Predicting E-commerce Revenue Trends: A Fusion of Big Data Analytics and Time Series Analysis

Nayantara Varadharajan
BL.EN.U4CSE21133
Amrita School of Computing, Bengaluru
Amrita Vishwa Vidyapeetham, India
BL.EN.U4CSE21133@bl.students.amrita.edu

Mukil L.D.
BL.EN.U4CSE21109
Amrita School of Computing, Bengaluru
Amrita Vishwa Vidyapeetham, India
BL.EN.U4CSE21109@bl.students.amrita.edu

Kartthikeyan N.
BL.EN.U4CSE21094
Amrita School of Computing, Bengaluru
Amrita Vishwa Vidyapeetham, India
BL.EN.U4CSE21094@bl.students.amrita.edu

***Abstract*—This paper explores the synergistic potential of big data analytics and time series analysis in unravelling intricate patterns within historical sales data to predict and understand e-commerce revenue trends. The amalgamation of these two methodologies provides a robust framework for businesses to gain actionable insights, enabling strategic decision-making and fostering revenue growth. The utilization of big data analytics enables the processing and analysis of vast datasets, encompassing customer behaviours, market trends, and transactional details. Coupled with time series analysis, which focuses on temporal patterns and trends, this fusion approach offers a comprehensive understanding of the dynamic nature of e-commerce revenue. Through the application of predictive models such as Catboost and XGboost, businesses can foresee future revenue trends, identifying peak sales periods, seasonal fluctuations, and potential market disruptions. This foresight empowers e-commerce platforms to optimize pricing strategies, capitalize on emerging opportunities, and mitigate risks. Furthermore, the integration of big data analytics and time series analysis facilitates the identification of hidden correlations and customer preferences. By discerning patterns in user interactions, businesses can tailor personalized customer experiences, enhancing satisfaction and loyalty. The strategic insights derived from this fusion approach go beyond mere trend identification. Businesses can implement targeted marketing campaigns, inventory management improvements, and website optimization strategies. This holistic understanding of the e-commerce landscape equips organizations to adapt swiftly to market dynamics and gain a competitive edge.**

***Index Terms*—Big Data Analytics, Time Series Analysis, E-Commerce, Revenue Trends, Catboost, XGboost.**

I. Introduction

In the rapidly evolving landscape of e-commerce, the proliferation of data has become both a challenge and an unprecedented opportunity for businesses seeking to gain a competitive edge. Harnessing the power of big data analytics and time series analysis has emerged as a pivotal strategy for uncovering hidden patterns within historical sales data, thereby illuminating the path to predictive insights and informed decision-making. This paper delves into the fusion of these two methodologies, specifically employing CatBoost and XGBoost predictive models, to predict and understand e-commerce revenue trends, providing a comprehensive framework for driving strategic insights and fostering sustainable growth.

The era of digital transactions has bestowed upon e-commerce platforms a wealth of data, encompassing customer behaviours, market dynamics, and transactional intricacies. In this context, big data analytics stands as a beacon, offering the means to process and derive meaningful insights from vast datasets. Paired with the temporal granularity provided by time series analysis, this fusion approach enables a nuanced exploration of historical sales data, unravelling intricate patterns that may hold the key to anticipating future revenue trends.

Within this paradigm, the deployment of advanced predictive models, such as CatBoost and XGBoost, becomes instrumental. These models excel in handling complex datasets, providing a robust foundation for predicting peak sales periods, identifying seasonal fluctuations, and anticipating potential market disruptions. The predictive prowess of these models empowers e-commerce platforms to optimize pricing strategies dynamically, seize emerging opportunities, and proactively mitigate risks, thereby laying the groundwork for sustained revenue growth.

Beyond the realm of predictions, the integration of big data analytics and time series analysis offers a holistic understanding of the e-commerce landscape. Uncovering hidden correlations and discerning customer preferences from historical data allows businesses to tailor personalized customer experiences. Armed with these insights, businesses can go beyond mere trend identification, implementing targeted marketing campaigns, optimizing inventory management, and refining website strategies to enhance customer satisfaction and loyalty.

In summary, this paper explores the fusion of big data analytics and time series analysis, employing CatBoost and XGBoost models, as a potent strategy for predicting and understanding e-commerce revenue trends. The insights derived from this fusion not only empower businesses to make informed decisions but also pave the way for dynamic pricing strategies, improved customer experiences, and strategic initiatives that position e-commerce platforms for sustained success in an ever-evolving market.

II. Related Works

This study is a significant contribution to the field of e-commerce forecasting, leveraging a structural time series model integrated with Google Trends data to predict sales.[1] The use of a structural time series model acknowledges the

inherent complexities in e-commerce sales patterns, offering a comprehensive approach. The incorporation of web search data as a predictor adds a new dimension by reflecting users' online behavior, capturing evolving consumer interests and choices in the dynamic e-commerce landscape. Notably, the focus on the Chinese context recognizes regional variations in e-commerce characteristics, considering cultural and economic factors. While the study provides valuable insights, considering the substantial changes in the e-commerce industry since 2014, future research should incorporate sophisticated analytical approaches aligned with the latest advancements in technology, consumer behavior shifts, and big data analytics.

This research explores e-commerce sales forecasting by employing a hybrid machine learning approach, emphasizing advanced techniques to address contemporary challenges.[2] The use of hybrid models, combining various algorithms, demonstrates a nuanced understanding of the intricate patterns within e-commerce sales data. Focusing on product sales forecasting aligns with the practical needs of businesses in the dynamic e-commerce landscape, crucial for effective inventory management and strategic planning. The study's integration of smart electronics and communication underscores the synergy between technology and e-commerce. Presented in 2023, it reflects the current technological and e-commerce landscape, highlighting the commitment to scholarly communication through the inclusion of a Digital Object Identifier (DOI) for easy access to the complete paper by researchers and practitioners.

This paper significantly contributes to the field of e-commerce analytics by investigating analytical methodologies for sales analysis and prediction, particularly focusing on the application of machine learning.[3] The incorporation of machine learning models aligns with the increasing demand for data-driven decision-making in the e-commerce sector. Notably, the study adopts a global perspective, recognizing the diverse market dynamics and technological landscapes shaping e-commerce practices worldwide. The emphasis on analytical methods underscores the commitment to deriving meaningful insights from extensive datasets, crucial for informed decision-making, optimized marketing strategies, and overall operational efficiency in e-commerce. The inclusion of a Digital Object Identifier (DOI) reflects the paper's dedication to scholarly communication, providing a standardized reference for researchers and practitioners seeking detailed information on the presented analytical methods. This commitment supports knowledge dissemination and the reproducibility of the research. Given the dynamic nature of the e-commerce landscape, the paper acknowledges the rapid advancements in machine learning and data analytics technologies, suggesting opportunities for future research to adapt methodologies to emerging technologies and industry trends.

This survey provides valuable insights into the application of predictive analytics in the realm of e-commerce annual sales [4]. The inclusion of predictive analytics in the context of e-commerce annual sales underscores the growing importance of leveraging data-driven approaches for strategic decision-making. By utilizing advanced analytics techniques, the authors offer a framework for forecasting and understanding the complex patterns inherent in e-commerce sales data. The collaborative authorship by experts in data analytics, management, and related fields enhances the interdisciplinary nature of the research. Such collaboration ensures a comprehensive exploration of predictive analytics within the specific domain of e-commerce, providing a holistic view that considers both technological and managerial perspectives. The publication in the Lecture Notes on Data Engineering and Communications Technologies series reflects a scholarly approach, contributing to the academic discourse on data science and analytics.

The paper delves into predicting Amazon sales through the application of time series modeling techniques, presenting valuable insights published in 2020.[5] By focusing on one of the world's largest e-commerce platforms, Amazon, the study addresses the critical task of sales forecasting. The choice of time series modeling reflects an understanding of temporal dependencies and patterns inherent in sales data, aligning with established forecasting best practices. Analyzing Amazon's sales adds practical relevance, given the platform's scale and product diversity, contributing insights that can impact both academic research and industry practices in e-commerce sales forecasting. This interdisciplinary approach recognizes the complexity of contemporary e-commerce systems, requiring expertise not only in data science but also in power and control domains. The inclusion of a Digital Object Identifier (DOI) ensures standardized referencing, facilitating accessibility and citation by researchers, practitioners, and academicians interested in the topic. In summary, the paper, titled "Sales Forecast for Amazon Sales with Time Series Modeling," provides valuable contributions to the application of time series modeling for predicting sales on Amazon, making it a pertinent resource for scholars and professionals engaged in sales forecasting and e-commerce analytics.

The paper proposes a data analytics approach for short-term sales forecasting in the e-commerce marketplace, utilizing Shopee Malaysia as a case study.[6] Three forecasting methods, namely Simple Moving Average (SMA), Dynamic Linear Regression (DLR), and Exponential Smoothing (ES), are evaluated using metrics such as Mean Absolute Deviation (MAD), Mean Absolute Percentage Error (MAPE), and Mean Squared Error (MSE). The results consistently indicate that SMA outperforms the other models, demonstrating the least error across various evaluation metrics. A detailed examination of the forecasting performance for a specific product, Macaron Stereo Wired Earphone, further confirms the superiority of SMA in terms of both absolute and relative error. The study suggests that SMA is the most suitable forecasting model for short-term sales predictions in the e-commerce marketplace, offering practicality and efficiency for sellers in a competitive environment. Future work is proposed to explore larger datasets and include additional product details for enhanced forecasting accuracy.

The paper introduces a novel approach to sales forecasting in e-commerce, employing Convolutional Neural Network

(CNN) for automated feature extraction from structured time series data.[7] The algorithm, complemented by techniques such as sample weight attenuation and transfer learning, significantly enhances prediction accuracy compared to traditional methods. Experimental results, conducted on a dataset provided by Alibaba Group and spanning various regions, demonstrate the superior performance of the proposed algorithm over other approaches, including ARIMA and a complex feature-based model. The CNN-based algorithm, particularly when integrated with sample weight attenuation and transfer learning, exhibits reduced mean square error (MSE) scores, highlighting its effectiveness in improving the precision of commodity sales forecasts in diverse e-commerce scenarios.

This study introduces a novel model for forecasting short-term goods demand in the E-commerce domain by integrating a Long Short-Term Memory (LSTM) approach with sentiment analysis of consumer comments.[8] Utilizing sales figures and comments from "taobao.com," the LSTM model is trained to predict future sales based on the time-series sequence of sales and sentiment ratings. Given the challenges of short-term goods with limited historical data, the study emphasizes the need for prompt reactions to market conditions. The research demonstrates that adjusting the weight of sentiment ratings can enhance forecasting accuracy. The proposed model achieves high accuracy in predicting sales for goods with short-term demands, supporting efficient decision-making in the E-commerce sector. The study further explores the impact of sentiment ratings on forecasting performance and introduces model calibration to accommodate shorter training time-series, confirming the consistency of results with the proposed claims. Overall, the research showcases the potential of AI, LSTM, and sentiment analysis in achieving maximal predictive accuracy with minimal historical data in E-commerce sales forecasting.

In this paper, the authors present a novel sales demand forecasting framework for E-commerce, utilizing Long Short-Term Memory (LSTM) networks.[9] Addressing challenges like non-stationary data and sparse sales patterns, the proposed methodology incorporates cross-series information within related products. The framework involves systematic preprocessing, LSTM network architecture with various learning schemes, and the inclusion of static and dynamic features. Empirical evaluations on Walmart.com datasets reveal that the LSTM framework, particularly when leveraging product grouping strategies, outperforms traditional methods, showcasing improved accuracy in demand forecasting for both category and super-department level datasets. The results highlight the effectiveness of LSTM networks in capturing non-linear relationships within E-commerce product hierarchies.

This article emphasizes the pivotal role of prediction in various facets of business, underscoring its increased complexity due to market competition, diverse production, and globalized supply chains.[10] Leveraging advanced digital technologies like cloud computing, IoT, and social media, it advocates for big data analysis to enhance sales predictions, customer behavior understanding, and supply chain management effectiveness. Focusing on the e-commerce sector, the paper highlights the challenges in predicting customer demands and stresses the multifaceted factors influencing sales predictions. The authors propose an integrated approach employing Support Vector Regression (SVR) with Auto-Regressive Integrated Moving Average (ARIMA) for revenue prediction, demonstrating superior accuracy compared to standalone ARIMA and artificial neural network (ANN) methods through empirical evaluations on Indian apparel and smartphone sales datasets from 2021. The results affirm the efficacy of the SVR-ARIMA hybrid model in optimizing sales forecasts, contributing significantly to financial risk management in the e-commerce industry.

III. Proposed Methodology

A. Data Preprocessing

The data preprocessing involves several crucial steps to ensure the dataset's cleanliness and suitability for time-series modeling. Firstly, missing values are addressed by dropping rows with undefined "CustomerID" and filling in "Description" gaps with a placeholder. Negative quantities, representing returned items, are removed, and rows with zero or negative unit prices are filtered out. The timestamp information is parsed, converting "InvoiceDate" into a datetime object, and additional features like day, month, and year are extracted.

Feature engineering introduces new dimensions, such as the length of "StockCode" and the count of numeric characters in it. Outliers in "UnitPrice" and "Quantity" are filtered, ensuring the removal of extreme values. The data is then structured for modeling through the creation of a pivot table, aggregating daily quantities and revenues for each product, with missing values appropriately filled.

Start Timepoint	2010-12-01 08:26:00
End Timepoint	2011-12-09 12:50:00
Number of Days	373 days

TABLE I
TIMELINE OF THE DATA

Overall, these preprocessing steps collectively handle data integrity, feature engineering, and outlier management, setting the stage for effective time-series analysis. The resultant dataset is well-structured and ready for subsequent stages, such as feature selection, model training, and evaluation. The preprocessing choices align with the goals of time-series analysis and are tailored to the specific characteristics of the dataset.

B. Exploratory Data Analysis

The exploratory data analysis (EDA) for this project begins with a comprehensive review of the dataset, encompassing both numerical and categorical features. Initial data summaries reveal key statistics, distributions, and unique values, while univariate analyses, including histograms and box plots, expose patterns and outliers. Subsequent bivariate and multivariate analyses delve into feature interactions, scrutinizing relationships through scatter plots.

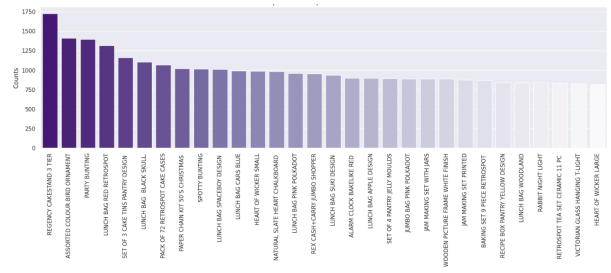

Fig. 1. Most common product descriptions

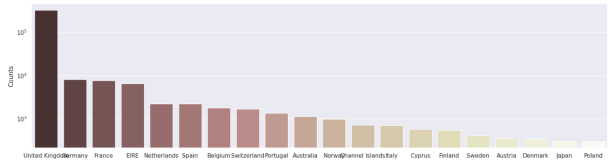

Fig. 2. Countries by transaction counts

EDA also involves handling missing values, detecting outliers, and assessing overall data quality. Visualizations, which include bar graphs, enhance the understanding of the dataset, enabling the identification of potential influencing factors on the target variable.

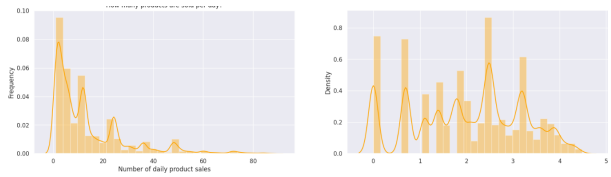

Fig. 3. Daily product sales distribution

The graphs illustrate the distribution of daily product sales quantities. In the first subplot, the untransformed distribution reveals a right-skewed pattern, indicative of a majority of products experiencing lower daily sales. Notably, the presence of multiple peaks at quantities 1, 12, and 24 suggests a multimodal distribution. The additional observation that these quantities often follow divisibility by 2 or 3 adds a layer of complexity, hinting at purchasing behaviors where products are acquired in pairs or triplets.

In the second subplot, the application of a logarithmic transformation aims to mitigate skewness and accentuate differences in the lower quantity range. Despite the transformation, the essential features of the distribution persist. The reduced right-skewness and clearer visibility of patterns post-transformation enhances the understanding of the dataset.

C. Model Building

The primary focus is on predicting daily product sales using the CatBoost regressor. This predictive model is well-suited for regression tasks, and the evaluation metric is the root mean square error (RMSE). The model training and validation strategy involve careful consideration of the temporal nature of the data, including a sliding window time series validation approach to account for the significant increase in sales during the pre-Christmas period. The root mean square error (RMSE) formula is given by:

$$E = \sqrt{\frac{1}{N}\sum_{n=1}^{N}(t_n - y_n)^2} \tag{1}$$

To streamline the experimentation and comparison of models, a series of classes have been developed, including the CatHyperparameter class for managing hyperparameters, the Catmodel class for individual model training and analysis, and the Hypertuner class for Bayesian hyperparameter search. The TimeSeriesValidationCatfamily class orchestrates the model training with sliding window validation, facilitating a comprehensive evaluation of the model's performance across different time periods. The model building process also incorporates feature engineering, including the creation of product types, exploration of temporal patterns, and the generation of lag features. These engineered features aim to capture underlying patterns and improve the model's predictive capabilities. The comprehensive approach to model building and evaluation ensures a thorough exploration of the data and the development of predictive models capable of capturing the nuanced patterns in daily product sales.

The proposed methodology seeks to provide a robust framework for predicting product quantities, contributing valuable insights for businesses to enhance sales forecasting strategies.

IV. Results and Discussion

The initial model evaluation yielded a root mean square error (RMSE) on the validation data of approximately 0.2071, indicating a relatively good fit. However, it's crucial to note that RMSE is susceptible to outliers. Therefore, a more comprehensive analysis of individual absolute errors was conducted, revealing a right-skewed distribution. The median absolute error was found to be 0.7965, suggesting the presence of higher errors for validation entries with true quantity values exceeding 20.

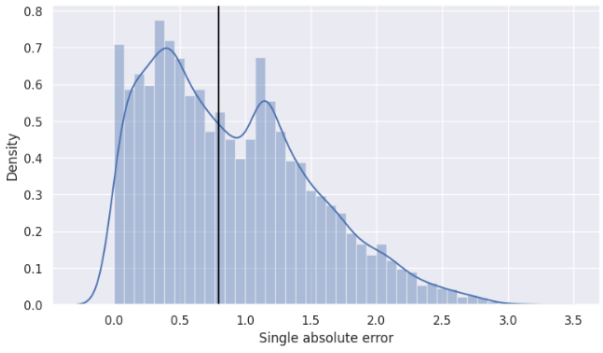

Fig. 4. Single absolute error versus density(after tuning)

To gain insights into the distribution of errors, a plot of target versus prediction was generated. It depicted higher errors for products with true quantities above 20, emphasizing the need for improved predictions in this range. Key features

TABLE II
MODEL EVALUATION METRICS BEFORE AND AFTER HYPERPARAMETER TUNING

Metric	Before Tuning	After Tuning
RMSE	0.2071	0.4787
Median Absolute Error	0.7965	0.6024
Mean Absolute Error	9.7208	7.7804
Median of Predicted Quantities	4.7128	4.0149

influencing predictions were identified, including stock code, product description, and weekday.

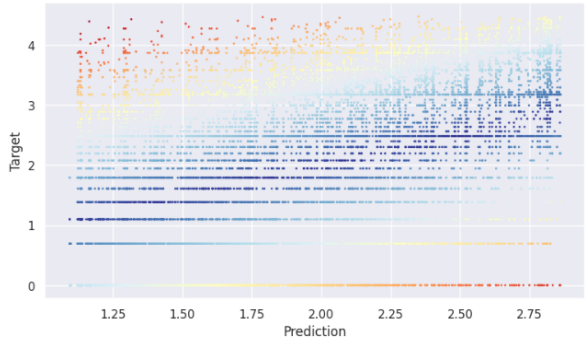

Fig. 5. Plot of target versus prediction(before tuning)

The temporal analysis of daily quantities sold revealed interesting patterns. The weekday emerged as a significant feature, aligning with earlier explorations that indicated higher product sales from Monday to Thursday. This correlation was visually confirmed in the plot, where low weekday values (Monday to Thursday) correlated with high product sales, while higher values (Friday to Sunday) corresponded to lower sales.

Thursday emerged as the day with the highest product sales, while Friday and Sunday exhibited significantly lower transactions. Saturdays showed no transactions at all. Additionally, the pre-Christmas season, starting in September and peaking in November, highlighted the importance of seasonality. February and April stood out as months with notably low sales.

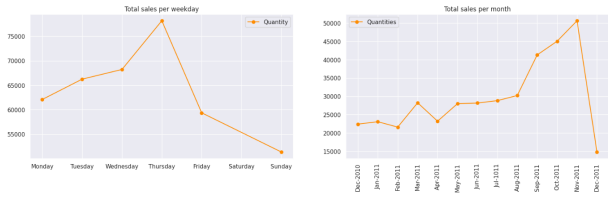

Fig. 6. Temporal Graph

The subsequent focus on stock code and description underscored their importance as features, albeit their complexity. With close to 4000 stock codes and numerous descriptions, the need for feature engineering to describe products more generically was emphasized.

The hyperparameter tuning process, employing Bayesian optimization with GPyOpt, aimed to enhance model performance. Post-tuning, the RMSE increased to 0.4787, suggesting a trade-off between the model's generalization and its ability to capture the training data. The median absolute error post-tuning was 0.6024, indicating a slight decrease compared to the untuned model. The mean absolute error declined to 7.7804, while the median absolute error decreased slightly to 4.0149.

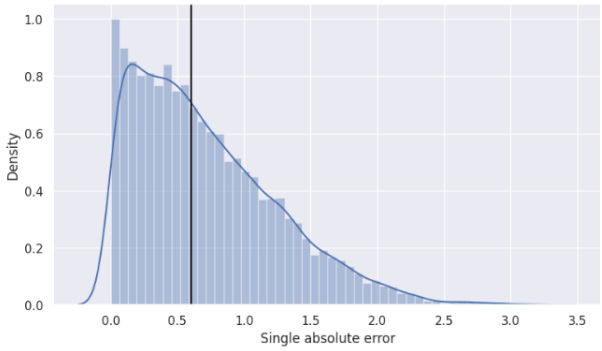

Fig. 7. Single absolute error versus density(after tuning)

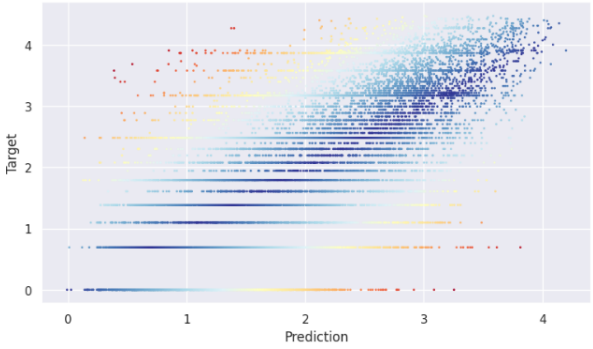

Fig. 8. Plot of target versus prediction(before tuning)

In summary, the model evaluation, hyperparameter tuning, and temporal analysis collectively provide a comprehensive understanding of the dataset and the predictive model's performance. The insights gained contribute to refining the model and enhancing its accuracy in capturing the nuances of product sales patterns.

V. CONCLUSION

This project presents a comprehensive approach to predictive modeling. It seamlessly integrates data exploration, advanced regression modeling with CatBoost, thoughtful validation strategies, hyperparameter optimization, and extensive feature engineering. This iterative process ensures a refined model that not only predicts sales quantities but also offers insights into the underlying dynamics of the dataset. In conclusion, big data analytics and time series analysis are indispensable tools for e-commerce businesses seeking to uncover hidden insights, make informed decisions, and drive revenue growth.

REFERENCES

[1] Wei, Dai & Peng, Geng & Ying, Liu & Shuaipeng, Li. (2014). A prediction study on e-commerce sales based on structure time series model and web search data. 26th Chinese Control and Decision Conference, CCDC2014.53465351. 10.1109/CCDC.2014.6852219.

[2] R. Kumar, "Hybrid Machine Learning Method for Product Sales Forecasting in E-Commerce," 2023 4th International Conference on Smart Electronics and Communication (ICOSEC), Trichy, India, 2023, pp. 781-787, doi: 10.1109/ICOSEC58147.2023.10276171.

[3] K. Anushka Xavier, C. Manjunath, M. Manohar, V. R. Gurudas, N. Jayapandian and M. Balamurugan, "Analytical Methods of Machine Learning Model for E-Commerce Sales Analysis and Prediction," 2023 IEEE International Conference on Contemporary Computing and Communications (InC4), Bangalore, India, 2023, pp. 1-6, doi: 10.1109/InC457730.2023.10263003.

[4] Predictive Analytics on E-commerce Annual Sales. In: Gupta, D., Polkowski, Z., Khanna, A., Bhattacharyya, S., Castillo, O. (eds) Proceedings of Data Analytics and Management. Lecture Notes on Data Engineering and Communications Technologies, vol 90. Springer, Singapore. (2022)

[5] "Sales Forecast for Amazon Sales with Time Series Modeling," 2020 First International Conference on Power, Control and Computing Technologies (ICPC2T), Raipur, India, 2020, pp. 38-43, doi: 10.1109/ICPC2T48082.2020.9071463.

[6] Chee, C.C.F., Leng Chiew, K., Sarbini, I.N.B., & Jing, E.K.H. (2022). Data Analytics Approach for Short-term Sales Forecasts Using Limited Information in E-commerce Marketplace. Acta Informatica Pragensia, 11(3), 309-323. doi: 10.18267/j.aip.196

[7] Pan, H., Zhou, H. Study on convolutional neural network and its application in data mining and sales forecasting for E-commerce. Electron Commer Res 20, 297–320 (2020). https://doi.org/10.1007/s10660-020-09409-0

[8] Shih, YS., Lin, MH. (2019). A LSTM Approach for Sales Forecasting of Goods with Short-Term Demands in E-Commerce. In: Nguyen, N., Gaol, F., Hong, TP., Trawiński, B. (eds) Intelligent Information and Database Systems. ACIIDS 2019. Lecture Notes in Computer Science(), vol 11431. Springer, Cham. https://doi.org/10.1007/978-3-030-14799-0_21

[9] Bandara, K., Shi, P., Bergmeir, C., Hewamalage, H., Tran, Q., Seaman, B. (2019). Sales Demand Forecast in E-commerce Using a Long Short-Term Memory Neural Network Methodology. In: Gedeon, T., Wong, K., Lee, M. (eds) Neural Information Processing. ICONIP 2019. Lecture Notes in Computer Science(), vol 11955. Springer, Cham. https://doi.org/10.1007/978-3-030-36718-3_39

[10] Sharma G, Patil S. Big Data Analysis for Revenue and Sales Prediction using Support Vector Regression with Auto-regressive Integrated Moving Average. sms [Internet]. 14Jan.2023 [cited 9Jan.2024];15(01):1-. Available from: https://www.smsjournals.com/index.php/SAMRIDDHI/article/view/3077