Transactions on Computer Science and Intelligent Systems Research
ISSN: 2960-1800, eISSN: 2960-2238 | Volume 5
AIDML 2024

Research on Product Demand Forecasting Based on Random Forest and ARIMA Time Series: Precision Forecasting Method for Data-Scarce Environments

Shuo Wu [*, #], Zeshuo Zhang [#], Yixuan Ru [#]

School of Computing Science, North China Institute of Science and Technology, Sanhe, China, 065201

* Corresponding author: zyjvsn263178@gmail.com

[#]These authors contributed equally.

Abstract. In the field of e-commerce, accurately predicting the future demand for specific products is crucial for efficient inventory management and supply chain optimization. Although traditional time series forecasting methods have shown reliability in handling rich data, they are significantly affected by seasonal fluctuations and sudden market events. This is especially true for newly launched products or long-tail products with insufficient sales data, which often results in poor performance, limiting the accuracy and feasibility of traditional models in practical applications. To address this issue, this paper proposes an innovative forecasting framework based on extensive historical sales data from e-commerce platforms. This framework employs random forests to deeply mine and learn from historical data, and uses the K-Means clustering algorithm and Euclidean distance to identify the most matching sales feature trends for target products with limited data. It further provides accurate predictions for these products through the ARIMA time series model, offering robust support for further optimization of the forecasting model. Ultimately, by comparing with traditional models, the innovative forecasting framework presented in this paper demonstrates a significant improvement in accuracy when predicting sales of products with limited data, overcoming the challenges of forecasting under data constraints, and fully proving its effectiveness and innovativeness.

Keywords: Random Forest; ARIMA Time Series Forecasting; Euclidean Distance; K-Means Cluster Analysis; Product Demand Prediction.

1. Introduction

With the rapid development of the socio-economy and the gradual increase in people's consumption levels, e-commerce platforms play an essential role as the core channel of modern consumption. Merchants rely on dedicated warehouses of e-commerce platforms for unified management of goods, and accurate prediction of product demand becomes key to ensuring maximum profit, reducing inventory costs, and ensuring timely delivery. By analyzing a large amount of sales data, e-commerce platforms can predict future product demands, thus preparing inventory in advance. Although analyzing a large volume of sales data can predict future product demands to a certain extent, traditional demand forecasting methods often have limited effects when facing real challenges such as data scarcity, large market fluctuations, and numerous influencing factors.

However, implementing this forecasting process is not simple. Especially for newly launched or long-tail products, the limited amount of historical sales data makes demand forecasting more complex. Additionally, product sales are easily affected by many market events, such as holidays and promotional activities, making demand forecasting more challenging. Supply chain management of e-commerce platforms is a complex and important task. To ensure that goods can be smoothly delivered from warehouses to consumers, e-commerce platforms need to continuously optimize and innovate to ensure their supply chain management is both efficient and economical [1].

Current methods for predicting demand mainly include three types: time series forecasting methods, machine learning predictions, and multi-model combination forecasts [2]. Traditional time series

forecasting models are mainly used for linear data predictions with small sales volumes and minor sales fluctuations and are difficult to adapt to today's e-commerce demand data, which is complex and variable. As methods like machine learning mature, many scholars apply machine learning to solve related problems [3]. However, due to many factors affecting the prediction results in practice, the prediction accuracy of a single model is generally low, and multi-model combination forecasting has gradually become a trend [4]. Scholar Wei Kaicong used machine learning and linear regression to establish a forecasting model to predict taxi demand, improving the accuracy of predictions [5]. Scholar Zhang Jingjing combined time series and BP neural network models to predict PM2.5 values in situations of large data volumes and discrete data, choosing the most suitable prediction model according to local conditions [6]. In recent years, multi-model combination forecasting methods have gradually received attention due to their ability to integrate the advantages of various models, but how to effectively combine different methods to address the challenge of data scarcity remains an open question.

This paper aims to fill this research gap by proposing an innovative framework that combines random forests, K-Means clustering algorithms, and ARIMA time series forecasting, specifically for accurate demand forecasting of products with limited historical sales data. The data of this article is based on http://www.mathorcup.org.Initially, this paper utilizes the random forest model to deeply analyze a large amount of historical sales data from e-commerce platforms for subsequent forecasting, identifying the sales sequence characteristics of different product combinations. Then, through the K-Means clustering algorithm, sales sequences are classified to discover product groups with similar sales feature trends. Finally, combining the ARIMA model, precise demand forecasting is conducted for specific products with scarce data. Moreover, this study particularly considers the volatility of product sales, market events, and other factors that may affect demand, to enhance the adaptability and accuracy of the forecasting model. Compared to existing methods, this study not only improves the accuracy of predictions under conditions of limited data but also provides a new intelligent decision-making tool for the supply chain management of e-commerce platforms.

2. Establishment of the Prediction Model

2.1. Establishment of the Random Forest Model

After preprocessing a large historical sales dataset, this paper performs descriptive statistics, calculating the characteristic values of product data across various dimensions.

This paper then conducts statistical analysis of the dataset and visualizes the results, as shown in Figure 1.

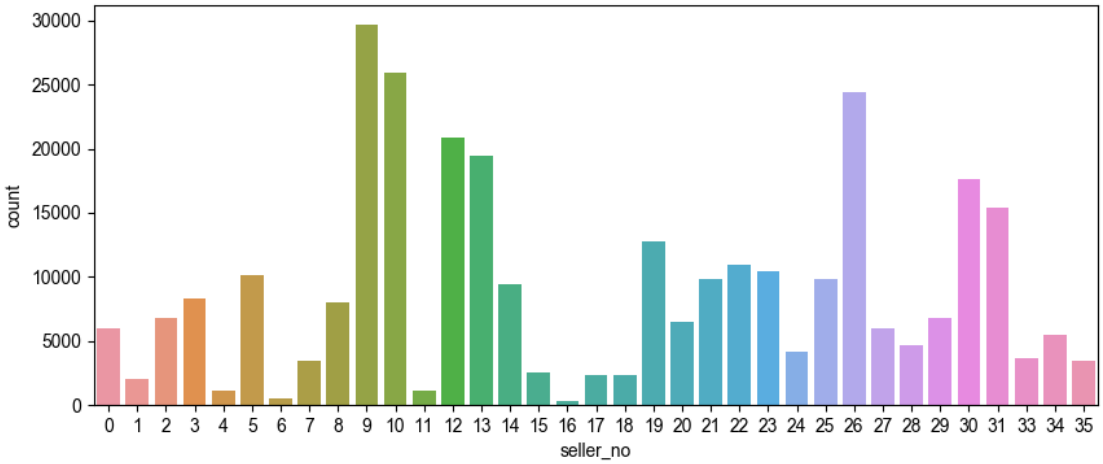

Figure 1. Post-Coding

The data characteristics of the combinations of various merchants, warehouses, and products in the dataset are too discrete, with significant fluctuations, and the data exhibit clear seasonal features. If

traditional time series models are directly used, although the models can roughly learn the basic feature engineering information of product sales trends, the error is large, and the model accuracy is too low.

Furthermore, considering that the real-life demand for products is influenced by holidays, promotions, and other activities, traditional time series models, which only use time as an analytical factor, cannot consider the impact of other factors on product demand. Therefore, they cannot accurately learn the multi-scale features of product time series. After exploring multiple models, this paper finds that the random forest model can handle large datasets and adapt to discrete data, while also considering the impact of holidays, promotions, and other activities. The random forest model also has the capability to select data characteristics of different merchants and products. For these reasons, this paper adopts the random forest model for simulation [7]. The random forest model was chosen for learning the time series features and preliminary prediction [8], as shown in Figure 2.

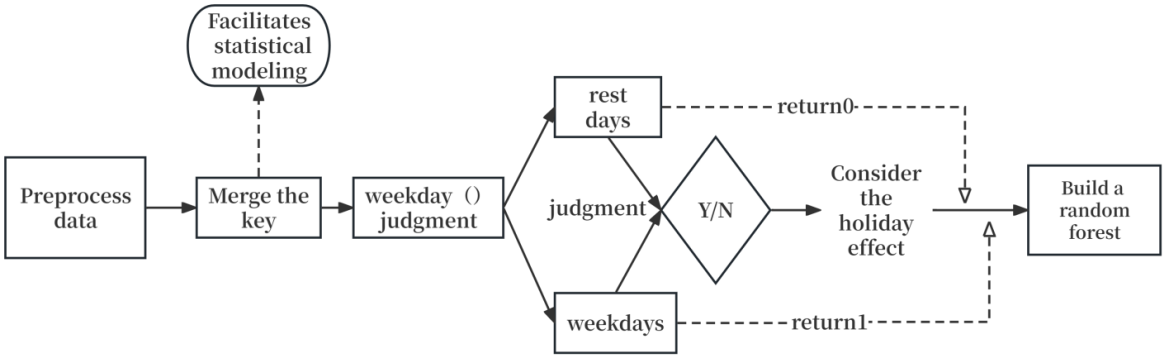

Figure 2. Random Forest Model Establishment Process

The mathematical model of Random Forest can be expressed by Formula (1):

$$F(x) = argmax\left(1 / N \cdot \sum F\_i(x)\right) \tag{1}$$

The prediction result of each decision tree is denoted as F(x), and the accuracy of the analysis is improved by combining multiple decision trees.

In summary, this paper applies the Random Forest model to deeply mine and analyze the large-scale historical sales data of the e-commerce platform, successfully identifying the basic features and multi-scale time series features of product sales trends. This step not only demonstrates the powerful capability of Random Forest in handling complex and sparse datasets but also lays a solid foundation for further analysis. Specifically, through the analysis with the Random Forest model, this paper can more accurately identify the key features and patterns in the sales data, providing crucial preliminary information and support for the subsequent identification of product groups with similar sales trends using the K-Means clustering algorithm. This step is a key link in establishing a high-accuracy prediction model, which will directly affect the performance and practical value of the prediction model.

2.2. K-Means Clustering Analysis

This study utilizes the preliminary analysis results from the Random Forest to further refine and categorize the sales feature trends in the dataset through K-Means clustering [9], aiming to identify those similar patterns that are particularly crucial for future product demand forecasting.

Initially, the data is merged, classifying and grouping different products from different warehouses of different merchants. Then, by extracting feature values for the target forecast products and each group, relevant features are calculated, such as mean, variance, skewness, kurtosis, as well as some features derived from time series data, like Fourier spectra and wavelet transforms. The data for each

group is then dimensionally elevated to enable a more comprehensive comparison and identify similarity in features.

After extracting the feature values for each group, the feature values are normalized. The feature values are transformed into a standard normal distribution with a mean of 0 and a variance of 1, thereby eliminating the scale differences between different features. This method can reduce the importance of features, increase the stability and reliability of the model, and improve the accuracy of the model.

Then, the Elbow Method is used to determine the K value. By classifying the data, a curve graph of the number of clusters and their compactness is drawn, as shown in Figure 3.

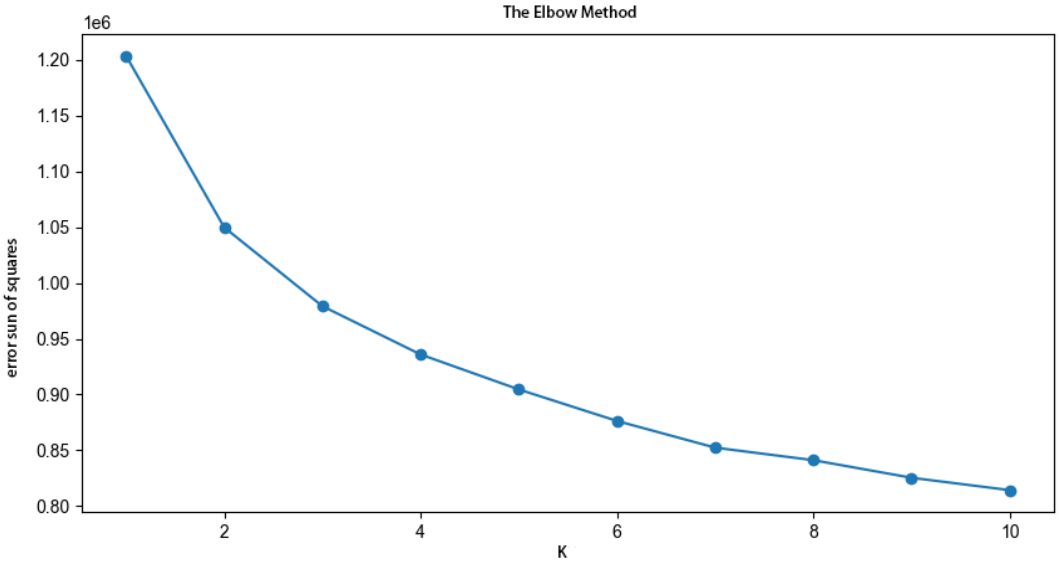

Figure 3. The Elbow Method

By observing the curve, it is found that as the number of clusters K increases, the sum of squared errors decreases. However, when more clusters are added, the rate of decrease becomes slower. The "elbow" is the point where the speed of the decrease in the sum of squared errors starts to slow down, so the value of k can be chosen at the "elbow".

The "elbow" is roughly located in the range of 2-4 clusters. When k>3, it may increase the noise of model feature values and the added computational complexity affects computational efficiency, so K is chosen for cluster analysis.

Finally, different features are divided into K groups, with K feature vectors randomly selected from each group as the initial cluster centers. Then, the distance between each feature and the cluster centers is calculated, and each feature vector is assigned to the nearest cluster center. The cluster centers and the high-dimensional feature vectors assigned to them represent a cluster. Each time a product data is assigned, the cluster center will be recalculated based on the existing data in the cluster. This process will iterate continuously until the condition that the data in this group is most similar in demand is met.

The final classification results are visualized, as shown in Figure 4. below.

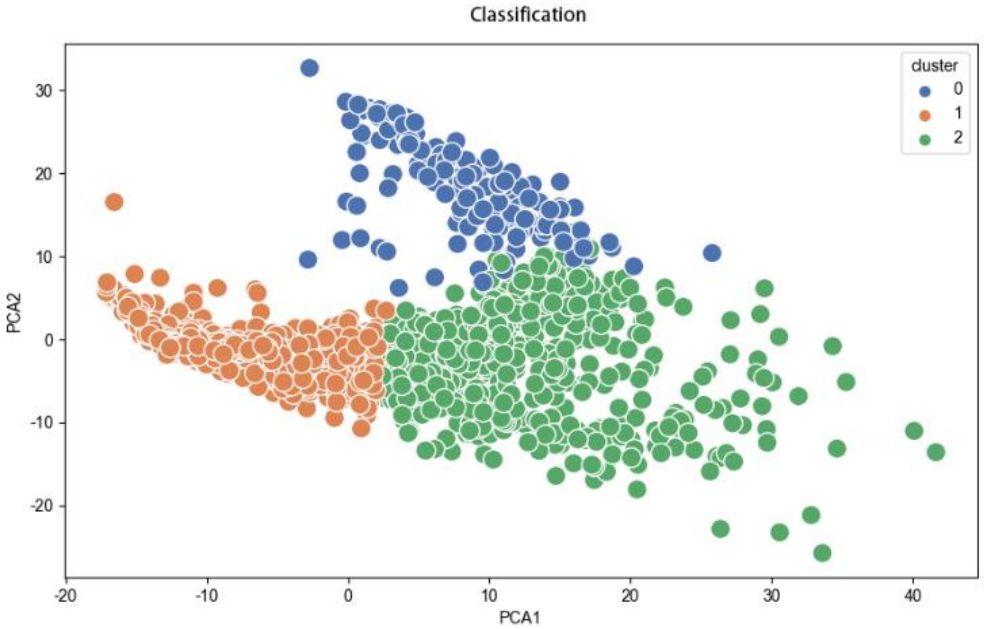

Figure 4. Classification

The classification results are partially shown in Table.1 below. Through K-Means clustering analysis, this paper has successfully divided the e-commerce sales data into three distinct categories, each representing a group of product sales sequences with similar features, while also finding the sales sequences most like the target forecast products. This classification not only reveals the intrinsic structure of the product sales data in the market but also provides a solid foundation for subsequent in-depth analysis and prediction of product demand. Especially, by determining the optimal number of clusters and identifying specific cluster centers, a clear starting point is set for the next step of the analysis.

Table 1. Partial Prediction Result Chart

Class 1	Class 2	Class 3
seller_11_wh_1_product_135	seller_21_wh_15_product_509	seller_11_wh_1_product_133
seller_14_wh_19_product_235	seller_10_wh_13_product_1930	seller_12_wh_19_product_321
seller_14_wh_20_product_222	seller_10_wh_1_product_1916	seller_14_wh_19_product_214
seller_5_wh_1_product_31	seller_11_wh_16_product_1781	seller_7_wh_1_product_300
seller_33_wh_51_product_1375	seller_19_wh_28_product_438	seller_6_wh_1_product_1296
seller_9_wh_1_product_102	seller_22_wh_35_product_616	seller_5_wh_6_product_31
seller_30_wh_23_product_1001	seller_11_wh_16_product_1580	seller_14_wh_1_product_202
seller_28_wh_1_product_776	seller_23_wh_36_product_584	seller_5_wh_1_product_1730
seller_33_wh_54_product_1375	seller_25_wh_14_product_1414	seller_3_wh_1_product_460
seller_2_wh_1_product_85	seller_27_wh_25_product_645	seller_9_wh_1_product_96

2.3. Establishment of the ARIMA Model

For newly launched products or long-tail products with insufficient sales data, by using Euclidean distance as the measurement standard [10], this study first calculates the similarity between product sales data points and each cluster center to accurately locate the product combination most similar in

demand volume, sales trends, and other key sales indicators. This process not only deepens the understanding of market dynamics but also improves the prediction accuracy of future demand for specific sales trend products.

To enhance the prediction accuracy and efficiency of the model, after determining the category most like the target forecast product, this study calculates the Euclidean distance between the long time series of products in that category and the time series of the target forecast product to find the most similar product sales sequence. To improve computational efficiency, this paper introduces the sliding window subsequence method for this sales sequence. Specifically, in the selected long time series, sliding window subsequences of 45 days each are established. These subsequences are arranged in sequence at certain time intervals, for example, from December 1 to January 14, from December 2 to January 15, and so on. By calculating the Euclidean distance between the target forecast sequence and these sliding subsequences to find the smallest Euclidean distance, this study can efficiently identify the most similar sales feature trend.

The formula for Euclidean distance is shown in Equation (2):

$$Distance(X,Y) = \sqrt{\sum_{i=1}^{n}(X_i - Y_i)^2} \tag{2}$$

In the equation, X and Y represent two time series, X_i and $Y_i$respectively are their values at the ith time point, with n represents the length of the time series, meaning the total number of observations.

Considering the substantial computational effort required for comparing all long time series in the entire category with sliding window subsequences, leading to low efficiency in practical forecasting for businesses, this study adopts a strategy of randomly extracting a total of 20 subsequences from a long time series, ensuring at least one sequence is extracted per month. It then uses Euclidean distance to filter out the subsequence most like the target forecast sequence. Following this, around the month of the most similar subsequence, the study continues to extract 15 more sliding window subsequences and repeats the calculation of Euclidean distances. Through an iterative squeezing method, the search range is gradually narrowed down until the most similar subsequence is found, whose Euclidean distance to the target forecast sequence is the closest and whose iterative Euclidean distance almost remains constant. The specific method is illustrated in Figure 5.

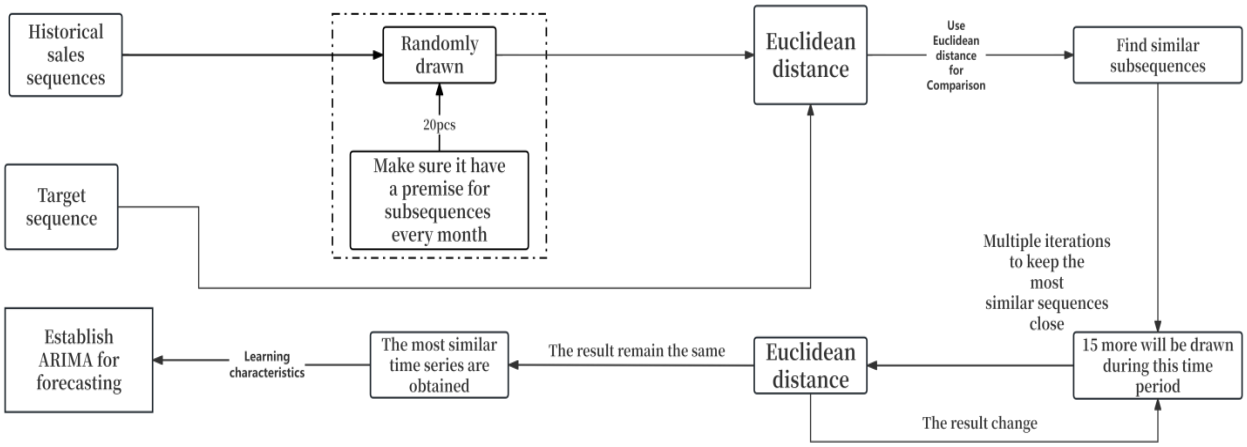

Figure 5. ARIMA Model Establishment Process

Finally, once the specific subsequence most like the target forecast sequence has been identified, this paper uses the features of that sequence to establish an ARIMA time series model for the final prediction [11]. The establishment of the ARIMA time series model is shown in Equations (3) to (5).

$$ARIMA(p,d,q) = AR(p) + I(d) + MA(q) \tag{3}$$

$$AR(p) = c + \varphi_1 Y_{t-1} + \varphi_2 Y_{t-2} + \cdots + \xi_t \tag{4}$$

$$MA(q) = \mu + \epsilon_t + \theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2} + \cdots + \theta_q \epsilon_{t-q} \tag{5}$$

This process not only significantly improved the accuracy of the forecast, targeting the forecasting challenges of newly launched or long-tail products with insufficient sales data, but also greatly reduced computation time, providing businesses with a faster forecasting capability. This offers strong decision support for the supply chain management and inventory optimization of e-commerce platforms, enabling them to maintain competitiveness in a rapidly changing market environment.

The final forecast results are presented, in part, as shown in Figure 6:

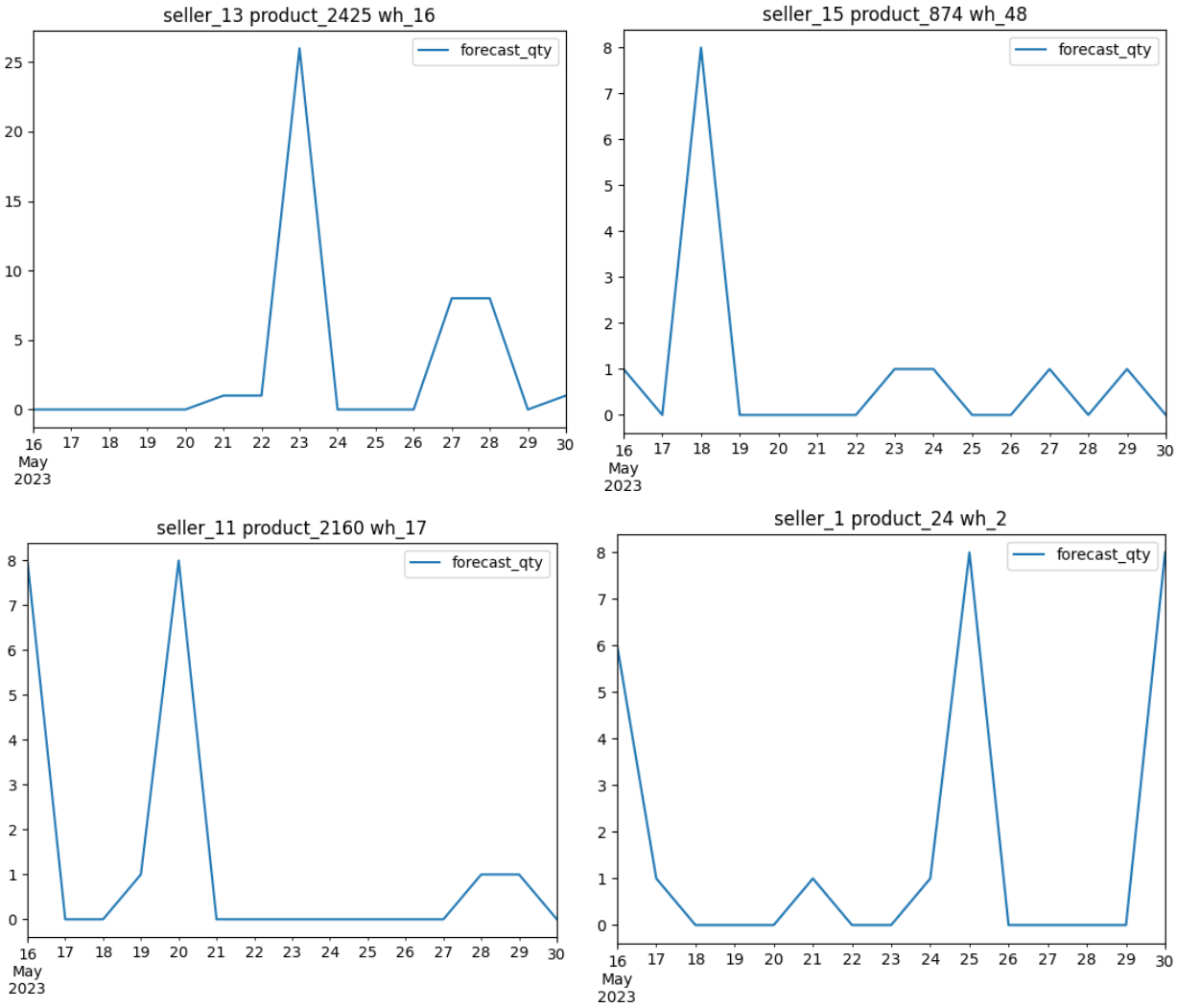

Figure 6. Partial Product Forecast Results

3. Results

3.1. Model Accuracy and Performance Analysis

① Evaluation of Forecast Accuracy

A commonly used metric for evaluating forecast accuracy is WMAPE (Weighted Mean Absolute Percentage Error). This is a measure of weighted average absolute error in percentage terms, used to assess the accuracy of model predictions. WMAPE can consider the different impacts of various times on forecast error, making it adaptable to models that account for seasonal variations, holidays, promotions, and other events affecting demand.

The larger the value of 1-wmape, the smaller the error between the model's predictions and the actual results, indicating better model accuracy.

The calculation formula (6) is as follows:

$$1 - wmape = 1 - \frac{\sum|y_i - \hat{y}_i|}{\Sigma y_i} \quad (6)$$

In the formula, y_i represents the actual demand for the ith sequence, and $\hat{y}_i$ represents the predicted demand for the ith sequence.

The larger the value, the smaller the error between the model's predictions and the actual results, indicating a better accuracy of the model.

After calculation, the average WMAPE value is 0.171, and the value of 1-WMAPE is 0.829. This indicates that the model has good prediction accuracy.

② Evaluation of Forecast Error and Stability

For the evaluation of model prediction error and stability, this paper uses R^2, MAE values, as shown in Equation (7).

$$R^2 = 1 - (Sum\ of\ Squared\ Residuals/Total\ Sum\ of\ Square) \quad (7)$$

R^2 is the proportion of the sum of squares of differences between actual and predicted values to the total variation. The range of R^2 is (0, 1), where a lower value indicates that the model can explain less of the data's variability, such as the impact of holidays on product demand.

MAE (Mean Absolute Error) is the average of the absolute errors between predicted values and actual values, as shown in Equation (8). It is used to measure the average deviation between model predictions and actual values. A smaller MAE indicates higher model accuracy.

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i| \quad (8)$$

N represents the number of samples, y_i represents the number of actual values, $\hat{y}_i$represents the number of forecast values.

By calculating, the R^2value is 0.8281, indicating that the model explains approximately 82.81% of the variability in the data. This is a good result, suggesting that the model can predict well in most cases. The MAE value is 2.798, indicating that the model predictions are relatively accurate.

In summary, the innovative forecasting model framework proposed in this paper demonstrates high accuracy. The model effectively captures internal patterns within the data, considering factors such as holidays, promotions, etc., and provides reliable forecasts for businesses.

4. Conclusion

This study proposes an innovative demand forecasting framework specifically tailored for newly launched products or long-tail products with limited sales data. By employing a combined approach of random forest, K-Means clustering algorithm, and ARIMA time series forecasting model, this research effectively addresses the limitations of traditional forecasting models when dealing with sparse data, market volatility, and multiple influencing factors. It also demonstrates the powerful capability of accurately predicting product demand on e-commerce platforms.

Through in-depth analysis of vast historical sales data from e-commerce platforms, this study successfully identifies sales sequence characteristics for different product combinations. The K-

Means clustering algorithm efficiently categorizes sales sequences into groups with similar sales trends, thereby enhancing prediction efficiency and providing businesses with faster forecasting capabilities, enabling them to gain a competitive edge in rapidly changing market dynamics. Additionally, this research fully considers sales volatility, market events, and other potential influencing factors in model design, further improving the adaptability and accuracy of the forecasting model.

In comparison with existing forecasting methods, this study not only significantly improves prediction accuracy under limited data conditions but also provides a novel intelligent decision support tool for e-commerce platforms. This tool aids in optimizing supply chain management, reducing inventory costs, mitigating stockout risks, and ensuring timely delivery of goods to consumers. The effectiveness and practical value of the multi-model fusion forecasting method in the rapidly changing e-commerce environment are successfully validated through WMAPE verification.

In summary, this study not only theoretically pioneers a new method for demand forecasting but also demonstrates its effective application in practice, particularly for newly launched products or those with limited sales data. The findings of this research are expected to provide substantial support for the operational management of e-commerce platforms and lay a theoretical and practical foundation for exploring new methods of demand forecasting in e-commerce for future research endeavors.

References

[1] WangXuan, ZhouYun. Research on the digital application of manufacturing enterprise procurement supply chain platform [J]. China Shipping Weekly, 2024 (13): 72 - 74.

[2] Yi Keke. Research on Q Company’s FMCG demand forecast based on machine learning [D]. Shanghai: Donghua University, 2023.

[3] Xu Dongdong. Research on data-driven retail network inventory replenishment[D]. Shanghai: Donghua University, 2022.

[4] Fang Kuangnan, Wu Jianbin, Zhu Jianping, et al. A review of research on random forest methods [J]. Statistics and Information Forum, 2011, 26 (3): 32 - 38.

[5] Wei Kaicong. Research on taxi demand forecasting based on machine learning methods [D]. Guangzhou University, 2020.

[6] Zhang Jingjing. PM2.5 combination forecast based on ARIMA-BP neural network [D]. Lanzhou University, 2023.

[7] Wen Jiaxuan, Wang Miao, Liu Ji. A Time Series Multi step Prediction Algorithm Based on Time Series Decomposition and Random Forest [J]. Journal of East China University of Science and Technology (Natural Science Edition), 2023, 49 (06): 873 - 881.

[8] Zhao Juanhe. Research on short-term demand forecasting method for e-commerce commodities based on multi-layer hybrid deep neural network [D]. Chang'an University, 2020.

[9] Zhang Yifan, Hu Jiahao, Li Yiqiao. Research on commodity clustering based on kmeans algorithm [J]. Digital Technology and Applications, 2020, 38 (04): 108 - 110.

[10] Li Kewen, Lin Yalin, Yang Yaozhong. An improved SDRSMOTE algorithm based on Euclidean distance [J]. Computer Engineering and Science, 2019, 41 (11): 2063 - 2070.

[11] Nie Yuxuan. Research on automatic pricing and replenishment strategy of fresh commodities based on ARIMA prediction optimization model - taking vegetable commodities as an example [J]. Trade Exhibition Economics, 2024 (05): 19 - 22.