Research on e-commerce retail demand forecasting based on multi-model fusion

Haozhi Mao *, Zihan Gao, Jiayi Hou

School of Medical Information Engineering, Shandong University of Traditional Chinese Medicine, Jinan, China, 250355

* Corresponding Author Email: MHz15864465283@163.com

Abstract. With the rapid refinement of the e-commerce retail sector, accurately predicting the demand for commodities has become the key for enterprises to optimize inventory management and improve supply chain efficiency. From current research, we propose an isodimensional new interest recurring GM model, ARIMA model, and K-Means clustering multi-model fusion method for e-commerce retail demand forecasting. Firstly, the etailing statistics was cleaned and preprocessed, and the 3-sigma criterion was used to identify and eliminate outliers, and secondly, the isodimensional new information compensatory GM model and the ARIMA model were combined for preliminary prediction, and the key features were extracted by K-means clustering analysis. The findings indicate that the model can accurately and robustly predict the demand in e - commerce retail. It shows high prediction precision and stability. The model developed in this study is able to offer technical assistance for the demand forecasting within the e - commerce retail domain.

Keywords: Demand Forecasting for E-commerce Retail, GM model, ARIMA model, K-Means clustering.

1. Introduction

As the Internet advances by leaps and bounds, e - commerce has assumed an obviously significant position in the global economic landscape. Statistical statistics indicates that latterly, the sales amount of e - commerce retail has been on a continuous upward trend. Simultaneously, consumers' shopping behaviors have become more personalized and diversified. Against this backdrop, for e - commerce retail merchants, accurately forecasting product demand has become of utmost importance. It directly impacts merchants' inventory costs, supply efficiency, and customer satisfaction.

In the sphere of e-commerce retail demand forecasting, there have been multitudinous related studies, among which the early studies mostly used simple statistical methods, such as the Delphi method [1] , analogy [2], simple moving average [3], multiple linear regression [4] . However, as the amount of statistics grows and the complexity of requirements increases, these methods gradually expose their limitations and inefficiency. With the recent years of machine learning and deep learning technology in the energy sector [5] and medical and health fields [6] , financial sector [7] and so on. They have certain advantages in statistics prediction with certain trends and seasonal characteristics, but they are not well adapted to nonlinear or abrupt change statistics, so they should be combined with other methods to further complete the requirements, among which the cluster analysis method can mine the potential structure of statistics, and it is tough to accomplish explicit prediction alone, so the combination of the two types of methods can achieve the accuracy of e-commerce retail demand forecasting. The current research still needs to be strengthened in the comprehensive application of multiple technologies to adapt to the complex and changeable scenarios of e-commerce retail demand forecasting.

The purpose of this study is to use the fusion technology of GM model, ARIMA and K-Means clustering method to dig deep into the rules in e-commerce retail statistics and construct a more accurate demand forecasting model, for the sake of furnishing strong decision support of e-commerce retailers, and at the same time provide new approaches for research in this sphere.

2. Materials and methods

2.1. Statistics integration and processing

2.1.1. Acquisition of statistics

The research statistics is sourced from the open - source website (http://mathorcup.org). The statistics encompasses historical shipment volumes of e - commerce retail merchants, product information, merchant information, and warehouse information, among others.

2.1.2. Statistics preprocessing

Use the pandas library in Python to extract the statistics use.The merge function is based on the merchant code (seller_no), product code (product_no), warehouse number (warehouse_noThe standard combines multiple statistics and collects statistics. As Figure 1, Figure 2 shown:

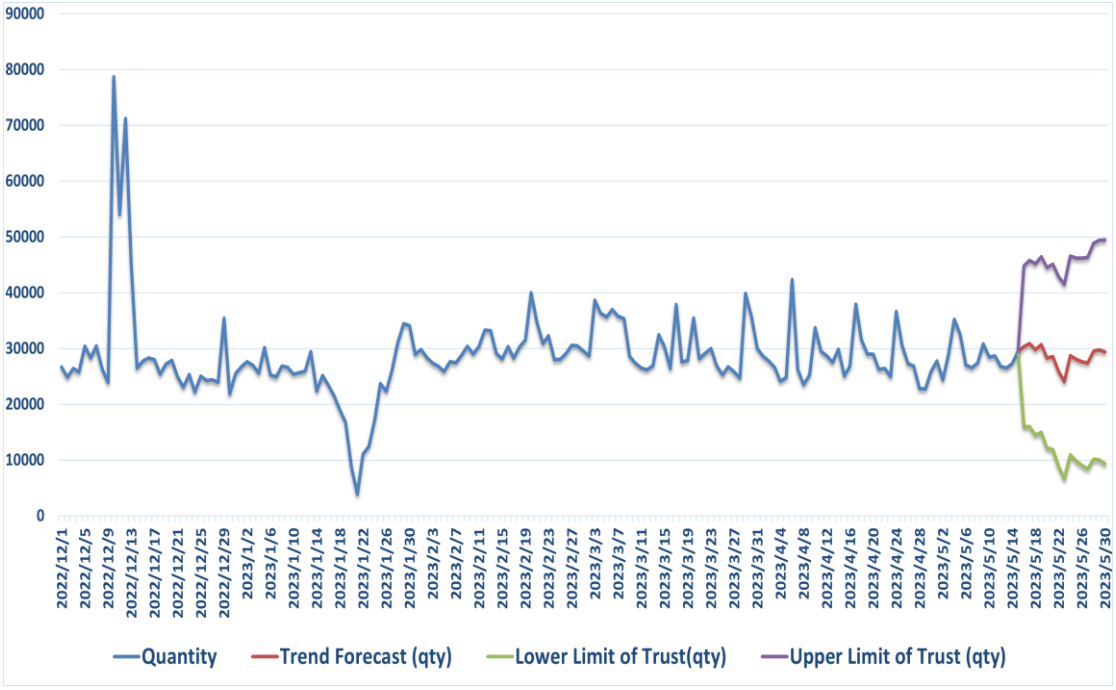

Figure 1. 2022.12 - 2023.5 Commodity Chart

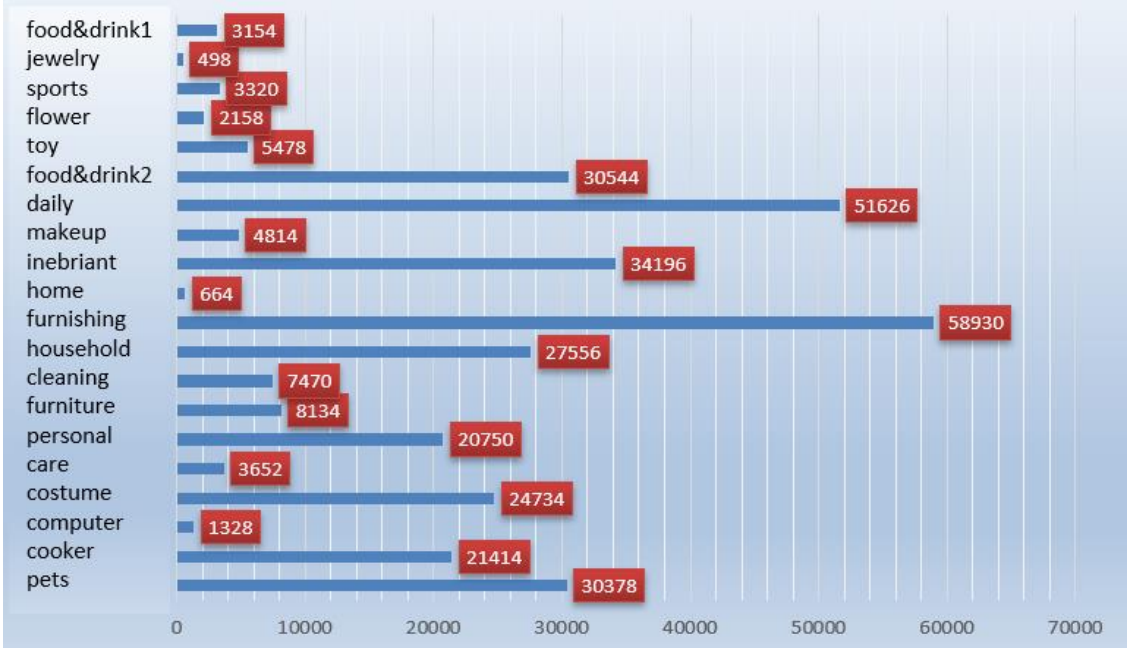

Figure 2. Commodity category demand statistics

Observing the statistics analysis after integration, it was found that there was a large difference in the demand of each group. In order to avoid the influence of special values (missing values, outliers) on the model results, the 3-Sigma criterion was used for detection. This criterion assumes that a set of pedagogical statistics includes solely random errors, and the standard deviation is calculated by means of computational processing, and subsequently, an interval is ascertained for the purpose of identifying errors outside the range of the interval as gross errors and eliminate them.

When the 3-sigma criterion satisfies the status of Gaussian Distribution, if the absolute value of the residual error vi of a certain measured value is greater than 3sigma, it is identified as a bad value and should be removed. Usually, ±3sigma is defined as the limit error. In the case of normal distribution, the probability of falling outside this range is only 0.27%, so the possibility of its occurrence in a limited number of measurements is very low.

Normal Distribution Formula:

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

On the normal distribution curve, the statistical likelihood that the values are distributed within the interval (μ-σ, μ+σ) is 0.6526.

The probability that the values are distributed within the interval (μ-2σ, μ+2σ) is 0.9544.

The probability that the values are distributed within the interval (μ-3σ, μ+3σ) is 0.9974.

It can be considered that the values of f(x) are almost all concentrated within the interval (μ-3σ, μ+3σ), and the possibility of exceeding this range accounts for less than 0.3%. From this, it can be inferred that this statistics is an outlier, as demonstrated in Table 1 below:

Table 1. Outlier detection

seller_no	product_no	warehouse_no	statistics	qty
seller_32	product_1091	wh_1	2022/12/10	17323
seller_32	product_1091	wh_1	2022/12/12	14148

As can be seen from the statistics, the commodity with the number 1091 is likely to be a special building material. There will be a significant increase in demand for it in December or from March to April due to its own properties and functions, which is in line with the actual situation. Therefore, the statistics remains valid.

As shown in Table 2 is the grouping standard:

Table 2. Grouping criteria

Indicator serial number	index
1	business
2	commodity
3	warehouse
4	Product Category
5	The type of business that the merchant operates
6	Inventory type
7	Merchant tier
8	Warehouse type
9	Warehouse area

Through the integrated statistics, it is known that there are 35 merchants, 1212 products, and 54 warehouses. A total of 1996 kinds of statistics of the equivalent business, the same warehouse and the same product are combined in the following table, and two sets of statistics are randomly selected according to the classification according to the following table, as shown in Figure 3 extracted statistics group 1 and Figure 4 extracted statistics group 2.

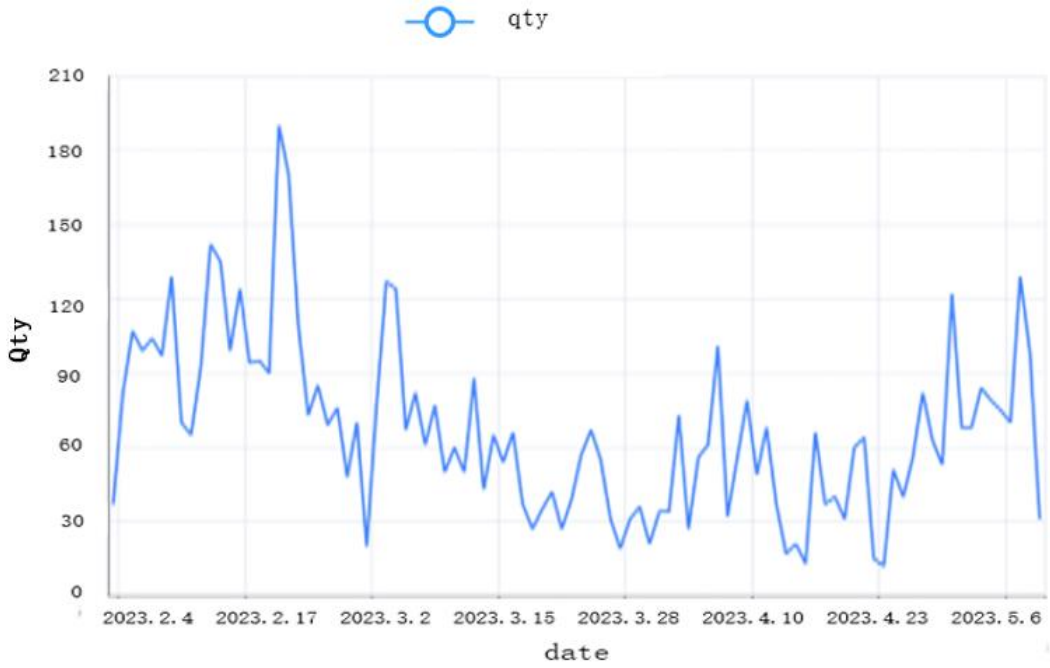

Figure 3. Extract statistics set 1

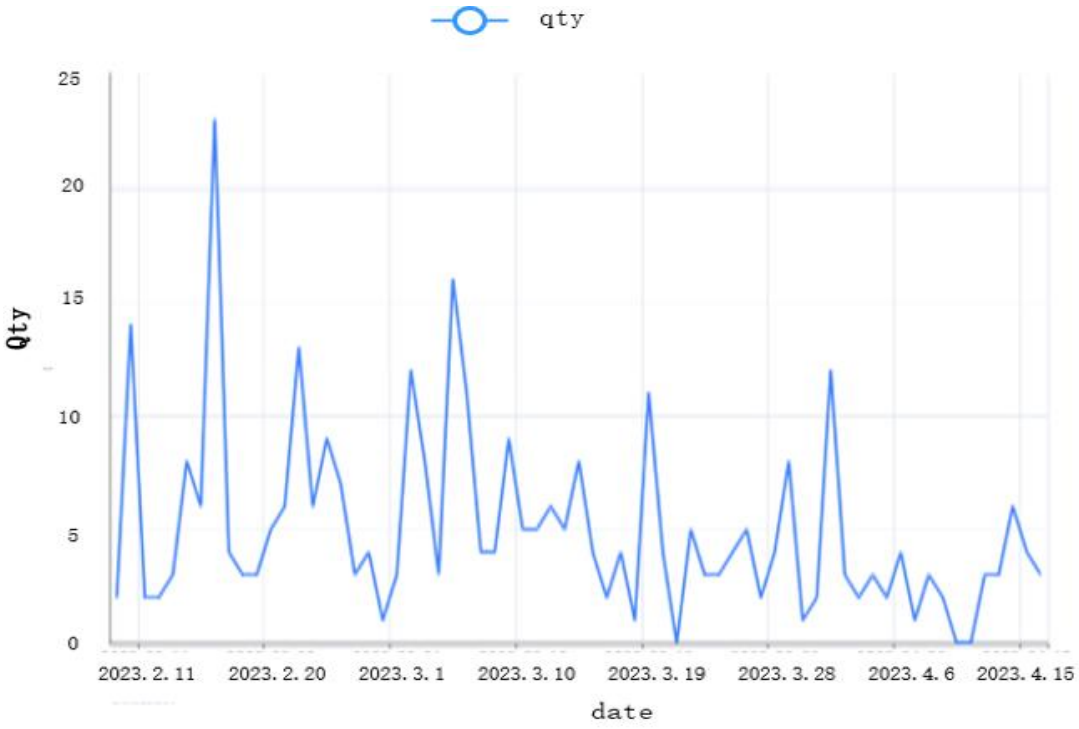

Figure 4. Extract statistics set 2

2.2. Research Methodology

This article focuses on solving the problem of e-commerce retail demand forecasting. Firstly, the pandas library in the Python language is used to scan statistics, and the merge function is employed to combine the statistics. The 3-sigma criterion is utilized to handle the outliers in the statistics through the normal distribution. Subsequently, the ARIMA time series model is used to analyze and interpret the statistics, making predictions about the demand quantities of distinctive products from dissimilar merchants in different warehouses for the next fifteen days. SARIMA is utilized for the analysis of the seasonality of the statistics, elevating the fidelity of the model and optimizing its integrity. Finally, K-means clustering analysis serves to obtain the eigenvalues of the required statistics.

3. Model establishment and solving

3.1. Numerical prediction of the combination of the isodimensional new interest recursive GM model and ARIMA

The ARIMA model belongs to the category of seasonal autoregressive integrated moving average models. Its core concept is to make use of the historical data of the statistical variable itself to forecast future trends. The value of the label at a certain moment within the timeline is impacted by both the label values from previous time intervals and the unforeseen events that took place during those earlier time periods. In essence, the ARIMA model supposes that the label value oscillates around the overall temporal trend. Here, the trend is generated by the cumulative effect of historical label values, and the oscillations are caused by the influence of random events within a specific time frame. Additionally, the overall trend doesn't necessarily maintain a stable state [8] .

The ARIMA model requires that there be no autocorrelation in its residuals, signifying that the model residuals ought to be white noise. To validate this assertion, an inspection of the model's comprehensive test result schedule can be carried out. The white-noise characteristic of the model can then be evaluated by referring to the P-value of the Q statistic. In detail, when the P - value exceeds 0.1, the residuals are regarded as white noise.

When conducting comparisons among multiple models, in accordance with the information - based guidelines, the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC) are utilized. In general, lower values of these criteria are preferred. R^2 serves as an indicator of the time series' degree of fit, with values approaching 1 signifying a better fit.

Regarding the variable "qty", an analysis of the Q statistic results reveals that Q6 does not demonstrate consequences at the given level. Thus, the hypothesis that the model residuals are white noise sequences cannot be rejected. Moreover, the goodness - of - fit R^2 of the model is 1.0, indicating outstanding model performance and that the model fundamentally meets the specifications.

Typically, the performance of an ARIMA model can be evaluated using metrics such as the Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and Mean Absolute Error (MAE). Both the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC) are applied during the model selection process. These criteria take into account factors like model complexity and goodness of fit. Generally, a lower AIC or BIC value implies a higher - quality model.

AR Model Expressions:

$$Y_t = c + \varphi_1 Y_{t-1} + \varphi_2 Y_{t-2} + \cdots + \varphi_p Y_{t-p} + \varepsilon_t \tag{2}$$

MA Model Expressions:

$$Y_t = \mu + \epsilon_t + \theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2} + \cdots + \theta_p \epsilon_{t-p} \tag{3}$$

ARIMA Model Expressions:

$$Y_t = \mu + \epsilon_t + \theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2} + \cdots + \theta_p \epsilon_{t-p} + c + \varphi_1 Y_{t-1} + \varphi_2 Y_{t-2} + \cdots + \varphi_p Y_{t-p} + \varepsilon_t \tag{4}$$

Combined, ARIMA models can capture trends in statistics and handle statistics that are temporary, abrupt, or noisy.

When the P - value of a time series is less than 0.05, it shows significance. This implies that the null hypothesis is refuted, and the time series can be regarded as a stationary one. In contrast, when the P - value is greater than or equal to 0.05, the time series is considered non - stationary.

In the context of the Augmented Dickey - Fuller (ADF) test result, the statistical value of the null hypothesis is compared against the critical values corresponding to the 1%, 5%, and 10% significance levels. When the ADF test result is simultaneously lower than the values corresponding to the 1%, 5%, and 10% significance levels, it strongly suggests that the null hypothesis is convincingly rejected.

The order of differencing essentially involves subtracting the previous value from the next value. This operation is mainly aimed at eliminating certain fluctuations, thus making the statistical data tend to be more stable. Non - stationary time series can be transformed into stationary ones via

differencing operations. The Akaike Information Criterion (AIC) is utilized as a metric to assess the goodness - of - fit of a statistical model. Typically, a lower AIC value indicates a higher quality of model fit. A critical value represents a fixed value associated with a particular significance level.

Based on the variable "qty", the outcomes of the series test demonstrate that when the differencing is set at the 0th, 1st, and 2nd orders, the significance P - value amounts to 0.000***. This evidences significance at the level, resulting in the null hypothesis being rejected. Consequently, it is inferred that the series represents a stationary time series.As shown in Table 3:

Table 3. Inspection form

ARIMA model (0,1,0) test table		
item	symbol	value
	Df Residuals	331334
Sample size	N	331336
Q statistic	Q6 (P-value)	0(1.000)
	Q12(P-value)	0(1.000)
	Q18 (P-value)	0(1.000)
	Q24 (P-value)	7.66(0.983)
	Q30 (P-value)	7.667(0.999)
Information guidelines	AIC	264160.402
	BIC	264181.824
Goodness of fit	R^2	1
Note: ***, **, and * represent the significance levels of 1%, 5%, and 10%, respectively		

The ARIMA model demands that its residuals exhibit no autocorrelation, meaning the model residuals should be white noise. To verify this, one can examine the model test table and assess the white noise property of the model based on the P - value of the Q statistic. Specifically, when the P - value is greater than 0.1, the residuals are considered white noise.

For multiple model comparisons, the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC) are utilized following the information guidelines. In general, lower values of these criteria are preferred. R^2 serves as an indicator of the time series' degree of fit, with values approaching 1 signifying a better fit.

Regarding the variable "qty", an analysis of the Q statistic results reveals that Q6 does not demonstrate significance at the given level. Thus, the hypothesis that the model residuals are white noise sequences are beyond rejection. Moreover, the goodness - of - fit R^2 of the model is 1.0, indicating excellent model performance and that the model essentially meets the requirements.

Typically, the performance of an ARIMA model can be evaluated using metrics such as the Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and Mean Absolute Error (MAE). In conjunction with the Akaike Information Criterion (AIC), the Bayesian Information Criterion (BIC) is harnessed during the model - selection protocol. These criteria take into account factors like model complexity and goodness of fit. Generally, a lower AIC or BIC value implies a higher - quality model.At the same time, applied to the LSTM model, this recurrent neural network can model the long-term dependence of time series and support multiple input characteristics of the model.

3.2. Lassco regression precisely selects the best features

Lasso regression is a variant of linear regression that attains characteristic selection by incorporating an L1 regularization term. The L1 regularization term causes the coefficients of some features to become 0, thus achieving feature sparsity, i.e., selecting only features that exert a substantial influence on the target parameter [9] .

The objective function of Lasso regression can be expressed as:

$$Loss = \frac{1}{m}\sum_{i=1}^{m}(y_i - \hat{y}_i)^2 + \lambda\sum_{j=1}^{n}|\beta_j| \quad (5)$$

Where Y stands for the target attribute, X represents the matrix composed of features , β is the regression coefficient, and λ is the regularization parameter.

However, due to the regularization parameter alpha of Lasso regression, the choice of the feature selection has a great influence on the result of feature selection. A larger alpha value results in more feature coefficients of 0, resulting in more sparse feature selection results. Therefore, the appropriate alpha value can be selected through methods such as cross-validation.

In addition, Lasso regression can also control the number of iterations by setting max_iter parameters to ensure model convergence. If the model does not converge, you can increase the value of the max_iter.

In summary, Lasso regression realizes feature selection by adding L1 regularization terms, and the best feature can be selected according to the coefficients of the feature. In actual applications, there is a requirement to adjust the value of the regularization parameter alpha according to the characteristics and requirements of the statisticsset, and perform appropriate model tuning.

By comparing the fit cases, we still chose K-Means clustering to select the features.

3.3. K-Means cluster analysis

The K - Means algorithm is designed with the intention of minimizing the within - cluster sum of squares. This is because a smaller E value demonstrates a more substantial degree of likeness within the set of samples within the "cluster". Nevertheless, this minimization procedure is an NP - hard problem, and there exists no efficient algorithm. As a result, one can only perform an iteration over all potential combinations (i.e., "clusters"), which is very resource-wasting or even unsolvable when the sample size is large[10] . Figure 5 -Means Number of Clusters is shown.

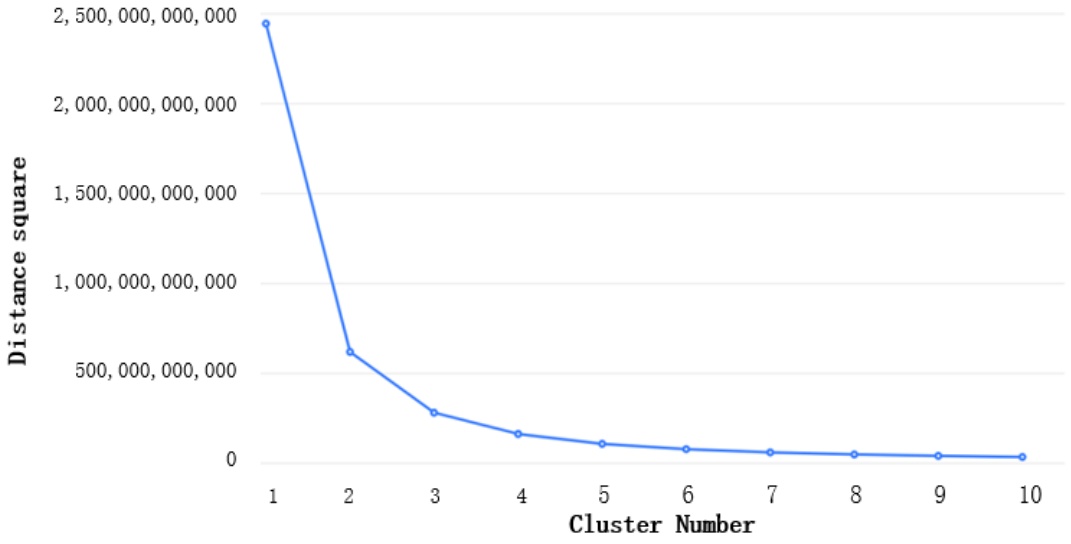

Figure 5. K-MeansThe number of clusters is displayed

ANOVA results indicated that for the variable “date,” the significance P - value was 0.954. This demonstrated no significance at the given level, and thus the null hypothesis could not be rejected. It implied that there was no substantial difference among the categories separated by cluster analysis. Regarding the variable “product_no” the significance P - value was 0.032**, signifying significance at the level. Consequently, the null hypothesis was rejected, suggesting a significant difference between the “product_no” categories derived from cluster analysis.

For the variable “category1,” the significance P - value was 0.000***, indicating significance at the level. As a result, the null hypothesis was rejected, signifying a significant difference among the categories divided by cluster analysis for this variable. For the variable “qty” the significance P -

value was 0.000***, showing significance at the level. The null hypothesis was rejected, meaning there was a significant difference between the categories separated by cluster analysis. For the variable “seller_no” the significance P - value was 0.000***, demonstrating significance at the level. By rejecting the null hypothesis, it was clear that there was a significant difference among the categories divided by cluster analysis for the “seller_no” variable.

For the variable “category2” the significance P - value was 0.002***, which attained significance at that particular degree. The null hypothesis was rejected, indicating a significant difference among the categories divided by cluster analysis in the “seller_category” variable. For the variable “inventory_category” the significance P - value was 0.000***, manifesting significance at the predefined threshold. Rejecting the null hypothesis implied a significant difference among the categories divided by cluster analysis for the “inventory_category” variable.

The outcomes of cluster analysis revealed that the clustering results were partitioned into two categories. The frequency of clustering category_1 was 15409, constituting 50.037% of the total. The frequency of clustering category_2 was 15386, accounting for 49.963%. As shown in Figure 6:

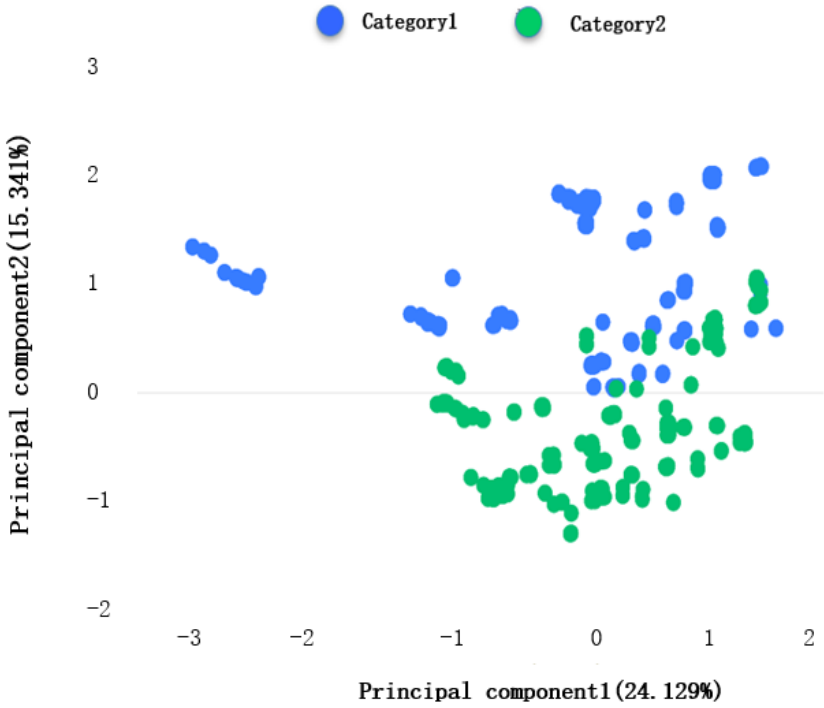

Figure 6. Distribution of major component dispersion points

Its profile coefficient can be described as:

Intra-cluster dissimilarity: the average distance of each sample centroid (prototype, center) Between - cluster dissimilarity: The distance of each sample to the mean value of all samples within a particular different cluster is referred to as the dissimilarity between the sample and that "cluster", and the "cluster" dissimilarity of the sample = $\min\{b_{ij}b_ib_{i1}b_{i2},.....\}$.

The silhouette parameter of Sample i was characterized on the basis of the intra - cluster disparity and between - cluster distinction of sample i.

4. Conclusions

This study comprehensively applied various methods to address and analyze the issue of demand forecasting in e-commerce retail. This model not only accurately forecasts the demand for various types of goods but also deeply analyzes the importance of statistics throughout the entire forecasting process. From multiple aspects such as statistics integrity, accuracy, and statistics dimensions, it reveals the crucial impact of statistics on the forecasting results. This series of achievements provides highly valuable reference for merchants in the e-commerce retail field in their daily operations and

strategic decision-making, enabling merchants to better grasp market dynamics, optimize inventory management, and enhance operational efficiency.

Although certain achievements have been obtained in this context, there remains substantial potential for enhancement. In subsequent research, more influencing factors such as the macroeconomic situation, changes in policies and regulations, and the orientation of social media public opinion can be introduced to enrich the input variables regarding the model, and refine the model's architecture to enhance the prediction precision.

References

[1] Klein J F, Stead S, Salge O T, et al. Forecasting the future of smart hospitals: findings from a real-time delphi study. [J]. BMC health services research, 2024, 24 (1): 1421.

[2] Savas S ,Mehmet A .Marine propeller underwater radiated noise prediction with the FWH acoustic analogy part 3: Assessment of full-scale propeller hydroacoustic performance versus sea trial statistics[J].Ocean Engineering, 2022, 266 (P2).

[3] Dudukcu V H, Taskiran M, Kahraman N. UAV instantaneous power consumption prediction using LR-TCN with simple moving average [J]. Concurrency and Computation: Practice and Experience, 2023, 36 (3).

[4] Örkényi L. A New Method for an Objective Measurement of the Judicial Workload–the Application of a Prediction Model Based on an Algorithm Formed by Multiple Linear Regression in Court Administration [J]. International Journal for Court Administration, 2022, 13 (1): 2-2.

[5] Chen X, Singh M M, Geyer P. Utilizing domain knowledge: Robust machine learning for building energy performance prediction with small, inconsistent statisticssets [J]. Knowledge-Based Systems, 2024, 294111774.

[6] M C E, J S M, J R W, et al. Unsupervised machine learning methods and emerging applications in healthcare. [J]. Knee surgery, sports traumatology, arthroscopy: official journal of the ESSKA, 2022, 31 (2): 376-381.

[7] Liu X, Salem S, Bian L, et al. Application of machine learning algorithms in the domain of financial engineering [J]. Alexandria Engineering Journal, 2024, 9594-100.

[8] Sharma V, Ghosh S, Mishra N V, et al. Spatio-temporal Variations and Forecast of PM2.5 concentration around selected Satellite Cities of Delhi, India using ARIMA model [J]. Physics and Chemistry of the Earth, 2025, 138103849-103849.

[9] Thao P P, Huong T T L. Unveiling key determinants of higher education expenditure in Vietnam: an adaptive LASSO approach in Tobit regression analysis [J]. Journal of Applied Research in Higher Education, 2025, 17 (2): 833-851.

[10] Zhao Y, Lin C, Ni A, et al. Interaction characteristics between rock burst risk region and neighboring fractural plane: a numerical investigation based on moment tensor inversion and Kmeans++ clustering [J]. Computational Particle Mechanics, 2024, (prepublish): 1-26.