E-Commerce Letters 电子商务评论, 2026, 15(3), 941-951
Published Online March 2026 in Hans. https://www.hanspub.org/journal/ecl
https://doi.org/10.12677/ecl.2026.153354

Hans汉斯

销量预测中四类模型的性能对比研究

王则林，张艺恒，张　林，周　洁*

南通大学人工智能与计算机学院，江苏　南通

收稿日期：2026年1月22日；录用日期：2026年2月5日；发布日期：2026年3月18日

摘　要

本研究系统评估了传统统计方法与现代集成学习、深度学习技术在零售销量预测中的适用性差异，为模型选型提供实证依据。基于M5 (沃尔玛日销量)数据集，在统一的数据切分、滚动回测与贝叶斯优化框架下，对SARIMA、Prophet、N-BEATS与XGBoost四类模型在店铺级与SKU级预测任务中的性能进行对比。店铺级以CA_1门店聚合销量为对象，在7天、14天与28天预测窗口下使用平均绝对误差(MAE)、均方根误差(RMSE)与平均绝对百分比误差(MAPE)评估模型表现；SKU级从CA_1门店随机抽取150个SKU，按历史销量三分位划分为高、中、低销量组，在28天预测窗口下通过3次滚动回测，并结合MAE、RMSE与均方根标度误差(RMSSE)进行综合评价，同时采用特征消融检验外生变量的作用。结果表明，XGBoost在两类任务中均表现出最优或近最优性能：店铺级28天预测中MAPE低至4.47%；SKU级平均RMSSE为0.933 ± 0.288，在低销量、高稀疏组中RMSSE为1.178 ± 0.332，相较N-BEATS提升约12.6%。各模型表现具有明显差异：SARIMA在强周度季节性场景下较为稳定；N-BEATS在高销量、低稀疏序列上具备竞争力但对稀疏性较敏感；Prophet对局部突变刻画不足。特征消融显示，价格与日历特征对稀疏SKU预测提升尤为显著。综上，本研究量化了预测跨度与稀疏度对模型性能的影响，验证了XGBoost在零售需求预测中的鲁棒性，并为零售领域的模型选型与预测优化提供了更为科学的决策支撑。

关键词

销量预测，M5数据集，集成学习，深度学习

Performance Comparison of Four Models in Sales Forecasting

Zelin Wang, Yiheng Zhang, Lin Zhang, Jie Zhou*

School of Artificial Intelligence and Computer Science, Nantong University, Nantong Jiangsu

Received: January 22, 2026; accepted: February 5, 2026; published: March 18, 2026

*通讯作者。

文章引用: 王则林, 张艺恒, 张林, 周洁. 销量预测中四类模型的性能对比研究[J]. 电子商务评论, 2026, 15(3): 941-951. DOI: 10.12677/ecl.2026.153354

Abstract

This study systematically evaluates the differences in applicability between traditional statistical methods and modern ensemble learning and deep learning techniques for retail sales forecasting, providing empirical evidence to support model selection. Using the M5 (Walmart daily sales) dataset, we compare the performance of four model families—SARIMA, Prophet, N-BEATS, and XGBoost—on store-level and SKU-level forecasting tasks under a unified framework of data splitting, rolling backtesting, and Bayesian optimization. For the store-level task, aggregated sales of the CA_1 store are used as the forecasting target, and model performance is evaluated over 7-day, 14-day, and 28-day horizons using mean absolute error (MAE), root mean squared error (RMSE), and mean absolute percentage error (MAPE). For the SKU-level task, 150 SKUs are randomly sampled from the CA_1 store and stratified into high-, medium-, and low-sales groups based on historical sales tertiles; under a 28-day forecasting horizon, three rolling backtests are conducted and a comprehensive evaluation is performed using MAE, RMSE, and root mean squared scaled error (RMSSE). In addition, feature ablation is employed to examine the role of exogenous variables. The results indicate that XGBoost achieves the best or near-best performance in both tasks: in the store-level 28-day forecasting, MAPE is as low as 4.47%; at the SKU level, the average RMSSE is 0.933 ± 0.288, and in the low-sales, highly sparse group, RMSSE is 1.178 ± 0.332, representing an improvement of approximately 12.6% compared with N-BEATS. Model performance differs markedly across methods: SARIMA is relatively stable in scenarios with strong weekly seasonality; N-BEATS is competitive on high-sales, low-sparsity series, but is sensitive to sparsity; Prophet is insufficient in characterizing local abrupt changes. Feature ablation shows that price and calendar features yield particularly significant improvements for forecasting sparse SKUs. Overall, this study quantifies the impact of forecasting horizon and sparsity on model performance, verifies the robustness of XGBoost in retail demand forecasting, and provides a more scientifically grounded basis for model selection and forecasting optimization in the retail domain.

Keywords

Sales Forecasting, M5 Dataset, Ensemble Learning, Deep Learning

1. 引言

随着电商数据维度快速增长，利用统计学、集成学习与深度学习[1]技术挖掘销售时间序列趋势，已成为提升零售决策效率的关键。从传统统计方法向现代集成学习范式及深度学习方法的演进，拓展了时序预测在电商领域的应用边界。面对非线性、高波动的销量数据，依据预测周期选择适配算法[2]是提升决策质量的核心。零售预测具备多粒度特征：管理层关注门店、品类聚合需求以支撑补货与资源配置[3]，一线运营则依赖 SKU 级精细化预测，应对长尾、缺货与促销带来的库存波动[4]。因此，在同一体系下对比聚合序列与 SKU 细粒度序列的预测效果，对模型选型至关重要。

过往研究中，统计模型长期占据主导。SARIMA 模型[5]可解析季节性与平稳性；Prophet [6]采用可分解结构，在趋势突变与节假日效应建模中优势显著。但下沉至 SKU 级后，销量常呈现零膨胀、间歇性需求[7]，外生因素影响更为突出，传统模型的线性与平稳性假设面临挑战。为刻画非线性交互关系，集成

学习与深度学习成为重要补充：XGBoost [8]可融合多类外部特征并刻画变量交互，在高噪声、稀疏数据下效率更高；深度学习模型 N-BEATS [9]通过残差结构提升长序列建模能力，但深度模型通常依赖充足样本[10]，在 SKU 级稀疏场景下的稳定性仍需验证。

现有研究较少在统一框架下系统对比统计、集成学习与深度学习模型在不同预测跨度与粒度下的性能差异，对“跨度–粒度–稀疏度”耦合影响的量化仍不足[11]。基于此，本文在 M5 数据集[12]上对 SARIMA、Prophet、N-BEATS (深度学习)与 XGBoost (集成学习)开展实证：设置店铺级与 SKU 级任务，店铺级比较 7/14/28 天窗口下的精度与适应性；SKU 级按稀疏度分层，评估间歇性需求下的中长期误差。通过对比各模型在波动、趋势与稀疏峰值捕捉上的差异，明确其适用边界，为零售销量预测的模型选型提供量化依据。

2. 方法

2.1. 数据来源与预处理

本研究采用 M5 Forecasting 数据集(Walmart 零售日销量)，包含 3 个州(CA/TX/WI)共 10 家门店、3049 个 SKU 的日销量序列以及对应的日历信息与历史价格信息。为保证实验一致性，我们以门店 CA_1 为核心研究对象，并构建两类预测任务：

(1) 店铺级预测：聚合全店 SKU 销量，预测窗口设定为 $H \in \{7,14,28\}$。(2) SKU 级预测：从 CA_1 门店随机抽取 150 个 SKU (覆盖 FOODS/HOUSEHOLD/HOBBIES)，仅在 $H=28$ 下进行多步预测，并按历史总销量三分位分为 High/Medium/Low 三组(每组 50 个)。

为避免信息泄露并与多步预测对齐，本文采用时间顺序切分。设全序列长度为 T。对每次回测，使用训练集 $[1, T_{train}]$、验证集 $(T_{train}, T_{val}]$ 与测试集 $(T_{val}, T_{test}]$，其中测试窗口长度为 H，验证窗口长度为 V。店铺级采用单次切分 $T_{test}=T$，$T_{val}=T-H$，$T_{train}=T-H-V$，SKU 级采用 3 次滚动回测。第 k 次回测 $(k=1,2,3)$的测试区间为 $[T_k+1, T_k+H]$，其中 T_k 以固定步长 Δ 向后滚动。

为了保证实验的针对性并控制计算复杂度，本文选取了加利福尼亚州 1 号门店(CA_1)的数据作为核心研究对象。本研究针对电商销量数据的特性，构建了两类特征系统：

(1) 时间日历特征(Calendar Features)

零售销量具有较强的周期性，本文从日期字段中提取并编码了以下特征：基础时间特征，包括年份、月份、日期、星期；周末标识，通过构建布尔变量以捕捉周末购物高峰效应；特殊事件特征，对节假日进行分类编码，并引入相关变量，以刻画食品券发放对食品类商品销量的拉动作用。

(2) 价格特征(Price Features)

价格是影响销量的最敏感因子，我们构建了相对价格特征，以反映当前价格与历史均价的差异。

价格变动率：$(\text{Current\_Price} - \text{Average\_Price}) / \text{Average\_Price}$，该特征能有效捕捉促销打折(负值)对销量的非线性刺激作用。

2.2. 模型介绍与参数设置

本研究选取了四种具有代表性的预测模型，分别覆盖了经典统计学、工业界广义加性模型、前沿深度学习以及集成集成学习范式。

SARIMA

SARIMA 适用于具有趋势和季节性的单变量序列。针对 M5 数据集中显著的“周”循环特性($s=7$)，记为 $\text{SARIMA}(p,d,q)(P,D,Q)_s$。对序列 y_t 做常规差分与季节差分：

$$\Phi\left(B^s\right)\phi(B)\nabla^d\nabla_s^D y_t=\Theta\left(B^s\right)\theta(B)\varepsilon_t \tag{1}$$

其中 B 为滞后算子，ε_t 为白噪声；Φ,Θ 为季节 AR/MA 多项式。参数通过极大似然估计，并以信息准则辅助选择阶数。

Prophet

Prophet 采用广义加性模型(GAM)将时间序列分解为趋势、季节与节假日三部分：

$$y(t)=g(t)+s(t)+h(t)+\epsilon_t \tag{2}$$

其中 $g(t)$ 为分段趋势项(含变点)。$s(t)$ 为季节项，$h(t)$ 为节假日/事件项。季节性用傅里叶级数表示：

$$s(t)=\sum_{n=1}^{N}\left[a_n\cos\left(\frac{2\pi nt}{P}\right)+b_n\sin\left(\frac{2\pi nt}{P}\right)\right] \tag{3}$$

P 为周期长度(如周周期)，N 为傅里叶阶数。节假日效应可表示为指示变量的线性组合：

$$h(t)=z(t)^{\top}\kappa \tag{4}$$

其中 $z(t)$ 为节假日/事件指示向量，κ 为对应权重系数。

N-BEATS

N-BEATS 是一种基于全连接网络的端到端预测模型，不依赖 RNN，而用堆叠残差块逐步分解序列并生成预测。模型由多个 Stack 组成，每个 Stack 含若干 Block。对第 l 个 Block：

(1) 特征编码(Feature Encoding)。输入是回望窗口的历史数据 x_l (初始为 $Sales_{t-w:t}$)。通过 4 层全连接层(FC)和 ReLU 激活函数提取非线性特征：

$$h_{l,1}=\mathrm{ReLU}\left(W_{l,1}x_l+b_{l,1}\right)\ \cdots\ h_{l,4}=\mathrm{ReLU}\left(W_{l,4}h_{l,3}+b_{l,4}\right) \tag{5}$$

(2) 基函数展开系数(Expansion Coefficients)。网络的最后层输出两个投射向量——“向后系数”θ_l^b 和“向前系数”θ_l^f：

$$\theta_l^b=\mathrm{Linear}\left(h_{l,4}\right),\quad \theta_l^f=\mathrm{Linear}\left(h_{l,4}\right) \tag{6}$$

(3) 波形合成与分解(Synthesis)。这是 N-BEATS 的核心。系数与预设的基函数 $g(t)$ 相乘，生成拟合曲线：

$$\hat{x}_l=g^b\left(\theta_l^b\right)=\sum_{i=1}^{M^b}\theta_{l,i}^b\,g_i^b(t),\quad \hat{y}_l=g^f\left(\theta_l^f\right)=\sum_{i=1}^{M^f}\theta_{l,i}^f\,g_i^f(t) \tag{7}$$

(4) 双向残差传递(Doubly Residual)

最终模型输出为所有 Block 的 forecast 累加：

$$\hat{y}_{final}=\sum_l\hat{y}_l \tag{8}$$

这种机制允许第一个 Stack 专注于捕捉宏观趋势，后续 Stack 专注于捕捉高频细节。

XGBoost

XGBoost 在本研究中作为集成学习的代表，不同于上述端到端的时序模型，XGBoost 采用了“特征工程 + 监督回归”的范式。通过逐轮叠加回归树来拟合残差，从而形成强预测器。其基本形式为加法模型：

$$\hat{y}_i=\sum_{t=1}^{T}f_t(x_i),\quad f_t\in\mathcal{F} \tag{9}$$

二其中 x_i 为输入特征，f_t 表示第 t 棵回归树，$\mathcal{F}$ 为所有树模型空间。

XGBoost 以目标函数最小化为训练准则，将数据拟合误差与模型复杂度进行联合约束：

$$\mathcal{L}=\sum_{i=1}^{n} l\left(y_i,\hat{y}_i\right)+\sum_{t=1}^{T}\Omega\left(f_t\right),\ \Omega\left(f\right)=\gamma K+\frac{1}{2}\lambda\left\|w\right\|^2 \tag{10}$$

其中 $l(\cdot)$ 为损失函数(回归任务常用平方误差或绝对误差)，K 为叶节点数，w 为叶节点权重，γ 与 λ 分别控制树结构与叶权重的正则化强度。训练过程中采用二阶泰勒展开近似损失并进行贪心分裂，从而在保证精度的同时提升计算效率。

2.3. 多步预测策略与模型参数

为保证基准模型比较的公平性与可重复性，本文对所有模型采用统一的超参数调优流程：基于滚动时间验证集，使用 Optuna 的 TPE 采样器进行贝叶斯优化；门店级以 MAPE，SKU 级以 RMSSE 作为目标函数。对每个预测跨度 $H\in\{7,14,28\}$ 分别调优，以避免不同跨度下误差分布差异影响结论。门店级与 SKU 级实验使用一致的搜索空间设置，在控制计算成本的前提下尽量保证不同规模门店之间的可比性。各模型搜索空间如下。SARIMA 的非季节部分的自回归阶数 p、滑动平均阶数 q 均取{0, 1, 2}，差分阶数 d 取{0, 1}；季节部分的自回归阶数 P、滑动平均阶数 Q 取{0, 1, 2}，季节差分阶数 D 取{0, 1}，季节周期固定为 m = 7 (日度数据)，趋势项在“无趋势/常数/线性”中自动选择。Prophet：变化点先验尺度{0.001, 0.01, 0.05, 0.5}，季节性先验尺度{0.01, 0.1, 1.0, 10.0}，节假日先验尺度{0.01, 0.1, 10.0}；节假日效应使用统一节假日表，并以日期哑变量输入。N-BEATS 的历史销量与日历/价格等外生特征拼接后经线性投影输入；模型堆叠层数取值{2, 4, 6}，每个堆叠 block 数固定为 1，隐藏宽度{128, 256, 512}；学习率[4×10^{-5}, 4×10^{-3}]，batchsize{32, 64, 128}；最多训练 50 轮，早停耐心 10；输入窗口长度按实验统一设为 H 的若干倍，并以验证集 MAPE/RMSSE 最小化选取配置。XGBoost：最大深度{3, 5, 7, 9}，学习率{0.01, 0.05, 0.1, 0.3}，基学习器数{100, 300, 500, 1000}；样本/特征采样比例均为{0.6, 0.8, 1.0}，最小叶子节点权重取{1, 3, 5}，L2 正则[0, 1.0]；采用早停，耐心 50。

上述设置在门店级与 SKU 级实验中保持一致，并对每个 H 独立寻优；文中所有指标均基于该流程得到的最优超参数计算。

3. 模型结果与分析

3.1. 店铺级销量预测

本节以 M5 数据集中的典型门店 CA_1 (加利福尼亚州 1 号店)为例，展示了 SARIMA、Prophet、N-BEATS 和 XGBoost 四种模型在 7 天(短期)、14 天(中期)和 28 天(长期)预测窗口下的可视化结果与量化评估。下面将展示四个模型在不同时间跨度下的预测曲线对比。图中灰色区域为历史训练数据，浅灰色阴影区域为预测窗口，黑色虚线代表真实销量，彩色实线代表模型预测值。图 1~3 分别展示了各模型在 7 天、14 天及 28 天窗口下的预测表现：

图 1~3 显示，7/14 天窗口波动强且存在短期突变，28 天窗口周度季节性更稳定。总体上 XGBoost 在三种跨度下误差最低，兼顾突变响应与周期稳定性；SARIMA 在 28 天场景下表现接近并优于 N-BEATS；Prophet 因平滑假设对峰值刻画偏弱。本文还记录了四个模型在训练过程中的评估指标，结果如表 1 所示。

综合分析 M5 数据集上不同预测窗口的(7 天、14 天、28 天)上的销量预测结果，XGBoost 凭借对滞后特征的高效利用，在兼顾短期突发波动的灵敏度与长期周期变化的稳定性上表现突出，在各预测步长中均取得了相对最低的 MAPE (4.47%~6.12%)，整体性能在本次实验中更具优势。随着预测窗口从 7 天扩展至 28 天，数据特征从以不规则震荡为主逐渐转向以规律的周度季节性为主，各模型的整体精度也呈现出逐步提升的趋势。在此过程中，传统统计模型 SARIMA 展现出较好的长周期适应性，在 28 天预测场

景下精度逼近甚至反超了深度学习模型 N-BEATS。尽管 N-BEATS 在处理大规模同质化数据时通常表现优异，但在本研究采用的 M5 数据集上，传统 ARIMA 模型展现出了更强的鲁棒性。这主要归因于 ARIMA 对单变量序列的独立建模方式，减少了不同商品间噪声的负迁移影响；在 M5 数据集呈现高度稀疏和随机波动的特点时，结构相对简单的统计模型遵循“均值回归”原则，一定程度上避免了深度学习模型过度拟合噪声的倾向；同时，统计方法的差分机制在处理非线性数据趋势外推时，表现往往优于未经过度

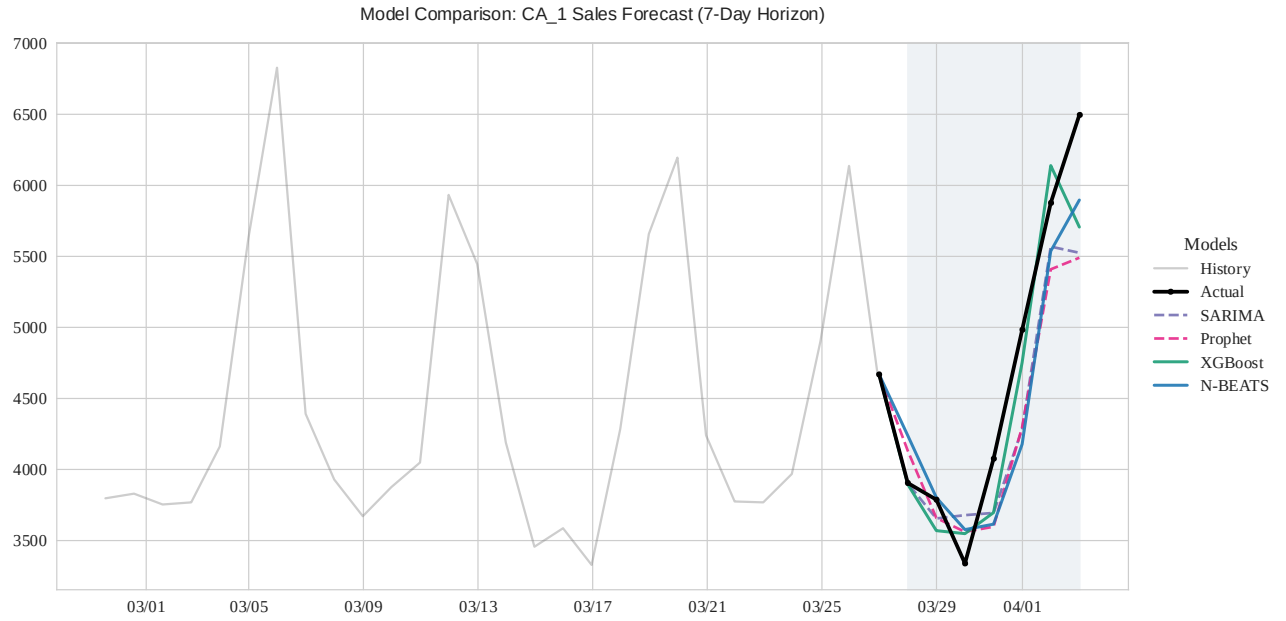

Figure 1. Comparison of predicted vs. actual results for each model in the 7-day forecast window
图 1. 7 天预测窗口中各模型预测结果与实际结果对比图

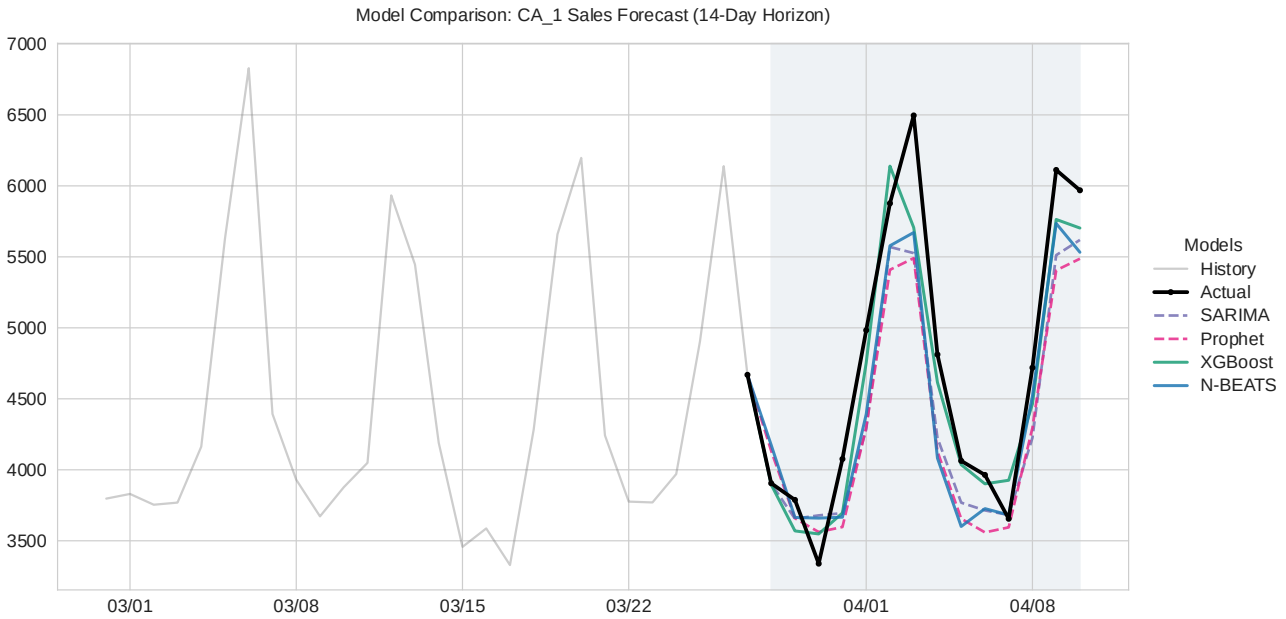

Figure 2. Comparison of predicted vs. actual results for each model in the 14-day forecast window
图 2. 14 天预测窗口中各模型预测结果与实际结果对比图

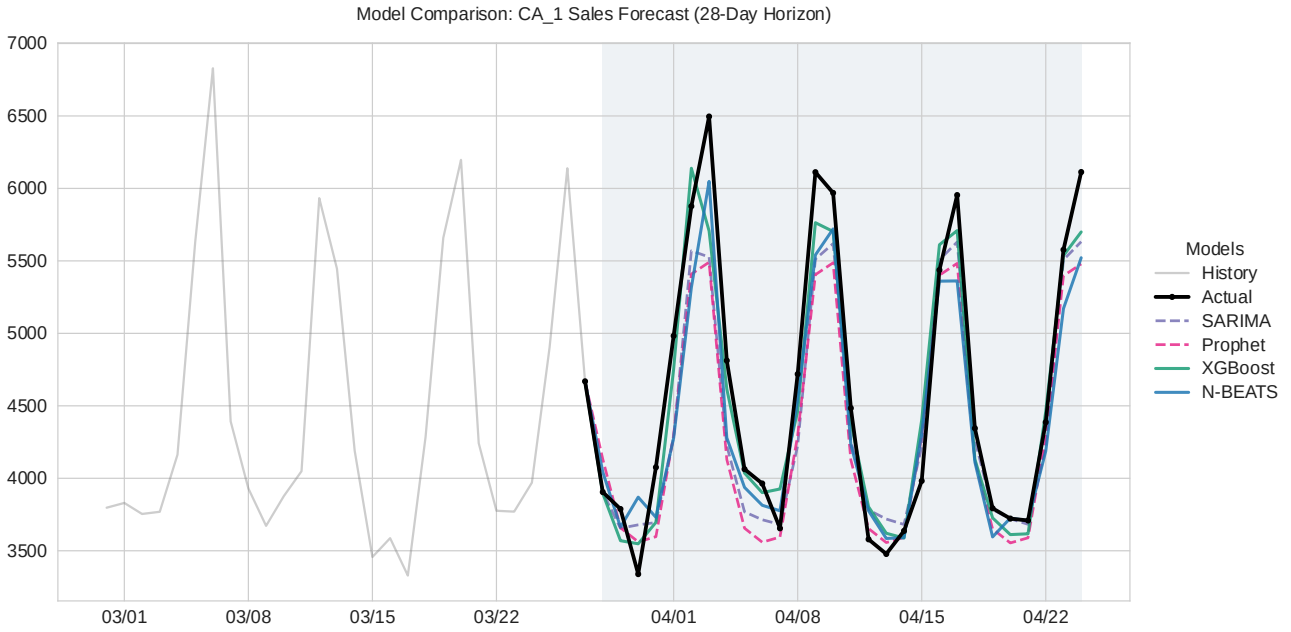

Figure 3. Comparison of predicted vs. actual results for each model in the 28-day forecast window
图 3. 28 天预测窗口中各模型预测结果与实际结果对比图

Table 1. Summary of model evaluation indicators
表 1. 模型评估指标汇总表

预测窗口	评价指标	SARIMA	Prophet	N-BEATS	XGBOOST
7-Day	MAE	404.41	459.41	348.14	299.69
	RMSE	507.16	540.45	413.51	374.21
	MAPE (%)	8.18	9.26	7.39	6.12
14-Day	MAE	388.04	455.9	325.85	251.74
	RMSE	464.53	517.85	374.44	311.01
	MAPE (%)	7.85	9.25	7.09	5.15
28-Day	MAE	271.77	332.77	306.79	210.41
	RMSE	359.03	412.78	383.07	264.14
	MAPE (%)	5.61	6.85	6.42	4.47

调优的深度学习模型。相比之下，Prophet 由于其平滑假设带来的结构刚性，较难刻画销量的微观波动与极值，在各时间跨度下的表现整体上不及其他模型。

3.2. SKU 级销量预测

为进一步验证模型在细分场景下的适应性，本节从 CA_1 店铺的 3000 余个 SKU 中随机抽取 150 个样本，按销量水平分层后开展 SKU 级销量预测实验，重点分析模型在不同稀疏度数据下的性能差异。这 150 个 SKU 覆盖所有三大品类 FOODS、HOUSEHOLD、HOBBIES。基于总销量，进行三分位分层，分为 High、Medium、Low 三组，每组 50 个 SKU。各组的 SKU 稀疏度如表 2 所示。

由表 2 可见，从 High 到 Low 组，平均日销量急剧下降(26.8→1.1)，零销售占比显著上升(14.8%→

79.6%)，平均需求间隔从 1.18 天延长至 5.89 天。Low 组呈现典型的间歇性需求特征(大量零值 + 偶发峰值)，这对模型处理零膨胀与非连续突变的能力提出了更高要求。

Table 2. Comparison of the sparsity indicators of the three sets of sales data
表 2. 三组销量数据稀疏度指标对比

销量组	SKU 数量	平均日销量	零销售占比	平均需求间隔
High	50	26.8	14.8%	1.18
Medium	50	9.2	42.5%	2.14
Low	50	1.1	79.6%	5.89

店铺级聚合后零值少且更关注相对误差，故以 MAPE 为主；而 SKU 级零值多使 MAPE 不稳定，改用更稳健的 RMSSE 并辅以 MAE/RMSE，同时为控制计算成本采用 3 次滚动回测(每次预测未来 28 天并留 28 天验证用于早停与调参)，最终报告三次结果的均值±标准差。RMSSE 的 scale 由训练集上 $\sqrt{\frac{1}{T-1}\sum_{t=2}^{T}\left(y_t-y_{t-1}\right)^2}$ 计算。各销量组的主要结果如表 3 所示。

Table 3. Average performance statistics table of each sales group for 3 consecutive rolling tests
表 3. 各销量组 3 次滚动回测平均性能统计表

销量组	模型	MAE	RMSE	RMSSE
High	XGBoost	18.4 ± 2.1	28.7 ± 3.4	**0.758 ± 0.104**
	N-BEATS	19.8 ± 2.6	30.9 ± 4.1	0.792 ± 0.108
	Prophet	22.6 ± 3.0	35.2 ± 4.7	0.908 ± 0.132
	SARIMA	23.8 ± 3.3	37.1 ± 5.0	0.955 ± 0.140
Medium	XGBoost	6.8 ± 1.4	11.2 ± 2.3	**0.862 ± 0.162**
	N-BEATS	7.4 ± 1.7	12.6 ± 2.8	0.918 ± 0.178
	Prophet	8.9 ± 2.0	15.1 ± 3.5	1.142 ± 0.212
	SARIMA	9.4 ± 2.2	16.3 ± 3.9	1.208 ± 0.228
Low	XGBoost	0.92 ± 0.31	2.1 ± 0.7	**1.178 ± 0.332**
	N-BEATS	1.12 ± 0.41	2.6 ± 0.9	1.348 ± 0.368
	Prophet	1.38 ± 0.48	3.3 ± 1.1	1.712 ± 0.418
	SARIMA	1.48 ± 0.52	3.6 ± 1.2	1.865 ± 0.445
Average	XGBoost	8.7 ± 1.5	14.0 ± 2.4	**0.933 ± 0.288**
	N-BEATS	9.5 ± 1.8	15.4 ± 2.9	1.019 ± 0.312
	Prophet	11.0 ± 2.3	17.9 ± 3.8	1.254 ± 0.378
	SARIMA	11.6 ± 2.5	19.0 ± 4.1	1.343 ± 0.402

表 3 的 3 次滚动回测结果显示，随着 SKU 销量从 High 降至 Low，四类模型的 MAE、RMSE 与 RMSSE 整体上升，表明零膨胀与间歇性增强显著提升预测难度。

High 组(日均 26.8、零占比 14.8%)序列更连续规律，XGBoost 获最低 RMSSE (0.758 ± 0.104)，N-BEATS (0.792 ± 0.108)紧随其后，Prophet (0.908 ± 0.132)与 SARIMA (0.955 ± 0.140)相对落后；Medium 组

(日均 9.2、零占比 42.5%)稀疏度上升，模型分化加剧，XGBoost 仍最优(0.862 ± 0.162)，N-BEATS (0.918 ± 0.178)与统计模型(Prophet 1.142 ± 0.212、SARIMA 1.208 ± 0.228)差距拉大；在 Low 组(日均 1.1、零占比 79.6%、需求间隔 5.89 天)高度间歇场景下，XGBoost (1.178 ± 0.332)较 N-BEATS (1.348 ± 0.368)提升约 12.6%，并大幅领先 Prophet (1.712 ± 0.418)与 SARIMA (1.865 ± 0.445)，体现树模型在处理大量零值与非线性特征交互时的鲁棒性，而传统统计模型受线性/加性假设限制，难以适配“长零 + 峰值”形态。三组加权平均下，XGBoost 以 RMSSE (0.933 ± 0.288)排名第一，N-BEATS (1.019 ± 0.312)第二，Prophet (1.254 ± 0.378)与 SARIMA (1.343 ± 0.402)随后。结论：在 SKU 级尤其稀疏零膨胀场景中，梯度提升树方法更稳定准确。随后本文还选取了 High 与 Low 组进行了特征消融证明时间日历特征与价格特征的重要性，结果如表 4 所示。

Table 4. Ablation experiments on price and calendar features
表 4. 价格特征与时间日历特征消融实验

销量组	模型	配置	MAE	RMSE	RMSSE	相对完整特征恶化% (RMSSE)
High	XGBoost	完整特征	18.4	28.7	0.758	-
	XGBoost	无价格特征	20.2	31.6	0.832	(+9.8%)
	XGBoost	无日历特征	19.5	30.4	0.802	(+5.8%)
	XGBoost	仅单变量	21.3	33.2	0.878	(+15.8%)
	N-BEATS	完整特征	19.8	30.9	0.792	-
	N-BEATS	无价格特征	21.1	32.9	0.845	(+6.7%)
	N-BEATS	无日历特征	20.5	32.0	0.822	(+3.8%)
	N-BEATS	仅单变量	21.6	33.7	0.868	(+9.6%)
Low	XGBoost	完整特征	0.92	2.1	1.178	-
	XGBoost	无价格特征	1.26	2.9	1.512	(+28.4%)
	XGBoost	无日历特征	1.08	2.5	1.342	(+13.9%)
	XGBoost	仅单变量	1.48	3.4	1.712	(+45.3%)
	N-BEATS	完整特征	1.12	2.6	1.348	-
	N-BEATS	无价格特征	1.24	2.9	1.492	(+10.7%)
	N-BEATS	无日历特征	1.18	2.8	1.412	(+4.7%)
	N-BEATS	仅单变量	1.29	3.1	1.548	(+14.8%)

消融实验表明外生特征对 SKU 级预测精度贡献显著，且在间歇性最强的 Low 组作用最大。去除价格特征后，XGBoost 在 Low 组 RMSSE 恶化 28.4% (High 组仅 9.8%)，说明低销量 SKU 的需求峰值更依赖促销价格驱动；去除日历特征的影响次之，体现周末/节假日对需求节奏的规律性作用。在仅使用单变量历史序列时，两类模型在 Low 组退化最明显(XGBoost-45.3%, N-BEATS-14.8%)，凸显稀疏场景下特征工程对弥补有效样本不足的重要性。在 28 天预测任务中，模型表现随销量下降而分化：High 组序列较平稳，N-BEATS 与 XGBoost 差距较小；Medium 组稀疏度上升后，XGBoost 优势开始扩大；Low 组高度零膨胀与偶发峰值主导时，XGBoost 更能利用树分裂捕捉非线性促销交互并控制误差，显著优于其他模型。Prophet、SARIMA 等统计模型在 Low 组受限，主要因线性/加性假设难以适配“长零 + 峰值”的非连续需求模式。

3.3. 深度学习模型在店铺级和 SKU 级预测表现的进一步讨论

本实验中，以 N-BEATS 为代表的深度学习模型整体不及 XGBoost，店铺级长期预测被 SARIMA 反超，SKU 级中低销量组随数据稀疏度上升差距扩大，根本原因是零售需求序列的数据生成特性，与各模型归纳偏置、特征利用方式及有效样本规模匹配度不足。

店铺级序列兼具稳定周周期与促销、节假日等引发的短时结构性突变(急跌–反弹)。N-BEATS 依赖基函数分解与端到端全局拟合，擅长连续规则趋势/周期外推，但对非周期冲击项易出现相位滞后与幅度低估，正则化会弱化局部尖峰；XGBoost 借助显式滞后、日历等强特征，通过树分裂捕捉突变更精准。预测跨度拉长后，误差由稳定季节性主导，SARIMA 以简洁季节差分结构高效外推周周期，参数少、估计方差低，长期预测更稳定。

SKU 级叠加零膨胀与间歇性需求，大量零值背景夹杂少量促销峰值，导致信噪比低、有效非零样本稀缺。深度学习模型需同时学习需求发生判别、幅度估计及与价格/事件的交互，数据效率要求高，易低估峰值或拟合不稳定；XGBoost 可自然表达阈值效应与高阶交互，在稀疏场景更稳健。

所以在“强周期 + 局部突变”且样本有限、或高稀疏间歇需求的零售序列中，树模型与传统统计模型更高效稳定；扩充样本并采用零膨胀损失、注意力架构、跨 SKU 预训练等策略，深度学习模型仍有提升空间。

4. 总结

本研究证实了 XGBoost 在零售销量预测中的综合优势。在店铺级聚合预测中，XGBoost 在各时间窗口下误差最低，整体优于 SARIMA、擅长平滑的 Prophet 与深度模型 N-BEATS。在 SKU 级预测中，随着数据稀疏度和间歇性增加，模型性能分化显著：传统统计模型 SARIMA 难以应对零膨胀数据，N-BEATS 受限于样本规模泛化性下降，而 XGBoost 凭借对稀疏特征和非线性交互的高效利用，展现出最佳鲁棒性。在资源有限的中小零售企业中，XGBoost 结合简单特征工程即可获得较好效果，优先采用 XGBoost 更为稳妥；在具备更强端到端建模需求或可进一步扩充训练数据的场景下，N-BEATS 可作为重要的深度学习备选方案，为后续模型优化与实际部署提供参考。

参考文献

[1] Bzdok, D., Altman, N. and Krzywinski, M. (2018) Statistics versus Machine Learning. *Nature Methods*, **15**, 233-234. https://doi.org/10.1038/nmeth.4642

[2] 袁瑞萍, 魏辉, 傅之家, 等. 融合 CNN 和 WDF 模型的电商企业商品销量预测研究[J]. 计算机工程与应用, 2025, 61(2): 335-343.

[3] 霍佳震, 徐骏, 陈铭洲. 基于 EEMD-HW-GBDT 模型的零售商品销量多步预测[J]. 工业工程与管理, 2024, 29(1): 133-141.

[4] 向易, 丛丽丽, 王玮鹏, 等. 层次时间序列预测方法与应用综述[J]. 计算机科学, 2025, 52(S2): 550-556.

[5] Dubey, A.K., Kumar, A., García-Díaz, V., Kumar Sharma, A. and Kanhaiya, K. (2021) Study and Analysis of SARIMA and LSTM in Forecasting Time Series Data. *Sustainable Energy Technologies and Assessments*, **47**, Article ID: 101474. https://doi.org/10.1016/j.seta.2021.101474

[6] Taylor, S.J. and Letham, B. (2018) Forecasting at Scale. *The American Statistician*, **72**, 37-45. https://doi.org/10.1080/00031305.2017.1380080

[7] 李扬, 肖勇波, 辛诚, 等. 信息不完全下基于关联匹配的工程物资需求预测[J/OL]. 系统管理学报, 2025: 1-20. https://link.cnki.net/urlid/31.1977.N.20250814.1553.002, 2026-02-06.

[8] 成耀, 张铎, 周宇, 何金凤, 程实. 基于模糊聚类的电商企业不平衡财务数据风险预测方法[J]. 电子商务评论, 2025, 14(1): 640-647.

[9] Oreshkin, B.N., Carpov, D., Chapados, N., *et al*. (2019) N-BEATS: Neural Basis Expansion Analysis for Interpretable

Time Series Forecasting.

[10] 李坤, 陈剑钧, 李国胜, 等. 小样本学习研究综述[J]. 机电工程技术, 2025, 54(6): 160-168.

[11] 范黎林, 杨凯, 毛文涛, 等. 融合结构化信息与时序演化信息的多变量间歇性时间序列预测[J]. 控制与决策, 2024, 39(1): 263-270.

[12] Qian, W., Rolling, C.A., Cheng, G. and Yang, Y. (2022) Combining Forecasts for Universally Optimal Performance. *International Journal of Forecasting*, **38**, 193-208. https://doi.org/10.1016/j.ijforecast.2021.05.004