Proceedings of the 3rd International Conference on Financial Technology and Business Analysis
DOI: 10.54254/2754-1169/128/2024.18262

Implementation of Bigdata Techniques in Sales Forecasting: Evidence from Retailing and E-commerce

Wenhan Zhang[1,a,*]

[1]*Department of Mathematics and Department of Economics, University of Wisconsin–Madison, 702 Langdon St, Madison, the United States*
a. wzhang599@wisc.edu
**corresponding author*

Abstract: Sales forecasting is a key part of business operations, providing essential insights for inventory management, financial planning, and strategic decision-making. With the advent of big data, traditional forecasting methods have evolved, leveraging advanced machine learning and different deep learning models to increase accuracy and efficiency. This paper reviews recent research on the use of big data techniques in sales forecasting, focusing on diverse industries such as retail, e-commerce, and automotive. Through an analysis of case studies and research findings, this study emphasizes the effectiveness of machine learning models like XGBoost, Random Forests, and RNNs in improving forecast accuracy in the context of big data. The review also addresses the challenges and limitations of current methodologies, offering insights into future research directions. The significance of this research lies in its comprehensive overview of the state-of-the-art in sales forecasting, providing a valuable resource for researchers and practitioners aiming to optimize business strategies through big data.

Keywords: Sales forecasting, Big Data, machine learning, predictive analytics.

1. Introduction

Sales forecasting has been a cornerstone of corporate strategy for decades, playing a critical role in guiding decisions related to inventory management, financial planning, and marketing. The practice of forecasting sales dates back to the early 20th century when businesses began systematically analyzing historical sales data to predict future demand. Early methods primarily depend on fundamental statistical techniques such as Moving Averages (MA) and Exponential Smoothing, which provided a basic level of accuracy but is often it is difficult to get very accurate data because they lag behind the real data [1, 2]. As businesses grew in scale and complexity, the need for more sophisticated forecasting methods became apparent. In response, the Autoregressive Integrated Moving Average (ARIMA) model and regression analysis emerged as more advanced statistical tools for sales forecasting [3]. These models introduced the ability to incorporate multiple variables and time-series data, offering improved accuracy over simpler methods. However, they still had limitations, especially when sales data have lots of outliers, missing data or other influencing factors [4]. The advent of the digital age marked a significant turning point in sales forecasting. The rise of computers and the proliferation of data collection technologies enabled businesses to gather and analyze data at an unprecedented scale. This era saw the introduction of more complex statistical

DOI: 10.54254/2754-1169/128/2024.18262

models and the early application of machine learning techniques, which began to outperform traditional methods by identifying hidden patterns and trends which cannot be predicted by experts [5]. Despite these advancements, challenges (e.g., data quality, model interpretability), and the computational demands of large-scale data analysis persisted, limiting the effectiveness of these methods.

In the 21st century, the landscape of sales forecasting has been revolutionized by the advent of big data and advanced analytics. The ability to gather extensive data from a variety of sources has opened new avenues for forecasting. Machine learning and deep learning models have become powerful tools for analyzing the data, offering the opportunity to greatly enhance the accuracy of sales forecasts. Researchers also combine two efficient regression models Support Vector Regression (SVR) and variable selection method to produce more efficient and powerful sales prediction methods. The experimental results prove that this new model is excellent in terms of prediction error and in predicting important variables [6]. These models can process large datasets and identify patterns that may not be obvious by means of traditional statistical methods. For instance, ForeXGBoost, a highly efficient and scalable implementation of gradient boosting, has been widely adopted for its ability to enhance predictive accuracy compared to Linear Regression model, Light Gradient Boosting, Logistic Regression and other models [7]. Recurrent Neural Networks (RNNs) and transformers, two important deep learning models have further advanced the field of sales forecasting. These models perform greatly in processing time-series data, letting them suited for capturing temporal dependencies in sales data [8, 9]. The ability of LSTM (Long Short-Term Memory) networks to keep information across extended sequences allows them to model complex temporal patterns, leading to more accurate forecasts in scenarios where traditional models struggle. In sales forecasting, cutting-edge applications of big data also include sentiment analysis and the incorporation of external sources of data. For example, sentiment analysis models that process social media data offer understanding of consumer behavior and market tendency, which can be used to refine sales forecasts. Additionally, the incorporation of macroeconomic indicators and other external factors into forecasting models has been shown to improve accuracy by accounting for broader market dynamics.

The incentive behind this paper arises from the necessary to understand how these advanced big data techniques are being applied in sales forecasting and to identify the challenges that remain. While significant progress has been made, issues such as data quality, model interpretability, and the computational demands of big data analytics continue to pose challenges. This paper aims to offer a thorough review of recent papers in this area, emphasizing the use of machine learning and deep learning models in retailing and E-commerce. The paper will perform in following order. Sec. 2 provides an overview of big data technologies and their applications in sales forecasting. The first part of Sec. 3 presents two detailed case study from the retail sector, examining how big data techniques have been applied to improve sales forecasting accuracy. The second part explores a case study from the E-commerce sector, concentrating on the application of deep learning models. Sec. 4 discusses the limitations of current methodologies and offers prospects for future study. Finally, Sec. 5 concludes the paper by summarizing the main findings and discussing their impact on the field.

2. Descriptions of Big Data

Big data encompasses the large quantities of organized and unorganized data produced by various sources, e.g., social media, transaction records, and sensors. The characteristics of big data (quantity, speed, diversity, and reliability) pose challenges and opportunities for data processing and analysis [10]. In the context of sales forecasting, big data give businesses opportunities to analyze customer behavior, market trends, and external factors more comprehensively. The integration of big data technologies such as Hadoop has enabled the handling of extensive datasets at scale, facilitating more accurate and timely forecasts [9].

DOI: 10.54254/2754-1169/128/2024.18262

Machine learning models are important in harnessing the power of big data to forecast the future demand. Techniques, such as ensemble learning, have been widely adopted, which integrates multiple models to perform more accurate prediction [10]. For instance, Random Forests and XGBoost are used to analyze large datasets, identifying patterns that inform sales predictions. Additionally, deep learning models like RNNs and LSTMs are employed to handle time-series data, capturing trends and seasonality that impact sales [11]. The use of big data in forecasting future demand has led to significant improvements in forecast accuracy, helping industries to make data-driven decisions and optimize their strategies.

3. Applications

3.1. Retail Sector

The retail sector has seen significant advancements in sales forecasting after using big data and machine learning techniques. Retailers are increasingly leveraging large datasets to predict consumer behavior and optimize inventory management. This section explores the application of machine learning models, specifically XGBoost and LSTM, in retail sales forecasting, using case studies and research findings published in 2020. The retail industries are highly volatile because consumer preferences and market conditions changing rapidly. Accurate sales forecasting is crucial for managing inventory, setting pricing strategies, and planning marketing campaigns. In the context of big data, since there are strong seasonal fluctuations in the retail industry, retailors generally use different ways to predict, such as time series and regression methods under neural network implementation. Then, they will use collected data to train machine learning models that can predict future sales with greater accuracy than traditional methods [12].

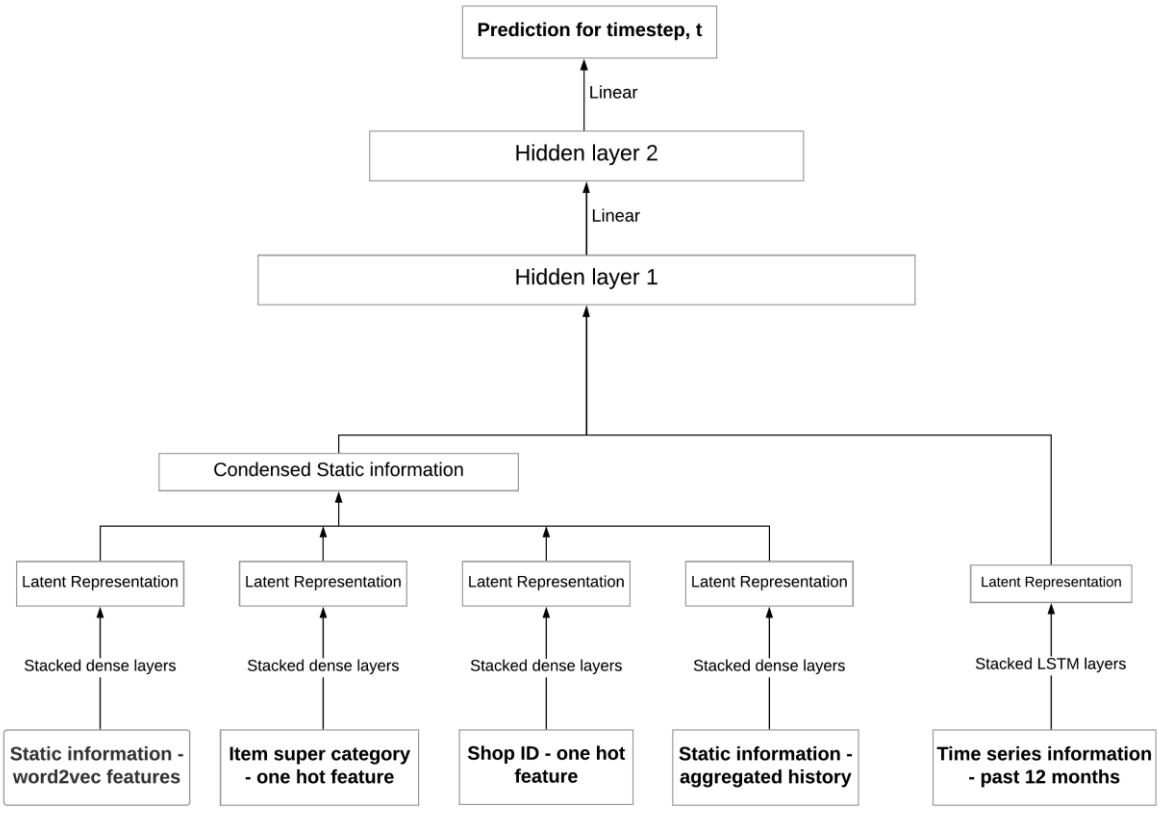

Figure 1: LSTM Neutral Network for Sales Forecasting in retailing [3].

DOI: 10.54254/2754-1169/128/2024.18262

One prominent study by Swami et al. utilized XGBoost and LSTM models to predict future demand for a major retail chain in Russia [3]. XGBoost, or Extreme Gradient Boosting, is a powerful ensemble learning technique in integrating different weaker learners to form a stronger predictive model. It is effective and efficient in managing big, sparse datasets and can capture complex relationships between variables. On the other hand, as one of the RNN, LSTM networks is useful in forecasting time-series data because of their capacity to retain information over extended periods [3]. The researchers used a dataset provided by 1C Company, one of the largest software firms in Russia, which included daily sales data for various products across multiple stores. The dataset was preprocessed to handle missing values, normalize the data, and create lagged features that could capture temporal dependencies. XGBoost was then used to predict sales based on a range of input features, including historical sales data, product information, and store characteristics. LSTM was employed to model the sales data which have the sequential nature, focusing on the time-dependent relationships between past and future sales [3]. A typical sketch is shown in Fig. 1.

The study found that XGBoost outperformed LSTM in terms of forecast accuracy, particularly for products with irregular sales patterns. XGBoost is great in managing data with many features and solve the non-linear relationships between variables made it more effective in predicting sales across different product categories and stores. The root mean squared error (RMSE)and other two related RMSE were used as performance metrics, with XGBoost achieving three lower RMSE compared to LSTM [3]. However, the researchers also noted that choosing the most effective parameters could determine which state-of-the-art performance approach to use. Thus, it may be possible to make LSTM more practical by tuning LSTM-based custom architectures using more rigorous hyperparameter selection methods. [3]. This application shows the potential of machine learning models like XGBoost and LSTM to improve sales forecasting accuracy in the retail sector. The integration of big data techniques allows retailers to predict more accurate future sales and enhance customer satisfaction by ensuring product availability.

3.2. E-Commerce Sector

The e-commerce sector has rapidly adopted big data analytics to forecast sales, given the dynamic nature of online shopping environments. This section discusses the application of LSTM, ARIMA, Facebook (FB) Prophet and new hybrid model in e-commerce sales forecasting. E-commerce platforms produce massive amounts of data every day, including purchase activities and browsing patterns to product reviews. Rapid technological breakthroughs, ever-changing industry trends and dynamic customer behavior characterize the e-commerce industry. In this environment, these companies find that the ability to correctly forecast sales become critical. In today's context under big data, machine learning technologies has become an advantage that can make the industry's sales prediction be more reliable and accurate. The implement of machine learning has changed the traditional methods to predict future sales in E-commerce industries, allowing these companies to utilize state-of-the-art algorithms to dig into the past and predict the future [13]. A study by Vavliakis et al. applied a combined model of ARIMA and LSTM to capture information, such as price reductions, special offers, and sales events by using the retail price, after discounts and trends and seasonality from ARIMA. The dataset used in the study is from an online pharmacy www.pharm24.gr in Greek (seen from Fig. 2) [14]. The ARIMA model is used to predict serial data in one dimension of time, LSTM neural network, this network on the other hand, is able to model the nonlinear residuals of the ARIMA model to form a new model that is better than both of them.

DOI: 10.54254/2754-1169/128/2024.18262

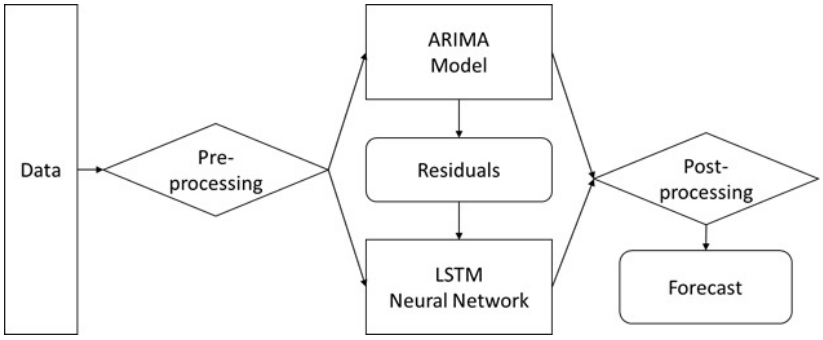

Figure 2: New hybrid of ARIMA and LSTM in e-commerce [14].

To illustrate, Vavliakis, Siailis and Symeonidis similarly used Mean SquareError (MSE), RMSE and Mean Absolute Error (MAE) to predict accuracy. The Table 1 shows that the new hybrid model was highly effective in forecasting E-commerce sales for random 50 products, resulting in much lower results in MSE, RMSE and MAE, which indicated a high level of accuracy [14]. Then, to test more, instead of choosing random 50 products, Vavliakis, Siailis and Symeonidis choose to pick 10 best seller products as the first case, 10 worst seller products as the second case, and 10 highly seasonal products. The results showed in Table 2. The study concluded that hybrid models offer a promising direction for improving sales forecasting accuracy in e-commerce. By combining the strengths of different machine learning techniques, businesses can achieve more reliable forecasts, leading to better inventory management, personalized marketing, and optimized pricing strategies. This application highlights the significance of adopting advanced machine learning models in the e-commerce sector, where sales forecasting is essential in maintaining competitiveness and meeting customer demands.

Table 1: The result of 500 random products (four decimal places) [14].

	MSE (unit:100)	RMSE (unit:100)	MAE (unit:100)
LSTM	5.4077	0.1326	0.0969
ARIMA	4.6606	0.1223	0.0922
Baseline (Zhang)	4.3852	0.1174	0.0876
Suggested approach	4.1544	0.1168	0.0888
Suggested approach incorporating retail price	4.1274	0.1152	0.0873

Table 2: Enhancement for 10 top-selling, 10 low-performing seller and 10 products with seasonal demand [14].

	Enhancement ratio for top-selling products	Enhancement ratio for low-performing products	Enhancement ratio for products with seasonal demand
MSE	0.0581	0.0027	0.0411
RMSE	0.0222	0.0015	0.0176
MAE	0.0171	0.003	0.0092

DOI: 10.54254/2754-1169/128/2024.18262

4. Limitations and Prospects

While the application of big data technology in sales forecasting has led to significant advancements, there are several limitations that researchers and practitioners must address to fully realize its potential. Data quality is an important challenge. Big data is often described by its quantity, speed, diversity, and reliability, which can lead to inconsistencies, missing values, and noisy data. These issues will make sales forecasting highly inaccurate, as the models rely heavily on the reliability of the input data. Partial or incorrect data can result in misleading predictions, which can negatively affect business decisions. Another limitation is the complexity of machine learning and deep learning models used in sales forecasting. While these models, such as XGBoost, RNNs, and LSTMs, have demonstrated exceptional ability in managing large datasets and identifying intricate patterns, they are intensive and demand significant expertise to develop and maintain. Small and medium-sized enterprises (SMEs) may be short on data and technical expertise needed to implement these advanced models effectively, leading to a gap in the adoption of big data technologies between large corporations and smaller businesses. Moreover, the interpretability of these models poses a significant challenge. Machine learning models, particularly deep learning networks do not provide clear explanations for their predictions. This lack of lucidity can pose challenges for businesses that need to understand the rationale behind forecasts to make informed decisions. The need for explainable AI (XAI) is increasingly recognized, but developing models that are both accurate and interpretable remains an ongoing challenge.

Looking to the future, there are several promising directions for research and development in sales forecasting using big data technology. One area of focus is improving data quality and preprocessing techniques. Advanced data cleaning, imputation, and normalization methods can help mitigate the impact of poor data quality on model performance. Additionally, integrating real-time data, such as public sentiment on social media, into forecasting models can offer a more complete perspective of the factors influencing sales and enhance forecast accuracy. Another area of interest is the development of hybrid models that integrate the advantages of various machine learning and deep learning methods. For example, integrating ARIMA with LSTM networks, as discussed in the studies by Vavliakis et al., has shown potential in improving forecast accuracy by leveraging the advantages of both models. Further research into hybrid approaches could lead to more robust and adaptable forecasting systems. Finally, the pursuit of explainable AI (XAI) remains a critical goal. As businesses increasingly rely on AI-driven forecasts for decision-making, the need for models that provide clear, understandable insights is paramount. Future study should concentrate on enhancing and creating models that balance accuracy with interpretability, making it easier for industries to trust and comply with the predictions produced by these advanced systems.

In conclusion, while there are challenges in the current application of big data technology in sales forecasting, ongoing research and technological advancements offer promising prospects for overcoming these limitations. By addressing issues related to data quality, model complexity, and interpretability, future developments in this field can make sales forecasts become more accurate, ultimately helping business make better decision and strategic.

5. Conclusion

To sum up, this study has explored the implementation of big data techniques in sales forecasting, highlighting the significant advancements made through the application of machine learning and deep learning models. The findings demonstrate that models such as XGBoost, ForeXGBoost, LSTM, ARIMA, and hybrid approaches effectively improve forecast accuracy across various industries, including retail and e-commerce. Despite the challenges related to data quality, model complexity, and interpretability, ongoing research continues to address these issues, leading to the development

DOI: 10.54254/2754-1169/128/2024.18262

of more resilient and flexible forecasting systems. Looking forward, the integration of more accurate data and the development of explainable AI are expected to further improve the utility of big data in sales forecasting. The significance of this paper is rooted in its thorough examination of current methodologies and its identification of key areas for future research, offering valuable insights for both academics and practitioners aiming to optimize sales forecasting strategies through advanced data analytics.

References

[1] *Karim, S.A. and Alwi, S.A. (2013) Electricity load forecasting in UTP using moving averages and exponential smoothing techniques. Applied Mathematical Sciences, 7(77-80), 4003-4014.*

[2] *Sinaga, H. and Irawati, N. (2020) A medical disposable supply demand forecasting by moving average and exponential smoothing method. In Proceedings of the 2nd Workshop on Multidisciplinary and Applications (WMA) 2018, 24-25.*

[3] *Swami, D., Shah, A. D. and Ray, S.K. (2020) Predicting future sales of retail products using machine learning. arXiv preprint arXiv:2008.07779.*

[4] *Pavlyshenko, B.M. (2019) Machine-learning models for sales time series forecasting. Data, 4(1), 15.*

[5] *Kulshrestha, S. and Saini, M.L. (2020) Study for the prediction of E-commerce business market growth using machine learning algorithm. In 2020 5th IEEE International Conference on Recent Advances and Innovations in Engineering (ICRAIE) pp. 1-6.*

[6] *Lu, C.J. (2014) Sales forecasting of computer products based on variable selection scheme and support vector regression. Neurocomputing, 128, 491-499.*

[7] *Xia, Z., Xue, S., Wu, L., Sun, J., Chen, Y. and Zhang, R. (2020) ForeXGBoost: passenger car sales prediction based on XGBoost. Distributed and Parallel Databases, 38, 713-738.*

[8] *Vallés-Pérez, I., Soria-Olivas, E., Martínez-Sober, M., Serrano-López, A.J., Gómez-Sanchís, J. and Mateo, F. (2022) Approaching sales forecasting using recurrent neural networks and transformers. Expert Systems with Applications, 201, 116993.*

[9] *Lim, B., Arık, S.Ö., Loeff, N. and Pfister, T. (2021) Temporal fusion transformers for interpretable multi-horizon time series forecasting. International Journal of Forecasting, 37(4), 1748-1764.*

[10] *Buyar, V. and Abdel-Raouf, A. (2019) A Convolutional Neural Networks-based Model for Sales Prediction. In Proceedings of the 2019 International Conference on Artificial Intelligence, Robotics and Control (pp. 61-67).*

[11] *Nazari, E., Shahriari, M.H. and Tabesh, H. (2019) BigData analysis in healthcare: apache hadoop, apache spark and apache flink. Frontiers in Health Informatics, 8(1), 14.*

[12] *Chu, C.W. and Zhang, G.P. (2003) A comparative study of linear and nonlinear models for aggregate retail sales forecasting. International Journal of production economics, 86(3), 217-231.*

[13] *Daulat Desale, I. (2024) E-commerce Sales Forecasting Using Machine Learning Algorithm (Doctoral dissertation, Dublin Business School).*

[14] *Vavliakis, K.N., Siailis, A. and Symeonidis, A.L. (2021) Optimizing Sales Forecasting in e-Commerce with ARIMA and LSTM Models. WEBIST pp. 299-306.*