Academic Journal of Science and Technology
ISSN: 2771-3032 | Vol. 10, No. 3, 2024

Research on E-Commerce Retail Demand Forecasting Based on SARIMA Model and K-means Clustering Algorithm

Yiding Zhao

School of Statistics, University of International Business and Economics, Beijing, China

Abstract: With the rapid development of e-commerce, precise demand forecasting and efficient inventory management have become essential for the success and profitability of retail businesses. This study focuses on demand forecasting for e-commerce retailers using the Seasonal Autoregressive Integrated Moving Average (SARIMA) model and the K-means clustering algorithm. The research utilizes a dataset containing 1996 time series of sales data from various products, merchants, and warehouses, aiming to predict demand changes for the next 15 days. The study initially evaluates three models—Linear Regression (LR), Autoregressive Integrated Moving Average (ARIMA), and SARIMA—by fitting them to historical sales data to forecast future demand. The SARIMA model is identified as the most effective through rigorous evaluation using 1-mWAPE (mean weighted absolute percentage error) and RMSE (root mean square error) metrics. To enhance homogeneity within demand categories, the K-means clustering method is applied to divide products into four distinct groups, further refining the forecasting process.The paper also addresses the challenge of integrating new sequences into the dataset by leveraging clustering results to classify sequences and using cosine similarity to identify analogous historical time series. These matched sequences serve as the basis for demand prediction using the established SARIMA model. The findings highlight the robustness of the SARIMA model in capturing trends and seasonality, providing a reliable framework for e-commerce demand forecasting that can significantly impact inventory strategies and operational efficiency.

Keywords: E-commerce, demand forecasting, SARIMA model, k-means clustering.

1. Introduction

In the context of the digital era, e-commerce has emerged as a significant driving force in the retail industry with its rapid growth. As online transaction volumes continue to soar, accurate demand forecasting and effective inventory management have become crucial for retailers. Precise demand forecasting enables retailers to align inventory levels with expected sales, thereby minimizing costs associated with overstock and stockouts while maximizing customer satisfaction and revenue[1-3].

This study aims to explore the application of advanced time series models for demand forecasting by e-commerce retailers[4], with a particular focus on the use of the Seasonal Autoregressive Integrated Moving Average (SARIMA) model. The SARIMA model, an extension of the ARIMA model[5], incorporates elements of autoregression and moving averages while also accounting for seasonal variations in the data. This makes it especially suitable for analyzing data that exhibits regular patterns over specific time periods[6], such as seasonal fluctuations in consumer purchasing behavior. The research is based on a comprehensive dataset containing 1996 time series, encompassing sales data for a diverse range of products across various merchants and warehouses. The goal of the study is to predict demand for the next 15 days, which is critical for inventory planning in e-commerce. By integrating the SARIMA model with clustering and similarity analysis techniques, this research offers a robust and nuanced approach to demand forecasting in the e-commerce domain.

2. Research Methodology and Model Development

2.1. Dataset

The dataset utilized in this study consists of 1996 time series that meticulously document the sales volume of various products across different merchants and warehouses. These data encompass a wide range of product categories, including food and beverages, home furnishings and building materials, toys and musical instruments, among others, ensuring the broad applicability and representativeness of the study's findings. Each time series contains sales information for a specific product under a particular merchant and warehouse, recording the quantity sold over a continuous period, providing rich historical information for demand forecasting.

Before proceeding with demand forecasting, the original dataset underwent a series of preprocessing steps to ensure the quality and suitability of the data for analysis. The following are the main preprocessing measures taken.

Initially, all null and duplicate records in the dataset were removed to prevent any adverse effects on the analytical results. For identified data outliers, the 3σ rule was applied for identification and treatment to ensure the accuracy of the data. The dataset was subjected to a validity check to ensure the consistency of key information such as merchant numbers, product numbers, and warehouse numbers, ensuring the integrity and accuracy of the data. Multiple tables provided based on the same field names were merged to form a comprehensive table, which included key information such as merchant numbers, product numbers, and warehouse numbers, laying the foundation for subsequent data analysis.

The merged table was arranged in chronological order, and

outliers were identified based on the 3σ rule. Interpolation methods were then applied to handle these outliers, ensuring the continuity of the time series. The 3σ rule, also known as the Laplace criterion, refers to the process of calculating the standard deviation of a set of data under the assumption that it contains only random errors. A certain interval is determined based on probability, and errors exceeding this interval are considered not random but significant, and data containing such errors should be discarded. The normal distribution formula is as follows.

$$f(x)=\frac{1}{\sqrt{2\Pi}\,\sigma}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \tag{1}$$

After integration calculations, we can determine that the area within the horizontal interval ($\mu-\sigma$, $\mu+\sigma$) is 68.268949%, the area within ($\mu-2\sigma$, $\mu+2\sigma$) is 95.449974%, and the area within ($\mu-3\sigma$, $\mu+3\sigma$) is 99.730020%. Therefore, data falling outside the 3σ interval can essentially be regarded as outliers and eliminated.

In the raw data, we performed separate counts for merchants, products, and warehouses, and found that there are a total of 35 merchants, 1,212 types of products, and 54 warehouses. The category with the highest total demand is food and beverages, followed by home furnishings and building materials, with the lowest demand being in toys and musical instruments. By using the three major dimensions of different merchants, products, and warehouses as classification indicators for grouping, we ended up with a total of 1,996 groups. This paper will select groups with the numbers 520, 1119, and 1129 as examples for visual presentation as follows.

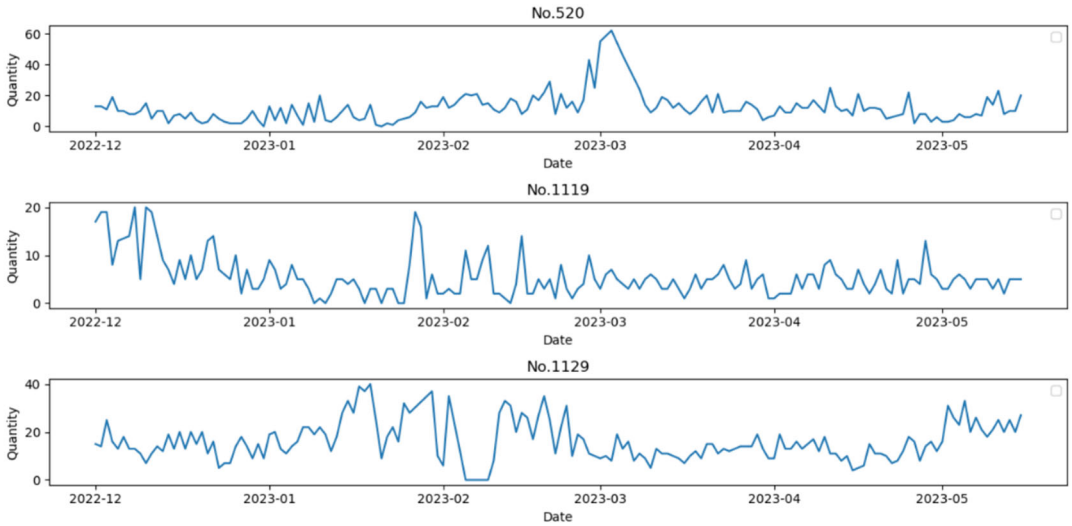

Figure 1. Time Series Charts for Products with Groups (520, 1119, 1129)

2.2. Establishment of the ARIMA Model

The ARIMA model, which stands for Autoregressive Integrated Moving Average model, is a synthesis of three components: Autoregression (AR), Integration (I), and Moving Average (MA). Essentially, it combines the predictive power of past values (autoregression), differencing to make the data stationary (integration), and smoothing out short-term fluctuations (moving average). The basic expression of the ARIMA model can be represented as follows.

$$\varphi(B)\,(1-B)^{d}y_t=\theta(B)\varepsilon_t \tag{2}$$

In the formula, y_t represents the time series data under consideration, y_t is the order of differencing, p and q correspond to the orders of the autoregressive and moving average parts, respectively, and ε_t is a sequence of independent and identically distributed white noise with a mean of zero and a constant variance. The lag operator B satisfies the following expression.

$$B^n y_t=y_{t-n} \tag{3}$$

$$\varphi(B)=1-\sum_{i=1}^{p}\varphi_i B^i \tag{4}$$

$$\theta(B)=1-\sum_{i=1}^{q}\theta_i B^i \tag{5}$$

The key to establishing an ARIMA(p,d,q) model lies in the selection of the three parameters: (p, d, q). Here, d represents the order of differencing, whose purpose is to transform the original observational series into a stationary time series. p denotes the order of the AR part, also known as the number of lags in the autoregression. q signifies the order of the MA part, which is the number of lags for the moving average. The crux of the ARIMA model is to determine these three parameters (p, d, q), with d being the order of differencing required to achieve stationarity.

Through the process of differencing, the ARIMA model converts the data to be more stationary and then constructs a model that regresses the lagged terms of the dependent variable and the current and lagged values of the random error term. In the formula, $y_{d(t-i)}$ and $\varepsilon_{d(t-i)}$ are multivariate

linear functions of the lags up to p and q periods, respectively, and ε_t is an independent and identically distributed white noise sequence with a mean of zero and a constant variance. The differencing performed to achieve stationarity is typically of the first order.

In the case of less model data, BIC, AIC or autocorrelation graph is generally used to select p and q values in ARMA respectively. First, this paper selects the ARIMA model to predict seller_19, product_448 and wh_30 products. The results of some models in group 1996 are shown as follows.

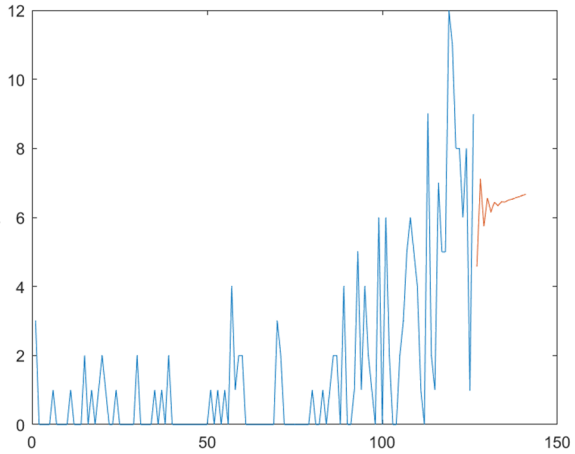

Figure 2. The result of selecting a product to predict

Since there are 1996 sets of data in this paper, considering the complexity of data and the efficiency of parameter solving, this paper selects the optimization function Auto-Arima to search for parameters. Since there are 1996 sets of data in this paper, considering the complexity of data and the efficiency of parameter solving, this paper selects the optimization function Auto-Arima to search for parameters. The average WMAPE of the three groups is 0.6142, so it needs to be improved.

2.3. Establishment of LR model

In order to predict the trend of demand this month, this paper constructs the following regression model.

$$\hat{y}_i = a + bx_i \tag{6}$$

$$a = \frac{1}{n}\sum_{i=1}^{n} y_i - b\frac{1}{n}\sum_{i=1}^{n} x_i \tag{7}$$

$$b = \frac{n\sum_{i=1}^{n} x_i y_i - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2} \tag{8}$$

Where y_i is the quantity demanded and x_i is the date. After calculating each parameter in the regression model, the predicted value of demand in the t+n period can be calculated by the following formula.

$$\hat{y}_{t+n} = a + bx_{t+n} \tag{9}$$

According to the predicted results, LR model is more sensitive to data volatility, which can better reflect the linearity of data, but it is difficult to accurately represent the volatility of different cycles. Therefore, the model is improved below.

2.4. Establishment of SARIMA model

SARIMA model is an improved version of ARIMA model, which can predict periodic time series data more scientifically and reasonably, while commodity demand has certain periodic characteristics. The SARIMA model performs period-based seasonal differences on the ARIMA model, and its expression is as follows.

$$\phi_p(B)\Phi_p(B^S)\nabla^d\nabla_S^D y_t = \theta_q\Theta_Q(B^S)\varepsilon_t \tag{10}$$

Where P is the order of seasonal autoregression, Q is the order of seasonal shift average, $\nabla^d = (1-B)^D$ represents D-order phase by phase difference, $\nabla_S^D = (1-B^S)^D$ represents D-order seasonal difference, $\Phi_p(B^S)$ and $\Theta_Q(B^S)$ are P-order autoregressive operators and Q-order moving average operators.

In this paper, the optimization function Auto-Arima is selected to search for parameters. The following is the result after selecting part of the 1996 group (520,1129) for prediction and fitting.

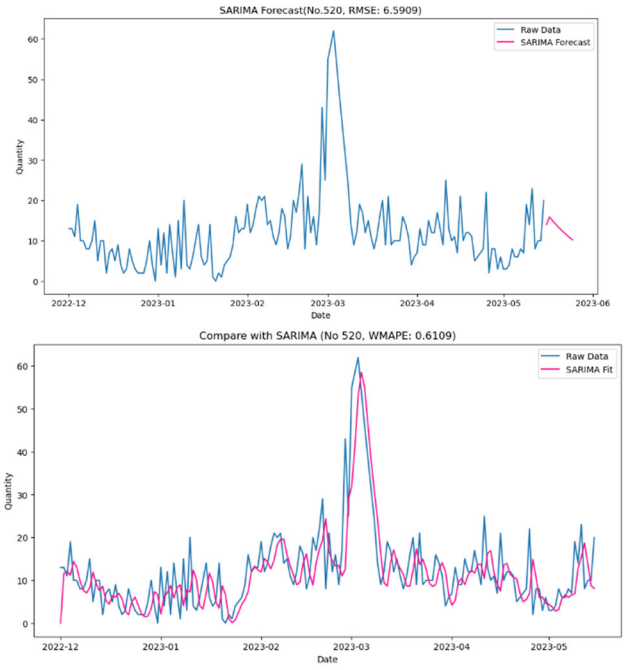

Figure 3. N0.520 Product fitting forecast results

It can be preliminatively concluded from the wmaped index of group 520 commodities that the SARIMA model has a better fit than the ARIMA model. Similarly, select 1129 data for viewing.

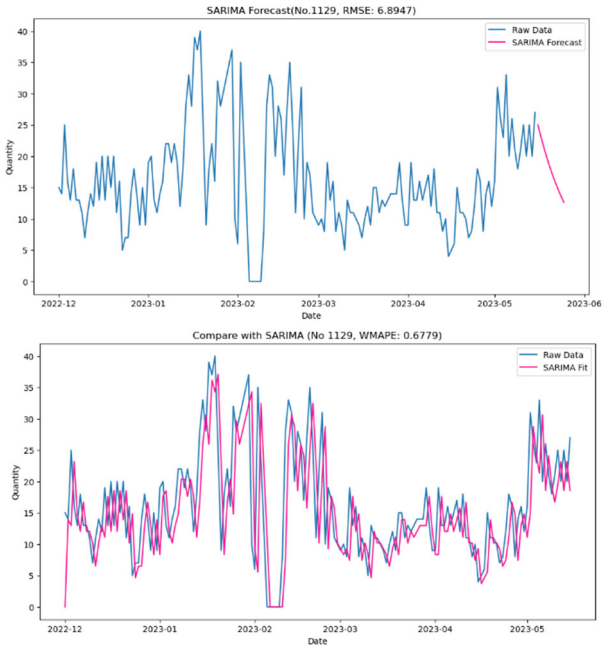

Figure 4. N0.1129 Product fitting forecast results

It can be intuitively seen from the figures of the above two groups that, compared with the ARIMA model, SAMRIMA had higher curve overlap when fitting based on known data, and the average WMAPE value was 0.6444. Therefore, the SARIMA model was chosen as the data prediction model.

3. Demand Forecasting and Clustering Analysis

3.1. Evaluation of Forecasting Results

In this paper, 1-wmape index and RMSE index are used to evaluate the accuracy of the three models. In this paper, the average value of the 1996 group of models is taken as the final reference, and the results are as Table 1.

Table 1. Evaluation index value of each model

Model	1-wmape	RMSE
LR	1-0.8119	10.8236
ARIMA	1-0.5944	4.4595
SARIMA	1-0.6278	6.7428

Wmape metrics can be used to assess the accuracy of various forecasting models, to monitor and evaluate the performance of forecasting systems, and to conduct confidence analysis of forecast results. According to the evaluation index of the model, SARIMA model was selected for prediction.

3.2. K-means Clustering Analysis

We classify the data series according to different merchants, delivery warehouses and goods, etc. In different dimensions, some raw data are strings, and map the data by way of recoding. There are as many as 330,000 data samples in this paper. If the original data is divided too carefully, it will lead to too many categories and too much calculation. Therefore, we use K-means algorithm to roughly reclassify the original data first. Therefore, this paper obtained 1996 rows of data after averaging each column in 1996 sets of data after coding, and carried out K-means clustering on the data to reduce the workload.

means classification is the most commonly used method in cluster analysis, it is based on Euclidean distance to divide a certain class of goods into the corresponding class with the smallest distance from a sample center. The general idea of the algorithm is to randomly select k sample center clusters from the sample, calculate the distance between the sample and all center clusters, and then merge them into the cluster with the smallest distance. The cluster centers of each cluster are recalculated and another sample is selected to repeat the process until all samples are classified. The basic formula is as follows.

$$E = \sum_{i=1}^{k} \sum_{x \in C_i} \|x - \mu_i\|_2^2 \tag{11}$$

The K-means algorithm operates iteratively to achieve local optimal solutions. The process begins by determining the total number of clusters, k, and selecting initial cluster centers for each. It then calculates the Euclidean distance of each sample from the cluster centers. Samples are assigned to the cluster with the shortest distance, and new cluster centers are recalculated accordingly. This process of reassigning and recalculating is repeated until the classification is complete, resulting in the final k clusters and their corresponding cluster centers. To prevent excessive computation time, a maximum number of iterations and a threshold value for cluster adjustment are typically set in the implementation of the K-means algorithm. Methods for choosing the optimal k value include prior knowledge, the elbow method, and validation techniques.

This article utilizes the elbow method and validation method for determination. The core indicator of the elbow method is the Sum of Squared Errors (SSE), with the formula as follows.

$$SSE = \sum_{i=1}^{k} \sum_{p \in C_i} |p - m_i|^2 \tag{12}$$

Where Ci is the ith cluster, p is the sample point in Ci, mi is the centroid of Ci (the mean of all samples in Ci), SSE is the clustering error of all samples, representing the quality of the clustering effect. The core idea is: the specific step is usually to gradually increase the k value, calculate the SSE of the clustering model under each k value, and then draw a graph of SSE and k value. The graph usually shows an elbow or bend point after k reaches a certain value, and the corresponding value of k is considered to be the best number of clusters, because increasing the value of k does not significantly reduce SSE.

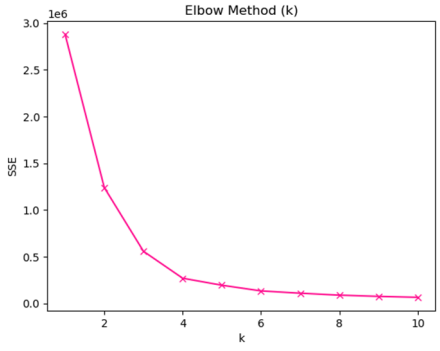

Figure 5. N0.1129 Product fitting forecast results

As can be seen from Fig. 6, the elbow method determines the value of k at the inflection point of the curve, so k=4 is selected, that is, the optimal clustering number is 4. The clustering results are as follows.

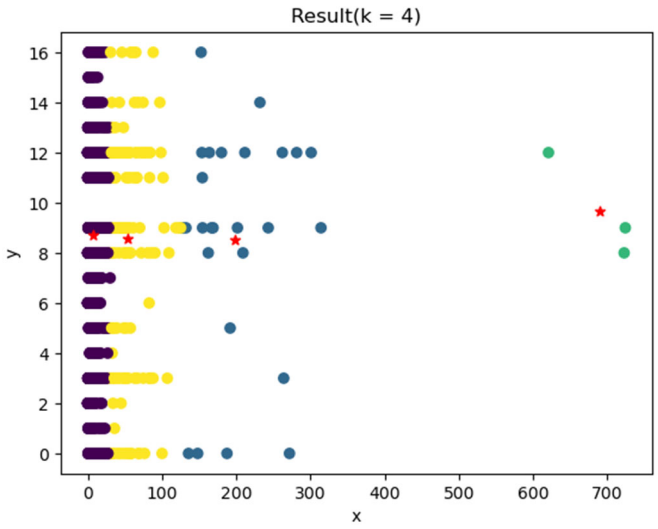

Figure 6. Clustering result

The centroid of the classification is as follows.

Table 2. Table of classification centers

Category	Demand Quantity	Merchant Classification	Inventory Classification	Merchant Scale	Warehouse Category	Warehouse Region
1	6.4919	8.690	8.6906	0.3552	0.3414	2.6796
2	199.004	8.5	8.5	0.3846	0.8077	2.1154
3	689.331	9.666	9.6667	1	1	2
4	53.9763	8.5478	1.1783	0.4268	0.6943	2.2675

As can be seen from Table 2, the demand for category 1 is relatively small, mainly concentrated in household appliances and home improvement building materials. The scale of merchants is medium and large, and the warehouse category is concentrated in regional warehouses, most of which are in South China. The demand of category 2 is medium, mainly concentrated in household appliances and home improvement building materials. The business scale is medium and large. The warehouse category is concentrated in the central warehouse, and the warehouse area is mostly in East China. The demand for category 3 is large, which is basically concentrated in the categories of home life and beauty and skin care. The business scale is medium-sized, the warehouse category is concentrated in the central warehouse, and the warehouse area is mostly in East China. The demand of category 4 is small, basically concentrated in household appliances and home improvement building materials, the scale of businesses is large, the warehouse category is concentrated in the central warehouse, and the warehouse area is mostly in South China.

4. Conclusion

This study has developed and evaluated an e-commerce retail demand forecasting system based on the SARIMA model and K-means clustering algorithm, providing retailers with a novel tool for inventory optimization. Through an in-depth analysis of 1996 time series datasets, it was found that the SARIMA model excels at capturing the seasonality and trends in sales data, demonstrating higher fit and predictive accuracy compared to traditional Linear Regression (LR) and Autoregressive Integrated Moving Average (ARIMA) models. This was substantiated by the 1-mWAPE and RMSE metrics, with the SARIMA model's average WMAPE value at 0.6278, which is lower than the LR model's 1-0.8119 and the ARIMA model's 1-0.5944, indicating its significant advantage in predictive precision. Furthermore, the application of the K-means clustering algorithm not only enhanced the homogeneity of demand forecasting but also improved the model's adaptability to new sequence data. By clustering analysis, products were divided into four categories, with more similar demand characteristics within each category, aiding the model in more accurately capturing potential changes in demand.

In summary, this research confirms the effectiveness and practicality of combining the SARIMA model with the K-means clustering algorithm in e-commerce demand forecasting. Future research could further explore the model's applicability in different e-commerce environments and consider incorporating additional influencing factors to enhance the comprehensiveness and accuracy of forecasts.

References

[1] Lalou P, Ponis S T, Efthymiou O K. Demand forecasting of retail sales using data analytics and statistical programming[J]. Management & Marketing, 2020, 15(2): 186-202.

[2] Bandara K, Shi P, Bergmeir C, et al. Sales demand forecast in e-commerce using a long short-term memory neural network methodology[C]//Neural Information Processing: 26th

International Conference, ICONIP 2019, Sydney, NSW, Australia, December 12–15, 2019, Proceedings, Part III 26. Springer International Publishing, 2019: 462-474.

[3] Leung K H, Mo D Y, Ho G T S, et al. Modelling near-real-time order arrival demand in e-commerce context: a machine learning predictive methodology[J]. Industrial Management & Data Systems, 2020, 120(6): 1149-1174.

[4] Shih Y S, Lin M H. A LSTM approach for sales forecasting of goods with short-term demands in E-commerce[C]//Intelligent Information and Database Systems: 11th Asian Conference, ACIIDS 2019, Yogyakarta, Indonesia, April 8–11, 2019, Proceedings, Part I 11. Springer International Publishing, 2019: 244-256.

[5] Dabral P P, Murry M Z. Modelling and forecasting of rainfall time series using SARIMA[J]. Environmental Processes, 2017, 4(2): 399-419.

[6] Dubey A K, Kumar A, García-Díaz V, et al. Study and analysis of SARIMA and LSTM in forecasting time series data[J]. Sustainable Energy Technologies and Assessments, 2021, 47: 101474.