Jestr

Journal of Engineering Science and Technology Review 16 (3) (2023) 74 - 83

Research Article

JOURNAL OF
Engineering Science and Technology Review

www.jestr.org

A Data-Driven Newsvendor Problem with Shifting Demand: A Deep Autoregressive Model with Attention Mechanism

Xin Li[1], Yongshi Hu[1], Xiaoli Su[2] and Bo Shao[3, *]

[1]School of Transportation, Fujian University of Technology, Fuzhou 350118, China
[2]School of Economics and Management, Fuzhou University, Fuzhou 350108, China
[3]Department of Industrial & Systems Engineering, University of Wisconsin-Madison, Madison, WI 53706, United States

Received 2 April 2023; Accepted 24 June 2023

Abstract

In a complex supply chain, product demand is characterized by non-linearity and instability. In particular, demand trajectory with intermittent or abnormal spikes exists, which makes it difficult for suppliers to accurately estimate the time-varying demand distribution and make the accurate inventory decision. Hence, a deep autoregressive model with a time attention mechanism (Attention-DeepAR model) was proposed to overcome this practical issue. In addition, a separated estimation and optimization approach was provided to solve a newsvendor problem with shifting demand. Temporal features were extracted from historical data through the Attention-DeepAR model to identify the long-term and short-term trends of the demand. Accordingly, the time-varying demand distribution was accurately fitted, which can assist in making precise inventory decisions. In addition, the rolling window design was proposed to introduce new demand data to update the demand. Finally, the validity of the Attention-DeepAR model was verified through the Monte Carlo simulations and a real case. Results show that the Attention-DeepAR model can effectively capture the temporal correlation between the regular and promotional demand values under volatile demands, improving the fitting accuracy of the time-varying demand distribution. This model can provide precise inventory decisions and significantly decrease average total costs.

Keywords: Attention mechanism, Deep autoregressive model, Newsvendor problem

1. Introduction

Inventory decision is one of the crucial decisions for suppliers [1]. Most suppliers aim to reduce inventory levels, improve the turnover rate of inventory, and develop a reasonable order strategy based on demand [2]. Traditional inventory management methods assume that the demand is continuous, stable, and relatively independent from other demands. However, with the rapid development of Internet technology and the widespread application of Internet platforms, multilateral supply and demand coupling with instantaneous transactions have been achieved, promoting the growth of a multi-channel supply chain [3]. In a multi-channel supply chain, suppliers can deal their products on various platforms online and offline, such as physical stores, websites, and markets, providing multiple touchpoints that can reach their prospective customers in a diverse, individualized, and timely manner to meet their fragmented order demand.

In the intricate demand scenario, traditional inventory management methods face numerous challenges, including slow responsiveness to demand fluctuations, excessive inventory pressure, and low inventory turnover rates. Thus, demand forecasts should be combined with inventory decision-making, and an inventory decision model should be developed based on demand forecasts to overcome these challenges [4]. With the continuous advancement of information science and computer technology, enormous amounts of multi-source demand data can be collected and stored. Machine learning [5] and data analysis technologies can be applied to observe and analyze time series demand data, opening new ideas for supply chain inventory management based on demand forecasting [6]. The rise of multi-channel supply chain has led to strong stock-keeping unit demand, resulting in a massive volume of inventory and demand data, including volatile demand that poses serious challenges for suppliers' product inventory management [7]. First, demand fluctuations are exceedingly erratic owing to promotional activities, resulting in uncertain and non-linear demand data. For instance, Adobe reported that Black Friday online demand hit a record-breaking $9.12 billion in 2022, soaring 221% compared with an average day in October. Second, promotional activities are frequently conducted in many forms, such as product price reductions, store coupons, direct price cuts, gifts, and shopping subsidies, each with different scopes and intensities. Finally, multi-channel supply chain is time-efficient and can generate variable demand fluctuations in a short time. Therefore, accurate forecasting of commodity demand based on available data is a challenge for suppliers.

Two major methods of forecasting product demand are currently used: single historical demand data and multi-source data. The traditional forecasting methods that rely on historic demand data are limited in their effectiveness. The reason is that they are susceptible to single factors and are not suitable for capturing demand fluctuations or nonlinear forecasting. Additionally, the accuracy of the forecast is restricted by demand stability. Conversely, multi-source data prediction methods can establish the relationship between demand and both internal and external factors, partly using

[*] E-mail address: bo.shao@wisc.edu

ISSN: 1791-2377 © 2023 School of Science, IHU. All rights reserved.
doi:10.25103/jestr.163.10

complex nonlinear mapping relationships to obtain more accurate matching results. However, these methods are only valid for relatively regular and stable demand. They are not suitable for intermittent commodities as they have low accuracy in capturing information related to demand changes.

Multi-channel demand has intermittent and momentary characteristics. Hence, changes in demand trends need to be captured, the hidden demand trends behind the data should be investigated, and reliable predictive methods that aid suppliers in making better order decisions must be found [8]. Furthermore, suppliers must consider how to integrate new data and capture future market demand. The reason is that up-to-date observation data can be an invaluable tool for detecting anomalies timely and enhancing suppliers' perception of future markets.

The need for time variance distribution in inventory management issues was examined in this study. Specifically, the structure design of the DeepAR network was improved, and time attention mechanisms were introduced to the decoding layer. A deep autoregressive model, known as the Attention-DeepAR model, was then created based on time attention mechanisms. The amount of time information contained in each coordinate of historical demand data was determined, and effective information was filtered out to enhance the present hidden vector. The model can accurately predict normal demand values, even in the presence of intermittent or abnormal peak values. This case enables the learning of global and effective time models from all-time series data while considering complex patterns (e.g., seasonality, data with time uncertainty growth) and long-term information with time-uncertain growth.

Suppliers can exploit multi-source demand data by employing the Attention-DeepAR model to adapt the distribution of product demand that changes over time, even in complex situations, such as intermittent and abnormal peaks. Consequently, the Attention-DeepAR model can aid suppliers to make more precise order decisions. Additionally, utilizing the Attention-DeepAR model can enable suppliers to predict future demand trends for multiple terms, anticipate market demand trends, and make swift responses to changes in demand. Furthermore, using the rolling window design permits new demand data to be introduced into the Attention-DeepAR model to update suppliers' perceptions of market demand.

The rest of the study is organized as follows: Section 2 briefly reviews research relevant to data-driven newsvendor problem. Section 3 builds up the Attention-DeepAR model to adapt to the distribution of time variable needs. Then, Section 4 uses numerical simulations and real-life cases to experiment to verify the validity of the model. Finally, Section 5 concludes the study and discusses future research prospects.

2. Literature Review

This study aims to investigate the data-driven newsvendor problem with multi-source demand data, contributing to the following two streams of literature: (1) demand prediction with machine learning and (2) the data-driven newsvendor problem.

2.1 Demand prediction with machine learning

Forecasting demand poses a challenge owing to several factors, including seasonality and external market fluctuations. These factors lead to nonlinear and uncertain changes in demand patterns thereby making it harder to predict accurately. Hence, this study focused on accurately predicting demand trends using historical data.

Previous studies have predominantly employed the eXtreme Gradient Boosting model (xgboost) [9-11], support vector machine [12-13], and Exponential Smoothing (ES) model [14-15] to forecast product demand. Given their simplicity, feasibility, and adaptability to the complexity of demand changes, these methods only had relatively satisfactory prediction results. Moreover, accurately predicting demand using a single historical demand time characteristic is challenging because product demand is influenced by various complex factors. Therefore, researchers have adopted several methods to identify key influencing factors, established complex nonlinear mapping relationships, and generated predictions based on them to achieve accurate forecasting results [16-18]. Singh et al. [19] employed Holt-Winters exponential smoothing, neural network autoregression model and Autoregressive integrated moving average (ARIMA) models to forecast Amazon's quarterly demand in 2019, where ARIMA was found to be the most effective. Weng et al. [20] developed a supply chain demand forecasting model using a Light Gradient Boosting Machine (LightGBM) and Long Short Term Memory (LSTM) networks, which offers a scientific and reasonable approach to predicting long-term product demand. Bandara et al. [21] utilized LSTM to examine Walmart's actual online market dataset and found that LSTM effectively predicted long-term and short-term dynamic changes in financial time series. Dong et al. [22] combined Autoregressive Recurrent Networks (DeepAR) models with a large amount of monitoring equipment data to anticipate slope displacement.

The encoder output of the above model can provide historical information about the demand time series that can help the model better predict future values. However, these models only use the output of the previous moment to predict long time series and ignore the output of the coding phase at each moment. Hence, they are prone to losing their memory and unable to capture long-term trends, seasonal information, and other key data. Moreover, they have difficulty accurately evaluating and weighing the temporal correlation between regular and promotional demands in the demand environment.

The attention mechanism enhances the importance of input features in time sequence [23]. This mechanism automatically learns the correlation between the hidden vectors generated by the decoder and encoder. Deep learning models are unable to differentiate the degree of correlation between the hidden layers of input and output sequence across multiple time steps. The attention mechanism compensates for this limitation. Additionally, the attention mechanism reduces the distortion rate after the input of real data. The attention mechanism is widely used in the field of demand prediction as evidenced by numerous studies [24-28].

Integrating the attention mechanism into the DeepAR model was developed as the Attention-DeepAR model. This model assigned weights based on the importance of time, enhanced the attention of critical timing input, and implemented effective learning of the global time model from all time-series data. The model accounted for the impact of complex patterns, such as seasonality and uncertain data growth over time, and long-term information. Consequently, the model improves the ability to predict complex scenarios, such as intermittency and abnormal peak values. This case can lead to higher prediction accuracy, particularly for e-commerce businesses with complex product demand forecast characteristics.

This study aims to enhance the DeepAR network structure design by introducing the time attention mechanism in the decoding layer. The mechanism determines the time information contained by each covariate in historical data, screens the pertinent information to enhance the currently used hidden vector, and allocates time weight based on the hidden vector information. These enhancements yield better time correlation ability between the forecast demand normal value and intermittent or abnormal peak value.

2.2 Data-driven newsvendor problem

Since its proposal, the newsvendor model has gained widespread application in operational management, including inventory, supply chain contracts, and procurement. The classic newsvendor model accounts for uncertainty in demand faced by decision-makers and aims to achieve optimal expected benefits or costs.

Chen et al. [29] proposed a joint pricing and ordering decision model for a single product, considering periodic inventory replenishment with a finite horizon. They introduced the concept of symmetric k-concave functions and provided methods to describe and construct optimal strategies. Levi et al. [30] examined the effectiveness of the sample mean approximation method in updating the demand distribution using independent historical demand data. They also established the theoretical boundary of the single-level inventory model. Building on the study, Levi et al. [31] reinforced the reliability of the sample-based approach by identifying the specific conditions that demand distribution satisfies, providing further upper bounds. Bertsimas et al. [32] developed a data-driven approach to establish a robust newsvendor model, without the assumption of an existing demand distribution. They introduced an adjustable parameter alpha to adjust the supplier's risk preference levels.

In practice, decision-makers frequently encounter significant demand fluctuations owing to rapid market changes and shortened product lifecycles [33]. When historical demand data and vast amounts of information related to demand are available, Tapiero et al. [34] assumed that demand and price are exogenous, and the joint distribution is known to determine newsvendor decisions. Pearson proposed a model and algorithm that optimizes inventory levels, establishing performance metrics and target revenue functions as constraints, assuming that the mean of the demand distribution is known [35]. Ban et al. [36] introduced demand characteristics and established theoretical performance bounds under various scale features through the linearization of the newsvendor model and the combination of machine learning with linear models. Ban et al. [36] and Oroojlooyjadid et al. [37] conducted a broad investigation of the newsvendor model by accounting for not only historical demand data but also customer demographics, weather patterns, seasonality, and economic indicators.

After reviewing the literature above, we conclude that an ideal newsvendor model should go beyond the assumption that the demand distribution is known, and consider historical data, including potential endogenous and exogenous factors, to develop multi-period decisions.

3. Methodology

3.1 Newsvendor problem based on multiple demand features

The supplier deals perishable goods and uses stochastic demand D to make ordering decisions. Considering that market demand is nonlinear and volatile, the supplier is unable to observe the true demand distribution $F(\cdot)$. Therefore, we investigate the supplier's ordering decisions during a finite selling season using stochastic demand and feature data (e.g., number of collectors and comments), particularly when the demand distribution $F(\cdot)$ is unknown. Notably, the supplier has access to historical demand observations over T periods (i.e., $\{(\boldsymbol{x}_t, z_t)\}_{t=1}^{T}$) before making ordering decisions. Each historical demand observation for the *t*-th period includes historical demand data z_t (where $z_t \in R$) and multiple features $\boldsymbol{x}_t = \left\{x_t^1, x_t^2, \ldots, x_t^P\right\}$. In particular, $X = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_T\}$, and $Z = \{z_1, z_2, \cdots, z_T\}$.

The supplier can observe demand-related information data $\boldsymbol{x}_{T+i}$ for period $T+i$ (where $i \in \{1, 2, \cdots N\}$) before making the next order decision. We developed a data-driven newsvendor model that can efficiently process multiple demand information and optimize the conditional expected cost function by including several features of observed demand data.

$$\min_{q \geq 0} E\left\{[b(D_{T+i} - q)^+ + h(q - D_{T+i})^+] \middle| \boldsymbol{x}_{T+i}\right\} \tag{1}$$

where b and h correspond to unit shortage cost and unit holding cost, respectively.

We can express the supplier's order decisions in period $T+i$ using Eq. (2):

$$q^*(\boldsymbol{x}_{T+i}) = \inf\{q \left| \hat{F}(q|\boldsymbol{x}_{T+i}) \geq b/(b+h)\right\} \tag{2}$$

Evidently, the optimal inventory $q^*(\boldsymbol{x}_{T+i})$ is a function that maps the collected information to the estimated demand distribution $\hat{F}(q|\boldsymbol{x}_{T+i})$. Therefore, the objective of this model is to find the best conditional distribution $\hat{F}(q|\boldsymbol{x}_{T+i})$ that minimizes the expected cost. In the next section, we discuss how to estimate the conditional distribution $\hat{F}(q|\boldsymbol{x}_{T+i})$ using historical demand observations $\{X, Z\}$ and the future feature value $\boldsymbol{x}_{T:T+N}$ to obtain the optimal estimate of external factors.

3.2 Deep autoregressive model with attention mechanism

This study proposed the Attention-DeepAR algorithm, which combines the DeepAR algorithm with the time attention mechanism. The algorithm utilizes historical demand data and exogenous feature vectors to capture the complex and abnormal information of intermittent and lumpy demand effectively. Further explanations of the algorithm's process are detailed in the following subsections.

3.2.1 Input layer

Historical product data from complex environments have multiple characteristics, which could result in gradient disappearance and explosion in Recurrent Neural Network (RNN) models, leading to reduced prediction performance.

To address this issue, the Attention-DeepAR algorithm uses a recurrent window to input historical demand and product characteristics. Assuming that the product time series window length is T, the external variable dimension is P, and the training BatchSize is B. B samples are collected from the product time series. The external variables (P dimension) and observed historical demand variables (1 dimension) are combined through the embedding method. Hence, the dimension of input sample data is $B\times T\times(1+P)$.

3.2.2 Time correlation layer

(1) LSTM

LSTM constitutes the core module of the DeepAR algorithm, achieving logical information association by retaining past information. LSTM resolves gradient issues commonly found in RNNs while capturing long-term temporal correlation. An LSTM neural network must be introduced to compute the hidden state vector h. The network's main working principle is to retain past information and use it to inform present decisions.

The LSTM architecture comprises the memory cell C, input gate i_t, forgetting gate f_t, and output gate O_t. C_t represents the cell state at time t in the LSTM model. The input gate controls the flow of information into the cell. Eqs. (3)-(8) show how the LSTM layer is calculated at each time *t*.

$$f_t=\sigma\left(W_f\left[h_{t-1},\boldsymbol{Ah_t}\right]+b_f\right) \tag{3}$$

$$i_t=\sigma\left(W_i\left[h_{t-1},\boldsymbol{Ah_t}\right]+b_i\right) \tag{4}$$

$$C't=\tanh\left(W_c\left[h_{t-1},\boldsymbol{Ah_t}\right]+b_c\right) \tag{5}$$

$$C_t=f_t\otimes c_{t-1}+i_t\otimes C't \tag{6}$$

$$o_t=\sigma\left(W_o\left[h_{t-1},\boldsymbol{Ah_t}\right]+b_o\right) \tag{7}$$

$$h_t=o_t\otimes tanh\left(C_t\right) \tag{8}$$

The forgetting gate uses the *Sigmoid* function σ to determine whether to discard information from memory cell C. The forgetting coefficient f_t is obtained by weighting input value h_t in layer t and the previous state h_{t-1} in layer $t-1$ according to Eq. (3). The input gate generates a coefficient i_t based on whether certain information in the current input vector $\boldsymbol{Ah_t}$ should be discarded using Eq. (4). The neuron updates the cell state candidate value $C't$ at the current moment using tanh activation Eqs. (5) and (6), and $\otimes$ is a multiplication operation. The output gate Eqs. (7) and (8) evaluates the current value of C_t to decide whether to discard information and generate the final cell state output h_t. Notably, W and b represent the weight matrix and bias vector, respectively.

(2) Deep autoregressive neural network

The DeepAR model is an autoregressive recursive neural network model proposed by David Salinas and other scholars [38]. This model is based on multiple time series training, enabling it to accurately estimate the probability distribution $P\left(D_{T+1:T+N}\middle|\ z_{1:T},\boldsymbol{x}^{p}_{1:T+N}\right)$ of future time series $D_{T+1:T+N}$ using existing time series $z_{1:T}$ and covariate $\boldsymbol{x}_{1:T+N}$. In this study, we outputted the probability distribution of each DeepAR phase in the future to obtain the demand distribution and improve the time sensitivity. Fig. 1 shows the basic structure of the DeepAR model.

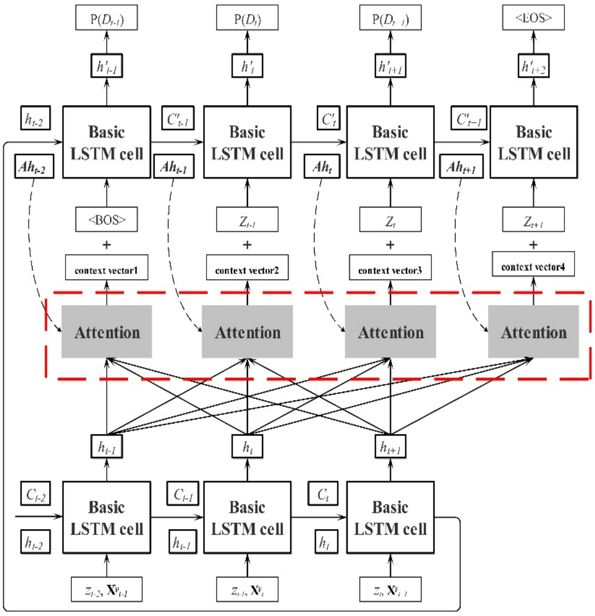

Fig. 1. The Structure of the DeepAR Model.

The DeepAR model considers the conditional probability distribution as the concatenation of likelihood factors over a sequence of timestamps:

$$\begin{aligned}P\left(D_{T+1:T+N}\middle|\ z_{1:T},\boldsymbol{x_{1:T+N}}\right)&=\prod_{t=T+1}^{T+N}Q_{\Theta}\left(D_t\middle|\ z_{1:T},\boldsymbol{x_{1:T+N}}\right)\\&=\prod_{t=T+1}^{T+N}l\left(D_t\middle|\ \theta\left(\boldsymbol{Ah_t},\boldsymbol{\Theta}\right)\right)\end{aligned} \tag{9}$$

$$\boldsymbol{Ah_t}=h\left(\boldsymbol{Ah_{t\text{-}1}},z_{t-1},\boldsymbol{x_t},\boldsymbol{\Theta}\right) \tag{10}$$

The LSTM cyclic network's implicit state after the calculation is represented by h, whereas $\boldsymbol{Ah_t}$ represents the implicit state of the autoregressive cyclic network after strengthening time attention. Fig. 1 illustrates that the model requires the previous time's network output $\boldsymbol{Ah_{t-1}}$, the target value $z_{1:T}$ at the previous time, the covariable $\boldsymbol{x_t}$ at the current time, and the model parameter Θ to obtain $\boldsymbol{Ah_t}$ at time *t*. Parameter $\boldsymbol{\Theta}$ comprises the RNN's $\boldsymbol{Ah}(\cdot)$ and $\theta(\cdot)$ parameters. The likelihood function ℓ plays a role in the model's noise. Typically, the Gaussian and negative binomial likelihood functions are utilized.

3.2.3 Time extraction layer

In the time correlation layer, the LSTM treats different marginal values of demand data equally, regardless of the historical moment in the input samples. This approach makes accurately capturing and weighing the temporal correlation between regular demand and promotional demand difficult. To overcome this limitation, the time attention mechanism is incorporated into the decoding layer of the DeepAR model (as highlighted in the red box in Fig. 1). Leveraging the superior self-learning ability of the attention mechanism in capturing time correlation, the

Attention-DeepAR algorithm considers the importance of relevant input sequences, captures time patterns across multiple time steps, and assigns different weights to important time nodes. Consequently, the Attention-DeepAR algorithm can capture long-term time correlation, intensify the degree of attention to important time nodes in the input, and achieve the profound key time characteristics of mining.

Fig. 2 illustrates the operational principle of the Attention-DeepAR model.

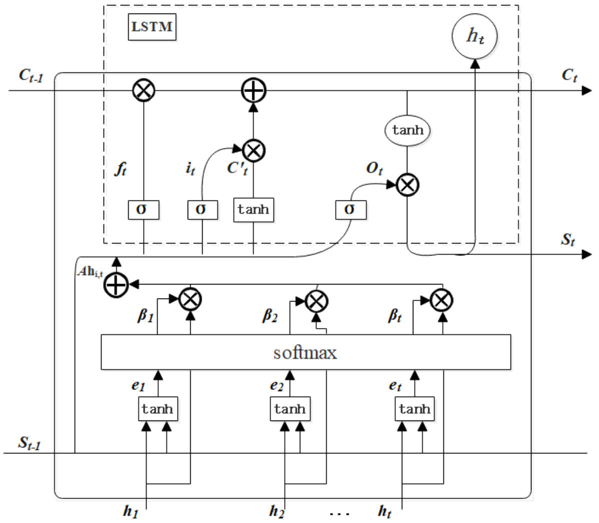

Fig. 2. The work principle of the Attention-DeepAR model.

First, the encoder processes historical demand data. The weight vector of attention time e_t is calculated for the encoder's output using Eq. (11). Then, the time information included in the covariate of historical demand data is comprehensively evaluated.

$$e_t = \tanh\left(W_d\left[h_t; s_{t-1}\right] + b_d\right) \tag{11}$$

where W_d is the weight matrix to be learned by the model, h_t is the hidden state of the encoder, S_{t-1} is the unit state of the encoder at the last moment, and b_d is the bias vector of the fully connected layer.

Second, the attention probability β_t of the input sequence at each time step is normalized using a softmax function.

$$\beta_t = \exp\left(e_t\right) \Big/ \sum_{t=1}^{T+N} \exp\left(e_t\right) \tag{12}$$

Third, relevant information is extracted to enhance the currently used hidden vector using weighted aggregation of the context vector $\boldsymbol{Ah}_t$ at t time, which yields the weight assigned by the hidden state at different points in time. Thus, the final output vector of the attention layer is defined as follows:

$$\boldsymbol{Ah}_t = \sum_{t}^{T+N} \beta_t h_t \tag{13}$$

Finally, time weights are assigned to the hidden vector information. This information is then integrated and output using an LSTM gated unit.

3.2.4 Output layer

The two-stage process of time correlation and time extraction allows for the in-depth mining of product demand distribution. Thus, the output layer is primarily responsible for building the Attention-DeepAR model to predict the final product demand. The model parameter $\boldsymbol{\Theta}$ is composed of parameters of $h_i'(\cdot)$, a cyclic neural network that integrates an attention mechanism, and parameters of $\theta(\cdot)$. This parameter is trained using the maximum likelihood estimation to learn. The loss function can be expressed as follows:

$$Loss = \sum_{i=1}^{I} \sum_{t=t_0}^{T} -\log \ell_{\mathrm{G}}\left(z_t \mid \theta\left(h_i'\right)\right) \tag{14}$$

The model arrives at a stable structure with a small deviation value through continuous iteration.

4. Numerical Experiments

In this section, Monte Carlo numerical simulation is used to verify the robustness and effectiveness of the Attention-DeepAR model. Moreover, the influence of the model on the optimal order quantity and the expected cost are analyzed from the time weight perspective.

4.1 Data generating

This section generates data by referring to the data generation method proposed by Ban et al. [36] to more accurately describe the intermittent or abnormal peak characteristics of product demand in the volatile demand environment. The data generation process consists of two steps as follows:

(1) Generate demand covariable samples

Product feature covariable X^p is independently generated, including two dynamic covariables and five static covariables. Covariables are mainly subject to the binomial, Bernoulli, continuous, and discrete uniform distribution. Each covariable has the same weight, which is $a_p = 1/7$.

$$X^1 \sim Bernoulli\left(h\right) \tag{15}$$

$$X^2、X^5 \sim Bernoulli\left(k\right) \tag{16}$$

$$X^3、X^4、X^6、X^7 \sim Normal\left(\mu_1, \sigma^2{}_1\right) \tag{17}$$

where, μ_1 , $\sigma^2{}_1$ is randomly generated and μ_1, $\sigma^2{}_1 \in U\left(0,2000\right)$, $h = 0.25$, $k = 0.5$.

(2) Generate historical demand data

Historical demand data D are generated, and noise data e that follow normal distribution are added. A random value f was added at common promotion time points to simulate the abnormal peak value during the promotion to get closer to the real demand. In addition, a time column was added, which was set from November 1, 2016, to July 28, 2019, to test the ability of time sensitivity of the model.

$$D = a + \sum_{p=1}^{7} a_p X^p + e + f \tag{18}$$

where $a = 3.64$, $e \sim N(0,5^2)$, $f \in U(800,1200)$.

Data are generated based on the above steps, with a sample size of 1000. Fig. 3 shows the historical demand data of a simulation.

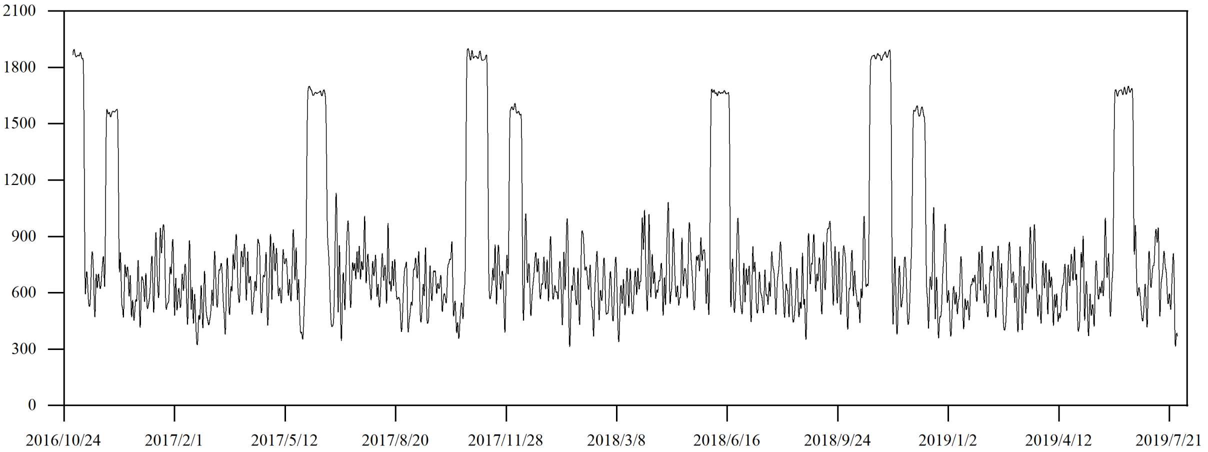

Fig. 3. Historical demand data graph for a simulation.

As shown in Fig. 3, when promotion nodes such as "Double 11 shopping carnival" and "618" exist, product demands deviate from the trend of daily demand and show an abnormal peak. In addition, demand volume before and after the promotion is affected, leading to intermittent fluctuations in demand. This case brings great difficulties to the fitting of demand distribution and the subsequent ordering decision.

4.2 Experimental setting

Step 1) Splitting data samples.

The demand forecast during "618" in 2019 was selected as the verification object to verify the timely and accurate capture of abnormal peak value by the Attention-DeepAR model. That is, the demand data from June 1, 2021, to July 28, 2021, were taken as the test sample ($T' = 58$), and the remaining data were taken as the training sample ($T = 942$).

Step 2) Reference model setup.

Li et al. [25] verified that the DeepAR model has higher prediction accuracy than the LSTM and Convolutional Neural Networks (CNN) models. In this part, the prediction accuracy of the Attention-DeepAR model relative to the LSTM and CNN models would not be discussed. Only the DeepAR model is used as the benchmark model to compare the fitting performance and ordering decision of the Attention-DeepAR model under complex conditions, such as the intermittent and irregular characteristics of demand.

Step 3) Model training.

In multi-step fitting, a rolling window scheme is adopted. T is selected as the fixed scrolling window size, and $1:T$ is the initial estimation period. After fitting the demand distribution of N cycles in each scrolling window, the estimated period is moved back one phase along the entire data set, that is, the first observation is discarded and the next new observation is added back.

Step 4) Performance evaluation.

In this experiment, two indicators were used to evaluate the out-of-sample performance of the Attention-DeepAR model.

(1) The optimal parameters and model are obtained to objectively evaluate the demand forecasting model during the promotion period. In this study, the relatively common mean absolute percentage error (MAPE) was used to evaluate the predicted and actual demand values of each model. Evaluation indicators are defined as follows:

$$MAPE = 1/N \sum_{i=1}^{N} \left| (z_t - D_t)/z_t \right| \times 100 \tag{19}$$

where D_t represents the predicted demand at time i (the DeepAR model represents the quantile of forecast sales), z_t represents the actual demand at time t , and N represents the total number of predicted days.

(2) Performance indexes of multi-step advance probability distribution fitting of order quantity, including prediction interval coverage probability (PICP) and prediction interval normalized average width (PINAW). The specific formula of PICP is as follows:

$$PICP = \frac{1}{N} \sum_{i=1}^{N} I_i \tag{20}$$

$$I_t = \begin{cases} 1, y_t \in \left[W_{t,1}, W_{t,Z}\right] \\ 0, y_t \notin \left[W_{t,1}, W_{t,Z}\right] \end{cases} \tag{21}$$

where $\left[W_{t,1}, W_{t,Z}\right]$ is the prediction interval within the predetermined coverage range (the confidence interval based on the given confidence level), and I_t is a Boolean variable. When the actual value of the *t*-th period is within the interval, it is assigned a value of 1, and 0 if otherwise. The larger the PICP, the more predicted value of the predicted interval coverage, and the more convincing the fitting results.

In addition to evaluating reliability, PINAW is also required to comprehensively evaluate the performance of interval prediction. The specific formula is as follows:

$$PINAW = \sum_{i=1}^{N} \left(\left(W_{t,Z} - W_{t,1}\right) / \left(N \times D\right) \right) \tag{22}$$

where D is the difference between the maximum and minimum values of the target value. The smaller the PINAW, the more accurate the fitting result. The relative ratio of the Attention-DeepAR model and DeepAR model PINAW values is defined as follows:

$$\Delta_p = \frac{PINAW_{Attention\text{-}DeepAR} - PINAW_{DeepAR}}{PINAW_{DeepAR}} \times 100\% \qquad (23)$$

(3) Out of the sample average total cost $MC = \frac{1}{T'}\sum_{i=T+1}^{T+T'} C_i$, where $C_i = b(y_i - \hat{q}_i)^+ + h(\hat{q}_i - y_i)^+$. Meanwhile, the disappointment level of the expected cost between the Attention-DeepAR model and the average total cost out of the sample of the DeepAR model is defined as follows:

$$\Delta_C = \left(\left(MC_{Attention\text{-}DeepAR} - MC_{DeepAR}\right)/MC_{DeepAR}\right) \times 100\% \qquad (24)$$

Step 5) Robustness validation.
We repeat the above four steps 100 times and then calculate the average values of probability distribution fitting, the order quantity, MAPE, and the out-of-sample cost.

4.3 Simulations analysis

In this subsection, numerical experiments are conducted based on the proposed data sample and simulation procedure in the previous subsections. We first analyze the results of the demand probability distribution estimator for the DeepAR and Attention-DeepAR algorithms. Second, we compare the optimal order quantity and the corresponding cost obtained with different learning algorithms.

4.3.1 Analysis of probabilistic forecasts

Table 1 presents the estimation sharpness of demand probability distribution for the DeepAR and Attention-DeepAR algorithms. Two algorithms exhibit a 100% empirical coverage exceeding the expected nominal one of a 90% confidence level. However, from the value of PINAW with a 90% confidence level, the Attention-DeepAR algorithm provides the lowest PINAW with different forecast horizons (*N*). In addition, the normalized widths of the Attention-DeepAR algorithm become lower as the forecast horizon (*N*) increases. Specifically, the difference in the PINAW value between the DeepAR and Attention-DeepAR algorithms is -27.11%. These results imply that incorporating the attention mechanism with the superior time correlation self-learning ability into the DeepAR algorithm (i.e., the Attention-DeepAR algorithm) can avoid the additional uncertainty of promotions on demand estimation and improve the sharpness of probabilistic forecasts in a complicated and intermittent environment.

Table 1. The estimation sharpness of demand probability distribution.

N	DeepAR		Attention-DeepAR		Δ_p
	PICP	PINAW	PICP	PINAW	
8	100%	9.53	100%	8.12	-14.80%
18	100%	8.41	100%	7.44	-11.57%
28	100%	7.34	100%	5.35	-27.11%

4.3.2 Comparison of out-of-sample order decisions

We compare the order decision and the incurred out-of-sample costs from June 1, 2021, to July 28, 2021, in this subsection to further evaluate the performance of the two algorithms. First, taking actual demand as the reference point, we compare order quantities obtained with two algorithms. The comparison results of MAPE in Table 2 can reflect the performances of two algorithms with different critical ratios.

Table 2. The comparison results of order decisions.

b/(b+h)	Attention-DeepAR		DeepAR	
	Mean	95% confidence interval	Mean	95% confidence interval
	0.61	[0.59,0.64]	0.65	[0.63,0.68]
0.3	0.31	[0.30,0.32]	0.31	[0.30,0.33]
0.5	0.52	[0.51,0.52]	0.54	[0.54,0.55]
0.7	0.32	[0.32,0.32]	0.35	[0.34,0.35]
0.9	0.06	[0.06,0.07]	0.08	[0.07,0.09]

From Table 2, we can learn that the Attention-DeepAR algorithm can achieve a higher level of order accuracy than the DeepAR algorithm. Specifically, the mean values of the MAPE difference become the smallest as the critical ratio is 0.9.

Then, Table 3 shows the statistical results of incurred out-of-sample costs with different critical ratios. Hence, the average cost of the Attention-DeepAR algorithm is significantly lower than the DeepAR algorithm. For example, for a critical ratio of 0.3, the mean cost using the DeepAR algorithm is reduced from 52.99 to 47.99 using the Attention-DeepAR algorithm, making the disappointment level of the cost Δ_c reach 9.44%. The results imply that improving the accuracy of demand estimation by using the Attention-DeepAR algorithm is the best option, and this algorithm can significantly reduce the supplier's cost.

Table 3. The comparison results of order decisions.

b/(b+h)	Attention-DeepAR		DeepAR		Δ_c
	Mean	95% confidence interval	Mean	95% confidence interval	
0.1	93.81	[70.78,116.84]	95.75	[71.57,123.92]	-2.03%
0.3	47.99	[36.75,58.11]	52.99	[42.06,63.94]	-9.44%
0.5	86.44	[50.14,122.74]	87.59	[53.43,123.75]	-1.31%
0.7	68.47	[41.39,95.56]	69.61	[42.61,96.61]	-1.64%
0.9	34.62	[24.67,47.08]	35.87	[28.37,50.86]	-3.48%

5. Empirical Study

From Table 3, we can learn that the Attention-DeepAR algorithm can achieve a higher level of order accuracy than the DeepAR algorithm. Specifically, the mean values of the MAPE difference become the smallest as the critical ratio is 0.9.

The effectiveness of the Attention-DeepAR model through numerical simulations is presented in Section 4. Additionally, the performance of the model is confirmed through experimental analysis using the open retail Kaggle dataset.

5.1 Data description

The Kaggle dataset is a historical demand dataset from 45 Walmart stores situated in diverse regions (https://www.kaggle.com/datasets/yasserh/walmart-dataset). The dataset contains weekly demand data of Walmart stores over the years 2010 and 2012, including relative internal (e.g., store ID, data, holiday indicator) and external (e.g., temperature, fuel price, CPI, unemployment index) features, as shown in Table 4.

Table 4. Structure of historical demand data.

Name	Description	Example
Store	the store 9number	1
Date	the week of sales	05-02-2010
Weekly_Sales	sales for the given store	1643690
Holiday_Flag	1 - Holiday week 0 - Non-holiday week	0
Temperature	Temperature on the day of sale	42.31
Fuel_Price	Cost of fuel in the region	2.57
CPI	Prevailing consumer price index	211.10
Unemployment	Prevailing unemployment rate	8.11

Fig. 4 illustrates the demand trend of products in Store 1 taking the historical demand data of Store 1 as an example.

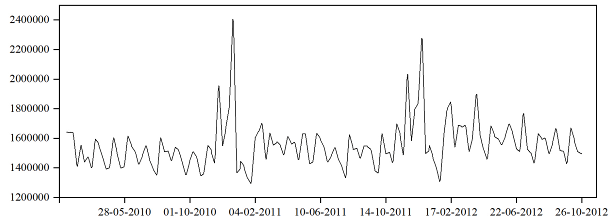

Fig. 4. Overall trend of product demand.

Fig. 4 shows that the demand data for this product displayed specific intermittent and abnormal peak characteristics during holidays, such as Super Bowl, Labor Day, Thanksgiving, and Christmas, which were influenced by promotional seasons. We conducted an experiment with the same settings as Section 4.2 to evaluate the performance of the Attention-DeepAR model in real-world scenarios. The training set comprised demand data from February 5, 2010, to September 6, 2012, whereas the test set comprised data from September 7, 2012, to November 1, 2012.

5.2 Result Analysis and Discussion

5.2.1 Analysis of probabilistic forecasts

Out-of-sample prediction intervals of the probability distribution using a 90% confidence interval were observed for the Attention-DeepAR and DeepAR models. Fig. 5 shows the results.

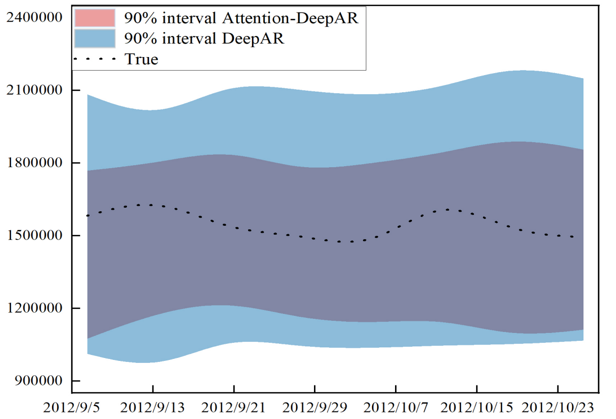

Fig. 5. Interval plot of the out-of-sample probability distribution for the Attention-DeepAR model and DeepAR model.

As shown in Fig. 5, the Attention-DeepAR and DeepAR models are capable of effectively capturing demand trends in most cases. However, the DeepAR model is less sensitive to demand fluctuations resulting from promotions, causing the prediction intervals to be wider than those of the Attention-DeepAR model. This result indicates a need for improvement in reducing the uncertainty of demand forecasts. Conversely, the Attention-DeepAR model effectively utilizes implicit time information and performs better in mitigating risks. Table 5 presents a detailed comparison of PICP and PINAW values of the two models.

Table 5. Evaluation of probability distribution fitting performance.

N	DeepAR		Attention-DeepAR		Δ_c
	PICP	PINAW	PICP	PINAW	
2	100%	27.90	100%	16.15	-42.11%
3	100%	25.67	100%	14.04	-45.31%
4	100%	25.73	100%	9.16	-64.40%
5	100%	9.25	100%	6.82	-26.27%
6	100%	9.26	100%	6.41	-30.78%
7	100%	9.39	100%	6.67	-28.97%
8	100%	9.41	100%	6.75	-28.27%

Table 5 shows that both models have a 100% predictive interval coverage, demonstrating that the true values fall within the predictive intervals. The improved model reduces interval width by 64.40%, particularly when $N=4$, emphasizing its ability in mitigating uncertainty risks. Over time, as the probability of the predictive interval approaches 100%, the predictive intervals of both models decrease, and the difference between them decreases. Nevertheless, good performance is still demonstrated in the improved model, indicating that the Attention-DeepAR model can effectively reduce the width of the predictive interval, improve the fitting accuracy by learning long-term trends, and accurately capture short-term fluctuations when dealing with complex situations.

5.2.2 Comparison of out-of-sample order decisions

The two algorithms' order decisions and out-of-sample costs incurred during the period of September 07, 2012, to November 01, 2012, were compared to assess their performance. In comparing the order quantity deviation of the two algorithms, we used actual demand as a reference point. These comparisons were made to evaluate the effectiveness of the algorithms with varying decimal points. Fig. 6 displays the results of the MAPE comparison, which provides insights into the performance of both algorithms.

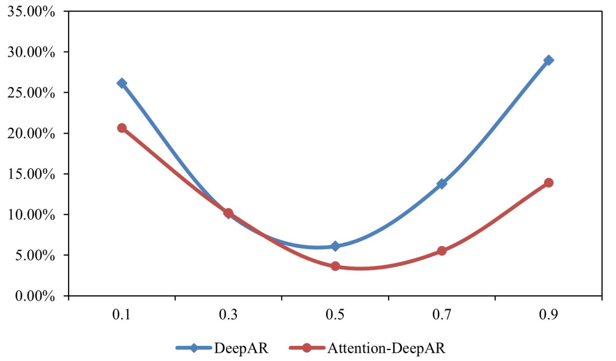

Fig. 6. MAPE values for different quantiles of out-of-sample order volume for Attention-DeepAR model and DeepAR model.

Fig. 6 presents the comparison of the improved and unimproved model's error values. The results show that the improved model consistently achieves lower error values except at a service level of 0.3 where both models exhibit similar errors. Although both models exhibit a similar trend and achieve the minimum MAPE at the median, our model outperforms the unimproved model with higher prediction accuracy. Fig. 7 illustrates the total cost at the quantile of each of the next eight phases for our model and the benchmark model. The results suggest that our model consistently incurs lower costs, indicating a cost advantage. A significant deviation in order quantities at sub-sites of the distribution acquisition in our model renders cost prediction at $N \geq 4$ ineffective at a service level of 0.3.

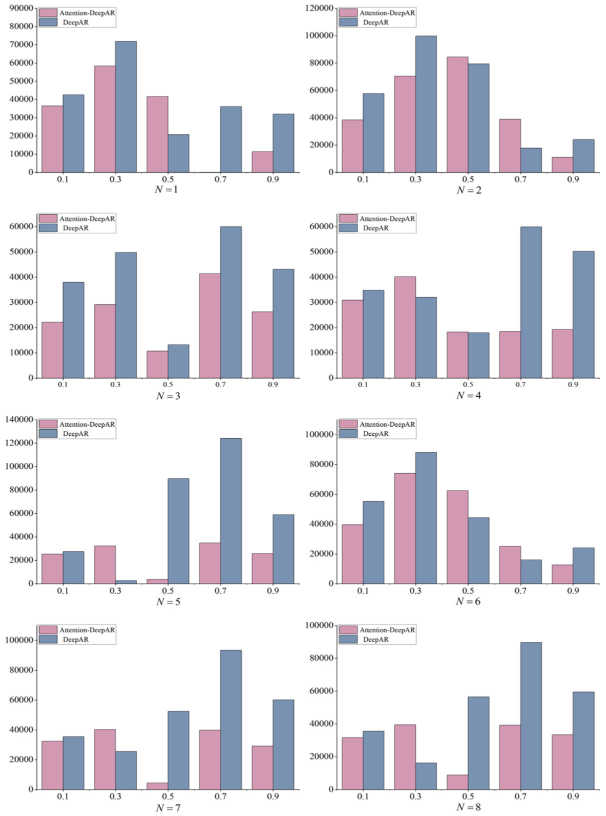

Fig. 7. Cost per period for different quantiles of out-of-sample order volume for Attention-DeepAR model and DeepAR model.

The Attention-DeepAR model has been able to significantly improve the cost of online suppliers with each operation by effectively capturing the time-varying demand distribution, leading to a reduction in the average total cost, as presented in Table 6.

Table 6. Average total cost under different service levels.

b/(b+h)	**Attention-DeepAR**	**DeepAR**	Δ_c
0.1	32109.29	40865.91	-21.43%
0.3	48035.05	48246.73	-0.44%
0.5	29342.31	46815.58	-37.32%
0.7	29720.82	62127.38	-52.16%
0.9	21114.87	44029.18	-52.04%

According to Table 6, the average cost of online suppliers varies across service levels. When the service level is 0.3, both models yield cost predictions equivalent to the MAPE value of the comparison model. However, our model's performance surpasses that of the DeepAR model by 0.44%. Furthermore, at a service level of 0.7, the Attention-DeepAR model reduces the average total cost by 52.16%. Our model demonstrated the narrowest decision deviation compared with the DeepAR model. Moreover, the comparison of our model with the benchmark model revealed that our model yielded the lowest cost of 40,865.91 when the service level is 0.1, representing a 21.43% reduction compared with the benchmark model.

In summary, the empirical findings indicate that the demand trajectory of intermittent or abnormal peaks is effectively identified and captured by the Attention-DeepAR model. Valuable support is provided for suppliers to make precise ordering decisions, resulting in a significant reduction in the average total cost.

6. Conclusions

6.1 Main findings

In the supply chain, with the intensification of market competition, the dynamic, diverse, and sudden nature of demand has led to frequent changes in plans. Inventory determined by the plan can no longer meet the agility and flexibility requirements of market demand. Consequently, external and internal resources within the organization need to be optimally used to optimize the inventory system. The current study examines a newsvendor problem, with reliance on data, in which we need to learn and analyze the trends from historical demand data and product characteristics, as the demand distribution trend remains uncertain. The Attention-DeepAR model based on the attention mechanism was constructed. The objective is to counter the difficulties of an uncontrollable and complex setting that comprise varying demand distribution, including an abnormal peak demand and heterogeneous demand distributions that are hard to fit accurately over time.

The Attention-DeepAR model that applies the attention mechanism to identify and learn crucial time nodes from historical data was proposed in this study. These important nodes and the inherent correlation between the states at different times in the decoding layer were captured by the model. As a result, the ability of the model to fit demand distributions in complex circumstances was improved.

Moreover, the leverage of multiple product feature data as input values to predict the probability distribution of product demand during specific periods was done in this study. Inventory management was optimized by this approach, which reduced demand forecast uncertainty and associated risks and enhanced the use of historical data.

Finally, the model's effectiveness is verified through Monte Carlo numerical simulations, including real case studies. The results demonstrate that the model enhances the accuracy of fitting time-varying demand distributions, leading to more accurate optimal ordering decisions. Consequently, this case reduces costs incurred by the suppliers by enabling them to respond to different service level demands more effectively.

6.2 Limitations and prospects

The newsvendor problem under shifting demand was explored in this study by constructing the Attention-DeepAR model.

Despite some progress, this study still has limitations. As an illustration, owing to constraints on the availability of data, this study relied only on product data from a small sample of 45 stores between 2010 and 2012. This limited sample size and short time span may decrease the generalizability of the empirical findings. To address this shortfall, the sources of data should be broadened, and the time span must be extended to enhance the dependability and generalizability of the study. Additionally, concerning model design, subsequent research can extend the investigation of the autoregressive model by incorporating the attention mechanism and determining the optimal order quantity directly through the use of historical demand data.

Acknowledgements

The authors are grateful for the support provided by the National Social Science Foundation of China (Grant No. 20BGL112).

References

1. Mamani, H., Nassiri, S., Wagner, M. R. “Closed-form solutions for robust inventory management”. *Management Science*, 63(5), 2017, pp.1625-1643.
2. Lackes, R., Siepermann, M., Vetter, G. “What drives decision makers to follow or ignore forecasting tools-A game based analysis”. *Journal of Business Research*, 106, 2020, pp.315-322.
3. Tian, X., Wang, H., Erjiang, E. “Forecasting intermittent demand for inventory management by retailers: A new approach”. *Journal of Retailing and Consumer Services*, 62, 2021, pp.102662.
4. Yu, Y., Qiu, R., Sun, M. “Joint pricing and ordering decisions for a loss-averse retailer with quantity-oriented reference point effect and demand uncertainty: A distribution-free approach”. *Kybernetes*, 52(4), 2023, pp.1294-1324.
5. Lu, X., Liu, C., Lai, K.K. Cui, H. “Risk measurement in Bitcoin market by fusing LSTM with the joint-regression-combined forecasting model”. *Kybernetes*, 52(4), 2023, pp.1487-1502.
6. Lolli, F., Gamberini, R., Regattieri, A., Balugani, E., Gatos, T., Gucci, S. “Single-hidden layer neural networks for forecasting intermittent demand”. *International Journal of Production Economics*, 183, 2017, pp.116-128.
7. Shi, J. “Application of the model combining demand forecasting and inventory decision in feature based newsvendor problem”. *Computers & Industrial Engineering*, 173, 2022, pp.108709.
8. Nikolopoulos, K. “We need to talk about intermittent demand forecasting”. *European Journal of Operational Research*, 291(2), 2022, pp.549-559.
9. Massaro, A., Panarese, A., Giannone, D., Galiano, A. “Augmented Data and XGBoost Improvement for Sales Forecasting in the Large-Scale Retail Sector”. *Applied Sciences*, 11(17), 2021, pp.7793.
10. Ramya, B. S. S., Vedavathi, K. “An advanced sales forecasting using machine learning algorithm”. *International Journal of Innovative Science and Research Technology*, 5(5), 2020, pp.342-345.
11. Choi, J., Yang, H., Oh, H. “Store sales prediction using gradient boosting model”. *Journal of the Korea Institute of Information and Communication Engineering*, 25(2), 2021, pp.171-177.
12. Güven, I., Şimşir, F. “Demand forecasting with color parameter in retail apparel industry using artificial neural networks (ANN) and support vector machines (SVM) methods”. *Computers & Industrial Engineering*, 147, 2020, pp.106678.
13. Di Pillo, G., Latorre, V., Lucidi, S., Procacci, E. “An application of support vector machines to sales forecasting under promotions”. *4OR*, 14, 2016, pp.309-325.
14. Taylor, J. W., Snyder, R. D. “Forecasting intraday time series with multiple seasonal cycles using parsimonious seasonal exponential smoothing”. *Omega*, 40(6), 2012, pp.309-325.
15. Nirmala, V. W., Harjadi, D., Awaluddin, R. “Sales forecasting by using exponential smoothing method and trend method to optimize product sales in pt. zamrud bumi indonesia during the covid-19 pandemic”. *International Journal of Engineering, Science and Information Technology*, 1(4), 2021, pp.59-64.
16. Liu, P., Ming, W., Hu, B. “Sales forecasting in rapid market changes using a minimum description length neural network”. *Neural Computing and Applications*, 33, 2021, pp.937-948.
17. Loureiro, A. L., Miguéis, V. L., da Silva, L. F. “Exploring the use of deep neural networks for sales forecasting in fashion retail”. *Decision Support Systems*, 114, 2018, pp.81-93.
18. Chandriah, K. K., Naraganahalli, R. V. “RNN/LSTM with modified Adam optimizer in deep learning approach for automobile spare parts demand forecasting”. *Multimedia Tools and Applications*, 80(17), 2021, pp.26145-26159.
19. Singh, B., Kumar, P., Sharma, N., Sharma, K. P. “Sales forecast for amazon sales with time series modeling”. In: *2020 first international conference on power, control and computing technologies (ICPC2T)*, Chhattisgarh, India: IEEE, 2020, pp.38-43.
20. Weng, T., Liu, W., Xiao, J. “Supply chain sales forecasting based on lightGBM and LSTM combination model”. *Industrial Management & Data Systems*, 120(2), 2020, pp.265-279.
21. Bandara, K., Shi, P., Bergmeir, C., Hewamalage, H., Tran, Q., Seaman, B. “Sales demand forecast in e-commerce using a long short-term memory neural network methodology”. In: *Neural Information Processing: 26th International Conference*, *ICONIP 2019*, Sydney, NSW, Springer International Publishing, 2019, pp.462-474.
22. Dong, M., Wu, H., Hu, H., Azzam, R., Zhang, L., Zheng, Z., Gong, X. “Deformation prediction of unstable slopes based on real-time monitoring and deepar model”. *Sensors*, 21(1), 2020, pp.14.
23. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Polosukhin, I. “Attention is all you need”. *Advances in neural information processing systems*, 30, 2017.
24. Zhang, D., Xiao, F., Shen, M., Zhong, S. “DNEAT: A novel dynamic node-edge attention network for origin-destination demand prediction”. *Transportation Research Part C: Emerging Technologies*, 122, 2021, pp.102851.
25. Li, D., Lin, K., Li, X., Liao, J., Du, R., Chen, D., Madden, A. “Improved sales time series predictions using deep neural networks with spatiotemporal dynamic pattern acquisition mechanism”. *Information Processing & Management*, 59(4), 2022, pp.102987.
26. Seyedan, M., Mafakheri, F. “Predictive big data analytics for supply chain demand forecasting: methods, applications, and research opportunities”. *Journal of Big Data*, 7(1), 2020, pp.1-22.
27. Zhang, M., Xu, H., Ma, N., Pan, X. “Intelligent Vehicle Sales Prediction Based on Online Public Opinion and Online Search Index”. *Sustainability*, 14(16), 2022, pp.10344.
28. DeYong, G. D. “The price-setting newsvendor: review and extensions”. *International Journal of Production Research*, 58(6), 2020, pp.1776-1804.
29. Chen, X., Simchi-Levi, D. “Coordinating inventory control and pricing strategies with random demand and fixed ordering cost: The finite horizon case”. *Operations research*, 52(6), 2004, pp.887-896.
30. Levi, R., Roundy, R. O., Shmoys, D. B. “Provably near-optimal sampling-based policies for stochastic inventory control models”. *Mathematics of Operations Research*, 32(4), 2007, pp.821-839.
31. Levi, R., Perakis, G., Uichanco, J. “The data-driven newsvendor problem: New bounds and insights”. *Operations Research*, 63(6), 2015, pp.1294-1306.
32. Bertsimas, D., Thiele, A. “A data-driven approach to newsvendor problems”. *Working Papere, Massachusetts Institute of Technology*, 51. 2005.
33. Snyder, L. V., Shen, Z. J. M. “Fundamentals of supply chain theory”. *John Wiley & Sons*, 2019.
34. Tapiero, C. S., Kogan, K. “Risk-averse order policies with random prices in complete market and retailers’ private information”. *European Journal of Operational Research*, 196(2), 2009, pp.594-599.
35. Pearson, M. A. “The incorporation of target performance measures and constrained optimisation in the newsboy problem”. *Journal of the Operational Research Society*, 51(6), 2000, pp.744-754.
36. Ban, G. Y., Rudin, C. “The big data newsvendor: Practical insights from machine learning”. *Operations Research*, 67(1), 2019, pp.90-108.
37. Oroojlooyjadid, A., Snyder, L. V., Takáč, M. “Applying deep learning to the newsvendor problem”. *IISE Transactions*, 52(4), 2020, pp.444-463.
38. Salinas, D., Flunkert, V., Gasthaus, J., Januschowski, T. “DeepAR: Probabilistic forecasting with autoregressive recurrent networks”. *International Journal of Forecasting*, 36(3), 2020, pp.1181-1191.