Proceedings of the 2nd International Conference on Mathematical Physics and Computational Simulation
DOI: 10.54254/2753-8818/36/20240621

Modeling e-commerce retailer demand using Stacking Ensemble and K-means clustering patterns

Ziyi Xie [1,3], **Qiyang Xie** [2]

[1]College of Business and Public Management, Wenzhou-Kean University, Zhejiang Wenzhou, 325060, China
[2]Overseas Education College, Jimei University, Fujian Xiamen, 361021, China

[3]xieziyi@kean.edu

Abstract. Demand forecasting plays a crucial role in e-commerce. By accurately forecasting product demand, e-commerce companies can better manage inventory levels and decrease excess or out-of-stock situations, thereby reducing inventory costs and increasing customer satisfaction. To address the issue of the traditionally low accuracy in market demand forecasting by practitioners who relied on manual experience, this paper proposed a demand forecasting model for e-commerce retailers based on a Stacking ensemble model. The paper successfully modeled and analyzed past user behavior data from JD.com platform, and the model achieved automated forecasting of future demand based on historical data. Upon testing, the proposed Stacking ensemble model demonstrated RMSE (Root Mean Square Error) is 256.34 and 1-WMAPE is 0.916 close to 1, indicating that the model performed well and was capable of accurately predicting the demand for these products. Additionally, this paper also employed a K-means clustering model based on a cost function to classify products into five distinct categories according to their demand characteristic indicators and product attribute indicators.

Keywords: Demand forecast, E-commerce, Ensemble Learning, Stacking, K-means clustering

1. Introduction

With the continuous progress of information technology, the scale of global e-commerce has been expanding and the value of online retail sales has been rising over the past decade or so, and e-commerce has gradually become the main driving force to promote socio-economic development and improve the quality of life of the people during continuous development. But at the same time, e-commerce platforms are also facing problems such as market demand forecasting and commodity inventory optimisation [1]. Demand forecasting is a key part of the decision-making process for e-commerce companies, and plays a vital role in improving operational efficiency, reducing costs, increasing revenues and enhancing the customer experience [2]. Traditional e-commerce operations often rely on manual judgement of market conditions, and it is difficult to achieve accurate prediction of market demand. However, with the development of big data technology, it has become possible to use machine learning modelling analysis to help e-commerce platforms achieve more efficient decision-making. Therefore, how to establish a demand forecasting model to solve the problem of intelligent demand forecasting for e-commerce enterprises has become a major issue at this stage [3]. To address the existing issues, this paper proposed a demand forecasting model for e-commerce retailers based on a Stacking ensemble model. The proposed demand forecasting model consisted of a first-layer learner that included time series, BP neural network, random forest, and Gradient Boosting Decision Tree (GBDT). Additionally, a second-layer learner was incorporated, which included a support vector machine (SVM) within the Stacking ensemble learning framework. Bedsides, this paper also utilized a K-means clustering model based on cost function to classify the commodities according to their demand characteristic indexes and commodity attribute indexes. The proposed approach was aimed at helping e-commerce retailers achieve accurate and intelligent demand forecasting, which enabled them to optimize inventory management and make better replenishment and pricing strategies.

2. Related work

In previous research, [4] used a decision tree algorithm to predict e-commerce customers repurchase behavior, based on real-time customer browsing data and historical behavioral data, exploring the application of machine learning algorithms being used in the field of demand forecasting. While [5] used artificial intelligence techniques such as XGBoost to predict the inventory demand of a cross-border e-commerce company, emphasising the need for models to be adapted to company-specific challenges and information needs.Luo et al.'s study utilized interpretable machine learning methods given to xgboost to study Australian customer needs [6] Meanwhile, the effectiveness of using deep learning in demand forecasting across multiple industries, including e-commerce were explored[7]. Specifically, it examines the effectiveness of Convolutional Neural Networks (CNN) for predicting demand in scenarios such as online car and taxi hailing, as well as the use of attention-based deep integration networks to improve prediction accuracy. The study highlights the ability of deep learning to deal with complex patterns and large datasets. Most of the past studies have used only a single model of machine learning algorithm for market demand prediction. However, the use of a single model often faces some problems, such as a single model tends to perform well on training data, but its ability to generalise to new data may be poor. Also, it may be difficult for a single model to capture all the complex patterns and relationships in the data, especially if the data is very high dimensional or the relationships are non-linear. Moreover, a single model may be very sensitive to small changes in the training data, which may lead to unstable prediction results. Therefore, this paper adopted the strategy of integrated learning to model past user data. Ensemble Learning is a machine learning approach in which the core idea is to combine several different models and use their individual predictions to form a more robust and accurate prediction. There is evidence to suggest that some joint machine learning algorithms seem to produce better generalization effects[8]. Shen et al.'s research also revealed that using multiple combinations of machine learning techniques can better reveal requirements with greater interpretability[9].In a previous study, [10] compared the application of single model, integrated model and comprehensive integrated machine learning techniques in stock market forecasting. The study concluded that integrated methods, especially those that use a combination of boosting or diverse algorithms e.g., SVM, logistic regression, and neural networks, tend to outperform single-model methods in terms of accuracy and robustness.

3. Methodology

The data used in this paper is sourced from the JD.com e-commerce platform. During the data preprocessing phase, the data was first cleansed, including handling missing values and outliers. Missing values were either deleted or filled, and outliers were identified and processed using the 3σ (sigma) rule. Subsequently, the data underwent a process of quantification, where categorical variables such as product categories, inventory classifications, and merchant scales were encoded to facilitate analysis. Additionally, the study employed canonical correlation analysis to explore the relationships between demand characteristics and product attributes for a better understanding of data distribution and correlation. In the model construction phase, the paper presented a demand forecasting model for e-commerce retailers based on a Stacking ensemble model, which consisted of two layers of learners. The specific structure and modeling process was illustrated in Figure 1. The first layer of learners included time series, BP neural networks, random forests and GBDT while the second layer comprised logistic regression models. The sample set was divided into a training set and a test set, with the training set further split into five parts. Four of these parts were used for training, and the remaining one was used for prediction, while the test set (Test Data) was also predicted. After five iterations, five sets of prediction results were produced. These five sets of prediction results formed a new dataset, represented as the 'New Feature' part in Figure 2 . Each base learner generated new data features, which served as input for the meta-learner. Five sets of predictions for the test set (Test Data) were also generated, and the average of these five predictions created a new Test Data, which was used as the test set for the meta-learner.

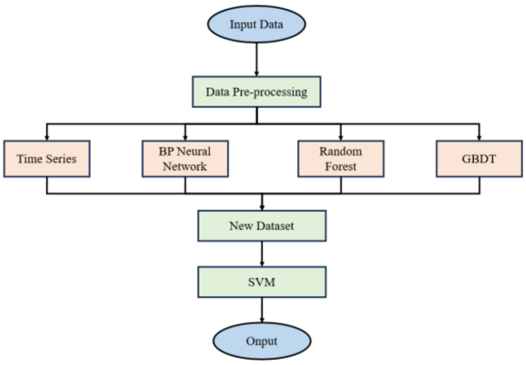

Figure 1. Detailed flow of the proposed methodology

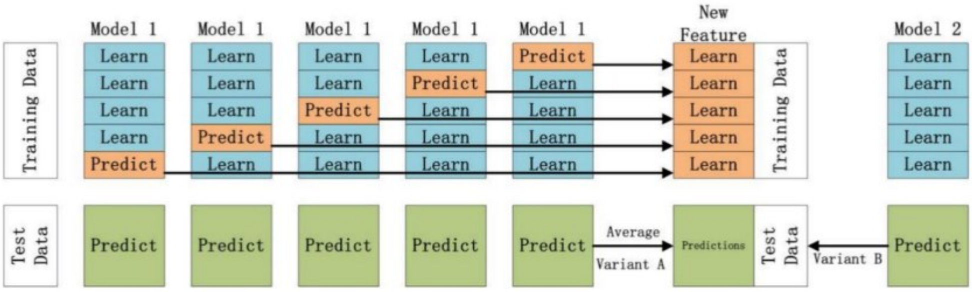

Figure 2. A diagram of the k-fold cross-validation process for a base learner

The K-means clustering model is an unsupervised machine learning algorithm that could classify unlabeled data without prior knowledge of the categories of all product demand data. The K value could be determined through a cost function by constructing a graph showing the relationship between the cost function and the number of clusters K. The inflection points on the relationship graph corresponded to the K value. The K-means clustering algorithm iteratively calculated the product demand data to derive the optimal solution for the distance from the dataset to the cluster centroids and the product demand data cluster centroids. The optimal solution results could represent a general classification of all product demand data characteristics. The basic flowchart of the algorithm was illustrated as shown in the following Figure 3.Date feature variables and rolling features were incorporated into the variables. The date feature variables included: average daily demand, maximum daily shipments, minimum daily shipments, median daily demand, standard deviation, and weekly demand, among others. If overfitting was detected, the date feature variables were removed. Rolling statistical features were calculated for the same day of each week. Subsequently, the optimal value of K was determined based on the criterion of the minimum distance cost function.

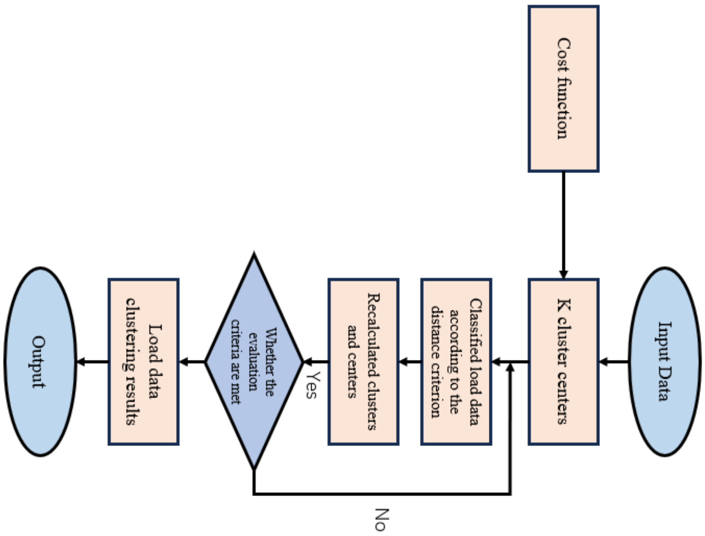
Cost function
Input Data
K cluster centers
Classified load data according to the distance criterion
Recalculated clusters and centers
Yes
Whether the evaluation criteria are met
No
Load data clustering results
Output

Figure 3. Flowchart of K-means clustering based on cost function

4. Experimental Results and Analyses

Based on the demand data for 1,994 categories of goods provided in the dataset from December 1, 2022, to May 15, 2023, the proposed model successfully predicted the demand for these goods from May 16, 2023, to May 30, 2023. As shown in Figure 3.1.1, it illustrates the actual and forecasted demand values for the category of goods 2-731-1. where the red line represents the actual value and the blue line represents the predicted value output by the model. Additionally, this paper utilized the demand data from December 1, 2022, to April 30, 2023, as the model training set, and the demand data from May 1, 2023, to May 15, 2023, as the model testing set. The e-commerce retail demand forecasting model based on the Stacking ensemble model was evaluated using the RMSE, MAE, WMAPE, and 1-WMAPE evaluation metrics. The evaluation results are presented in the following Table 1. As can be seen from the table, the evaluation indexes RMSE, MAE, and WMAPE of the Stacking integrated model are all smaller than the single prediction model, and the 1-WMAPE of the Stacking integrated model is closer to 1, which indicates that the Stacking integrated model performs better and can better predict the demand of these commodities. It can be seen that the evaluation indices RMSE, MAE, and WMAPE of the Stacking integrated model are all smaller than the single prediction model, and the 1-WMAPE of the Stacking integrated model is closer to 1, which indicates that the Stacking integrated model has better performance and can better predict the demand for these goods. From Figure 5, it can be seen intuitively that the value of the cost function decreases rapidly before the inflection point 5, and slows down after the inflection point 5, so it can be judged that the optimal number of clusters is 5.

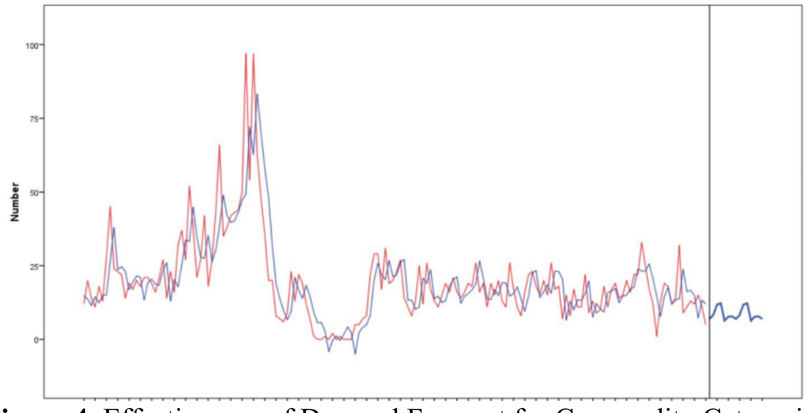

Figure 4. Effectiveness of Demand Forecast for Commodity Categories

Table 1. Calculated values of evaluation indicators for each forecasting model

Model	RMSE	MAE	WMAPE	1-WMAPE
Time Series Model	285.53	154.21	0.13	0.892
BP Neural Network	356.23	179.35	0.16	0.854
Random Forest	342.59	180.74	0.15	1.124
GBDT Model	295.41	150.96	0.13	0.902
SVM	372.84	192.08	0.18	1.205
Stacking Model	**256.34**	**135.52**	**0.09**	**0.916**

Analysis of the clustering results reveals the following characteristics for each category of goods: the quantity of goods in Category 1 is the smallest, with an average daily demand ranging between 0 to 40 units, primarily including product categories such as Food and Beverages, Mobile Communications, and Computer Office Supplies, and the warehouses are spread throughout the country. Category 2 consists of goods with a relatively low daily demand, averaging between 30 to 110 units, mainly including categories like Food and Beverages, Home Decoration and Building Materials, and Skin Whitening and Skin Care, with warehouses also located nationwide. Category 3 encompasses goods with an average daily demand in the range of 130 to 320 units, with main product categories being Home Decoration and Building Materials, Household Appliances, and Pet Health, and the warehouses are distributed across North China, South China, Central China, and East China, with merchant scales mainly being New and Medium. Category 4 includes goods with a higher daily demand, averaging between 190 to 630 units, primarily consisting of categories such as Food and Beverages, Computer Office, and Digital products, with warehouses located in North China, South China, and East China, and the merchant scale is predominantly Medium and Large. Lastly, Category 5 represents goods with the highest demand, with an average daily demand between 700 to 900 units, mainly including Home Decoration and Building Materials, and Household Appliances, with warehouses concentrated in East China and merchant scales primarily being New and Medium.

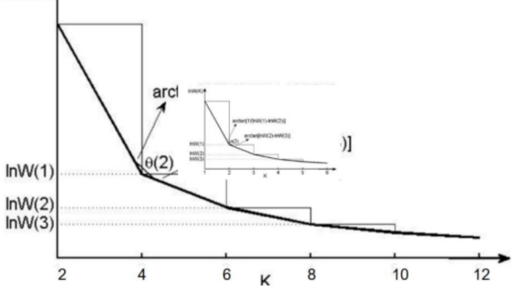

Figure 5. Plot of cost function versus K-value

5. Conclusion

In summary, this paper proposed a demand forecasting model for e-commerce retailers based on a Stacking ensemble model, which successfully automated the process of predicting future demand using historical data. Through testing and comparison with the results of single-model forecasts, the proposed Stacking ensemble model demonstrated superior performance with an RMSE of 256.34 and an MAE of 135.52, both of which were lower than those of the single models. Additionally, a 1-WMAPE value close to 1 (0.916) indicated that the Stacking ensemble model could integrate the strengths of various single predictive models, enhancing the accuracy of the forecast results. Moreover, this paper also employed a K-means clustering model based on a cost function to categorize products into five distinct categories according to their demand characteristic indicators and product attribute indicators.Since this study utilized only a subset of data mining models for the Stacking ensemble and did not optimize the combination of learners, there is room for improvement in future work. Different model combinations could be explored to achieve the optimal ensemble learning model. Furthermore, optimization algorithms could be employed to fine-tune model parameters, such as ant colony optimization, genetic algorithms, simulated annealing, and whale optimization algorithms, potentially leading to further improvements in forecasting accuracy. The integration of these advanced techniques may provide e-commerce retailers with an even more robust tool for demand forecasting, allowing for more strategic and effective business operations.

References

[1] "Ecommerce Demand Forecasting (2024 Guide) - 10XSheets." In 10XSheets Blog. [Online]. Available: https://www.10xsheets.com/blog/ecommerce-demand-forecasting. [Accessed: 26-04-2024].

[2] Li J, Cui T, Yang K, Yuan R, He L, Li M. Demand forecasting of e-commerce enterprises based on horizontal federated learning from the perspective of sustainable development. Sustainability. 2021;13(23):13050.

[3] A. Jain, V. Karthikeyan, S. B, S. BR, S. K and B. S, "Demand Forecasting for E-Commerce Platforms," 2020 IEEE International Conference for Innovation in Technology (INOCON), Bangluru, India, 2020, pp. 1-4, doi: 10.1109/INOCON50539.2020.9298395.

[4] Liu C-J, Huang T-S, Ho P-T, Huang J-C, Hsieh C-T (2020) Machine learning-based e-commerce platform repurchase customer prediction model. PLoS ONE 15(12): e0243105. https://doi.org/10.1371/journal.pone.0243105

[5] Tang, Y.M.; Chau, K.Y.; Lau, Y.-y.; Zheng, Z. Data-Intensive Inventory Forecasting with Artificial Intelligence Models for Cross-Border E-Commerce Service Automation. Appl. Sci. 2023, 13, 3051. https://doi.org/10.3390/app13053051

[6] Luo, Y., Zhang, R., Wang, F., & Wei, T. (2023, October). Customer Segment Classification Prediction in the Australian Retail Based on Machine Learning Algorithms. In 2023 4th International Conference on Machine Learning and Computer Application (pp. 498-503).

[7] Singh, G., & Yogi, K. K. (2023). Performance evaluation of plant leaf disease detection using deep learning models. Archives of Phytopathology and Plant Protection, 56(3), 209-233.

[8] Dai, W., Jiang, Y., Mou, C., & Zhang, C. (2023, September). An Integrative Paradigm for Enhanced Stroke Prediction: Synergizing XGBoost and xDeepFM Algorithms. In Proceedings of the 2023 6th International Conference on Big Data Technologies (pp. 28-32).

[9] Shen, X., Luo, S., & Zhang, M. (2023). House quality index construction and rent prediction in New York City with interactive visualization and product design. Computational Statistics, 38(4), 1629-1641.

[10] I. K. Nti, A. F. Adekoya, and B. A. Weyori, "A comprehensive evaluation of ensemble learning for stock-market prediction," J. Big Data, vol. 7, no. 20, 2020. [Online]. Available: https://doi.org/10.1186/s40537-020-00299-5 [Accessed: 26-04-2024].