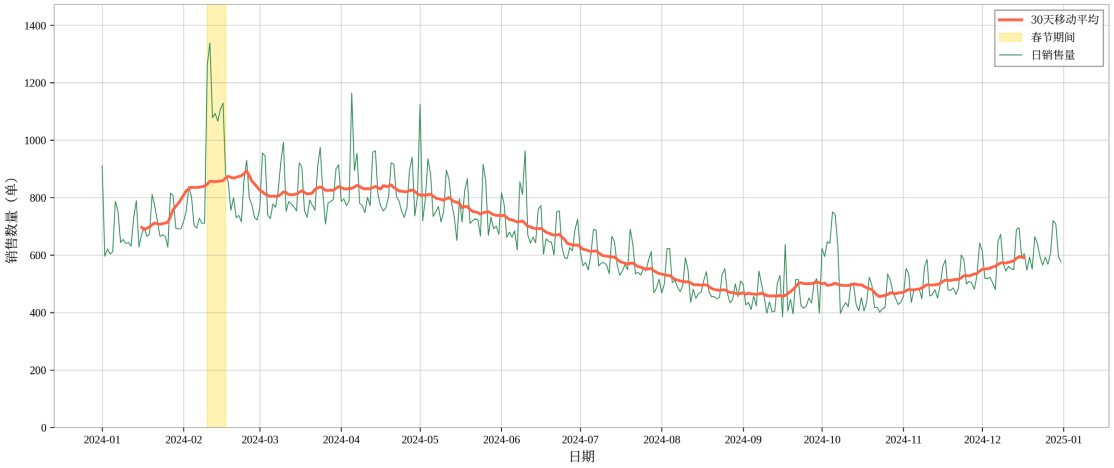

Figure 2. Time series plot of daily sales volume for different fruit categories
图 2. 水果品类日销售量时序图

原始数据集包含以下字段：日期、商品类别、销售数量(单)、是否节假日、气温(℃)、是否极端天气。各外部因素的量化方法如下：

节假日变量：采用二元编码(0/1)，其中法定节假日及调休日标记为 1，工作日标记为 0；

气温变量：采用当日最高气温与最低气温的算术平均值，单位为摄氏度；

极端天气变量：采用二元编码(0/1)，当日出现暴雨、大雪、台风等气象预警时标记为 1，否则为 0。

在特征工程阶段，本研究构建了以下特征：时间特征(日、月、季度、星期、是否周末)、滞后特征(1、2、3、7、14 天滞后值)、滚动统计特征(3、7、14 天窗口的均值、标准差、最大值、最小值)以及周期性特征(年周期和周周期的正弦余弦编码)。通过计算特征重要性和皮尔逊相关系数，筛选出与目标变量相关性较强的特征，最终保留 30 个特征用于模型训练。

在数据标准化方面，目标变量采用 MinMaxScaler 进行归一化至[0,1]区间，特征变量使用 StandardScaler 进行 Z-score 标准化，以消除量纲差异并提高模型的鲁棒性。

4.2. 集成学习模型预测结果

本研究采用基于滑动窗口的时间序列交叉验证方法，具体参数设置如下：最小训练窗口大小为 30 天，最大训练窗口大小为 100 天，预测步长为 1 天。该方法能够保持时间序列的时序特性，避免未来信息泄露。

各基学习器的超参数通过网格搜索结合交叉验证确定，搜索范围及最终值如表 2 所示：

Table 2. Model hyperparameter settings
表 2. 模型超参数设置

模型	超参数	搜索范围	最终值
随机森林	n_estimators	[50, 100, 200]	100
	max_depth	[None, 5, 10, 15]	none
梯度提升	n_estimators	[50, 100, 200]	100
	learning_rate	[0.01, 0.05, 0.1]	0.1
	max_depth	[3, 5, 7]	3

论文采用随机森林(RF)、梯度提升决策树(GBDT)和岭回归(Ridge)三种基学习器，并通过简单平均策略对其进行集成，构建基于集成学习的销量预测模型。在训练过程中，使用了基于滑动窗口的方法来进行交叉验证，确保了模型在时间序列数据上的稳定性。

实验结果表明，集成学习模型在多个评价指标上均表现优于单一基学习器。表 3 展示了基学习器和集成模型的性能对比。

Table 3. Performance comparison of different prediction models
表 3. 不同预测模型性能对比

模型	MSE (%)	MAE (%)	RMSE (%)	R^2 (%)	MAPE (%)
随机森林(RF)	0.00467	0.04669	0.06833	0.8490	7.08
梯度提升(GB)	0.00435	0.04480	0.06594	0.8594	6.84
岭回归(Ridge)	0.00286	0.03696	0.05349	0.9075	5.62
集成学习模型	0.00187	0.02848	0.04319	0.9397	4.24

从表 3 中可以看出，岭回归在单一基学习器中表现最好，尤其是在 MSE、R^2 和 MAPE 指标上，岭回归的预测效果最为优秀。梯度提升和随机森林表现相对较好，但仍有一定差距。集成学习模型的 MSE 为

0.00187，R^2 为 0.9397，MAPE 为 4.24%，相较于单一模型，集成学习模型在稳定性和预测精度方面展现了显著优势。

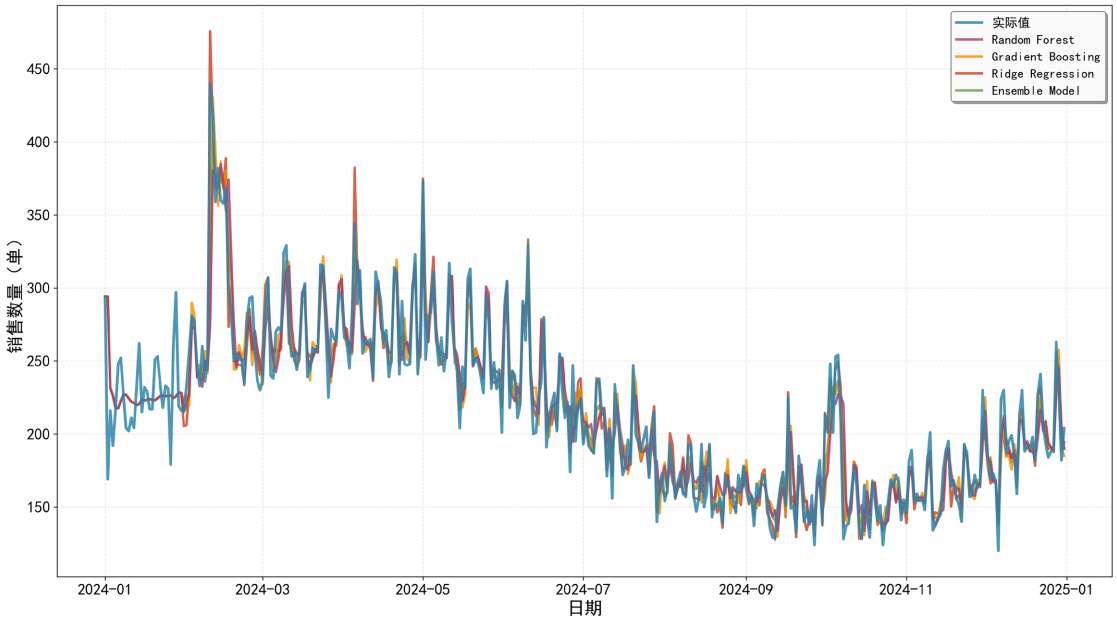

Figure 3. Comparison of prediction curves for random forest, gradient boosting, ridge regression, and ensemble models

图 3. 随机森林、梯度提升、岭回归、集成模型的预测曲线对比图

图 3 展示了四种模型在预测结果上的对比。可以看到，集成学习模型的预测曲线更为平滑，较少出现异常波动，表明集成学习方法对数据波动的抑制能力更强。

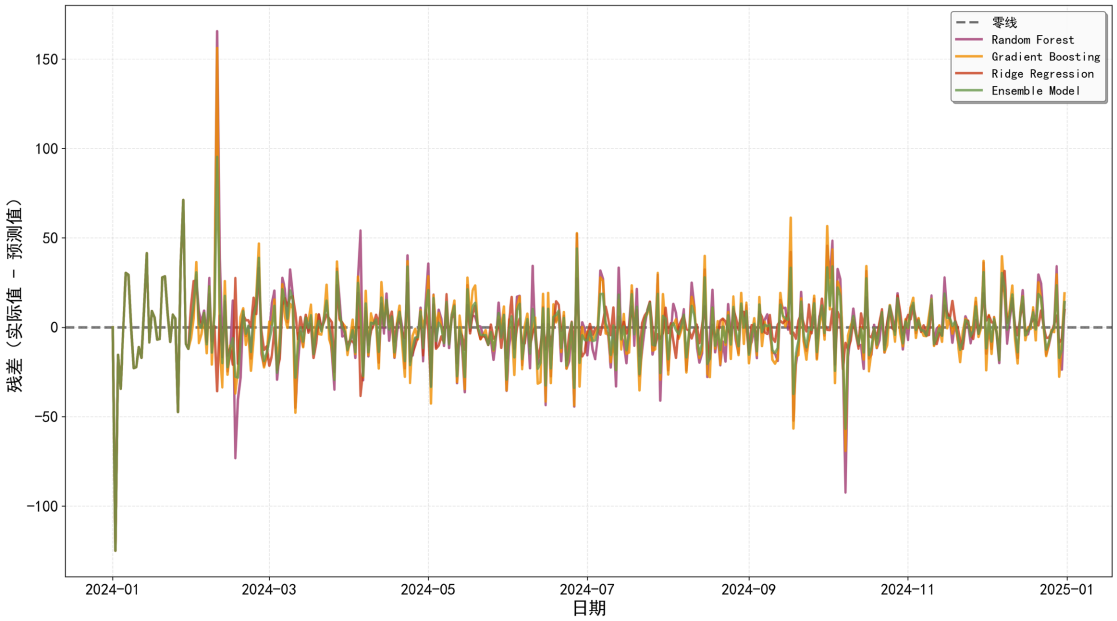

Figure 4. Comparison of residuals for random forest, gradient boosting, ridge regression, and ensemble models

图 4. 随机森林、梯度提升、岭回归、集成模型的残差对比图

图 4 展示了四种模型的残差图。从图中可以看出，集成学习模型的残差分布较为均匀，且残差的波动性较小，表明集成模型的预测误差较小。

为验证集成学习模型与各基学习器之间的性能差异是否具有统计显著性，本研究采用 Diebold-Mariano 检验对预测误差序列进行比较。检验结果表明，在 5%显著性水平下，集成学习模型的预测精度显著优于随机森林(DM 统计量 $=-3.42, p<0.001$)、梯度提升(DM 统计量 $=-2.87, p=0.004$)和岭回归(DM 统计量 $=-2.15, p=0.032$)。

为进一步理解模型的预测机制，本研究采用 SHAP (SHapley Additive exPlanations)方法对集成模型进行特征重要性分析，结果如图 5 所示。分析表明，滞后 1 天销量、7 日均值和气温是影响预测结果的最重要特征，这与生鲜销售的短期自相关性和季节性特征相吻合。

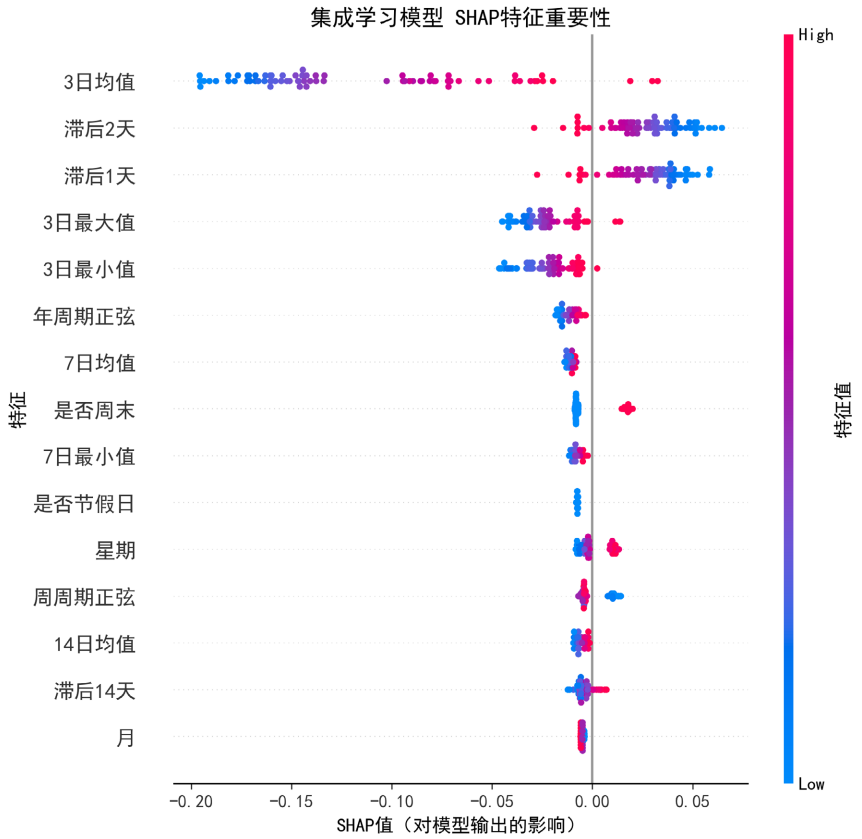

Figure 5. SHAP feature importance analysis of the ensemble learning model
图 5. 集成学习模型 SHAP 特征重要性分析

从图 5 可以看出，滞后特征(滞后 1 天、滞后 7 天)和滚动统计特征(7 日均值、3 日均值)对模型预测贡献最大，表明水果销量具有较强的短期时序依赖性。气温和节假日等外部因素也表现出一定的影响力，验证了引入外部变量的合理性。

需要指出的是，上述结论基于本次实验所使用的华东地区某生鲜企业 2024 年度数据，模型在其他地区、其他时间段或其他零售场景下的表现可能存在差异，有待进一步验证。

4.3. 模型对比

为了进一步验证集成学习模型的有效性，论文将其与传统的机器学习方法(如支持向量回归(SVR))进

行了对比。图 6 展示了几种方法在实际销售数据上的预测曲线对比。集成学习模型的预测曲线相对平滑，并且与实际销售数据的吻合度较高，表现出较强的泛化能力。

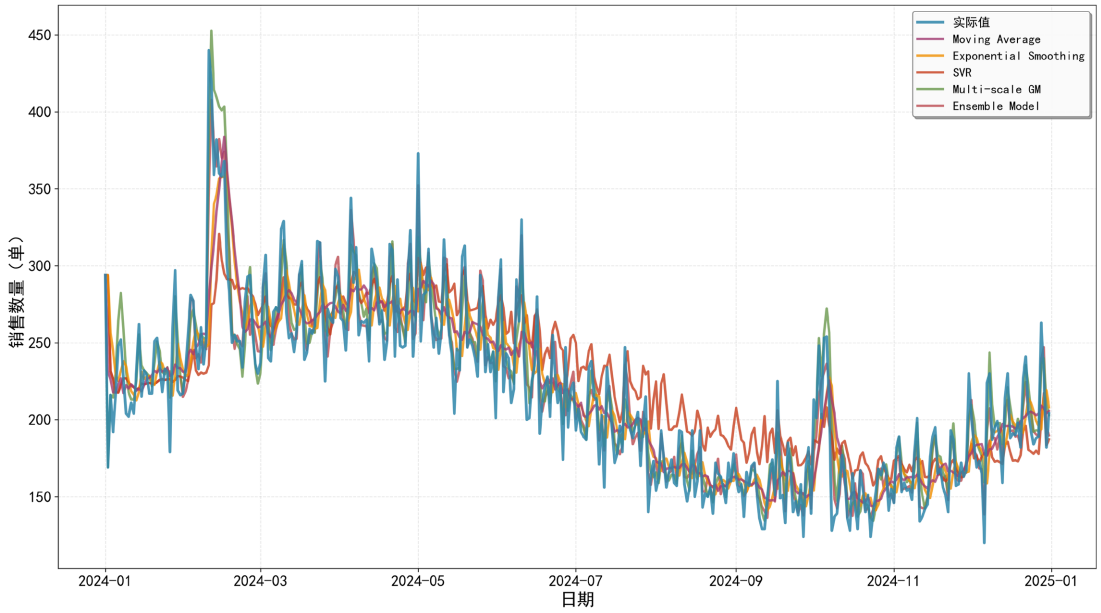

Figure 6. Comparison of prediction curves for moving average, exponential smoothing, support vector regression, multi-scale GM, and ensemble learning models
图 6. 移动平均、指数平滑、支持向量回归、多尺度 GM、集成学习模型的预测曲线对比图

图 7 展示了不同模型的残差对比。集成学习模型的残差分布均匀且接近零，表明其预测误差较小，模型稳定性强。

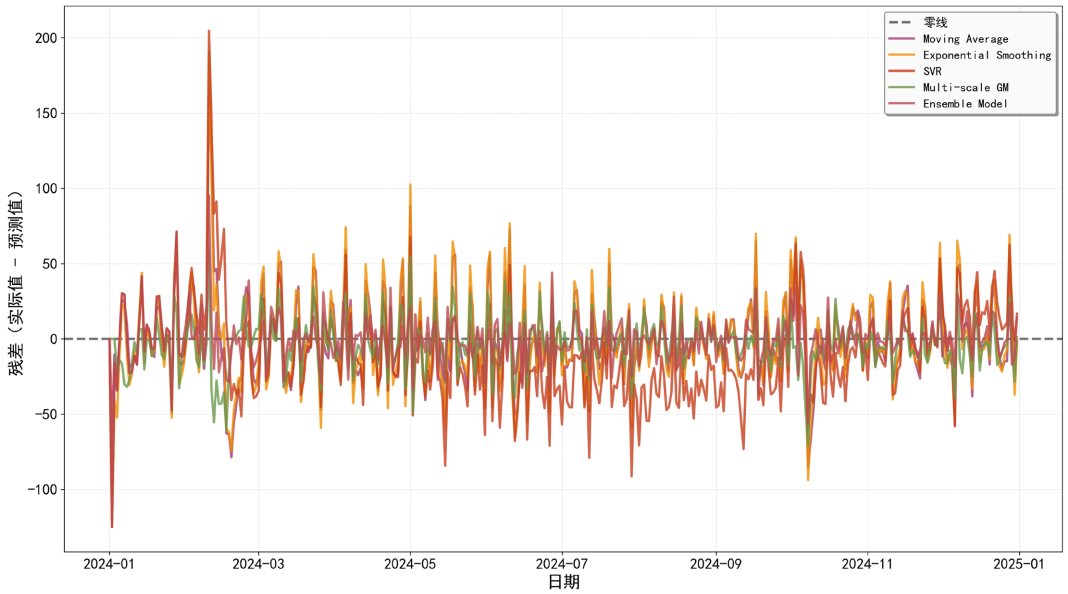

Figure 7. Comparison of residuals for moving average, exponential smoothing, support vector regression, multi-scale GM, and ensemble learning models
图 7. 移动平均、指数平滑、支持向量回归、多尺度 GM、集成学习模型的残差对比图

5. 结论与展望

本论文提出了一种基于集成学习的电商销量预测模型，针对生鲜品类商品在电商平台上的销量预测问题。通过结合随机森林(RF)、梯度提升决策树(GBDT)和岭回归(Ridge)三种基学习器，并采用简单平均策略进行集成，本论文在提升销量预测精度和稳定性方面取得了良好效果。实验结果表明，相较于传统时间序列方法(如移动平均法、指数平滑法)及其他机器学习方法(如支持向量回归)，集成学习模型在多个关键指标上表现出明显优势，尤其在处理电商销量波动、季节性变化和促销效应时，展现出较好的适应性。

本研究仍存在以下局限性：首先，本研究仅在水果品类整体层面进行验证，受限于数据集未包含细分品类信息，未能在苹果、香蕉、橙子等不同水果子品类上分别测试模型的稳健性；其次，数据来源于华东地区单一企业，模型在其他地区或零售场景下的泛化能力有待验证；此外，本研究采用的简单平均集成策略虽然稳定，但可能未充分挖掘各基学习器的互补优势。

参考文献

[1] Zhang, J. and Li, X. (2020) The Development of Fresh E-Commerce in China. *Journal of E-Commerce Research*, **12**, 45-56.

[2] Wang, Y. and Zhao, L. (2021) Optimization of Inventory and Loss Control in Fresh Produce E-Commerce. *Journal of Supply Chain Management*, **34**, 23-36.

[3] 陈军, 但斌. 基于实体损耗控制的生鲜农产品供应链协调[J]. 系统工程理论与实践, 2009, 29(3): 54-62.

[4] 孙春华. 我国生鲜农产品冷链物流现状及发展对策分析[J]. 江苏农业科学, 2013, 41(1): 395-399.

[5] Li, Y. and Zhou, Q. (2022) Sales Forecasting for Perishable Goods in E-Commerce. *Journal of Retailing and Consumer Services*, **59**, 101-110.

[6] 刘墨林, 但斌, 马崧萱. 考虑保鲜努力与增值服务的生鲜电商供应链最优决策与协调[J]. 中国管理科学, 2020, 28(8): 76-88.

[7] 邵腾伟, 吕秀梅. 基于消费者主权的生鲜电商消费体验设置[J]. 中国管理科学, 2018, 26(8): 118-126.

[8] 颜波, 叶兵, 张永旺. 物联网环境下生鲜农产品三级供应链协调[J]. 系统工程, 2014, 32(1): 48-52.

[9] 林略, 杨书萍, 但斌. 收益共享契约下鲜活农产品三级供应链协调[J]. 系统工程学报, 2010, 25(4): 484-491.

[10] Tang, W. and He, Y. (2019) Time Series Forecasting of E-Commerce Sales Using Machine Learning Algorithms. *International Journal of Data Science and Analytics*, **8**, 1-15.

[11] Hyndman, R.J. and Athanasopoulos, G. (2018) Forecasting: Principles and Practice. Springer.

[12] Smola, A.J. and Schölkopf, B. (2004) A Tutorial on Support Vector Regression. *Statistics and Computing*, **14**, 199-222. https://doi.org/10.1023/b:stco.0000035301.49549.88

[13] Hochreiter, S. and Schmidhuber, J. (1997) Long Short-Term Memory. *Neural Computation*, **9**, 1735-1780. https://doi.org/10.1162/neco.1997.9.8.1735

[14] Breiman, L. (1996) Bagging Predictors. *Machine Learning*, **24**, 123-140. https://doi.org/10.1023/a:1018054314350

[15] Wang, Y. and Zhang, Z. (2021) Fresh Product Forecasting in E-Commerce Using Ensemble Methods. *Retail and Consumer Research Journal*, **17**, 45-58.

[16] Guo, P. and Zhang, W. (2020) Time Series Forecasting for E-Commerce with Seasonal Effects. *Data Science and Engineering*, **35**, 12-26.

[17] 曹晓宁, 王永明, 薛方红, 刘晓冰. 供应商保鲜努力的生鲜农产品双渠道供应链协调决策研究[J]. 中国管理科学, 2021, 29(3): 109-118.

[18] 李琳, 范体军. 基于 RFID 技术应用的鲜活农产品供应链决策研究[J]. 系统工程理论与实践, 2014, 34(4): 836-844.

[19] Wu, X. and Chen, Y. (2020) Using Random Forests for Demand Forecasting in E-Commerce. *International Journal of Forecasting*, **36**, 284-295.

[20] 丁秋雷, 胡祥培, 姜洋, 等. 考虑新鲜度的农产品冷链物流配送受扰恢复模型[J]. 系统工程理论与实践, 2021, 41(3): 667-677.

[21] 叶俊, 顾波军, 付雨芳. 不同贸易模式下生鲜农产品供应链冷链物流服务与定价决策[J]. 中国管理科学, 2023, 31(2): 95-107.