

基于用户行为序列的电商需求销量预估研究

田珈瑞

贵州大学数学与统计学院, 贵州 贵阳

收稿日期: 2026年7月8日; 录用日期: 2026年7月22日; 发布日期: 2026年9月9日

摘要

在电商场景下精准预估供应链销量, 对于优化冷链配置、提升配送效率与降低库存周转成本具有重要的现实意义。本文基于淘宝开源电商数据集, 突破传统仅依赖历史销量的局限, 将用户在线行为(点击、加购、收藏)纳入销量预估框架。研究发现, VAR模型有效地揭示了流量转化过程中各个层级的动态传导机制, 而Transformer模型则精准捕捉了由大促活动引发的非线性需求脉冲。二者结合提升了短期订单量的预估精度。经验证, VAR模型能有效捕捉变量间动态联动关系, Transformer模型在非线性的拟合上表现更优, 二者形成互补。模型结果显示, “双十二”类促销活动能有效促进电商市场发展, 相关企业应提前布局应对物流高峰。本文旨在为电商平台与物流企业探索轻量化、可落地的需求预估方案, 为供应链协同决策提供数据支撑。

关键词

淘宝开源数据, 用户行为分析, 销量预估, VAR模型, Transformer模型

Research on E-Commerce Demand and Sales Forecast Based on User Behavior Sequences

Jiarui Tian

School of Mathematics and Statistics, Guizhou University, Guiyang Guizhou

Received: July 8, 2026; accepted: July 22, 2026; published: September 9, 2026

Abstract

Accurately predicting sales volume and supply chain throughput in e-commerce scenarios is of great practical significance for optimizing cold-chain allocation, improving distribution efficiency, and reducing inventory and turnover costs. Based on an open-source Taobao e-commerce dataset, this study breaks through the limitations of traditional sales forecasting by incorporating users' online behaviors (clicks, add-to-cart, favorites) into the forecasting framework. It is found that the

文章引用: 田珈瑞. 基于用户行为序列的电商需求销量预估研究[J]. 电子商务评论, 2026, 15(9): 206-215.

DOI: 10.12677/ecl.2026.159988

VAR model effectively reveals the dynamic transmission mechanism across all levels in the traffic conversion process, while the Transformer model accurately captures nonlinear demand shocks caused by livestreaming e-commerce or major promotion campaigns. Their combination improves the accuracy of short-term order volume prediction, providing a data-driven decision paradigm for omni-channel inventory allocation and front warehouse operations of e-commerce commodities. Verification shows that the VAR model effectively captures the dynamic linkage between variables, whereas the Transformer model performs better in nonlinear fitting, and the two models are complementary. The results indicate that shopping carnivals such as Double 12 effectively promote the positive development of the e-commerce market. Relevant enterprises and authorities should make advance arrangements to cope with logistics peaks during promotions. This study provides a lightweight and implementable solution for demand and sales forecasting for e-commerce platforms and logistics enterprises, and also offers data support for collaborative decision-making in the e-commerce supply chain.

Keywords

Taobao Open-Source Data, User Behavior Analysis, Sales Forecast, VAR Model, Transformer Model

Copyright © 2026 by author(s) and Hans Publishers Inc.

This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

<http://creativecommons.org/licenses/by/4.0/>



Open Access

1. 引言

1.1. 研究背景与问题提出

在“互联网+”深入推进与数字技术快速迭代的背景下，网络购物已全面融入日常生活。促销活动的新业态深刻改变了商品流通模式。电商消费需求呈现高频波动、强季节性与突发性特征，尤其在“双十二”、节假日等节点，流量与订单常出现脉冲式增长，对供应链响应速度提出极高要求[1]。

以生鲜等易腐商品为例，其线上交易规模扩大倒逼供应链从“人找货”向“货找人”转变。销量预估不仅是物流量预测的关键，更是库存周转、冷链配置的决策基础。然而，传统预测方法(如移动平均、简单线性回归)多基于平稳性假设，难以捕捉多行为指标间的动态传导与突发非线性波动。因此，依托海量历史交易与行为数据，构建精准、稳健的销量预估模型，推动供应链从“经验驱动”向“数据驱动”转型，具有重要价值。

然而，消费者需求的不确定性、数据的高噪声与强波动性，使得传统预测方法(如移动平均、简单线性回归)难以适配电商复杂场景。这类方法多基于平稳性假设，无法捕捉多行为指标间的动态传导关系与突发非线性波动，预测偏差易导致库存积压、缺货、冷链资源浪费、配送延迟等问题，既抬高运营成本，也影响用户体验。因此，依托海量历史交易与行为数据，构建精准、稳健的销量预估模型，推动供应链从“经验驱动”向“数据驱动”转型，具有重要理论与实践价值[2]。

1.2. 研究方法与技术路线

针对商品销量建模的复杂性，单一的销量预估模型存在着较大局限。例如线性模型难以捕捉突发的非线性波动，而复杂的非线性模型有时会缺乏对变量间长期依赖关系的可解释性。

为了解决这个矛盾，我们提出了双模型对比与互补的研究思路，分别构建 VAR(向量自回归)模型与

Transformer 模型。

1. VAR 模型是一种经典的计量经济学模型，它更加擅于处理多变量的时间序列数据。当数据呈现出某种周期性变化规律时通过引入滞后变量，来有效的捕捉电商消费指标(如订单量、点击量)与供应链销量间的动态联动关系和滞后效应，对于分析宏观趋势和政策效应具有非常重要的意义。VAR 模型不依赖先验的经济理论假设，而是对数据本身进行假设检验，来判断模型的具体形式[3]。

2. Transformer 模型基于 Encoder-Decoder 架构与自注意力机制，具备强大的非线性拟合能力。在高频波动场景下，能挖掘数据中潜在规律，适合短期精准预估与峰值捕捉。在电商订单这种高频波动的场景下，Transformer 可以通过自注意力机制结合多层感知机模型，挖掘数据中潜在的隐藏规律，适合进行短期的精准销量预估与峰值捕捉[4]。

1.3. 研究目的与意义

本文通过 VAR 模型和 Transformer 模型的交叉验证，探索电商消费驱动下的区域商品供应链销量最优预估方案。理论上，验证了统计模型与机器学习模型在处理电商时序数据上的适用边界；实践上，为物流企业提供轻量化工具，通过“趋势与波动”分解处理，辅助企业科学进行冷链仓储布局与运力调度。通过“趋势与波动”分解处理的销量预估，使得企业能够更加科学地进行冷链仓储布局、运力调度与库存管理，从而降低库存流通和周转成本，提升供应链的整体效率[5]。

2. 数据来源与预处理

本文选取淘宝开源电商数据集作为核心来源。该数据集涵盖了从 2014 年 11 月 24 日到 2014 年 12 月 24 日期间的真实用户消费行为。虽然数据年份较早，但其包含的“用户行为漏斗”逻辑(点击 - 收藏 - 加购 - 购买)在电商运营中具有普适性，仍能有效模拟电商环境下的流量转化特征，为模型训练提供坚实的数据基础。

2.1. 数据来源与研究价值

我们在构建销量预估模型的过程中，需要获取高质量的多维度数据。考虑到商业机密与用户的数据隐私问题，我们很难获取真实、完整的电商交易与物流链路数据，而这在一定程度上制约了相关领域的研究深度。

在本文中，我们选取了淘宝的开源电商数据集作为核心的数据来源[6]。这个数据集有着较高的可信度与研究价值，它在保护用户和商家隐私的前提下，涵盖了约 1.2 亿条从 2014 年 11 月 24 日到 2014 年 12 月 24 日期间的真实用户消费行为。通过这个数据集，我们能够有效地模拟电商运营环境，为模型训练性能测试与验证提供较为坚实的数据基础。

2.2. 指标体系

高质量多维度数据是提升预测精度的前提。受商业机密与隐私保护限制，真实完整的电商全链路交易与物流数据较难获取，一定程度制约了相关研究开展。

本文构建了基于“用户行为漏斗(Funnel)”的销量预估指标体系：

1. 用户消费行为指标：包括点击(Click)、收藏(Favorite)、加购(Cart)和购买(Buy)。这些指标反映了市场的潜在需求与显性需求。
2. 流量层(Awareness & Interest)对应了“点击(Click)”数据，能够反映商品的曝光度与市场引流能力；
3. 意向层(Search & Action)对应着“收藏(Favorite)”与“加购(Cart)”数据，这是最核心的前置指标(Leading Indicator)，直接决定了转化率(CVR)的上限；

4. 成交层(Purchase)对应的是“购买(Buy)”数据，即最终的 GMV 产出。

我们通过挖掘这四个层级的动态关系，模型能更敏锐地感知电商市场的供需变化。该数据集样本规模充足，时间跨度完整，能够真实模拟电商环境下产品的流通特征[7]。

2.3. 数据分布特征可视化分析

在建模之前，对原始数据进行探索性数据分析至关重要。通过对用户行为的统计分析，我们能够直观地观察到电商消费的周期性规律与波动特征，这为后续选择 VAR 和 Transformer 模型提供了直观依据。

1. 用户行为偏好分析

首先，我们对用户日均各类行为占比进行了统计。如图 1 所示，“购买”(Buy)行为虽然直接驱动物流，但在总交互行为中占比较低；而“点击”(Click)行为占比最高，反映了用户广泛的浏览兴趣。此外，“加购”与“收藏”行为占据了相当比例，这些行为往往是销量预估未来供应链销量的重要前置指标(Leading Indicators)。

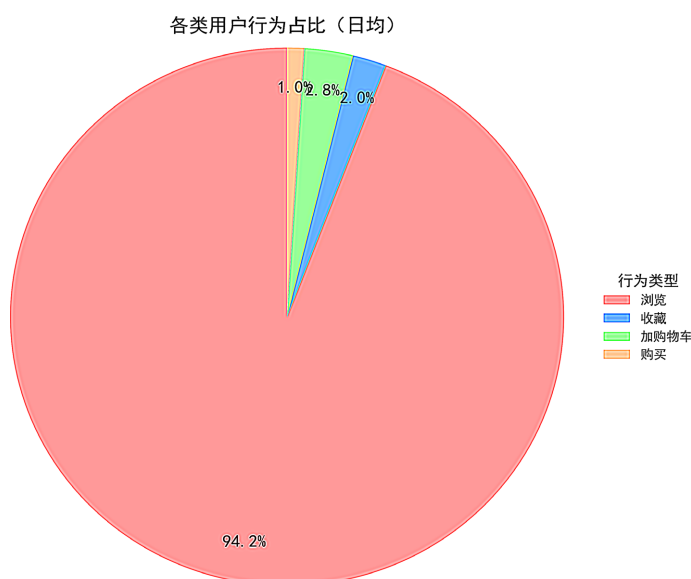


Figure 1. Pie chart of daily average user behavior distribution

图 1. 用户日均行为占比的饼图

2. 时间序列趋势分析

为了揭示需求的时间维度特征，我们绘制了每日用户行为总数趋势图与每小时用户行为趋势图。

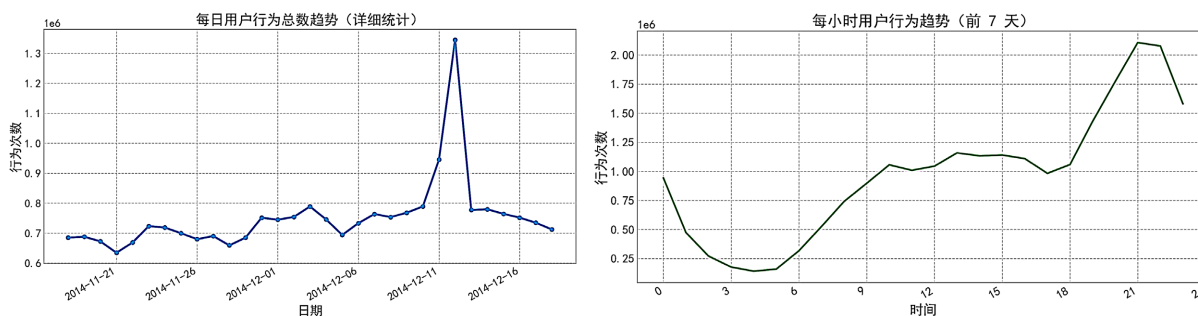


Figure 2. Trend chart of total daily and hourly user behaviors

图 2. 每日与每小时用户行为总数趋势图

如图 2 所示，用户行为呈现出显著的“脉冲式”特征。在特定时间节点(如双 11 预热期或平台大促日)，数据出现爆发式增长。这种“尖峰”特征是电商区别于传统零售的典型标志，反映了流量集中爆发对供应链的冲击。

图 3 的小时级趋势则完美契合了“晚高峰消费”场景。20:00~22:00 的流量峰值与当前主流电商及促销活动的流量逻辑一致，证明了数据具有极强的电商场景代表性[8]。

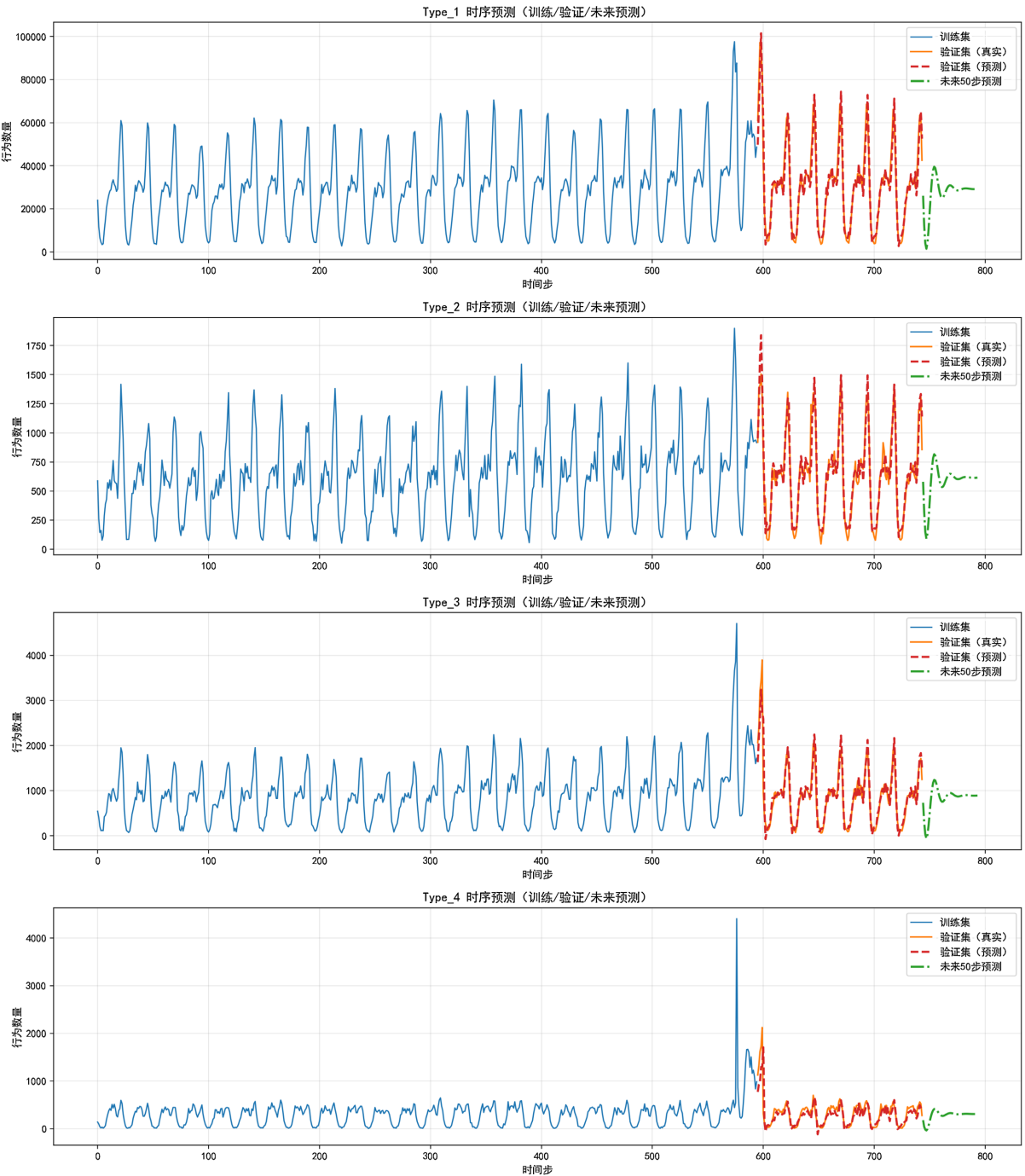


Figure 3. Sales estimation results of time series model (VAR)

图 3. 时间序列模型(VAR)的销量预估结果图

2.4. 数据预处理

由于原始电商数据往往存在缺失、异常以及量纲不一的问题,且 VAR 模型要求序列平稳,Transformer 模型要求输入数据归一化,因此必须进行严格的数据预处理。具体流程如下:

1. 数据清洗

采用线性插值填补缺失值,保障时序连续;人工甄别异常值,仅剔除记录错误噪声,保留大促等真实业务峰值。

2. 数据划分

按时间顺序拆分训练、验证集。VAR 使用 595 条训练、149 条验证小时数据;Transformer 因滑动窗口构造样本,数据集调整为 588 条训练、148 条验证数据。

3. 平稳性处理与标准化

经 ADF (Augmented Dickey-Fuller)检验,部分原始序列非平稳,对其差分至满足 VAR 平稳前提;针对行为、订单数据量纲悬殊问题,采用 Min-Max 归一化将特征映射至[0, 1]区间,加速网络收敛、提升预测精度,预测输出后逆缩放还原真实销量量级。

3. 销量预估模型构建

3.1. VAR 模型构建与检验

VAR 模型由 Sims 于 1980 年提出,是面向多维平稳时序的无约束线性模型,无需预设经济理论约束,将各内生变量表示为自身与其余变量滞后项的线性组合,适用于本文电商行为指标与供应链销量的联动关系分析,基础形式为:

$$Y_t = A_1 Y_{t-1} + A_2 Y_{t-2} + \cdots + A_p Y_{t-p} + \varepsilon_t$$

其中, Y_t 是 $k \times 1$ 维的内生变量列向量(本文中 $k = 4$, 代表四种供应链销量指标); p 是滞后阶数; A_i 是 $k \times k$ 维的系数矩阵; ε_t 是白噪声序列向量。

在构建模型前,必须对数据进行严格的统计检验,以确保模型的有效性。

平稳性检验(ADF Test): 浏览、加购序列原始非平稳,对其差分直至通过单位根检验,剔除趋势与季节成分,满足 VAR 平稳前提[9]。

滞后阶数确定(Lag Order Selection): 滞后阶数过小易引发残差自相关,过大会降低参数估计精度;结合 AIC、BIC 准则确定最优滞后阶($p = 2$)。管理学层面,滞后 1 阶对应用户即时冲动转化,滞后 2 阶反映加购比价冷静期与仓储履约时滞。

残差检验: 对各变量残差进行正态性检验,结果显示残差序列不存在显著的自相关性,满足建模要求。

3.2. Transformer 模型构建

VAR 擅长拟合线性关系,但电商数据存在大量非线性规律,因此本文引入带自注意力机制的 Transformer 时序模型。该模型采用编码器-解码器架构,依托多头自注意力捕获时序长距离依赖,无需循环单元;由多头注意力、前馈网络、残差连接、层归一化模块组成,搭配位置编码注入时序信息。编码器提取特征上下文表征,解码器实现预测。本文选用编码器型时序 Transformer,可挖掘点击、加购、收藏、下单多变量动态非线性关联,对销量峰值与短期波动预测效果更优。

3.3. 关键参数设定与技术细节

结合本文数据特征,Transformer 超参数设置如下:输入时序窗口长度设为 8,依托历史 8 个时步数

据捕捉短期波动；输入特征与标签采用 Z-score 标准化，数据集按时序划分为训练、验证集。网络每层隐藏神经元 256 个，隐藏层激活函数选用 ReLU 以缓解梯度消失、加速收敛；优化器采用 Adam，学习率 0.001，适配高噪声电商时序数据；损失函数为 MSE，最小化预测值与真实销量误差平方和。模型总参数量 1,062,916，通过正则化与早停机制抑制过拟合；训练以反向传播迭代更新权重，最小化损失函数实现供应链销量高精度拟合[10]。

4. 模型训练与结果分析

为了验证基于淘宝开源数据构建的 VAR 与 Transformer 模型在区域商品供应链销量预估中的有效性，本节将对模型进行实证训练，并对销量预估结果进行深入的对比分析。通过观察销量预估曲线的拟合度与损失函数的变化趋势，评估模型在处理线性趋势与非线性波动时的表现。

4.1. VAR 模型训练结果

VAR 是线性统计模型，优势在于刻画多变量系统长期均衡关系。本文构建 VAR (2)模型，含 4 个核心变量，总参数 36；全部特征根落在单位圆内，系统平稳。模型残差呈偏态尖峰分布，契合电商数据特征：平日销量集中，大促、节假日易出现峰值。

为检验 VAR 残差是否服从白噪声，本文进行了残差 ACF、PACF 检验。滞后 1~20 阶内，四类商品残差相关系数大多落在 95%置信区间内；仅 Type_1、Type_4 滞后 1 阶、加购滞后 2 阶小幅超出区间，无系统性衰减或周期特征，符合白噪声特征。其中 Type_1、Type_4 滞后 1 阶溢出系数绝对值小于 0.1，高阶迅速回落，属于随机扰动，不影响白噪声判定；收藏、加购全阶相关系数均稳定落在置信区间内。结果说明 VAR 已充分提取序列线性关联，残差无剩余可建模自相关，模型诊断通过。综上，四类商品 VAR 模型残差序列基本满足白噪声假设，模型设定合理，可用于后续的脉冲响应分析与预测。

如图 4 所示，VAR 模型的销量预估曲线能够较好地跟随真实值的整体走势。这表明 VAR 模型成功捕捉到了电商消费指标与供应链销量之间的动态联动关系。虽然在部分波动剧烈的节点(如突发的流量高峰)存在一定的滞后性，但其整体趋势销量预估的稳定性极高，非常适合用于宏观层面的物流资源规划。

4.2. Transformer 模型训练过程与结果

相比于 VAR 模型的线性解释，Transformer 模型致力于通过非线性拟合来逼近数据的本质规律。本节将重点展示 Transformer 模型的训练收敛过程及其最终的销量预估表现。随着训练的进行，训练集损失值迅速下降并趋于平稳，表明模型正在有效学习数据中的特征，Loss 验证集损失值同样呈现下降趋势，且与训练 Loss 趋于一致，未出现明显的发散现象。

早停机制(Early Stopping)，模型在第 76 轮训练时触发了早停机制，此时验证 Loss 达到最小值 0.1751，训练 Loss 为 0.0324。这说明模型在未发生过拟合(Overfitting)的情况下找到了最优解，此时得到的模型泛化能力较强。

Transformer 模型在“点击”和“加购物车”的波峰处表现优异。在电商大促场景中，这对应着对“爆款单品(Hit Product)”或“直播间引流款”销量爆发的精准预判。这种对“长尾效应”中极端值的捕捉能力，能有效防止电商平台出现“爆单缺货”或“流量浪费”的运营事故。

4.3. 模型对比与互补性分析

对比训练结果可见两类模型各有侧重、优势互补：VAR 模型参数少、可解释性佳，擅长挖掘变量因果关系，适配中长期宏观销量趋势预测，但难以捕捉短期剧烈波动；Transformer 依托百万级参数(1062916)，

非线性拟合能力突出，短期预测精度更高，却存在黑盒、难以解释其经济逻辑的短板。实操中可将 VAR 作为销量预测基准，再用 Transformer 修正短期波动偏差，二者联用能兼顾宏观趋势与微观波动，为区域物流企业提供更可靠的销量预测与决策依据。

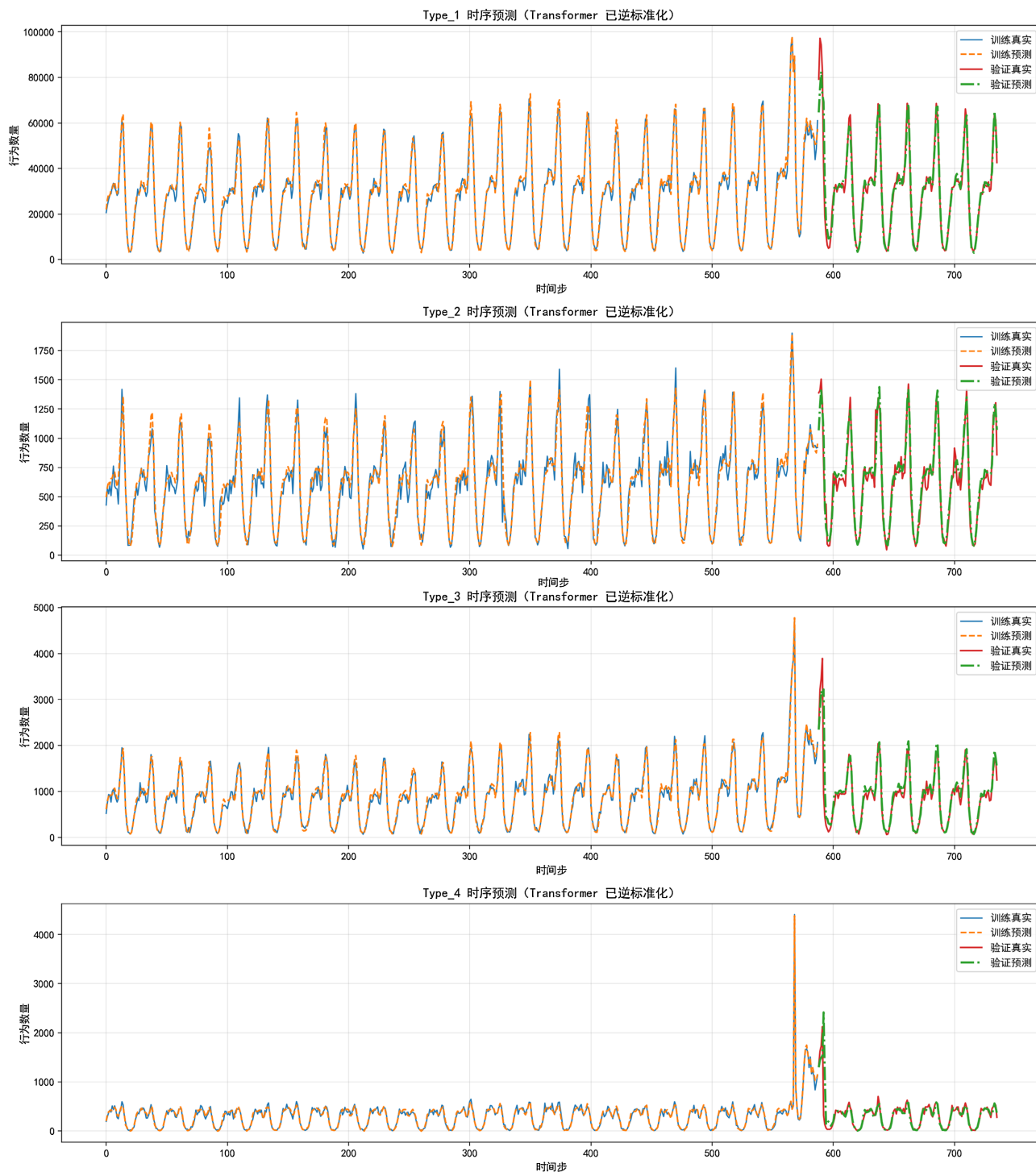


Figure 4. Sales forecast result graph of the transformer model

图 4. Transformer 模型的销量预估结果图

5. 结论

5.1. 研究结论

本文基于淘宝开源电商数据集，围绕电商用户行为序列驱动的商品销量预估问题，构建 VAR 模型与 Transformer 模型并开展实证对比，形成以下主要结论：

本章系统归纳研究结论并阐述供应链应用价值：点击 - 收藏 - 加购 - 购买的用户行为漏斗可完整刻画需求演变，收藏、加购是销量预测核心先行指标；小时、日度销量存在显著时段与大促波动特征，可作为模型预测先验。商家可依托收藏、加购走势提前备货，优化人货场匹配。Transformer 依托自注意力机制，非线性拟合与峰值捕捉能力突出，短期预测、大促脉冲需求识别效果更佳，企业可据此提前开展预售下沉、前置仓补货，提升供应链响应效率[11]。

模型结果显示，在“双十二”这类消费购物狂欢节能够有效促进电商市场的积极发展，相关企业及部门应提前布局，以应对促销期间的物流高峰。数据支撑的协同决策，研究结果为电商平台与区域物流企业的协同决策提供数据支撑。通过预测未来的供应链销量高峰，企业可以更科学地进行冷链仓储布局、动态调整运力配置、优化库存管理，从而显著降低流通库存周转成本，提升供应链整体效率。“双十二”等大型促销活动可显著刺激电商需求增长，预测结果可为库存决策提供参考输入，应对高峰履约压力，能够为电商全渠道库存管理、供应链协同决策提供数据支撑[12]。

本文采用基础 VAR 与 Transformer 结构，未来可从三方面进一步优化：一是引入 LSTM、GRU、Informer 等更适用于长时序的模型，提升复杂波动拟合能力；二是融合天气、节假日、促销强度等外部特征，增强模型鲁棒性；三是探索加权融合、残差修正等组合预测策略，进一步提升预测精度与泛化能力。

5.2. 局限性与展望

本文使用的数据集采集于 2014 年，虽然涵盖了完整的“双十二”大促周期且数据质量较高，但电商环境在近年来发生了巨大变化。因此，本文结论主要验证了方法论的可行性，在实际应用中，建议结合近年的实时数据进行模型微调，以确保预测结果的现实指导意义。

未来研究可从三方面优化：一是引入 LSTM、Informer 等更适用于长时序的模型；二是融合天气、节假日等外部特征；三是探索加权融合策略，进一步提升预测精度。未来研究可进一步通过消融实验量化各行为特征(如收藏、加购)对预测精度的边际贡献。

参考文献

- [1] 艾乐怡, 刘珉熙, 李浩, 杨新平. 基于贝叶斯模型与 VAR 模型的蔬菜销量与定价预估[J]. 数学建模及其应用, 2024, 13(2): 57-65.
- [2] 周正. 基于 VAR 模型的气候变化对水资源的影响研究[J]. 黑龙江环境通报, 2024, 37(8): 47-49.
- [3] 祝元浩, 常国宾, 杨木森. 应用赤池信息量准则优选惯性元器件的随机误差模型[J]. 中国惯性技术学报, 2024, 32(2): 180-186.
- [4] Wen, Q., Zhou, T., Zhang, C., Chen, W., Ma, Z., Yan, J., et al. (2023) Transformers in Time Series: A Survey. *Proceedings of the 32nd International Joint Conference on Artificial Intelligence*, 19-25 August 2023, 6778-6786. <https://doi.org/10.24963/ijcai.2023/759>
- [5] 刘电威, 牛龙龙. 基于混沌 RBF 神经网络的电商市场销量预测模型技术研究[J]. 湘潭大学学报(自然科学版), 2023, 45(5): 57-64.
- [6] 淘宝用户行为数据集[EB/OL]. <https://www.heywhale.com/mw/dataset/5ef7024363975d002c9235d3>, 2026-04-06.
- [7] 周文. 基于 Transformer-LSTM 算法的电子商务平台用户行为预测研究[J]. 电子商务评论, 2025, 14(7): 1902-1910.
- [8] 徐浩森. 基于时间序列模型的动态零售环境下产品月销量预测——基于多个模型的比较分析[J]. 统计学与应用, 2024, 13(2): 180-186.

-
- 2023, 12(6): 1746-1762.
- [9] 武玉英, 蒋国瑞. 融合 LSTM 与时间序列模型的电商销量预测方法[J]. 计算机工程与应用, 2019, 55(15): 177-184.
- [10] Lim, B., Arık, S.Ö., Loeff, N. and Pfister, T. (2021) Temporal Fusion Transformers for Interpretable Multi-Horizon Time Series Forecasting. *International Journal of Forecasting*, **37**, 1748-1764. <https://doi.org/10.1016/j.ijforecast.2021.03.012>
- [11] Zhou, H., Zhang, S., Peng, J., Zhang, S., Li, J., Xiong, H., et al. (2021) Informer: Beyond Efficient Transformer for Long Sequence Time-Series Forecasting. *Proceedings of the AAAI Conference on Artificial Intelligence*, **35**, 11106-11115. <https://doi.org/10.1609/aaai.v35i12.17325>
- [12] Yi, S. and Zhao, W. (2026) A Hybrid LSTM and Transformer Model for Analyzing E-Commerce User Behavior and Path Optimization. *PeerJ Computer Science*, **12**, e3707. <https://doi.org/10.7717/peerj-cs.3707>