Knowledge-Based Systems 258 (2022) 109956

Contents lists available at ScienceDirect

Knowledge-Based Systems

journal homepage: www.elsevier.com/locate/knosys

Predictive analytics for demand forecasting: A deep learning-based decision support system

Sushil Punia [a,*], Sonali Shankar [b]

[a] Vinod Gupta School of Management, Indian Institute of Technology Kharagpur, India
[b] Jindal Global Business School, O. P. Jindal Global University, Sonipat, India

ARTICLE INFO

Article history:
Received 19 September 2021
Received in revised form 26 September 2022
Accepted 26 September 2022
Available online 3 October 2022

Keywords:
Demand forecasting
Predictive analytics
Deep learning
Optimization
Demand sensing
Retail

ABSTRACT

The demand is often forecasted using econometric (regression) or statistical forecasting methods. However, most of these methods lack the ability to model both temporal (linear and nonlinear) and covariates-based variations in a demand series simultaneously. In this context, a novel forecasting model is proposed that combines a state-of-the-art sequence modeling method and a machine learning method in an ensemble model. The proposed model can handle both types of variations in demand data, and thus, enhances forecasts' accuracy. A big sample of 4235 demand series consisting of structured and unstructured data (could be referred to as "big data") related to packaged food products is used for experimentation. Data contain point-of-sales, promotion, weather, regional economy, internet media, and economic activity index related variables. Some of these variables and their combinations is probably used for the first time in a demand forecasting model. The forecasting results are evaluated through multiple error metrics (i.e. mean error, mean absolute error, mean squared error), and it has been observed that proposed method outperformed the benchmarking methods. A demand sensing algorithm is also proposed to forecast demand in real-time.

1. Introduction

Demand estimates act as a primary input for effective planning and decision making in any organization. A firm's marketing, production, distribution, and finance departments use short-to-long term forecasts to support different decisions. Being such a pivotal input to business decision-making, the quality of forecasts is very important.

Demand forecasting needs historical demand data and forecasting methods to forecast the future demand. The first step is to collect relevant data on various factors i.e., product features, promotional activities, calendar events, meteorological and general economic contexts that influence the demand for retail goods [1,2]. Understanding the impact of these factors on demand provides the needed business intelligence to the retailers for effective sales planning and management. Next, modeling and forecasting of demand data requires suitable forecasting methods/models.

In this paper, authors proposes a big data predictive analytics model capable of handling a large amount of demand data and provide short, medium, and long-term demand forecasts to a retailer. As per the classification of forecasting methods based on data characteristics by Punia et al. [3, p. 4965], the proposed model could be placed in the category of medium to a large dataset with multiple input variables. Thus, machine and deep learning techniques are used for forecasting.

1.1. Demand forecasting in retail

The research on forecasting models started with univariate models with sales series as the input data. These models detect and use the temporal patterns to predict the sales for the future [3]. However, the sales pattern is influenced by multiple factors and forecasts from multivariate models are often better than forecasts from univariate models [4,5].

In multivariate models, a set of factors, broadly categorized into point-of-sales, promotion, weather, and general economics context variables, have been used as the independent variables in various studies. Geurts and Patrick Kelly [4] highlighted the importance of using point-of-sales and promotion data for sales forecasting. Choi and Li [6] reported that the forecasting models could achieve better performance using autoregressive components of sales variables. Choi [7] reported that market information for pre-seasonal products would lead to better forecasts and recommended the continuous update of forecasts based on available near real-time information. Au et al. [8] used point-of-sales information such as pricing, discounts, and product features to predict the sales in apparel fashion retail; and reported multivariate model are better than univariate models for regular products

* Corresponding author.
E-mail address: spunia@vgsom.iitkgp.ac.in (S. Punia).

https://doi.org/10.1016/j.knosys.2022.109956
0950-7051/

with low demand uncertainty and weak seasonal trends. Kumar and Patel [9] revealed that performing the clustering based on features leads to better performance of the forecasting methods.

In the past decade, researchers discovered several new external indicators to estimate future demand for products. The factors related to weather, economic indicators, internet social media, and sentiments indices are found to be most prominent. Since data on external indicators are not directly available to the retailer and ascertaining the veracity of the external data is challenging too. For these reasons, limited research is available on the use of multiple external factors for demand forecasting.

Osadchiy et al. [10] used financial market information, such as financial index, returns achieved on equity, etc., to better the forecasting performance. Ferreira et al. [5] used the point-of-sales, promotion, and time-based indicators such as *day of the week, day of the month,* etc., to forecast demand for an online retail company. They reported that the use of additional market and temporal information helped in better accuracy. Papanagnou and Matthews-Amune [11] used the internet social media indicators such as Google index, YouTube index for forecasting the sales of a pharmaceutical drug, and revealed a positive correlation between social media indicators and sales prediction. Verstraete et al. [12] used the weather information to forecast low-margin high-volume products for short-term and long-term forecasts. They inferred that weather indicators significantly influence the sales of low-margin, high-volume products.

With the increase in number of variables, the complexity of the model increases. It requires the management of massive data and complex models to handle, process, and generate a quality forecast. Thus, only a few studies tried to incorporate multiple types of indicators in a single forecasting model. For example, [13] developed a model that included product features, promotion, and economic indicators. However, they also did not include weather information and social media indicators.

This paper addresses the demand forecasting problem by incorporating data on factors related to product features, promotion, weather, regional economy, and internet social media. The data for three years for the 4235 demand time series with independent variables are used for the study. The details of the independent variables and factors are explained in Section 2. To manage the huge data, a big data framework using Apache Spark is designed. This framework is used to process, model, and analyze the data to generate forecasts efficiently. In addition, the relative importance (ranking) of the independent variables are provided to understand the impact of these variables on sales.

1.2. Methods for demand forecasting

The time-series data based demand forecasting methods can be classified into three categories: statistical methods, machine learning methods, and hybrid methods. In time-series methods, the exponential smoothing method, Auto-Regressive Integrated Moving Average (ARIMA), various decomposition models are used for forecasting. Further, the use of some multivariate time-series methods such as ARIMAX has also been used. The details of these methods are available in Hyndman & Athansopoulos [14].

In machine learning methods, artificial neural networks (ANNs) are widely used for demand prediction. Alon et al. [15] used the ANN to predict the aggregate sales in retail stores and reported that ANN could capture the dynamic nonlinear trend and seasonality in the sales data. Au et al. [8] used the evolutionary neural network for sales forecasting in fashion retail and reported improvements in the accuracy of forecasts. Ferreira et al. [5] predicted the sales using the random forest (RF) method based on regression trees and bagging algorithms and reported the advantages of RF over NNs in terms of interpretability and accuracy. The latest addition to the repository of multivariate methods is a deep neural network, and results are encouraging to use of deep learning for the sales forecasting and decision making [16,17].

In hybrid models, both time-series and regression (or machine learning methods in recent) are used to model the demand patterns. The hybrids of ARIMA-ANN [18], ARIMA-regression [19], ARIMA-SVM [20], and seasonal ARIMA and wavelet transformations [21] were proposed for forecasting. It may be noted that ARIMA is widely used with other methods to develop hybrids. Because sARIMA can efficiently handle and model the linear temporal (level and seasonality) part of the time series, and the remaining nonlinear temporal and regression part is taken care of by method [22].

To further the research on hybrid models, this paper proposes a novel forecasting model, which combines deep learning-based long–short-term-memory (LSTM) networks and random forests (RF) method. LSTM networks are the state-of-the-art techniques to predict the linear and nonlinear sequential data [23], and RF is a machine learning technique to model relationships among sales and independent variables [5]. Both methods are combined using a genetic algorithm into an ensemble model. The forecasts from the proposed model are tested on three error metrics for bias, accuracy, and variance. The results are compared with forecasts from other demand forecasting methods. The proposed method outperformed all other methods on all three metrics. Further, this paper proposes two auxiliary algorithms to generate daily and long-term forecasts. These algorithms use temporal aggregation and temporal disaggregation methods to convert the forecasts from the proposed method to daily and long-term demand forecasts for retailers [14]. The study will significantly contribute to the literature on forecasting and forecasting applications in retail industry.

The remainder of this paper is organized as follows. The proposed forecasting model for demand planning is presented in Section 2. In Section 3, demand data and its characteristics are discussed. Section 4 contains the data preparation, data analysis, and results. Section 5 analyzes and discusses the results and provides managerial insights for the effective use of the proposed model. Finally, conclusions and future work are described in Section 6.

2. Proposed demand forecasting framework

The proposed demand forecasting framework provides the forecasts for short, medium, and long-term planning horizons. The short-term model can be used for operational decision-making while the medium and long-term model has usage for tactical decisions such as handling products with long lead times in retail. The factors related to point-of-sale, promotion, time, store, and external indicators related to weather, social, and economy are used input data. An overview of the whole methodology is described in Fig. 1.

2.1. Methodology

To eliminate reduce the high dimensionality of data and multicollinearity among variables, Principal Components Analysis (PCA) is used. PCA is a dimensionality reduction multivariate statistical technique first introduced by Pearson and later in 1933, developed by Hotelling. PCA uses linear combinations to generate non-overlapping components retaining maximum original information. Therefore, it can help in reducing the multicollinearity and dimensionality of retail data. Further, to avoid poor fitting of traditional linear time-series model to nonlinear sequences and incapability of machine learning methods on sequential data, we

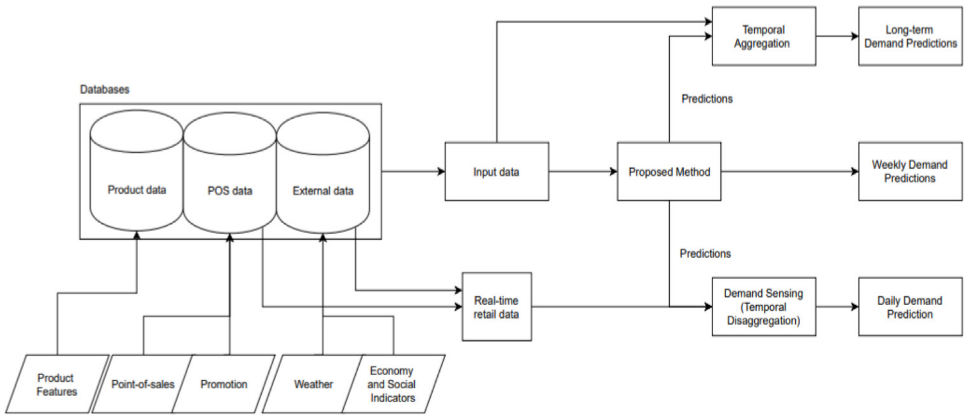

Fig. 1. Flow diagram for Proposed Demand Forecasting Framework.

propose to use the long–short-term memory (LSTM) networks. The LSTM networks are state-of-the-art deep learning techniques to handle sequential data. LSTM networks will help to address the linear and nonlinear relationships among dependent and independent variables. Further, being a deep learning technique, it can also handle big data efficiently than other traditional methods.

However, LSTM uses time windows (look back) for making future predictions. Due to this approach, LSTM will get fewer data per independent variable in one step than a machine learning method. This approach restricts the scope of LSTM networks to model the impact of the interaction among independent variables on sales. Therefore, we propose to make a hybrid method in which a separate machine learning method will be combined with LSTM networks. Based on the literature, it is found that random forest (RF) performs better than linear regressions and neural networks for demand forecasting and also provides more interpretable models. This will help the retail managers to understand the influence of independent variables on sales.

To combine predictions from LSTM networks and RF, we proposed an ensemble technique. We used the LSTM only for univariate sales time-series data and the RF for the multivariate model. Further, a genetic algorithm is used to assign the weights analytically, eliminating any bias in weights. In these ways, the proposed methodological framework provide a better alternative to existing ones. The schematic diagram of the proposed framework is shown in Fig. 2. Morever, a detail look at the working of LSTM, RF and optimization for ensemble is provided in the below sub-sections.

2.1.1. Time-series data modeling the LSTM networks

The base for LSTM networks is the Recurrent Neural Networks (RNN), which are a better version of the traditional neural network (NN) to handle the sequential (time-series) data. To forecast the time series data, methods are required to memorize the temporal patterns. The nodes in the hidden layers of the RNN are connected, whereas, in traditional NN, only hidden layers are connected. This accounts for the efficient memory of the RNN over NN and its ability to memorize the information from previous time steps. Unlike NN, the hidden layer in RNN at time-step t receives the information from the input layer and receives the information transmitted from the hidden layer at time-step *t-1*. Therefore, RNNs use their memory to remember the dependencies among elements of the sequence and can process long sequences. However, RNN requires overcoming the vanishing gradient problem, i.e., non-convergence of the network towards the end, to learn long-term dependencies. Gradient clipping is one of the methods to handle this problem, which is applied by modifying RNNs to the long–short memory (LSTM) networks (see Fig. 3).

Notations.

- X_t: the input vector at time step t
- h_t : the output vector at time step t
- c_t : the vector for cell state t
- $\tilde{c}_t$: the vector for a candidate value for input gate
- f_t, i_t, o_t : vectors of values obtained after activation of the gates (forget gate, input gate, and output gate)
- b_f, b_i, $b_{\tilde{c}}$, b_o : bias vectors

$W_{f,x}$, $W_{f,h}$, $W_{\tilde{c},x}$, $W_{\tilde{c},h}$, $W_{i,x}$, $W_{i,h}$, $W_{o,x}$, $W_{o,h}$: weight matrices for input and outputs for the three gates

The LSTM networks are proposed by Hochreiter and Schmidhuber [24] and are further enhanced by Graves and Schmidhuber [23]. The LSTM handles the vanishing gradient problem by an improved structure of a memory cell. The memory cell is shown in Fig. 4. The LSTM adds and forgets information using three gates, namely, input gate (i), forget gate (f), and output gate (o). Let, C_{t-1} and C_t are the cell states at time t-1 and t respectively. The input and hidden states are represented by X_t and h_t respectively. The cell states and gates are updated through the following equations:

$$f_t = \sigma\left(W_{f,x}X_t + W_{f,h}h_{t-1} + b_f\right) \tag{1}$$

$$\tilde{c}_t = tanh\left(W_{\tilde{c},x}X_t + W_{\tilde{c},h}h_{t-1} + b_{\tilde{c}_t}\right) \tag{2}$$

$$i_t = \sigma\left(W_{i,x}X_t + W_{i,h}h_{t-1} + b_i\right) \tag{3}$$

$$c_t = f_t * c_{t-1} + i_t * \tilde{c}_t \tag{4}$$

$$o_t = \sigma\left(W_{o,x}X_t + W_{o,h}h_{t-1} + b_o\right) \tag{5}$$

$$h_t = o_t * \tanh\left(c_t\right) \tag{6}$$

$\sigma(.)$ denotes the sigmoid and *tanh(.)* hyperbolic tangent function. $W_{,x}$ and $W_{,h}$ indicate input weights and recurrent weights, respectively, b represent the bias. The backpropagation algorithm updates weights. The forget gate remove the information from the cell state c_{t-1} using Eq. (6). After that input gate adds information to the cell state c_t as shown in Eqs. (2) and (3). The output gate through Eqs. (6) and (7) decides how much information will be carried forward to the next cell of the LSTM network.

The LSTM networks require the variable in to form of sequences for training. The target and input sequences are generated for sales data, as illustrated in Fig. 5. The sales variable time-series sequence is used as the target variable, and for the creation of independent variables, the sales variable is ordered as time-series, and the autoregressive series are generated by time-periods shifting.

2.1.2. For multivariate data modeling:Random Forest (RF)

The Random Forest (RF) algorithm is used to develop a model between principal components and sales. RF works on the

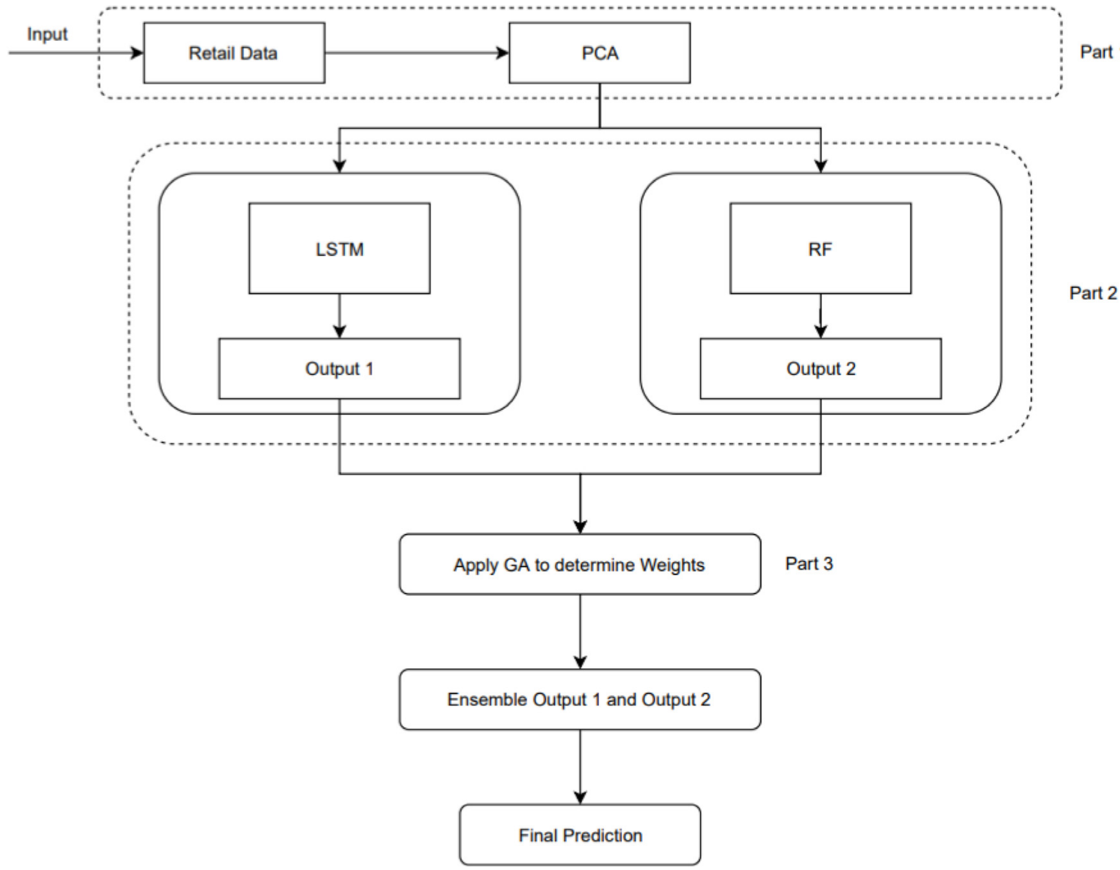

Fig. 2. Flowchart of the Proposed Ensemble Method.

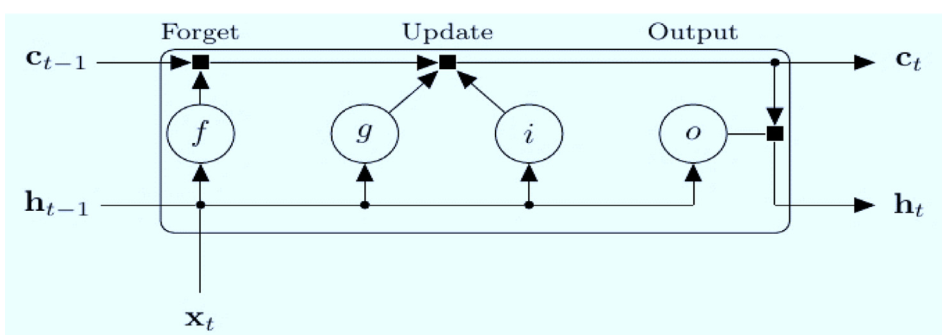

Fig. 3. The LSTM Memory Cell Architecture.

	Product 1									Product 2			
Week	1	2	3	...	101	102	103	104	...	1	2	3	...
Units	27	36	28	...	48	44	30	27	...	14	23	17	...

	Product 1						Product 1			
Seq. 1	1	2	3	...	101	102	1	2	3	...
	27	36	28	...	48	44	14	23	17	...

	Product 1						Product 1		
Seq. 2	2	3	...	101	102	103	2	3	...
	36	28	...	48	44	30	36	28	...

Fig. 4. Input sequences for the LSTM networks (all sequences are transposed).

principle of *bagging*. The predictions from many decision trees are combined to provide a final prediction. The bootstrap sample S_n is randomly selected from the n observations with equal probability and replacement. Let the variable Φ_q, be an i.i.d. random variable. Several such samples $(S_n^{\Phi 1}, \ldots, S_n^{\Phi q})$ are selected from the data. Afterward, the classification and regression trees

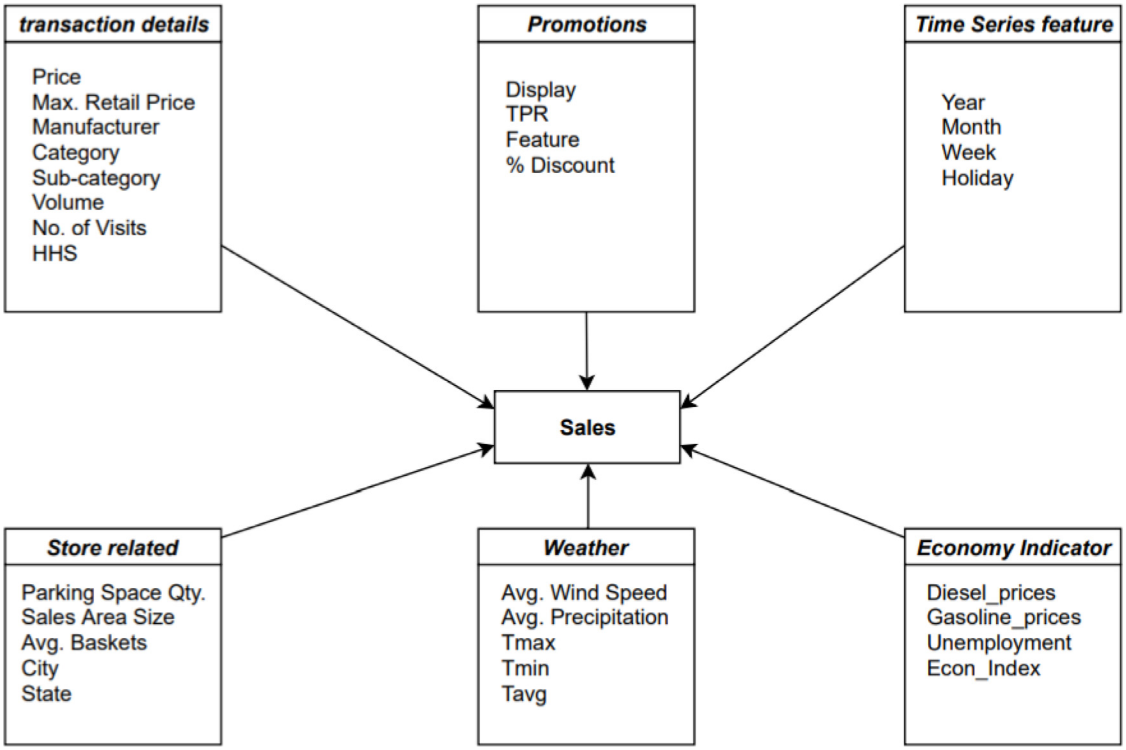

Fig. 5. Groups of Independent Variables Used for Demand Forecasting.

(CART) algorithm is applied to these samples to obtain q predicting trees $(\hat{h}(X, S_n^{\Phi 1}), \ldots, \hat{h}(X, S_n^{\Phi q}))$. The outputs are aggregated through a bagging algorithm to provide the result. The output, Y' for the unknown input, X,' is estimated with the help of Eq. (8). The Y' is calculated by averaging the output from all the tress.

$$Y' = \frac{1}{q}\sum_{l=1}^{q}\hat{h}(X, S_n^{\Phi_l}) \tag{7}$$

2.1.3. Ensemble model

After getting the predictions from the LSTM networks and RF, the weighted aggregation of the prediction is performed. Suppose for a time-series vector (Y_t), the predictions from the LSTM networks and RF are $\hat{y}_t$ and y_{rf} vectors respectively, then the aggregation vector equation is given by:

$$\hat{Y}_t = \alpha_1\hat{y}_t + \alpha_2 y_{rf} \tag{8}$$

The weights (α_1 and α_2) for the aggregation are calculated by an optimization problem. The minimization of the mean absolute error $(\sum_{i=1}^{n}\left|Y_t - \hat{Y}_t\right|/n)$ is considered as a cost function, and the summation of weights is restricted equal to 1 with both weights can vary between 0-1. The Genetic Algorithm (GA) is used for the solution of this optimization problem.

In GA, initially, a set of the random population of weights (known as *chromosomes*) is generated. Over the next several steps, these potential solutions are modified using selection operator (selection of a set of solutions for further operation), crossover operations (swapping of genetic contents in chromosomes), mutation operations (point changes in the chromosome), and elitism (retaining the top solution) operations in the algorithm [25]. The steps are repeated in each subsequent generation until the convergence or the user-defined maximum number of generations is achieved. The best weights for the ensemble are the output of the optimization problem. As no experts' opinions are used to decide the weights, this process is free from any bias, and the whole demand forecasting process is automated.

2.2. Methodologies for generating short-term and long-term forecasts

2.2.1. Long-term forecasts

The weekly time series are aggregated to lower frequency (say monthly or quarterly) from a higher frequency (weekly) series. After that, the proposed forecasting method is applied to monthly and quarterly data, and separate forecasts are generated for these horizons. Then forecasts are reconciled using the hierarchical reconciliation [26]. It makes the forecasts coherent, i.e., the sum, average, start, and end of all forecasts are consistent. For reconciliation, the algorithm proposed by Athanasopoulos et al. [26] is used. The benefits of using this algorithm are that it extracts both high and low-frequency components and provides a chance to model seasonality and trend/cycle separately. These components help to understand better and model the data and tend to give accurate, coherent, and robust forecasts. In this way, the monthly and other low-frequency forecasts were generated.

2.2.2. Short-term Forecasts (demand sensing)

We propose the use of real-time point-of-sales data to generate the daily sales forecasts from the weekly forecast. In this way, first, all types of variables (point-of-sales, promotions, economic indicators, etc.) will be utilized to generate the weekly forecasts. When additional daily point-of-sales information is available, the daily sales forecasts can be generated. This type of forecasting is also known as *demand sensing*. The concept of temporal disaggregation (TD) is used to accomplish this task. TD uses the available high-frequency indicator series to disaggregate the low-frequency series to high-frequency series. TD is a well-known technique in other research areas for time-series forecasting [27, 28]; however, it has not been used much in retail.

For TD, [29], [30], [31] are some of the most prominent available methods. Denton [29], Dagum and Cholette [30], mainly concerned with movement preservation, while the other mentioned methods used the Generalized Least Squares (GLS) regression among indicators and target variable to generate the low-frequency series. Chow and Lin [31] is more suitable for stationary series [32]. We use the [33] where weekly forecasts series with daily point-of-sales variables as the high-frequency indicator will be the input to [33] method. The daily sales forecasts will be generated after fitting the GLS regression among low-frequency and high-frequency independent variables.

3. Demand data and its characteristics

3.1. Factors influencing the sales

The factors that influence sales can be categorized into point-of-sales, promotion, time, store, weather, and external indicators. The point-of-sales consists of variables related to product specification (category, color, manufacturer, volume, size, etc.) and customer's transactions (units sold, price, discount, no. of customer visits, etc.). Then, there promotional events and marketing related variables, such as temporary price reduction (TPR), featuring in the display area, etc. These variables represent special attention given to a product for a short period to boost its sales. Further, store-related variables such as location, parking space, total sales area, city name, type of store, etc., may not be directly used but can be helpful to extract the information on external indicators on regional weather and economy.

Weather related factors such as temperature, and wind speed are included to incorporate the impact of weather on the sales of products and to improve forecasting of seasonal products [34]. Further, economic indicators such as fuel prices, unemployment rate, and metropolitan economic index are selected as the proxy for economic indicator of the geographical region in which the store is located. Here, the economic index represents the average economic growth in the metropolitan area based on a set of dynamic factors [35]. Therefore, it can indicate the economic condition of any store's nearby area and its consumers.

Finally, a weekly Google search index is obtained from the Google trends website and used to predict sales. The name and category of the product are used as the keywords. Google search index indicates the volume ratios of keywords searched in specific geographical areas related to the word of interest. The formula for obtaining the Google search index is given by:

$$GI_{t,i} = \left[\frac{S_{t,i}/R_{t,i}}{\sum_{t=1}^{N_w} (S_{t,i}/R_{t,i})} \right] * 100, \quad t \in [1, N_w] \tag{9}$$

where $S_{t,i}$ represent the number of searches with selected keywords ($k_{t,i}$), $R_{t,i}$ denotes the total number of search queries in geographical area i in total N_w weeks. The search volume index (keyword search ratio) obtained from google trends is given by ($k_{t,i}/GI_{t,i}$). In the Google search occurring over multiple geographical regions, the google search index is simply aggregated over all geographical regions.

In recent literature, researchers ignored the time-series nature of the retail data to deploy machine learning techniques for sales forecasting. These studies included the proxy variables such as day of sales, week of sales, month of sales, etc. as independent variables to account for the temporal variations in the sales. In this study, we keep the data as time series, yet the time-series features are extracted for the input to benchmarking methods. All the independent variables used for the current study are summarized in Fig. 5.

3.2. Data characteristic

Espousing on the recent approach of researchers and practitioners, "5Vs" big data framework is used to explain the characteristics of the retail data. The 5Vs stand for the big data's innate characteristics and refer to Volume, Variety, Velocity, Veracity, and Value. In a retail chain, data on hundreds of thousands of products are available, which adds to a massive *volume* of data (e.g., ∼24 million data points in this study). The structured (e.g., historical sales), semi-structured or unstructured data (e.g., Google search index) represent the *variety* characteristic. The data is continuously generated through various retail operations, thus, representing the *velocity* dimension. The *value* of retail data is high as it is used for determining the impact of predictor variables on the forecasting accuracy, and *veracity* is also high because data is obtained through in-house or reliable sources.

To summarize, the following characteristics are identified for retail data:

- multivariate time series data
- it exhibits 5Vs of the big data
- nonlinear relationships between dependent and independent variables
- interaction among independent factors
- multicollinearity among some independent variables

4. Data analysis

4.1. Data and summary statistics

The data is taken from a large retailer who operates multiple retail store. The data consists of weekly sales for 55 food items sold through 77 stores and is available for the duration from January 2009 to January 2012. The total sample size adds up to 4235 demand time series with multiple independent variables. The independent variables are divided into five categories: point-of-sales, promotions, store, geographical weather, and other economic indicators. In addition, the sixth category is derived from the time dimension and will only be used with a machine learning method, which ignores the time-series nature of the data. The summary statistics of the variables from the retailer's dataset are provided in Table 1.

Many independent variables are self-explanatory from variables names, yet, descriptions of some of the new variables are as follows. The "no. of visits" and HHS means the number of customers and the number of purchasing households that visited the store in the given week, respectively. The Display, TPR, and Feature stand for whether the product was part of the in-store display, temporary price reduction, and appeared in in-store circular, respectively. Finally, the economic activity index (Econ_Index) measures the average economic growth in a metropolitan area, as described in Section 3.1. The weather and economic indicators data are collected through the National Oceanic and Atmospheric Administration (NOAA), USA, and www.data.gov public websites, respectively, so these indicators are not summarized here. For the model development, two years of data is used for training, and the remaining one year of data is used for testing purposes.

4.2. Big data framework for data management and modeling

For the pre-processing of the data, two freeware and open-source platforms, R and Python, are used. Further, to model the big data, the *Apache SystemML* library is used in integration with Keras [36] to apply the machine learning and deep learning algorithms. The *Apache SystemML* is the workspace that works on top of the big-data framework Apache Spark to efficiently run machine learning with big data [37]. Apache Spark is a popular framework to process big data through distributed operations. For benchmarking, *forecast* package [38] for S/ARIMAX, *neuratnet* [39] for backpropagation neural network, and *caret* [40] for data modeling was used. The flowchart for the big data analysis framework is presented in Fig. 6.

Table 1
Summary statistics of input variables for demand forecasting.

Variable	2009		2010		2011	
	Median	Mean (SD)	Median	Mean (SD)	Median	Mean (SD)
Sales	11	16.88 (31.30)	11	17 (30.55)	10	16.04 (27.88)
MRP	3.00	3.50 (1.57)	3.05	3.55 (1.65)	3.29	3.64 (1.66)
Price	2.92	3.29 (1.54)	2.99	3.32 (1.55)	3.19	3.40 (1.57)
% Discount	0	4.98 (11.89)	0	4.86 (11.23)	0	4.83 (11.20)
Visits	10	14.93 (25.01)	9	15.04 (25.41)	9	14.27 (23.73)
HHS	9	14.57 (24.42)	9	14.67 (24.83)	9	13.96 (23.19)
Avg. Baskets	24766	24324 (8915)	24766	24345 (8874)	24766	24357 (8899)
Parking Qty.	0	128.32 (342.5)	0	127.85 (344.7)	0	127.81 (341.9)
Sales Area	48632	49892 (13434)	48632	49925 (13384)	48632	49974 (13480)
		% of cases		*% of cases*		*% of cases*
Feature		0.062 (0.28)		0.061 (0.28)		0.060 (0.27)
Display		0.094 (0.32)		0.082 (0.30)		0.091 (0.32)
TPR		0.129 (0.35)		0.109 (0.33)		0.111 (0.33)
Discount		0.273 (0.45)		0.243 (0.44)		0.251 (0.44)
No. of Samples	4235		4235		4235	
No. of Obs.	164145		179136		177894	

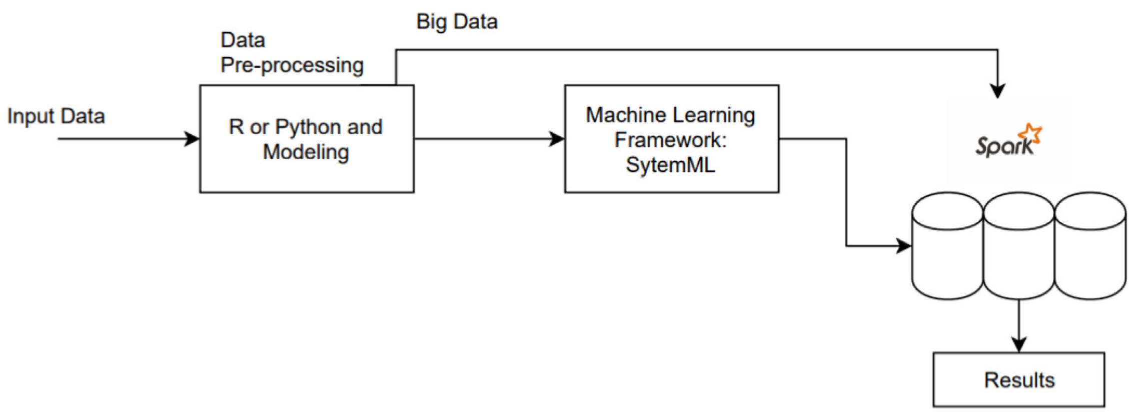

Fig. 6. Big Data Framework Used for Modeling and Analysis.

Table 2
Principal Component Analysis Results for Point-of-sales variables.

Component	Total	% of var	Cumulative %	variable	Explanation of components	Test
1	2.061	0.515	0.515	VISITS	0.9991	Bartlett's test
2	1.903	0.476	0.991	HHS	0.9992	$p < 0.0001$
3	0.034	0.008	1.000	PRICE	0.9829	Chi-sq. (n = 2)
4	0.002	0.000	1.000	MRP	0.9834	$p < 0.0001$

4.3. Data conversion: Principal component analysis (PCA)

Taking the example of point-of-sales variables (Table 2), it can be observed that the first two components explain over 99.1 percent of the variance of variables and at least 98 percent of each of the variables. The Bartlett test of sphericity with p-value < 0.0001 means the null hypothesis that the correlation matrix is an identity matrix can be rejected. It means the variables in the point-of-sales set were suitable for the PCA. For selecting the number of components, the chi-square test is conducted. For $n = 2$, with p < 0.0001, null hypothesis is rejected. Therefore, selected two principal components can replace the four variables in our model.

Similarly, the PCA is applied to variables from promotion, weather, and economic factors, and results are presented in Tables 3–5. Based on the chi-square test, the optimal number of principal components are selected from each PCA, and these components will be used as input to the forecasting model.

4.4. The benchmarking methods and performance metrics

The RF and backpropagation neural networks are widely used for demand forecasting in retail. Further, the ARIMA method is also used for time-series prediction; and ARIMAX is the multivariate version of ARIMA. All these methods are used as first set of Benchmarking Methods (BM) in the study. Further, the multivariate LSTM networks are also used to benchmark the proposed combination of LSTM and RF. We also compared our results with other hybrid methods i.e., ARIMA+NN, ARIMA+RF. The OLS regression is chosen as the baseline to scale the errors to a relative scale.

The parameters selection and hyperparmeters optimizations were performed for the proposed and benchmarking methods. For LSTM networks, in Keras, input data is provided as [samples, timesteps, features] and normalized to the range of [−1,1]. Further, grid search method was used for hyperparameter optimization. In grid search, optimizers ('adam", and "sgd"), activation functions ("linear" and "relu") and different values of LSTM layers for the network were taken [3]; and an optimization algorithm

Table 3
Principal Component Analysis Results for Promotion-related variables.

Component	Total	% of var	Cumulative %	variable	Explanation of components	Test
1	1.468	0.367	0.367	Discount	0.9789	Bartlett's test
2	1.234	0.309	0.676	Feature	0.9992	$p < 0.0001$
3	1.149	0.287	0.963	TPR	0.9833	Chi-sq. (n = 3)
4	0.149	0.037	1.000	Display	0.9959	$p < 0.0001$

Table 4
Principal Component Analysis Results for Weather-related variables.

Component	Total	% of var	Cumulative %	variable	Explanation of components	Test
1	2.916	0.583	0.583	Wind Speed	0.9997	Bartlett's test
2	1.041	0.208	0.791	Precipitation	0.9998	p < 0. 0001
3	1.004	0.201	0.992	Tmax	0.9759	Chi-sq. (n = 3)
4	0.039	0.008	1.000	Tmin	0.9816	$p < 0.0001$
5	0.000	0.000	1.000	Tavg	0.9999	

Table 5
Principal Component Analysis Results for Economy related variables.

Component	Total	% of var	Cumulative %	variable	Explanation of components	Test
1	2.149	0.537	0.537	Gasoline price	0.9998	Bartlett's test
2	1.042	0.260	0.798	Diesel price	0.9998	$p < 0.0001$
3	0.809	0.202	1.000	Unemployment	0.9999	Chi-sq. (n = 3)
4	0.000	0.000	1.000	Econ_Index	0.9999	$p < 0.0001$

Bergstra et al. [41] was used for hyperparameter optimization. Also, regularization and early stopping techniques were utilized to prevent the overfitting. For RF and NN methods, a similar approach of grid search was implemented by choosing different set of parameters and selecting the best configuration based on minimum error metrics.

To evaluate the forecasting performance of our proposed method mean error (ME) for bias, mean absolute error (MAE) for accuracy, and mean squared error (MSE) for variance are included in the study. Further, the errors metrics were made relative to make them scale-free and east-to-interpret, thus the relative mean error (RME), relative mean absolute error (RMAE), and the relative mean squared errors (RMSE) were used. The relative errors were calculated by dividing the sum or mean of errors from the evaluated method with a benchmark method [42]. As the study has thousands of time series and multiple methods, average relative forecast errors are used for better readability and interpretability of the results.

4.5. Results and discussions

Table 6 presents the performance of the proposed and benchmarking methods. It can be observed that the proposed method outperformed the other methods on all three error metrics. Notably, MAE and RMSE for the proposed methods are better than the other methods by a significant margin. The low MAE and RMSE mean predictions are closer to the actual values. The better accuracy of the proposed method can be attributed to its capability to capture linear and nonlinear temporal features. Because of these capabilities, the proposed method can capture the trend and seasonal peaks in the sales better than other methods. For ME, the proposed methods, ARIMA, ARIMA+RF, and LSTM, are competitive. The negative bias of proposed methods and other neural network-based suggests that these methods are slightly over-forecasting the sales. However, in absolute terms, the bias is lower than other methods. It means that the sales prediction from the proposed method is not only more accurate but also over- and under-forecasted more equally, which is important for decision-makers.

Table 6
Relative errors for One-week ahead predictions (with ranking in brackets).

	ME		MAE		RMSE	
OLS	1.000		1.0000		1.0000	
ARIMA	0.1553	(2)	0.6757	(7)	0.7918	(8)
ARIMAX	0.4018	(6)	0.6367	(5)	0.7009	(6)
RF	0.3146	(5)	0.5106	(4)	0.5796	(4)
NN	−0.4479	(8)	0.7511	(8)	0.7551	(7)
LSTM	−0.1950	(4)	0.4954	(3)	0.5588	(2)
ARIMA+NN	−0.2452	(7)	0.6586	(6)	0.6686	(5)
ARIMA+RF	0.1704	(3)	0.4221	(2)	0.5689	(3)
LSTM+RF	−0.1216	(1)	0.3569	(1)	0.4638	(1)

Further, based on ME, MAE, and RMSE (Table 6), it can be said that hybrid methods are performing better than their constituent methods. In one instance, the MEs of the ARIMA and its hybrid with RF are competitive. However, poor MAE and RMSE with good ME of ARIMA suggest that the ARIMA model is predicting around the mean values and due to which the bias of the ARIMA is lower, but the accuracy is poor. The multivariate version of ARIMA, which is ARIMAX, performed better than ARIMA because it can model the relationships among independent variables and sales. This phenomenon also highlights the importance of including contextual business variables in sales prediction in retail. The random forest is performing best among the non-hybrid methods on all three metrics. However, the performance of the hybrid method of RF is further better. This highlights the relative importance of keeping the time-series nature of the data.

The impact of PCA is also analyzed on the forecasting performance. For that purpose, the error metrics of the LSTM+RF without PCA and with PCA are calculated. The error metrics with PCA are reported in Table 6 and found to be slightly better than errors metrics for LSTM+RF without PCA (ME = −0.1291, MAE = 0.3657, and RMSE = 0.4782). In addition, the PCA made the results of benchmarking methods more accurate, as it is widely accepted that PCA reduces the number of predictors in linear regression and other regression methods and improves their performances. Also, PCA helps better explain factors (by

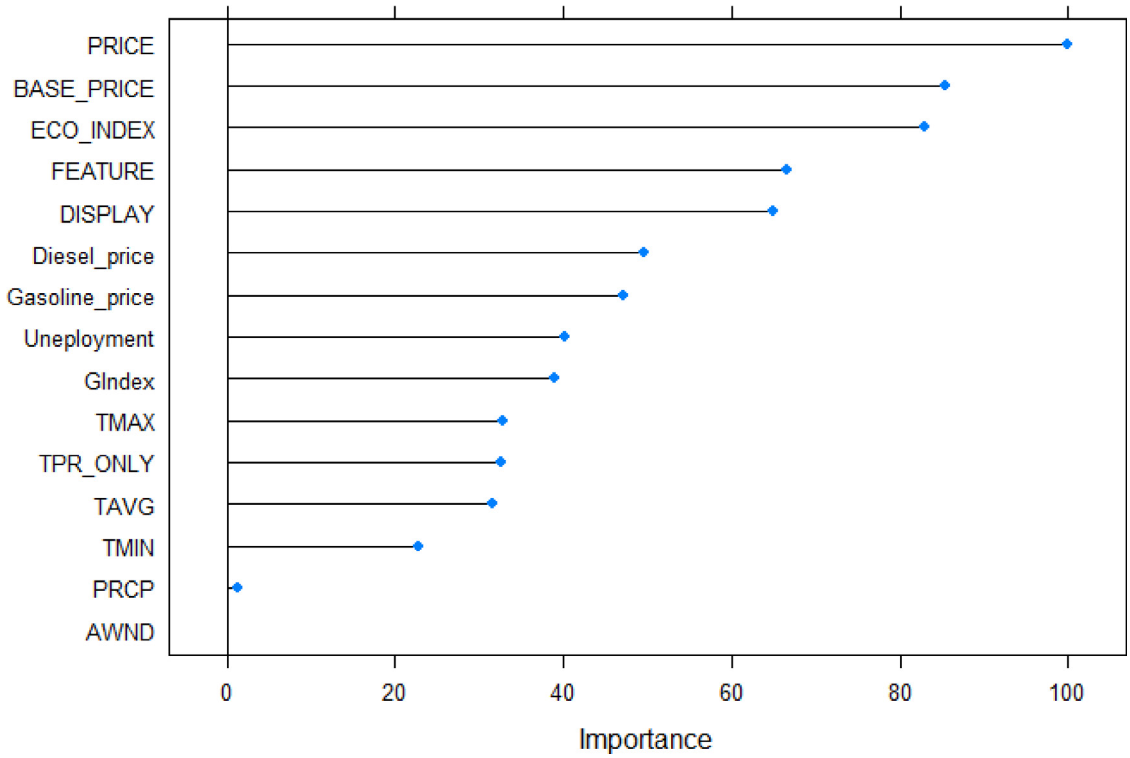

Fig. 7. Relative importance of input variables for sales prediction.

Table 7
Real-time daily point-of-sales data.

Day	1	2	3	4	5	6	7
PRICE	1.17	0.99	0.98	1.17	1.18	1.17	1.17
Discount	1	1	1	0	1	0	0
FEATURE	0	1	0	0	0	0	0
DISPLAY	1	0	0	0	1	1	1
TPR_ONLY	0	0	1	0	0	0	0

Table 8
Error results for the short term forecasting.

	ME	MAE	RMSE
Proposed Approach	0.1248	1.7428	2.2741
Mean Forecast	−0.1457	2.5715	2.8732
Relative Errors	0.8565	0.6777	0.7914

aggregating the different variables) and reducing the number of input variables.

Following the proposed methodology mentioned in Section 2.2.1, the weekly data is aggregated to generate long-term forecasts. Using the proposed method as the base forecast method, the separate forecasts are made for monthly and quarterly sales. Following the proposed approach from Section 2.2.2, the weekly forecasts are disaggregated to daily demand forecasts using daily point-of-sales data as indicators. For example, the weekly forecast for one sample week was 68, and real-time daily point-of-sales data is shown in Table 7. Then using the daily data, the weekly forecast was disaggregated into daily sales series of {7,5,5,8,8,14,14}. The actual aggregate sale for the week was 62, with daily sales distribution series of {11,7,5,5,8,14,12}, with a mean of 8.85 (∼9). A comparison of mean forecast and demand sensing results from our proposed approach reveals a significant improvement in the accuracy of the daily forecasts. The overall results for all demand series are summarized in Table 8.

5. Managerial implications

The proposed demand forecasting framework can be used for effective tactical and operational planning decision by retailers. The accurate and real-time estimates of daily, weekly, and monthly demand will help the retailer to plan the right products mix to procure from the distributors. This will lead to a significant saving of inventory and transportation costs. This will also help in the optimal planning of assortment and assortment promotions at the stores.

The proposed model provides the relative importance of the factors that influence sales (Fig. 7). Understanding these factors that drive the store and product sales is essential for the retailer to design the promotional events, plan the assortment display and shelf-space optimization, etc. Fig. 7 shows that Price and maximum retail price (base price) are the most critical factors, i.e., customers consider both types of prices for buying. This might be because customers want to buy quality products by avoiding artificially heavily discounted products. In the top three important variables, the presence of economy-related variable — Economy Index (Econ_Index) gives more support to the claims made in the literature that external variables can be good indicators for demand prediction in retail. Econ_Index is a combination of 12 economic indicators, and it is used for the first time in demand forecasting in retail. Therefore, we suggest the retailer and researchers include this variable in place of other economy and consumer-related individual variables for better results and concise data. Promotion-related variables are also coming in the top five crucial variables, suggesting promotion and assortment optimization is also essential for retail. Google index and weather-related variables have weak prediction power for sales, but it can be observed that maximum temperature is more important among weather-related variables.

The proposed demand forecasting framework also generated long-term and real-time updated daily demand forecasts. As the forecasts are coherent across planning horizons, it will remove the inconsistencies in planning and decisions across various departments of retail companies. Thus, demand forecasting framework can serve as decision support system for forecasting and operations planning for retailers. The proposed methods can also be used in improving forecasting in other industries like logistics & transportation [43,44], energy forecasting [45] hierarchical forecasting [46,47] etc.

6. Conclusions

A novel demand forecasting method is developed to generate accurate short, medium, and long term demand forecasts in retail stores. The proposed method combines LSTM networks and random forests into an ensemble model. The principal component analysis is used to reduce dimensionality of the input data. The proposed forecasting method is benchmarked against time-series, machine learning, and hybrid methods. The dataset of 4235 demand series with independent variables is used for analysis. The new factors related to product features, point-of-sales, promotion, weather, economy, and social media were incorporated into the model. Using an apache-spark based big data framework, the data is pre-processed and modeled. Based on the three-error metrics, which measure the bias (ME), accuracy (MAE), and variance (RMSE) of the forecasts, the proposed method is found to be the best performing method.

For short-term and long-term monthly forecasts, two auxiliary algorithms were also proposed. These algorithms use the concepts of temporal aggregations and temporal disaggregation to convert the forecasts from proposed methods to monthly and daily forecasts, respectively. The algorithm for daily forecasts enables the retailer to integrate real-time information in demand planning and thus acts as a demand sensing algorithm. Further, the relative importance (ranking) of factors influencing the sales was calculated. The point-of-sales variable and the external economic indicators are coming out to be found to most influencing factors for sales predictions.

The present study has some limitations. It used the data from food products, and in order to generalize the proposed method, the data from other product categories could be considered. Also, meta-heuristics other than a genetic algorithm could be tested to find weights for combining the forecasting methods. The current work can be extended in several possible ways. The authors have restricted the forecasting horizons to three months because longer-horizon forecasts require incorporation of human opinions and judgments in the forecasts. Therefore, this work can be extended by using judgment methods. Further, different deep learning methods and neural networks could also be used to extend the present study.

CRediT authorship contribution statement

Sushil Punia: Conceptualization, Methodology, Data curation, Analysis, Writing – Original draft, Writing – reviewing and editing. **Sonali Shankar:** Analysis, Revision, Writing – reviewing and editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] F. Badorf, K. Hoberg, The impact of daily weather on retail sales: An empirical study in brick-and-mortar stores, J. Retail. Consum. Serv. 52 (2020) http://dx.doi.org/10.1016/j.jretconser.2019.101921, Scopus.
[2] S. Thomassey, Sales forecasts in clothing industry: The key success factor of the supply chain management, Int. J. Prod. Econ. 128 (2) (2010) 470–483, http://dx.doi.org/10.1016/j.ijpe.2010.07.018.
[3] S. Punia, K. Nikolopoulos, S.P. Singh, J.K. Madaan, K. Litsiou, Deep learning with long short-term memory networks and random forests for demand forecasting in multi-channel retail, Int. J. Prod. Res. 58 (16) (2020) 4964–4979, http://dx.doi.org/10.1080/00207543.2020.1735666.
[4] M.D. Geurts, J. Patrick Kelly, Forecasting retail sales using alternative models, Int. J. Forecast. 2 (3) (1986) 261–272, http://dx.doi.org/10.1016/0169-2070(86)90046-4, Scopus.
[5] K.J. Ferreira, B.H.A. Lee, D. Simchi-Levi, Analytics for an online retailer: Demand forecasting and price optimization, Manuf. Serv. Oper. Manag. 18 (1) (2016) 69–88, http://dx.doi.org/10.1287/msom.2015.0561.
[6] T.-M. Choi, D. Li, Optimal two-stage ordering policy with Bayesian information updating, J. Oper. Res. Soc. 54 (8) (2003) 846–859, http://dx.doi.org/10.1057/palgrave.jors.2601584, Scopus.
[7] T.-M. Choi, Pre-season stocking and pricing decisions for fashion retailers with multiple information updating, Int. J. Prod. Econ. 106 (1) (2007) 146–170, http://dx.doi.org/10.1016/j.ijpe.2006.05.009, Scopus.
[8] K.-F. Au, T.-M. Choi, Y. Yu, Fashion retail forecasting by evolutionary neural networks, Int. J. Prod. Econ. 114 (2) (2008) 615–630, http://dx.doi.org/10.1016/j.ijpe.2007.06.013, Scopus.
[9] M. Kumar, N.R. Patel, Using clustering to improve sales forecasts in retail merchandising, Ann. Oper. Res. 174 (1) (2010) 33–46, http://dx.doi.org/10.1007/s10479-008-0417-z, Scopus.
[10] N. Osadchiy, V. Gaur, S. Seshadri, Sales forecasting with financial indicators and experts' input, Prod. Oper. Manage. 22 (5) (2013) 1056–1076, http://dx.doi.org/10.1111/poms.12022, Scopus.
[11] C.I. Papanagnou, O. Matthews-Amune, Coping with demand volatility in retail pharmacies with the aid of big data exploration, Comput. Oper. Res. 98 (2018) 343–354, http://dx.doi.org/10.1016/j.cor.2017.08.009, Scopus.
[12] G. Verstraete, E.-H. Aghezzaf, B. Desmet, A data-driven framework for predicting weather impact on high-volume low-margin retail products, J. Retail. Consum. Serv. 48 (2019) 169–177, http://dx.doi.org/10.1016/j.jretconser.2019.02.019, Scopus.
[13] Z.X. Guo, W.K. Wong, M. Li, A multivariate intelligent decision-making model for retail sales forecasting, Decis. Support Syst. 55 (1) (2013) 247–255, http://dx.doi.org/10.1016/j.dss.2013.01.026, Scopus.
[14] R.J. Hyndman, G. Athanasopoulos, Forecasting: Principles and Practice, OTexts, 2018.
[15] I. Alon, M. Qi, R.J. Sadowski, Forecasting aggregate retail sales: A comparison of artificial neural networks and traditional methods, J. Retail. Consum. Serv. 8 (3) (2001) 147–156.
[16] A.L.D. Loureiro, V.L. Miguéis, L.F.M. da Silva, Exploring the use of deep neural networks for sales forecasting in fashion retail, Decis. Support Syst. 114 (2018) 81–93, http://dx.doi.org/10.1016/j.dss.2018.08.010.
[17] S. Punia, S.P. Singh, J.K. Madaan, From predictive to prescriptive analytics: a data-driven multi-item newsvendor model, Decision Support Syst. 136 (2020) 113340.
[18] T. Taskaya-Temizel, M.C. Casey, A comparative study of autoregressive neural network hybrids, Neural Netw. 18 (5–6) (2005) 781–789.
[19] N.S. Arunraj, D. Ahrens, A hybrid seasonal autoregressive integrated moving average and quantile regression for daily food sales forecasting, Int. J. Prod. Econ. 170 (2015) 321–335, http://dx.doi.org/10.1016/j.ijpe.2015.09.039.
[20] P.-F. Pai, C.-S. Lin, A hybrid ARIMA and support vector machines model in stock price forecasting, Omega 33 (6) (2005) 497–505.
[21] T.-M. Choi, Y. Yu, K.-F. Au, A hybrid SARIMA wavelet transform method for sales forecasting, Decis. Support Syst. 51 (1) (2011) 130–140, http://dx.doi.org/10.1016/j.dss.2010.12.002.
[22] G.P. Zhang, Time series forecasting using a hybrid ARIMA and neural network model, Neurocomputing 50 (2003) 159–175.
[23] A. Graves, J. Schmidhuber, Framewise phoneme classification with bidirectional LSTM and other neural network architectures, Neural Netw. 18 (5) (2005) 602–610, http://dx.doi.org/10.1016/j.neunet.2005.06.042.
[24] S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural Comput. 9 (8) (1997) 1735–1780.
[25] K. Deb, A. Pratap, S. Agarwal, T. Meyarivan, A fast and elitist multiobjective genetic algorithm: NSGA-II, IEEE Trans. Evol. Comput. 6 (2) (2002) 182–197.
[26] G. Athanasopoulos, R.J. Hyndman, N. Kourentzes, F. Petropoulos, Forecasting with temporal hierarchies, European J. Oper. Res. 262 (1) (2017) 60–74, http://dx.doi.org/10.1016/j.ejor.2017.02.046.
[27] M. Al-Osh, A dynamic linear model approach for disaggregating time series data, J. Forecast. 8 (2) (1989) 85–96.

[28] T. Di Fonzo, The estimation of m disaggregate time series when contemporaneous and temporal aggregates are known, Rev Econ. Statist. (1990) 178–182.
[29] F.T. Denton, Adjustment of monthly or quarterly series to annual totals: An approach based on quadratic minimization, J. Amer. Statist. Assoc. 66 (333) (1971) 99–102.
[30] E.B. Dagum, P.A. Cholette (Eds.), The Cholette-Dagum Regression-Based Benchmarking Method–The Additive Model, in: Benchmarking, Temporal Distribution, and Reconciliation Methods for Time Series, Springer, 2006, pp. 51–84, http://dx.doi.org/10.1007/0-387-35439-5_3.
[31] G.C. Chow, A. Lin, Best Linear Unbiased Interpolation, Distribution, and Extrapolation of Time Series By Related Series, Princeton University, 1971.
[32] C. Sax, P. Steiner, Temporal disaggregation of time series, 2013.
[33] R.B. Litterman, A random walk, Markov model for the distribution of time series, J. Bus. Econom. Statist. 1 (2) (1983) 169–173.
[34] K.B. Murray, F. Di Muro, A. Finn, P. Popkowski Leszczyc, The effect of weather on consumer spending, J. Retail. Consum. Serv. 17 (6) (2010) 512–520, http://dx.doi.org/10.1016/j.jretconser.2010.08.006.
[35] M.A. Arias, C.S. Gascon, D.E. Rapach, Metro business cycles, J. Urban Econ. 94 (2016) 90–108, http://dx.doi.org/10.1016/j.jue.2016.05.005.
[36] F. Chollet, Keras, 2015.
[37] M. Boehm, M.W. Dusenberry, D. Eriksson, A.V. Evfimievski, F.M. Manshadi, N. Pansare, B. Reinwald, F.R. Reiss, P. Sen, A.C. Surve, Systemml: declarative machine learning on spark,, Proc. VLDB Endow. 9 (13) (2016) 1425–1436.
[38] R.J. Hyndman, G. Athanasopoulos, S. Razbash, D. Schmidt, Z. Zhou, Y. Khan, et al., forecast: Forecasting functions for time series and linear models, R Package Version 6 (6) (2015) 7.
[39] F. Günther, S. Fritsch, neuralnet: Training of neural networks, R J. 2 (1) (2010) 30–38.
[40] M. Kuhn, Caret package, J. Stat. Softw. 28 (5) (2008) 1–26.
[41] J.S Bergstra, R. Bardenet, Y. Bengio, B. Kégl, Algorithms for Hyper-Parameter Optimization, 2011, pp. 2546–2554.
[42] K. Nikolopoulos, S. Punia, A. Schäfers, C. Tsinopoulos, C. Vasilakis, Forecasting and planning during a pandemic: COVID-19 growth rates, supply chain disruptions, and governmental decisions, Eur. J. Oper. Res. 290 (1) (2021) 99–115.
[43] S. Shankar, S. Punia, P.V. Ilavarasan, Deep learning-based container throughput forecasting: A triple bottom line approach, Ind. Manag. Data Syst. 121 (10) (2021) 2100–2117, http://dx.doi.org/10.1108/IMDS-12-2020-0704.
[44] S. Shankar, P.V. Ilavarasan, S. Punia, S.P. Singh, Forecasting container throughput with long short-term memory networks, Ind. Manag. Data Syst. 120 (3) (2019) 425–441, http://dx.doi.org/10.1108/IMDS-07-2019-0370.
[45] Z. Hu, Y. Bao, R. Chiong, T. Xiong, Mid-term interval load forecasting using multi-output support vector regression with a memetic algorithm for feature selection, Energy 84 (2015) 419–431, http://dx.doi.org/10.1016/j.energy.2015.03.054.
[46] S. Punia, S.P. Singh, J.K. Madaan, A cross-temporal hierarchical framework and deep learning for supply chain forecasting, Comput. Ind. Eng. 149 (2020) 106796, http://dx.doi.org/10.1016/j.cie.2020.106796.
[47] K. Sanguri, S. Shankar, S. Punia, S. Patra, Hierarchical container throughput forecasting: the value of coherent forecasts in the management of ports operations, Comput. Indus. Engrg. (2022) 108651.