Sales Forecasting Based on Transformer-LSTM Model

Yi Liu *

Department of Computer Science, Jilin University, Changchun, China

* Corresponding Author Email: liuyi0919@mails.jlu.edu.cn

Abstract. As a matter of fact, sales forecasting becomes a hot topic in recent years. The LSTM model is a classic model in the field of sales forecasting. Recently, Transformer has gained a lot of popularity, and many researchers are applying it to time series forecasting. To be specific, this study proposes a novel Transformer-LSTM framework for the one-step sales prediction task. This study retained the classical Transformer structure, using it as the model's encoder, and then combined attention mechanisms with the LSTM model in the decoder section. This hybrid approach allows the model to leverage the advantages of Transformer in handling long sequences while benefiting from LSTM's unique strengths in processing time series data. Ultimately, the model's performance was evaluated using prediction curve comparisons and RMSE as validation metrics, achieving significant results. In conclusion, these results present a novel model for sales forecasting and time series analysis, contributing to improved prediction accuracy in these domains.

Keywords: Transformer-LSTM Model, sales forecasting, prediction curve comparisons, time series analysis.

1. Introduction

Sales forecasting, often modeled as a traditional time series forecasting problem, is widely used in industries such as manufacturing, retail, marketing, wholesale, and supply chain. Sales forecasting has an important impact on managers' inventory management, supply and demand analysis, business strategy and economic decision-making. Sales forecasting has a profound history. Armstrong showed that purchase intentions can produce more accurate forecasts than a simple extrapolation of historical sales trends [1], according to research that compared the relative accuracy of four approaches for predicting sales from intents. Liu did a thorough assessment of the literature and chose a number of studies on fashion retail sales forecasting [2]. It was presented how each forecasting technique has changed over the past 15 years. Kerkkanen explains how to evaluate the effects of incorrect sales forecasts through a case study [3]. Tsoumakas did a study on machine learning methods for predicting food purchases. He talked about the crucial choices a data analyst making predictions about food sales would make [4]. Falah design a computerized sales forecast management application that would generate sales targets, total expenses, gross profit, and gross margin for budgeted and projected sales, and automatically generate expense reports, budgeting reports, and reports summarizing final reports [5].

Time series forecasting (TSF) problems, such as the forecasting of product sales, can be resolved in one of two ways: linearly or nonlinearly. Traditional time series forecasting methods fit historical time trends by establishing appropriate mathematical models, relying on relatively simple data. For example, Ramos conducted research on market demand in the retail industry using both autoregressive moving average models and state-space models [6]. Sindia used the weighted moving average method to develop a system that can forecast sales in the future to increase a business's income [7]. When predicting the wholesale price of food, Menculini Considered Prophet, a scalable forecasting tool developed by Facebook [8]. Yu pointed out that the traditional simple forecasting models have developed theory and are easy to use, but their prediction of long-term time points tends to be mean value [9]. With the development of computer science, Artificial intelligence neural network methods leveraged the self-learning capabilities of individual neurons to capture nonlinear relationships between input and output data in samples. They possess a certain degree of nonlinear mapping capability in forecasting but have lower generality. For instance, Yuan combined the model of prophet and GRU [10]. In this composite model, linear and nonlinear features were captured using the Prophet model and the GRU model.

Among the various methods, LSTM is particularly well-suited for time series data. It can capture temporal dependencies and long-range relationships in sequences. Additionally, the Transformer model marks a substantial advancement in the domains of machine learning and natural language processing. Transformers have many benefits, but one that makes them particularly appealing for time series modeling is their ability to capture long-term dependencies and interactions. This has led to fascinating developments in a variety of time series applications [11].

According to the daily sales data of a French bakery, a composite LSTM and Transformer model is proposed in this work. Originally, the Transformer model was designed for machine translation tasks as a Seq2Seq (sequence-to-sequence) model. Time series forecasting can also be viewed as a type of Seq2Seq task. However, because the Transformer model architecture includes many neural network layers specifically designed for the "<Source> to <Target>" sequences in translation tasks, such as the Masked Multi-Head Attention layer, adjustments and modifications are required for tasks related to time series data. The paper primarily focuses on modifying the Decoder structure in the Transformer model. The LSTM model introduced into the Transfomer decoder, which facilitates the learning of long-range dependencies in the data sequence by the model, improving forecasting effectiveness. The remainder of this essay is structured as follows. Sec. 2 details the experiment setup and details the forecasting method. Sec. 3 provides results and describe the performance of the model. Sec. 4 provides limitations and proposes future research directions. Sec. 5 provides the conclusion of the research.

2. Data and method

This essay chooses from the Kaggle platform's daily sales information from a bakery. The dataset, which comprises about 136,000 transactions and 234005 entries, provides the daily transaction details of clients from 2021-01-01 to 2022-09-30. To provide a more direct representation of the data characteristics related to the product's sales volume, the aforementioned data was categorized based on daily sales statistics (seen from Fig. 1), revealing significant fluctuations in the sales data.

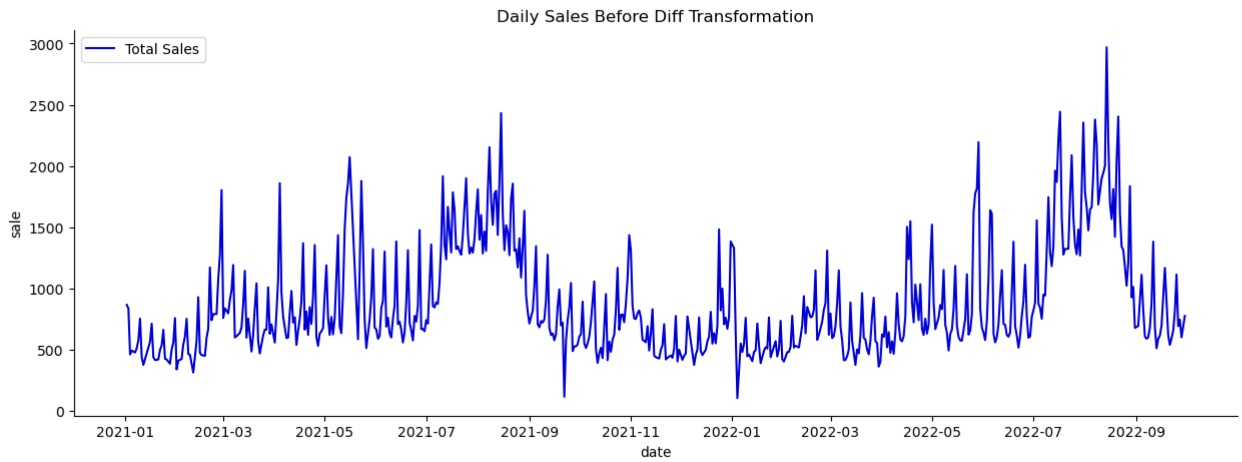

Figure 1. Time series trend of daily sales volume from 2021 to 2022 (Photo/Picture credit: Original).

Time delay embeddings" (TDEs) is a commonly used technique in time series analysis to transform time series data into a feature representation that captures temporal patterns and dynamic information more effectively. Each scalar value of a scalar time series data set xt is embedded into a time-delay space to create the time delay embedding (TDE):

$$\boldsymbol{TDE_{d,\tau}(x_t) = (x_t, x_{t-\tau}, \dots, x_{t-(d-1)\tau})} \tag{1}$$

Oertzen find that time delay embedding compared to conventional independent rows of panel data, increases the precision of parameter estimates and subsequently statistical power [12]. The issue of gradient vanishing in RNNs can be mitigated by introducing Long Short-Term Memory (LSTM) networks, which incorporate a specialized gating mechanism and internal memory management to regulate the flow of information. In contrast to the original RNN's hidden layer, which has parameters that are highly sensitive to short-term input, LSTM introduces a dedicated long-term state that similar to how the human brain's long- and short-term memory systems operate together. The attention mechanism is a resource allocation mechanism that enhances the precision of a model. Its fundamental idea is to focus solely on relevant information. Information's significance is specifically influenced by its probability distribution. Fig. 2 demonstrates the attention structure.

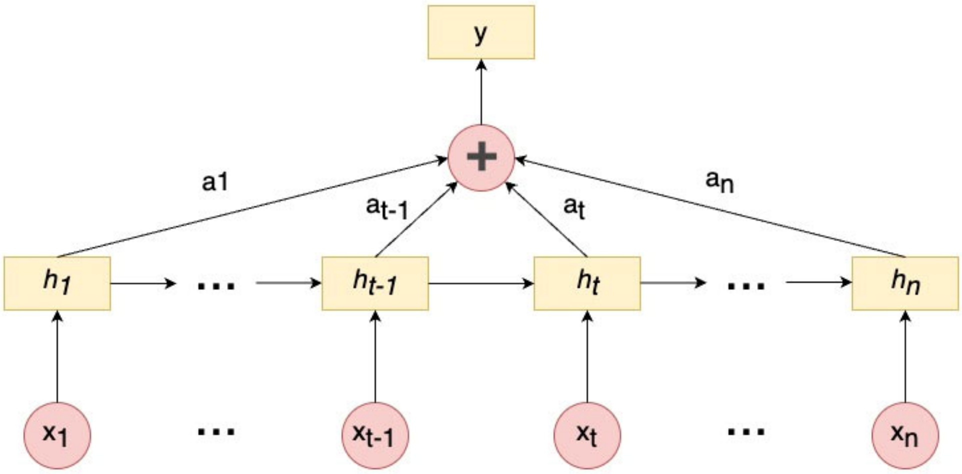

Figure 2. Attention Mechanism (Photo/Picture credit: Original).

The traditional Transformer, introduced in the seminal paper "Attention Is All You Need," is a groundbreaking deep learning architecture that has transformed the field of natural language processing and sequence-to-sequence tasks. At its core lies the self-attention mechanism, which permits long-range dependencies to be captured by attending to different elements within a sequence. The Transformer's architecture is characterized by stacked encoder and decoder layers, with positional encodings, residual connections, and layer normalization, facilitating the learning of hierarchical representations. Multi-head attention, along with scaled dot-product attention, enhances its ability to capture context and relationships in data. This study uses an encoder architecture that resembles the original Transformer architecture. The encoder consists of three main components: an input layer, a positional encoding layer, and four identical encoder layers. The input layer utilizes a fully connected network to map the input time series data into a model vector with a dimension of d. To capture the time series data's sequential information, one employs positional encoding, combining it with the input vector through element-wise addition. The resulting vector is then passed through the four identical encoder layers. Each layer comprises two sub-layers: a self-attention sub-layer and a fully connected feed-forward sub-layer. The decoder also consists of an input layer and four identical decoder layers. The final data point from the encoder's input is the first data point in the decoder's input. The decoder includes an additional third sub-layer, which applies a self-attention mechanism over the encoder's output (seen from Fig. 3). This study has employed a decoder design consisting of a LSTM model, typically used for sequence-to-sequence tasks, particularly for generating predictions in sales data. The decoder comprises an input layer, a stack of four identical decoder layers, and an output layer. The input layer maps the decoder input to a d-dimensional vector. The resulting vector is then passed through the decoder layers. Each decoder layer consists of three sub-layers: a dot-product attention sub-layer, an LSTM sub-layer and a fully-connected feed-forward sub-layer. The output layer, which ensures that the prediction of the sales data point solely relies on earlier data points, translates the output from the final decoder layer to the target time sequence.

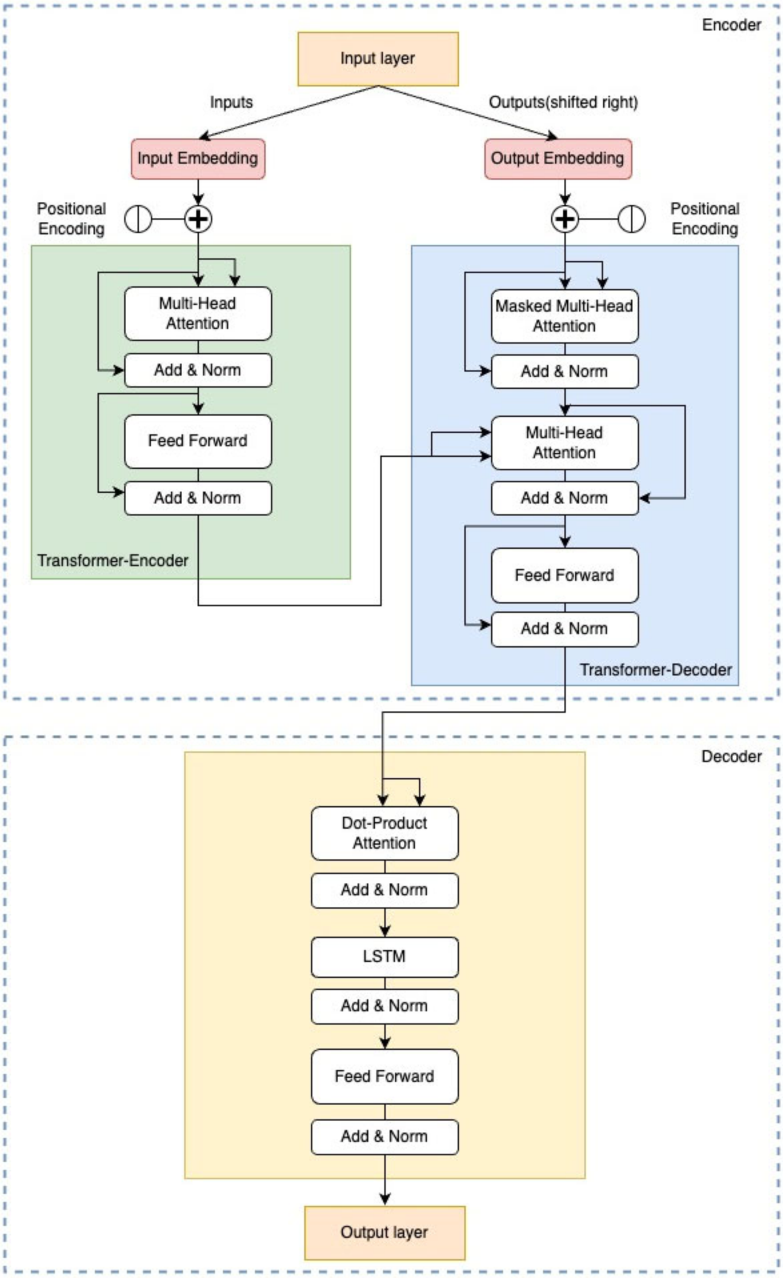

Figure 3. Transformer-LSTM Model (Photo/Picture credit: Original).

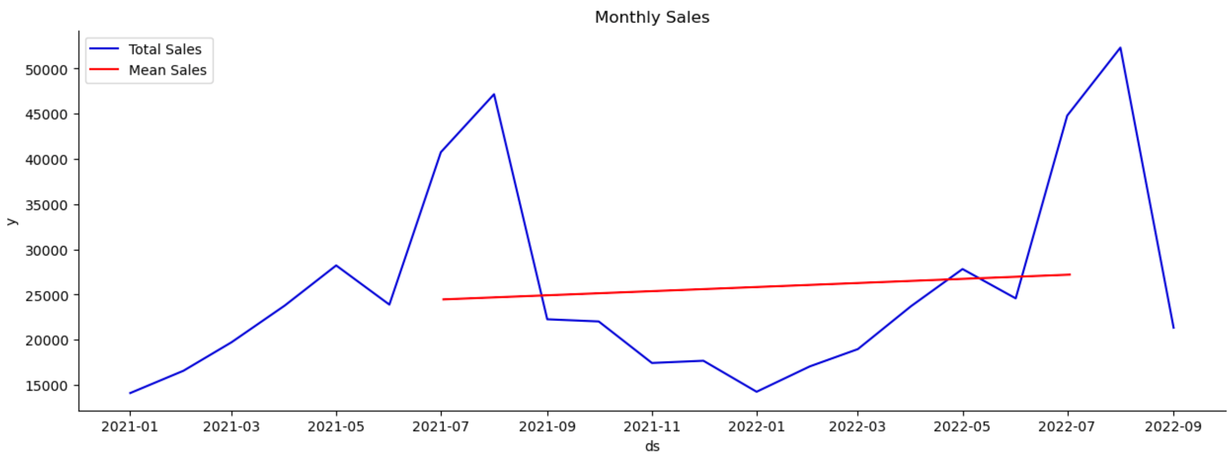

Figure 4. Monthly sales before difference transformation (Photo/Picture credit: Original).

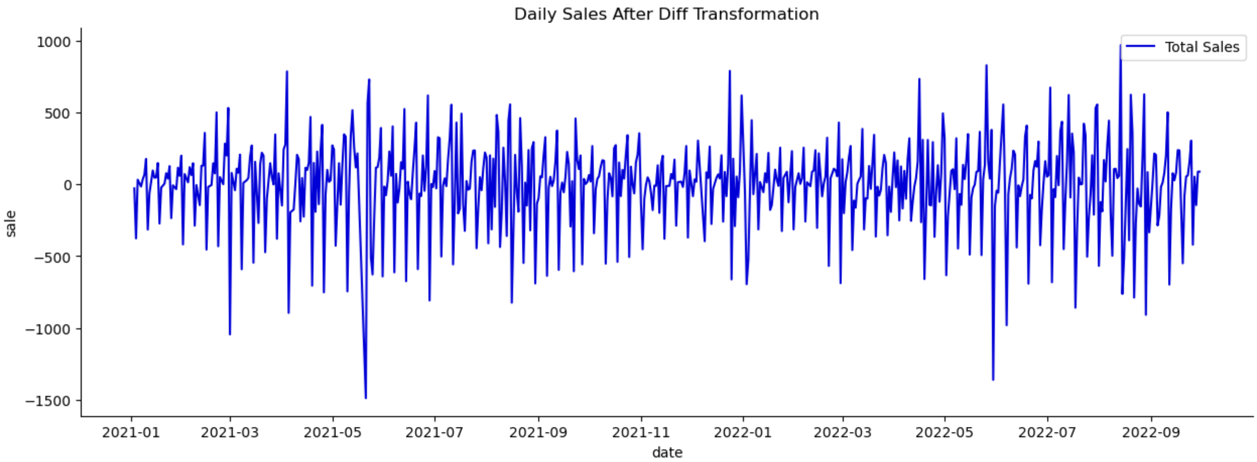

Figure 5. Daily sales after difference transformation (Photo/Picture credit: Original).

3. Results and discussion

In order to get a more holistic view of the data, the daily sales data was aggregated into monthly data. Fig. 4 shows a curve with less fluctuation, but a more intuitive monthly sales trend. The data over time showed some linear regularity. Additionally, there was randomness, which largely showed up as a general fall in sales volume and a decreasing tendency of partial seasonality. In time series forecasting, differencing data preprocessing is a common technique used to make time series data more stable, making it easier to model and predict. By applying first-order differencing, one obtains smoother daily sales data. First-order differencing calculates the difference between each time point and the preceding one, effectively removing the linear trend and randomness, making the data more suitable for time series analysis and forecasting. In order to enhance prediction accuracy and mitigate the influence of non-standard data, it is necessary to normalize all data points, meaning they are transformed into values falling within a specified range. This study employed the following normalization formula:

$$x_{normalization} = \frac{x-Min}{Max-Min} \tag{2}$$

To produce a supervised dataset, this study used a fixed-length sliding time window approach and time delay embeddings technique. This study restructures the data into a 5-column matrix with the first four columns for 4-dimension input encoder data X and the last column for the label y. It means that the length of the sliding window is 4, which is used for one-step forecasting. After that, the train and test split ratio is 8:2. It means that this study used the sales data from January 1, 2021, to July 1, 2022, as the training set, comprising a total of 480 days. This study uses the data from July 2, 2022, to September 30, 2022, as the test set, consisting of a total of 120 days (seen from Fig. 5). Fig. 6 shows the predictions from linear regression models and the LSTM. The blue curve represents the actual sales data, showing the sales volume of the bakery from 2022-07-02 to 2022-09-30. The yellow curve represents the forecasted data, demonstrating the predictive capability of the model. The composite model achieved higher prediction accuracy than any individual component model. This integrated model worked more effectively because it addressed the issues of the Linear Regression model merely shifting the original input data and the LSTM model struggling to adapt to frequent data fluctuations. Figures in Figure 6 demonstrate that the Transformer-LSTM model effectively captured the underlying patterns in sales prices, with only minor deviations from the actual values.

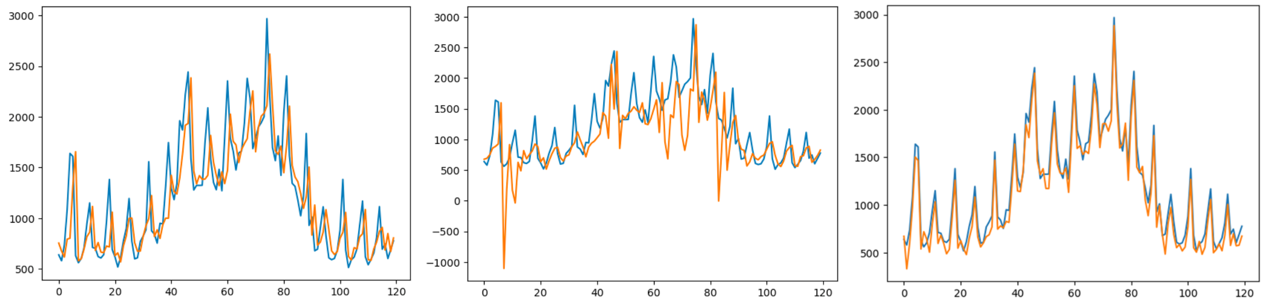

Figure 6. Performance of difference models: Linear, LSTM and transformer-LSTM (right) (Photo/Picture credit: Original).

RMSE is the root mean square error, which is calculated as:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n} |x - x^*|^2} \tag{3}$$

This study used it to indicate forecast accuracy, smaller values indicate more accurate predictions. RMSE of different model are specified in Table 1. It shows that Linear Regression and LSTM model have the similar performance in this data set, while Transformer-LSTM has an overwhelming advantage in predicting sales volume.

Table 1. RMSE of different model

Model	RMSE
Linear Regression	0.283
LSTM	0.273
Transformer-LSTM	0.091

For the Transformer-LSTM model, one sets n_heads = 4, learning_rate = 0.00001, feature_size=4, dropout=0.1, num_layer=4 and set the number of LSTM layer=1. Besides, this study applies the Adam optimizer to accelerate model training, enhance stability, and adaptively adjust learning rates in sales forecasting. This study trained the model to predict 1 future sales volume from 4 trailing daily datapoints. That is, given the encoder input (x1, x2, ..., x4), the decoder aims to output x5. The composite model has advantages in multiscale modeling: a Transformer-LSTM hybrid model can simultaneously capture information at multiple time scales. Transformers can handle longer time ranges, while LSTM can handle shorter-term information. This enables the model to consider the influence of different time scales simultaneously, enhancing prediction accuracy.

4. limitations and prospects

This study primarily investigates a prediction model based on LSTM and Transformer, providing new methods for sales prediction. It has achieved significant improvements in prediction accuracy and contributes to the field of time series forecasting by offering novel approaches. However, there are still some shortcomings in this study that need to be addressed, as detailed below.

● Feature Engineering Optimization. In this study, the model only utilizes the sales volume sequence as a feature indicator. In future research, one should consider incorporating additional relevant indicators with reference value, such as company performance metrics, unit prices of sold products, special promotional activities, or even some less quantifiable indicators like macroeconomic policies and shifts in consumer preferences. Furthermore, one can enhance the feature indicators through various processes, including quantifying the indicators, dimensionality reduction, and feature selection. Optimizing feature engineering may lead to improved model performance and better results.

- Due to limitations in time and space, this study primarily focused on single-step forecasting using a sliding window of size 4 within the Transformer-LSTM model. In terms of model architecture, only the combination of Transformer and LSTM was explored. However, future research opportunities include varying the sliding window size to investigate multi-feature input forecasting performance, exploring multi-step forecasting for extended prediction horizons, and experimenting with the combination of the Transformer-LSTM model with other architectures like CNN or GRU, potentially enhancing its applicability in sales forecasting.
- The dataset used was relatively small, which may raise concerns about the model's limited generalization ability. In the future, one can work with larger datasets to enhance the model's predictive capabilities.

5. Conclusion

In summary, this study proposes a novel Transformer-LSTM framework for the one-step sales prediction task. This paper retained the classical Transformer structure, using it as the model's encoder, and then combined attention mechanisms with the LSTM model in the decoder section. This hybrid approach allows the model to leverage the advantages of Transformer in handling long sequences while benefiting from LSTM's unique strengths in processing time series data. Ultimately, the model's performance was evaluated using prediction curve comparisons and RMSE as validation metrics, achieving significant results. However, the study still has limitations, including insufficient feature extraction dimensions and a relatively small dataset. These shortcomings are planned to be addressed in future research. In conclusion, this paper presents a novel model for sales forecasting and time series analysis, contributing to improved prediction accuracy in these domains.

References

[1] J. S. Armstrong, V. Morwitz and V. Kumar, Sales Forecasts for Existing Consumer Products and Services: Do Purchase Intentions Contribute to Accuracy (Rochester, New York, 2008), pp. 1 - 221.
[2] N. Liu, S. Ren, T. M. Choi, C. L. Hui, and S. F. Ng, Mathematical Problems in Engineering, 2013, 738675 (2013).
[3] A. Kerkkänen, J. Korpela, and J. Huiskonen, Inter. J. Prod. Eco., 118 (1), 43 – 48 (2009).
[4] G. Tsoumakas, Artif Intell Rev, 52 (1), 441 – 447 (2019).
[5] Y. M. Falah. Ahmed, R. A. Kanakarathanm, M. D. Saleh, and K. L. T. Aik, ICSET 9,162 – 165 (2019).
[6] P. Ramos, N. Santos, and R. Rebelo, Robotics and Computer-Integrated Manufacturing, 34, 151 – 163 (2015).
[7] R. Sindia, I. K. Siregar, and M. Ihsan, JURTEKSI, 8 (3), 285 – 294 (2022).
[8] L. Menculini, Forecasting, 3 (3), 644 – 662 (2021).
[9] Y. Han, ISCTT 5, 336 – 339 (2020).
[10] Y. Li, Y. Yang, K. Zhu, and J. Zhang, IEEE Trans. Ind. Inf., 17 (12), 8335 – 8344 (2021).
[11] Q. Wen "Transformers in Time Series: A Survey." arXiv preprint arXiv. 2202. 07125 (2023).
[12] T. von Oertzen and S. M. Boker, Psychometrika, 75 (1), 158 – 175 (2010).