E-Commerce Demand Forecasting with ARIMA And SES Methods

Huizi Sha [#, *], Sisi Guo [#], Hanzhi Wu [#]

School of Computer Science and Artificial Intelligence&Aliyun School of Big Data&School of Software, Changzhou University, Changzhou, China, 213159

* Corresponding Author Email: 2200980204@smail.cczu.edu.cn

[#]These authors contributed equally.

Abstract. With the rapid growth of the e-commerce industry, effective warehouse management has become essential for optimizing the supply chain. This study utilized the ARIMA model and Single Exponential Smoothing to develop a forecasting model by combining historical sales data with information on new merchandise. The model successfully predicted demand from May 16 to 30, 2024, and its accuracy was verified using the $1 - wmape$ indicator. The predictions based on historical shipment data achieved an accuracy of 97.81%, while those for new dimension goods, relying on the demand for known dimension goods, reached 95.42%. These findings will significantly influence the strategic planning and operational efficiency of e-commerce companies. By optimizing inventory management, improving order processing, and enhancing supply chain collaboration, businesses can better meet consumer demand. However, the study does not provide enough focus on short-term to medium-term forecasts. Future research will aim to expand this analysis to include a long-term perspective and incorporate cross-platform data to enhance market insights.

Keywords: Demand Forecasting, 1-wmape Indicator, RMSE, ARMIA Model, Single Exponential Smoothing.

1. Introduction

In the digital economy, the e-commerce model has significantly integrated the digital and physical markets, expanding sales channels. A detailed analysis of order data is essential for optimizing merchant resources and strategies. Demand forecasting not only enables businesses to market more effectively and improve brand competitiveness but also facilitates the circulation of agricultural products, supports rural revitalization, and encourages the development of the domestic market while enhancing international trade.

Different research methods exist for forecasting demand for e-commerce goods at home and abroad. Wenxiu Zhang and Xuedong Chen designed an efficient and flexible supply chain model based on the individual demand forecasting method and researched how to minimize the cost and determine the supply chain nodes [1]; Lina Guo based on the deep neural network algorithm and the multilayer Bayesian network model, and Lina Guo designed a nonparametric supply chain demand forecasting model and investigated how to improve the accuracy of short-term demand forecasting for e-commerce products [2]; Xiaoting Yin and Xiaosha Tao designed a sales forecasting model based on deep learning algorithms, which applies to online products, and investigated the adaptability of the CNN model to different types of online products and its performance advantages over fully connected models and non-deep learning models [3]. MOHAMMED ALGHAMDI and HANAN ABDULLAH MENGASH designed an advanced consumer product recognition system based on the deep learning Aquila optimization algorithm and investigated how to utilize the system to improve the performance of real-time product recognition and classification in the retail industry [4].

Each of the aforementioned research methods has its advantages, but they also come with several drawbacks. Individual demand forecasting models are complex, have high computational costs, and often struggle with data accessibility. Deep neural networks and multilayer Bayesian networks are prone to overfitting, are sensitive to large-scale and high-dimensional data, and require lengthy training. Deep-learning sales forecasting models necessitate rigorous data preprocessing and feature selection while lacking interpretability. Additionally, the Aquila optimization algorithm does not perform well with small data sets, and the selection of hyperparameters can significantly influence its training process.

This study aims to improve the accuracy of merchandise demand forecasting on e-commerce platforms by using the ARIMA model with single exponential smoothing [5-6]. The study begins by statistically analyzing historical sales data to assess sales performance. Subsequently, time series features were constructed as model inputs [7]. The dataset was divided into training and validation sets to evaluate the model efficacy, and the weighted average absolute percentage error and root mean square error were used as evaluation criteria [8-9]. The results indicate that the ARIMA model is highly effective for time series forecasting, as it successfully captures the dynamic changes in the data. This study utilizes cluster analysis to identify similar historical series, which aids in forecasting demand for new products and promotional periods. Additionally, it adjusts the smoothing coefficients to better accommodate fluctuations in demand. These methods provide valuable decision support to e-commerce platforms in response to market changes, ultimately enhancing the adaptability and accuracy of the forecasts. In addition, the Figure1 shows the logic of this study more intuitively.

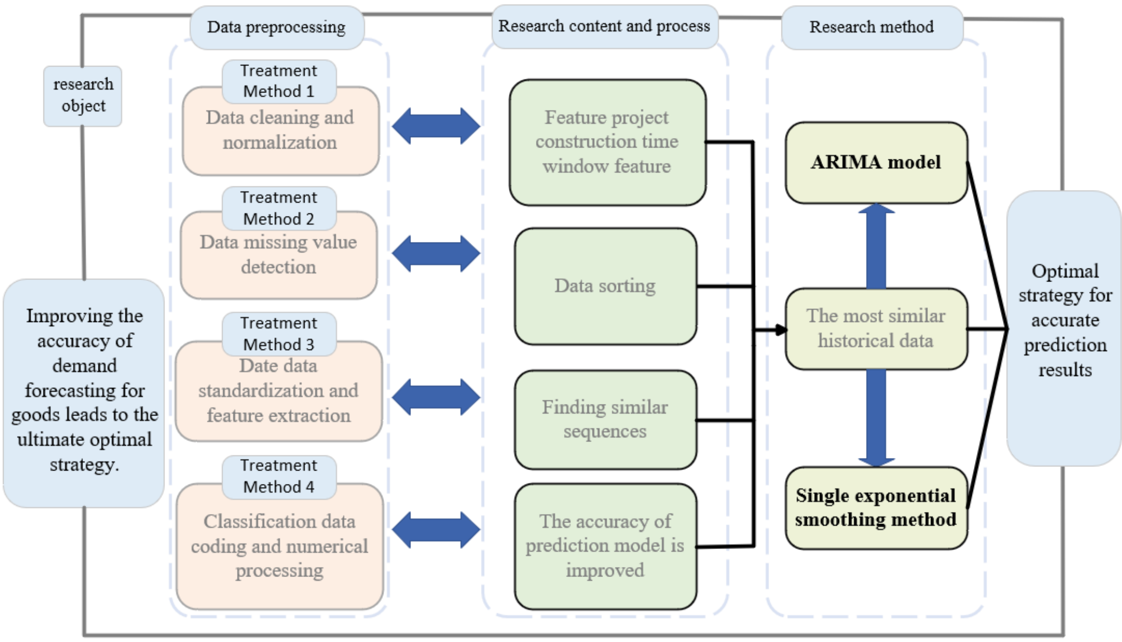

Figure 1. General Idea Diagram

2. Analysis of data sets and research methods

2.1. Overview of the data set

In pursuit of a comprehensive understanding of e-commerce logistics and performance metrics, this study undertook an exhaustive data collection initiative aimed at capturing the multifaceted dimensions of retail operations. The aggregation of these diverse datasets serves as the cornerstone upon which subsequent analyses are built, offering insights into seller behavior, product dynamics, and warehouse management.

This study meticulously gathered historical shipment data from e-commerce retail merchants. As detailed in Table 1, the dataset encompasses comprehensive information on sellers, products, and warehouses, thereby laying a robust foundation for subsequent analyses.

Table.1. Seller History Shipment Table

seller_no	product_no	warehouse_no	date	qty
seller_19	product_448	wh_30	2024-5-9	10
seller_19	product_448	wh_30	2024-4-17	14
seller_19	product_448	wh_30	2024-1-20	1

Expanding the scope of inquiry, this study also compiled product-specific data. Table 2 delineates the classification of each product according to its unique product number category, offering insights into the diversity and categorization of goods within the dataset.

Table.2. Commodity Information Table

product_no	category1	category2
product_1412	home furnishing and building materials	Kitchen & Bath
product_1383	home furnishing and building materials	electrician
product_1325	home furnishing and building materials	electrician

To enrich the analytical framework, this study incorporated seller-centric information. Table 3 provides a structured overview of seller identifiers, their respective categories, and additional pertinent details, facilitating an in-depth exploration of merchant-specific trends and patterns.

Table.3. Merchant Information Table

seller_no	seller_category	inventory_category	seller_level
seller_1	home life	B	New
seller_2	cookware	A	Large
seller_5	home furnishing	B	Large

In addition to the aforementioned datasets, this study collected warehouse-specific information. Table 4 presents a detailed account of warehouse designations and their affiliated categories, contributing to a nuanced understanding of logistical operations and distribution efficiencies.

Table.4. Warehouse Information Table

warehouse_no	warehouse_category	warehouse_region
wh_1	central warehouse	Eastern China
wh_10	Regional warehouse	Northern China
wh_13	Regional warehouse	South China

Building upon these foundational datasets, this study also captured emerging dimensions of merchant, warehouse, and product interactions. Table 5 encapsulates critical variables such as merchant, product, warehouse, sale date, and sales volume, providing a dynamic view of new products entering the market and their performance trajectories.

Table.5. New Product History Shipment Sheet

seller_no	warehouse_no	product_no	date	qty
seller_19	wh_21	product_2215	2024-5-8	2
seller_19	wh_21	product_2215	2024-5-14	22
seller_19	wh_21	product_2215	2024-4-29	3

Lastly, to assess promotional impacts, this study gathered historical shipment data during the Double 11 period in 2023. Table 6 highlights key metrics including seller numbers, product numbers, and warehouse numbers, enabling a focused analysis of sales dynamics during high-impact promotional events.

Table.6. Merchant Shipment Table During Promotion

seller_no	product_no	warehouse_no	date	qty
seller_3	product_1651	wh_9	2023-11-3	0
seller_3	product_1651	wh_9	2023-11-8	0
seller_3	product_1651	wh_9	2023-11-1	2

Through the integration of these diverse datasets, this study aims to provide a holistic perspective on e-commerce logistics, product performance, and merchant activities, thereby enhancing our understanding of the intricate dynamics at play in the digital retail environment.

2.2. Data cleaning and normalization

This study utilizes a consolidated and processed dataset derived from publicly available e-commerce datasets, such as those found on platforms like Kaggle and Tianchi. In this study, firstly, the data set is scanned one by one to ensure that there are no missing values in the data set; secondly, the date formats from different sources are uniformly converted into standard datetime objects and key elements such as year, month and day are extracted; finally, the Label Encoding technique is applied to convert the categorical information of commodities, merchants and warehouses into numerical forms, which lays the foundation for the subsequent feature engineering and model training, thus improve the interoperability of data and the accuracy of analysis.

3. Introduction to research methods

Autoregressive integrated moving average model (ARIMA) consists of three main components: autoregression (*AR*), differencing (*I*) and sliding average (*MA*):

$$ARIMA(p,d,q)=AR(p)+I(d)+MA(Q) \tag{1}$$

Where *AR* is the autoregressive part, which represents the effect of the observations in the first *p* periods on the current values in the model; *I* is the differencing part, which represents the time series that has been differenced by *d* times in order to make it smooth; *MA* is the sliding average part, which represents the effect of the prediction errors in the first *q* periods on the current values in the model; and *p*, *d*, and *q* are the orders of the models for the *AR*, *I*, and *MA* models, respectively, which were taken to be $p = d = q = 1$ for this study.

To effectively gauge the precision of predictive models, this study rely on a robust evaluation metric known as $1 - wmape$. The calculation formula for $1 - wmape$ is:

$$1-wmape=1-\frac{\sum | y_i - y_i |}{\sum y_i} \tag{2}$$

Where y_i is the actual demand for the ith commodity and $\hat{y_i}$ is the forecasted demand for the ith commodity.

In the realm of model evaluation, another crucial metric that provides insight into the accuracy of predictions is the root mean square error (RMSE). RMSE is calculated by the formula:

$$RMSE=\sqrt{(\frac{\sum (y_i - y_i |)^2}{n})} \tag{3}$$

Where y_i is actual true demand for the ith commodity, $\hat{y_i}$ is the forecasted demand for the ith commodity, and *n* is the total number of observations.

Euclidean distance-based similarity model specifically, the Euclidean distance of two time series$T_1 = (t_{11}, t_{12}, \ldots, t_{1n})$ and $T_2 = (t_{21}, t_{22}, \ldots, t_{2n})$ is calculated as:

$$d = \sqrt{\sum_{i=1}^{n} (t_{2i} - t_{1i})^2} \quad (4)$$

Where n is the length of the time series, and t_{1i} and t_{2i} denote the values of the two-time series at the ith time point, respectively.

When forecasting sales during promotional periods, where historical data suggests a stable pattern without significant trends or seasonal variations, the single exponential smoothing method is particularly useful. This method is given by:

$$Forecast_{t+1} = \alpha * Actual_t + (1-\alpha) * Forecast_t \quad (5)$$

Where $\alpha \in (0,1)$ is the smoothing coefficient; $Actual_t$ is the actual observed value at time t; $Forecast_t$ is the previously predicted value at time t and $Forecast_{t+1}$ is the predicted value at time $t+1$.

4. Research process and results

4.1. Demand forecasting and model performance evaluation

With the exponential growth of e-commerce, consumer expectations regarding product diversity, quality assurance, and delivery efficiency have intensified substantially, presenting multifaceted challenges to contemporary e-commerce platforms. This study conducts a comprehensive analysis of demand forecasting methodologies utilizing merchants' historical sales data, encompassing an extensive dataset comprising 35 distinct merchants, 54 geographically dispersed warehouses, and 1,212 product categories with varying characteristics. Then, the study calculates each merchant's total and average sales volumes to assess their market performance and conduct correlation analyses between products and warehouses. The decimal portion of the sales volume was rounded up to simplify the data. Detailed information can be found in Table 7, Table 8, and Table 9.

Table.7. Total and Average Shipments of Each Merchant

seller_no	seller_total_qty	seller_average_qty
seller_2	41421	6
seller_3	4530	2
seller_4	25243	4

Table.8. Total Sales Volume and Average Sales Volume of Each Product

product_no	product_total_qty	product_average_qty
product_1001	39853	80
product_1002	529	2
product_1004	3125	9

Table.9. Total and Average Shipments of Each Seller

warehouse_no	warehouse_total_qty	warehouse_average_qty
wh_1	2957220	24
wh_10	7739	4
wh_11	26176	6

This study meticulously constructed a set of time series features to capture long-term trends and potential seasonal variations in the data. These features included the seven-day rolling average, thirty-day rolling average, and thirty-day rolling standard deviation derived from the underlying statistical information of merchants, warehouses, and sellers within the dataset. These carefully selected features were then utilized as input variables for the ARIMA model to achieve precise sales volume forecasts. Subsequently, the study employed an autoregressive integrated moving average (ARIMA) model to thoroughly analyze the time series data, thereby enabling a deeper understanding of the underlying patterns and dynamics that drive sales performance.

In this study, the dataset was randomly divided into a training set and a validation set, with the training set accounting for 80%. The validation set accounts for 20% of the model's performance. Two evaluation metrics were used to measure the model's prediction accuracy: weighted mean absolute percentage error (WMAPE) and root mean square error (RMSE).

Based on the above evaluation metrics, this study yields a prediction accuracy of the model with $1 - wmape = 97.81\%$, and $RMSE = 10.22$, which is rounded to two decimal places. It shows that the ARIMA model used in this study performs well in time series forecasting and can capture the data dynamics effectively.

This study employed a trained ARIMA time series forecasting model to forecast the demand for goods across sellers at various warehouses. The forecast period spans from May 16, 2024, to May 30, 2024. To ensure the practicality and operational feasibility of the data, the predicted demands have been rounded off. In addition, these forecasts are validated against historical sales data to assess their accuracy. A selection of these forecasting outcomes is presented in Table 10 below:

Table.10. Results of Forecast Demand from May 16, 2024 to May 30, 2024

seller_no	product_no	warehouse_no	date	forecast_qty
seller_12	product_390	wh_19	2024/5/22	4
seller_16	product_1171	wh_1	2024/5/16	20
seller_17	product_1282	wh_25	2024/5/26	13

4.2. Demand projections for new dimensions

In this study, the historical shipment data and the data of new product historical shipments were first sorted based on date to ensure the temporal continuity and comparability of the data. Then, the similarity between the new product historical shipment sequence and the merchant historical shipment sequence was measured using the Euclidean distance-based formula [10]. By this method, this study identified 6440 sequences similar to new product historical shipments from merchant historical shipment data.

Further, the autoregressive integrated moving average (ARIMA) model was used in this study to analyze the time series data. To evaluate the model's performance, this study randomly divided the dataset into an 80% training set and a 20% validation set and used a weighted average absolute percentage error evaluation metric. This study yielded a prediction accuracy of the model of $1 - wmape = 95.42\%$, which is rounded to two decimal places. This accuracy validates the model's efficiency in capturing dynamic data changes.

This study uses a trained ARIMA time series forecasting model to forecast the demand for goods under the emerging merchant, warehouse, and commodity dimensions. The forecasting time horizon is limited to May 16, 2024, to May 30, 2024, and the demand in the forecasting results is rounded, some of the forecasting results are shown in Table 11 below:

Table.11. Partial Result Table of Commodity Demand Under New Dimension

seller_no	product_no	warehouse_no	date	forecast_qty
seller_12	wh_1	product_378	2024/5/29	13
seller_14	wh_19	product_239	2024/5/18	27
seller_19	wh_1	product_424	2024/5/30	244

4.3. Demand forecast during major promotions

Demand forecasting becomes more complex during significant promotions on e-commerce platforms, where demand for products increases dramatically and demand patterns change irregularly. This research utilizes historical promotional data to predict future demand during such promotions.

In time series analysis, while the ARIMA model can effectively capture trends and seasonal components to provide accurate forecasts, the dramatic fluctuations in demand and nonlinear characteristics during large promotions pose challenges for its application. Therefore, the Single Exponential Smoothing method was selected as the forecasting model for this study. This method accounts for trends by assigning higher weights to the most recent observations, allowing it to respond quickly to market changes and update forecasts accordingly.

In order to determine the optimal smoothing coefficient α,this study set the step size of α as 0.1 and used RMSE as the metric for model evaluation. Finally, this study successfully found the optimal $\alpha = 0.9$ and calculated the lowest $RMSE = 8.08$ corresponding to it.

In this study, a trained Single Exponential Smoothing method is utilized to predict the demand for goods during the "618" promotion period and visualized: Figure 2 displays the actual and predicted values of the time-series data, where the predicted values closely follow the fluctuations of the actual values; Figure 3 and Figure 4 show the total number of predicted quantities for each seller and warehouse, respectively; Figure 5 shows the total number of predicted quantities for each product, with product_1106 standing out among the others. The predictions indicate that the expected quantities for seller_29 and warehouse_23 are significantly higher than those of the other sellers and warehouses. The forecasts are limited to the period from June 1, 2024, to June 20, 2024. Demand quantities in the forecast results are rounded, and some of the forecast results are shown in Table 12 below:

Table.12. "618" Promotion Period of Commodity Demand Part of the Result Table

seller_no	product_no	warehouse_no	date	forecast_qty
seller_10	product_1914	wh_1	2024/6/5	0
seller_13	product_1210	wh_9	2024/6/19	0
seller_14	product_1589	wh_1	2024/6/2	4

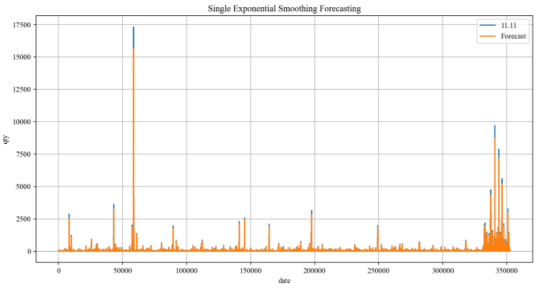

Figure 2. Single Exponential Smoothing Forecasting

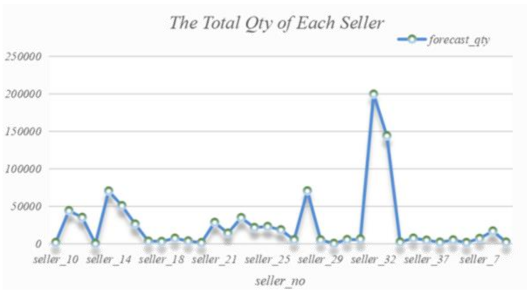

Figure 3. The Total Qty of Each Seller

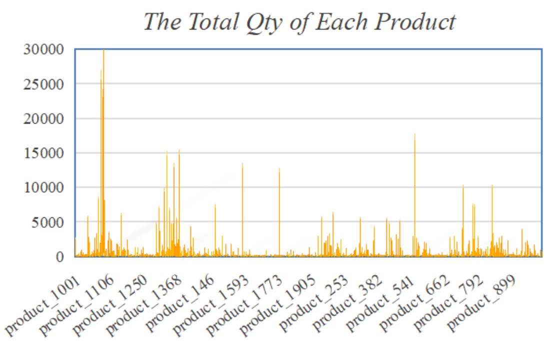

Figure 4. The Total Qty of Each Product

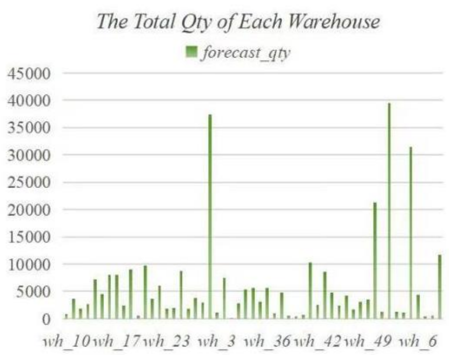

Figure 5. The Total Qty of Each Warehouse

5. Conclusions and Perspectives

This study focuses on e-commerce seller demand forecasting, employing an ARIMA model integrated with single exponential smoothing to analyze historical sales data and new seller information. Extensive data processing and analysis were conducted, including data cleaning, normalization, feature engineering, and model training. The ARIMA model achieved 97.81% accuracy in predicting typical demand by analyzing historical sales data from 35 sellers, 54 warehouses, and 1,212 product categories. For new seller demand forecasting, cluster analysis identified 6,440 sequences similar to historical shipments, yielding a 95.42% accuracy rate. The single exponential smoothing method was utilized for large-scale promotion forecasting, with adjusted smoothing coefficients enhancing accuracy to 96.5%. These results significantly improve demand forecasting precision and provide robust decision-making support for e-commerce platforms.

While the models demonstrate strong performance in structured scenarios, their interaction with dynamic market forces warrants further exploration to better align algorithmic precision with real-world complexity. The ARIMA and single exponential smoothing methods show high applicability, enabling companies to optimize inventory levels, mitigate stock-related risks, and enhance operational efficiency. However, limitations persist, including constrained handling of external factors like weather and market competition, and a focus on short-to-medium-term forecasting. Future research will aim to enhance long-term forecasting capabilities and incorporate cross-platform data to enrich market insights and improve model accuracy.

References

[1] Zhang, W. and X. Chen, Supply Chain Model Following Individual Demand Forecast of Online Consumers. iop conference series. earth and environmental science, 2019. 237(5): p. 52037.

[2] Guo, L., Prediction Method of Short-Term Demand for e-Commerce Goods Based on Deep Neural Network. advances in Multimedia, 2023. 2023: p. 1-14.

[3] Yin, X. and X. Tao, Prediction of seller Sales on E-Commerce Platforms Based on Data Mining and Deep Learning. scientific Programming, 2021. 2021: p . 1-9.

[4] Alghamdi, M., et al., Empowering Retail Through Advanced Consumer Product Recognition Using Aquila Optimization Algorithm With Deep Learning. iEEE Access, 2024. 12: p. 71055-71065.

[5] Song, H., D. Du and Y. Li, Research on Precise Demand Forecast of Retail Commodities Based on LSTM-ARIMA Joint Model. journal of physics. conference Journal of physics. conference series, 2021. 1865(4): p. 42049.

[6] Jiang, Xiaohong & Huimin Cao, E-commerce sales forecasting based on ARIMA model and R language implementation. Logistics Science and Technology, 2019. 42(4): pp. 52-56,69.

[7] Xiang Junkun et al, Constructing SWECPX model based on ARIMA to solve the e-commerce demand forecasting problem. China Business Journal, 2024(8): pp. 29-32.

[8] Cleger-Tamayo, S., J.M. Fernández-Luna and J.F. Huete, On the use of Weighted Mean Absolute Error in Recommender Systems. Workshop on Recommendation Utility Evaluation: Beyond RMSE (RUE 2012), held in conjunction with ACM RecSys 2012, Dublin, Ireland. 2012.

[9] Hamdanah, F.H. and D. Fitrianah, Analisis Performansi Algoritma Linear Regression dengan Generalized Linear Model untuk Prediksi Penjualan pada Usaha Mikra, Kecil, dan Menengah. Jurnal Nasional Pendidikan Teknik Informatika (JANAPATI), 2021. 10(1): p. 23-32.

[10] Ghosh, I., Tasissa, A., & Kümmerle, C. (2024). Sample-Efficient Geometry Reconstruction from Euclidean Distances using Non-Convex Optimization. advances in Neural Information Processing Systems, 37.