DAYANANDA SAGAR BUSINESS SCHOOL
Realising the Dream

Sales Demand Forecasting in Retail Sector using Transfer Learning

Dr. Rajashekhar Karjagi

Principal Data Scientist, Accenture, Bangalore

Rahul Ranjan

Analytics Advisory Analyst, Accenture, Bangalore

DOI: doi.org/10.34293/3108-1436.vimarsha.v1i2.007

Abstract

For many retail businesses, sales forecasting is critical. It is particularly important in the fashion retail service industry, where product demand is extremely variable and product-life cycles are short. This paper conducts a comprehensive literature review and selects a set of papers in the literature on fashion retail sales forecasting. Different types of analytical methodologies for fashion retail sales forecasting are reviewed for their benefits and downsides. Over the last decade, the progress of the respective forecasting systems has been revealed. Issues relating to real-world implementations of fashion retail sales forecasting models, as well as crucial future re- search areas, are highlighted, as well as how we may employ transfer learning in retail forecasting. The present work aims to propose a forecasting method for sales demand forecast in retail industry using transfer learning techniques to reduce the duration of historical data. A novel forecasting approach is proposed in this work for increased demand accuracy reduced time complexity. The dataset used for the experiment consists of sales data of retail industry and the base model is prepared using DNN approach. Further, the forecasting model is adapted to the target location for the new store. The proposed model shows high correlation coefficient values (R^2) and significantly low mean absolute percentage error (MAPE) than that of other baseline models. Thus, the proposed model helps in reducing the time complexity and increased forecasting performance.

Keywords: *Deep Neural Network, Transfer Learning, Retail Forecast, Evaluation Metrics.*

Introduction

In the fashion retailing industry, which is defined as the retailing business of fashion products including apparel, shoes, and fashion beauty products, forecasting itself can be treated as a "service" that represents the set of analytical tools which facilitate the companies to make the best decisions for predicting the future. Undoubtedly, a good forecasting service system can help to avoid understocking or over- stocking in retail inventory planning, which further relates to other critical operations of the whole supply chain such as due date management, production planning, pricing

[1], [2] and achieving high customer service level [3]. In order to achieve economic sustainability in a highly competitive

environment, a company should adopt a consumer-demand- driven "pull" operational strategy which means forecasting becomes a critically important task. Compared to other retailing service industries, it is well argued that sales forecasting is a very difficult task in fashion with the ever-changing taste of the consumers and the fashion product's life cycle is very short [4], [5].

Seasonal influences, fashion trend factors [6], and a slew of other nebulous variables all have a major "stochastic" impact on fashion product sales (e.g., weather, marketing strategy, political climate, item features, and macroeconomic trend). These factors, combined with the fact that fashion shops carry a large number of SKUs with limited historical sales data, make sales forecasting difficult and need the use of more sophisticated and versatile analytical techniques. The fashion clothing supply chain, on the other hand, is well-known to be a long one, encompassing upstream cotton plants, fiber manufacturers, apparel factories, distributors, wholesalers, and retailers. As a result, the well-known bullwhip effect [7] will have a disproportionately large impact on the fashion supply chain. Because forecasting is such an important component in determining the presence and importance of the bullwhip effect, better forecasting can help lessen the bullwhip effect, hence boosting the fashion supply chain's efficiency.

It is obvious from the preceding literature that fashion retail sales forecasting is a critical topic in practice. A number of research studies have been published in the literature throughout the last decade. Each forecasting method, however, has its own set of limitations and drawbacks. Traditional statistical methods, for example, rely heavily on the charac- teristics of time series data, which has a significant impact on forecasting accuracy. Artificial intelligence (AI) technologies can outperform standard statistical forecasting models in terms of accuracy, but they usually take significantly longer and demand a lot more computing capacity. To achieve an efficient and effective forecasting task, numerous scholars recommend combining different methodologies to develop a new "hybrid method."

Statistical methods have traditionally been used to forecast fashion sales. In fact, for sales forecasting, a variety of statis- tical methods have been utilized, including linear regression, moving average, weighted average, exponential smoothing (where a trend exists but is not linear), exponential smooth- ing with the trend, double exponential smoothing, Bayesian analysis, and so on. In sales forecasting, statistical time series analysis tools like ARIMA and SARIMA are commonly used [8]. These methods are simple and easy to implement, and the results may be computed rapidly because they use a closed-form expression for predicting. Green and Harrison [9] use a Bayesian technique to investigate forecasts for a mail- order company that sells ladies' gowns in the literature. Following that, Thomassey et al. [10] analyze the accuracy of sales forecasts for new items using item classification. They discover that in order to improve forecasting precision, a larger number of item families and relevant categorization

criteria are necessary in the respective forecasting technique. They conclude that forecasting for a product family and aggregated items are more accurate than predicting for individual items. Mostard et al. [11] have recently looked into the forecast- ing challenge using a case study of a mail-order fashion company. They suggest a ”top-flop” classification method, which they claim outperforms existing approaches. Further- more, they discover that for a restricted group of products, expert judgment approaches outperform advanced demand information methods. Another recent study [12] looks at the applicability of a Bayesian forecasting model for anticipating fashion demand. When compared to many other methods, the suggested hierarchical Bayesian methodology produces superior quantitative findings. Despite their popularity as a result of their ease of use and quickness, statistic approaches are generally recognized to have a number of flaws. To begin with, selecting the appropriate statistical methods is a diffi- cult undertaking. It necessitates ”expert” expertise. Second, they usually do not produce particularly promising results in terms of performance. Statistical models, in particular, perform poorly when compared to more advanced methods such as AI methods. Third, fashion sales are influenced by a variety of elements, including fashion trends and seasonality, and follow a highly irregular pattern [13], implying that pure statistical methods may fail to produce a good forecasting result.

Pure statistical models, as stated above have limitations in conducting fashion retail forecasting, which must be ad- dressed in order to increase forecasting accuracy. With the advancement of computer technology, AI methods develop.

In fact, AI models may quickly derive ”arbitrarily nonlinear” approximation functions from data. In the literature, popular methods such as artificial neural network (ANN) models [14] and fuzzy logic models are frequently used, and they are the first models used for fashion retail sales forecasting. ANN models, in particular, have been constructed and have shown to be effective in a variety of areas [15]–[17]. These all papers have limitations with the accuracy as well as historical data. Frank et al. [3] investigate the application of an ANN model for doing fashion retail sales forecasting in the literature on fashion sales forecasting. When compared to two other statistical methods in terms of forecasting results, the ANN model outperforms the others. Following that, the evolving neural network (ENN) model was employed in fashion sales forecasting, as it is a promising global searching strategy for feature and model selection. Au et al. [18] use ENN to find the best network structure for a forecasting system, and then they create an optimal neural network structure for fashion sales forecasting. They claim that for products with low demand uncertainty and weak seasonal trends, the performance of their proposed ENN model outperforms the traditional SARIMA model. But in these methods, if some new store or a new class at a fine level is added then the accuracy of this model also diminishes further. Despite the fact that ANN and ENN models produce high

forecasting accuracy (as evidenced by performance measures like the mean-squared error), they take a long time to accomplish the forecasting assignment. To put it another way, they take a lot of time. These models all use gradient-based learning techniques like the backpropagation neural network, which is why they have such a disadvantage (BPNN). Extreme learning machine (ELM)-based models have been developed as a solution to this challenge. In fact, ELM is known for being a lightning-fast method that can avoid issues such as halting criteria, learning rate, learning epochs, local minima, and over-tuning. ELM has been used in the literature to forecast fashion sales, and its performance has been shown to be superior to many backpropagation neural network- based approaches [19], [20]. Sun et al. [21] were the first to apply ELM for fashion sales forecasting. They look at the relationship between sales volume and the major elements that influence demand (e.g., design factors). However, ELM's most serious flaw is that it is "unstable," as it can produce different results in each run. To address this problem, [22] proposes an extended ELM approach (EELM), which computes the forecasted result by running the ELM multiple times. EELM's number of repeating times is, of course, a significant quantity that can be calculated. This method is not in use because of its unstable behavior. Hybrid forecasting methods are typically established because of their ability to combine the strengths of various models to create a new forecasting method. As a result, many of them are thought to be more efficient than pure statistical or artificial intelligence models.

The methods discussed above to forecast the sales for the same store performed well at the same store at which they are trained. The deployment of these models at target stores may degrade the performance as the models are not location class independent. Therefore, in the present work, a forecasting method is proposed which is developed at one store location (source location s) for a class d and can be adapted to deploy at a target store (target location s') for a class d and d'.

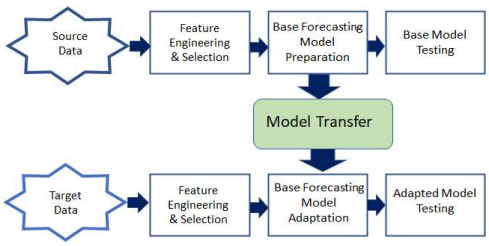

Figure 1 The Steps Performed to Obtain the Proposed Transfer Model

First the base forecasting model M using machine learning deep learning algorithms is developed at s using large training dataset of three years for a class d. After that, the M is adapted at s' using a lesser amount of training dataset for a class d and d'. Hence, the requirement of historical data can be significantly reduced. To achieve the objective, the transfer- learning based method is implemented in the present work. To the best of the authors' knowledge, transfer-learning based forecasting of a new store has not been explored yet. This paper contributes the following facet.

- The contribution of the proposed work is

that the de- veloped model is adaptable at (s, d') with two weeks of training dataset of the (s, d').

- The contribution of the proposed work is that the de- veloped model is adaptable at (s', d) with two weeks of training dataset of the (s', d).
- The other contribution of the proposed work is that the developed model is adaptable at (s', d') with two weeks of training dataset of the (s', d').

The paper's remainder is as follows: Section II describes data preprocessing and features extraction, forecasting models, transfer learning, and performance evaluation criteria of the forecasting model. Section III describes the case study. Section IV provides the conclusion and future work.

Methodology

This work presents a methodology for the forecasting of sales. The overall method is developed in three phases. In the first phase, the M for a d is developed using different machine learning techniques at s. Finally, the M is adapted for d and d' at s' using the transfer learning methods with a shorter historical data. The steps carried out for developing the calibration model are shown in Fig. 1, and described as follows:

Data Pre-Processing and Feature Preparation

The dataset contains features that have different ranges. We use standardization because different machine learning and deep learning-based algorithms are affected by the scale of the input. Hence, we perform feature standardization to reduce the training sensitivity to the inputs' range and make the fea- tures well-conditioned for optimization. The standardization is performed as follows [23]:

$$N_f = \frac{z - \bar{z}}{\sigma_z} \quad (1)$$

Where Nf and z denote the standardized input features and actual input features, the z and σz represent the mean and standard deviation of the actual input features.n this paper, effective time-lags [24], and time based features are used for developing the model. The effective values of lags for sales is found using the auto-correlation coefficients [25]. The other features are also selected using cross-correlation coefficients. The Pearson's correlation of the features w.r.t sales at time (t) is computed as follows:

$$r = \frac{\sum_{t=1}^{K}(x_t - x)(y_t - y)}{\sqrt{\sum_{t=1}^{K}(x_t - x)^2 \sum_{t=1}^{K}(y_t - y)^2}} \quad (2)$$

Where x_t and y_t denote the value of t^{th} feature and sales at time (t), respectively. The x, and y represent the mean value of the features and sales at time (t), respectively. During the feature selection process, the feature with the highest correlation coefficient is considered and is fed as input of the calibration model and the R^2 value is computed. Further, more features are added, one by one, based on their correlation coefficient, until the improvement in R2 becomes negligible. In this way, we have selected 21 features in the present work.

Techniques used to Develope the Forecasting Model

The various techniques have been adapted in the development of sales demand forecasting model, which are as follows:

- **Deep Neural Network (DNN)**

  Artificial Neural Network is a widely used tool in various time series regression and forcasting problems [24], [26]. It approximates the non-linear relationship between inputs and output for developing the calibration model. It has an input layer, output layer, and a hidden layer between the input and output layers [27]. The DNN consists of more than one hidden layers. The DNN has been used to develop the base forecasting model by training and testing for a particular class for same store.

- **Transfer Learning**

  In the present work, a forecasting model *M* is developed at (*s, d*) and adapted at (*s, d′*), (*s′, d*), and (*s′, d′*) using a shorter duration of training data.

This kind of approach is suitable at those stores where the inventory (sales) data are available for a short duration. As the direct deployment of *M* at (*s, d′*), (*s′, d*), and (*s′, d′*) may not show good performance due to spatial shift [28]. transfer learning-based techniques are required to overcome these issues. The transfer learning refers to adapt the knowledge from one store location to work to the new store location [28], [29]. The steps of the transfer learning method, *M* is adapted at (*s′, d′*) in the following two steps:

Algorithm 1: Transfer Learning method Input: A labelled source domain dataset DS = {XS, YS}, a small amount of labelled target domain dataset DT1 = {XT1 , YT1 }, and large amount of unlabelled target domain dataset DT2 = {XT2 }. Output: Labels YT2 of the unlabelled data XT2 in the target domain 1. Standardize the source and target domain features using (1). 2. Learn a regression model f: XS → Y for S using (3) and (4). 3. Save the models learned at source domain as base model M. 4. Freeze the layer of M and add a new layer at the top of the M. 5. Use the DT1 to train the newly added top layer with a higher learning rate using (5). 6. Fine-tune the entire model with a lower learning rate by using the DT1 using (6). 7. Utilize the new learned regression models for predicting the labels of DT2 as Y for T2 = f new (XT2). 8. Finally, evaluate the performance of the proposed model using the metrices such as, R2, MAE, MAP E, and SMAP E as shown in (7)-(11). 1. In step one, a new layer is added on the top of the layers of M learned at (s, d). The layers trained at (s, d) are frozen and only the newly added layer is trained with a much smaller dataset at (s ′, d′). 2. In this step, the entire model is finetuned with a much smaller dataset at (s ′, d′). The proposed algorithm for the adaptation of M is summarised in Algorithm 1. The dataset of source (DS) and target (DT) domain are splitted in training and testing data. The size of training dataset at DT is much smaller as compared to DS. We consider M with parameters θ at (s, d), that maps sales observations at previous days (x) to forecasted sales at time t (yˆ). We use a DNN model to implement M. The parameters of M at (s, d) are initialized randomly. The architecture of M consists of one input layer, seven hidden

layers, and one output layer. The parameter θ of M is updated using the three years of data at (s, d) using gradient descent algorithm over DS as follows, θ ′ ← θ − α∇LDS (θ) (3) Here, α ∈ R is the learning rate. The θ and θ ′ represent the randomly initialized weights and learned weights of M. The ∇ denotes the gradient. The used loss function is the mean absolute error, which is defined as LDS (θ) = X (x,y)∈DS |fθ(x) − y| (4) Now, to adapt M at (s ′ , d′), we discard the output layer of M and add a layer on the top of M. The weights of the layers trained at (s, d) are freezed and the weights (ϕ) of the added layer are trained using two weeks of data at (s ′ , d′): ϕ ′ ← ϕ − β∇ϕLDT (θ ′ , ϕ) (5)

Table 1 Average performance of baseline and proposed models for different tasks.

Task	Baseline Method			Proposed Method		
	R^2	MAPE (%)	SMAPE (%)	R^2	MAPE (%)	SMAPE (%)
Task-I	0.72	10.72	12.6	0.72	10.72	12.6
Task-II	-0.12	27.5	32.46	0.76	11.89	12.75
Task-III	-0.49	34.89	39.87	0.67	13.45	14.78

Here, β ∈ R is the learning rate which is kept higher than α ∈ R. The ϕ and ϕ ′ are the randomly initialize weights and the learned weights for the new layer added on top of M. Finally, the parameters (θ ′ , ϕ ′) of the entire model are finetuned using two weeks of data at (s ′ , d′) as follows, Θ ← (θ ′ , ϕ′) − γ∇LDT (θ ′ , ϕ′) (6) Here, γ ∈ R is the learning rate which is kept lower than α ∈ R. The Θ denotes the learned weight after the adaptation of the calibration model at (s ′, d′). The adapted model is tested using the rest of the data at (s ′, d′) over DT . C. Forecasting model's performance evaluation criteria The coefficient of determination R2 , the mean absolute error (MAE), the mean absolute percentage error (MAPE), and symmetric mean absolute percentage error (SMAPE) are computed to evaluate the performance of the forecasting model. These are expressed as follows: z = 1 K X K t=1 zt (7) R 2 = 1 − P K t=1 |zt − zˆt(x)| 2 P K t=1 |zt − z| 2 (8) MAE = 1 K X K t=1 |zt − zˆt(x)| (9) MAP E = 1 K X K t=1 |zt − zˆt(x)| |zt| × 100% (10) SMAP E = 1 K X K t=1 2|zt − zˆt(x)| |zt| + |zˆt(x)| × 100% (11) where, K denotes total number of observations. The zt and zˆt(x) are the sales actual and forecasted values for observation t, respectively.

Experimental Results & Discussion

We have experimented with three kinds of tasks. In the first task, we have trained M using the data at (s, d), and test its performance for the same (s, d). In the second task, the performance of M at (s, d ′) with and without transfer learning are compared. In the third task, the performance of M at (s ′, d ′) with and without transfer learning are compared. The dataset has used to simulate this work from kaggle. The training and testing period for task-I is three years and one month respectively. The training and testing period for task-II and task-III are two weeks and one month respectively. The evaluation metrics for the three different tasks are summarised in Tables I. The overall performance of the proposed algorithms is better as compare to the other baseline models. We have used scikit-learn and Tensor Flow libraries of Python programming language.

Conclusion & Future work

In this work, a novel transfer learning based forecasting method for sales demand forecasting in retail sector is proposed. This method makes use of deep learning for high performance and transfer learning for reducing the historical data samples. We compare the performance of the proposed transfer learning based forecasting method w.r.t the other baseline methods and find improvements with the proposed method. In future, we will also study the transfer learning based approach using meta-learning for sales demand forecasting in retail sector.

References

1. C.-H. Chiu, T.-M. Choi, and D. Li, "Price wall or war: The pricing strategies for retailers," IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans, vol. 39, no. 2, pp. 331–343, 2009.
2. C.-H. Chiu and T.-M. Choi, "Optimal pricing and stocking decisions for newsvendor problem with value-at-risk consideration," IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans, vol. 40, no. 5, pp. 1116–1119, 2010.
3. C. Frank, A. Garg, L. Sztandera, and A. Raheja, "Forecasting women's apparel sales using mathematical modeling," International Journal of Clothing Science and Technology, 2003.
4. T.-M. Choi, C.-L. Hui, and Y. Yu, Intelligent fashion forecasting systems: models and applications. Springer, 2013.
5. T.-M. Choi and S. Sethi, "Innovative quick response programs: a review," International Journal of Production Economics, vol. 127, no. 1, pp. 1– 12, 2010.
6. S. Thomassey, "Sales forecasts in clothing industry: The key success factor of the supply chain management," International Journal of Production Economics, vol. 128, no. 2, pp. 470–483, 2010.
7. H. L. Lee, V. Padmanabhan, and S. Whang, "Information distortion in a supply chain: The bullwhip effect," Management science, vol. 43, no. 4, pp. 546–558, 1997.
8. G. E. Box, G. M. Jenkins, G. C. Reinsel, and G. M. Ljung, Time series analysis: forecasting and control. John Wiley & Sons, 2015.
9. M. Green and P. Harrison, "Fashion forecasting for a mail order company using a bayesian approach," Journal of the Operational Research Society, vol. 24, no. 2, pp. 193–205, 1973.
10. T. Sebastien, H. Michel, and C. Jean-Marie, "Mean-term textile sales ´ forecasting using families and items classification," Studies in Informatics and Control, vol. 12, no. 1, pp. 41–52, 2003.
11. J. Mostard, R. Teunter, and R. De Koster, "Forecasting demand for single-period products: A case study in the apparel industry," European Journal of Operational Research, vol. 211, no. 1, pp. 139–147, 2011.
12. P. M. Yelland and X. Dong, "Forecasting

demand for fashion goods: a hierarchical bayesian approach," in Intelligent fashion forecasting systems: Models and applications. Springer, 2014, pp. 71–94.

13. T.-M. Choi, C.-L. Hui, and Y. Yu, "Intelligent time series fast forecasting for fashion sales: A research agenda," in 2011 international conference on machine learning and cybernetics, vol. 3. IEEE, 2011, pp. 1010– 1014.
14. L. M. Sztandera, C. Frank, and B. Vemulapali, "Predicting women's apparel sales by soft computing," in International Conference on Artificial Intelligence and Soft Computing. Springer, 2004, pp. 1193–1198.
15. D. Olson and C. Mossman, "Neural network forecasts of canadian stock returns using accounting ratios," International Journal of Forecasting, vol. 19, no. 3, pp. 453–465, 2003.
16. H. Yoo and R. L. Pimmel, "Short term load forecasting using a self-supervised adaptive neural network," IEEE transactions on Power Systems, vol. 14, no. 2, pp. 779–784, 1999.
17. L. Zampighi, C. Kavanau, and G. Zampighi, "The kohonen selforganizing map: a tool for the clustering and alignment of single particles imaged using random conical tilt," Journal of Structural Biology, vol. 146, no. 3, pp. 368–380, 2004.
18. K.-F. Au, T.-M. Choi, and Y. Yu, "Fashion retail forecasting by evolutionary neural networks," International journal of production economics, vol. 114, no. 2, pp. 615–630, 2008.
19. G.-B. Huang, Q.-Y. Zhu, and C.-K. Siew, "Extreme learning machine: theory and applications," Neurocomputing, vol. 70, no. 1-3, pp. 489–501, 2006.
20. Q.-Y. Zhu, A. K. Qin, P. N. Suganthan, and G.-B. Huang, "Evolutionary extreme learning machine," Pattern recognition, vol. 38, no. 10, pp. 1759–1763, 2005.
21. Z.-L. Sun, T.-M. Choi, K.-F. Au, and Y. Yu, "Sales forecasting using extreme learning machine with applications in fashion retailing," Decision Support Systems, vol. 46, no. 1, pp. 411–419, 2008.
22. Y. Yu, T.-M. Choi, and C.-L. Hui, "An intelligent quick prediction algorithm with applications in industrial control and loading problems," IEEE Transactions on Automation Science and Engineering, vol. 9, no. 2, pp. 276–287, 2011.
23. Zheng and A. Casari, Feature Engineering for Machine Learning: Principles and Techniques for Data Scientists. O'Reilly Media, Inc., 2018.
24. S. K. Jha, C. L. Dewangan, and N. K. Verma, "Multi-step load demand forecasting using neural network," in 2019 20th International Conference on Intelligent System Application to Power Systems (ISAP), 2019, pp. 1–6.
25. N. M. Pindoriya, S. N. Singh, and S. K. Singh, "One-step-ahead hourly load forecasting using artificial neural network," in 2009 International Conference on Power Systems, 2009, pp. 1–6.
26. C. L. Dewangan, S. N. Singh, and S. Chakrabarti, "Solar irradiance forecasting using wavelet neural network," in 2017

IEEE PES AsiaPacific Power and Energy Engineering Conference (APPEEC), 2017, pp. 1–6.

27. K. Alexandridis and A. D. Zapranis, Wavelet neural networks: with applications in financial engineering, chaos, and classification. John Wiley & Sons, 2014.
28. X. Xu, X. Zhou, R. Venkatesan, G. Swaminathan, and O. Majumder, "dSNE: Domain adaptation using stochastic neighborhood embedding," in 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 2492–2501.
29. Q. Sun, Y. Liu, T. Chua, and B. Schiele, "Meta-transfer learning for few-shot learning," CoRR, vol. abs/1812.02391, 2018. [Online]. Available: http://arxiv.org/abs/1812.02391