Forecasting Sales in e-Commerce: Insights from Exploratory Data Analysis and Machine Learning Techniques

Roxane Elias Mallouhy

College of Engineering, Al Yamamah University, Khobar, Saudi Arabia
Email: r mallouhy@yu.edu.sa
*Corresponding author

Manuscript received November 27, 2024; revised February 1, 2025; published August 27, 2025

***Abstract*—Sales forecasting in e-commerce is essential for optimizing inventory management, enhancing customer satisfaction, and supporting strategic decision-making. This study investigates sales forecasting using exploratory data analysis (EDA) and advanced feature engineering techniques. Utilizing rich data from a prominent household items retailer in Saudi Arabia, key patterns and trends influencing sales performance are identified. Several predictive models are developed and their performances are evaluated, with a particular focus on the impact of domain- specific features and state-of-the-art machine learning algorithms on forecast accuracy. The findings demonstrate that incorporating domain-specific features and advanced machine learning techniques significantly improves sales forecasting precision. This research provides valuable insights and practical methodologies for practitioners and researchers aiming to enhance their e- commerce sales forecasting capabilities. The implications of the results are discussed, providing a solid foundation for future work in the field.**

***Keywords*—E-commerce, feature engineering, machine learning algorithms, sales forecasting**

I. Introduction

E-commerce has revolutionized the retail landscape, offering significant advantages over traditional face-to-face shop- ping. It offers unparalleled convenience, a broader range of products, competitive pricing, and the ability to shop 24/7 from anywhere. According to recent statistics, global e-commerce sales reached $4.28 trillion in 2020 and are projected to grow to $6.38 trillion by 2024 [1]. This rapid growth under- scores the importance of effective sales forecasting to manage inventory, meet customer expectations, and drive business strategy. Furthermore, the number of digital buyers worldwide is expected to reach 2.14 billion by 2024, highlighting the growing reliance on e-commerce platforms [2].

Incorporating advanced machine learning techniques into e-commerce operations has dramatically enhanced the industry's efficiency and effectiveness. Machine learning algorithms can process and analyze massive datasets, identifying hidden patterns in consumer behavior, purchase trends, and product preferences. These insights allow businesses to implement personalized product recommendations, optimize pricing in real time, and forecast future sales more accurately. This helps e-commerce platforms not only meet customer expectations but also manage inventory efficiently by predicting demand fluctuations, reducing overstock and stockouts.

Moreover, machine learning is instrumental in improving the overall customer experience in e-commerce. Personalized recommendations powered by algorithms significantly boost conversion rates and customer satisfaction by offering relevant products and services. E-commerce businesses are also lever- aging machine learning to refine marketing strategies, enabling them to segment their audiences more precisely and deliver targeted ads that resonate with specific consumer groups [3]. Machine learning algorithms also contribute to reducing operational costs by automating key processes such as dynamic pricing, customer support through AI-powered chatbots, and efficient supply chain management. Fraud detection is another critical area where machine learning plays a pivotal role by identifying suspicious transaction patterns and preventing fraud in real time. As the e-commerce sector continues to expand, the integration of machine learning into business operations is not only driving growth but also providing a competitive edge to those who effectively harness its potential.

This study leverages data from a leading household items retailer in Saudi Arabia, referred to anonymously for confidentiality. Founded in 2011 as a part of a reputable Holding Company, this company has established itself as a pioneer in the retail homeware sector. With a presence in six major cities across the Central, Eastern, and Western provinces, its showrooms offer a unique shopping experience. The company is dedicated to providing a diverse range of products, from kitchen tools to furniture and décor, all designed by Saudi designers and a multi-cultural purchasing team. By delivering high-quality products at reasonable prices, it helps families furnish their homes elegantly and affordably. The data from this household items retailer is particularly valuable for the study because it encompasses diverse consumer purchasing behaviors across different regions, offering a comprehensive view of the e-commerce market in Saudi Arabia.

The integration of domain-specific features and advanced machine learning algorithms in this study demonstrates how e-commerce companies can significantly improve their sales forecasting accuracy. This research offers valuable insights and practical methodologies that can be applied by practitioners and researchers to enhance sales forecasting processes in the e-commerce sector. The implications of these findings are far-reaching, providing a solid foundation for future work and innovation in this dynamic field.

The remainder of this article is structured as follows: Section II reviews the state of the art in sales forecasting techniques, highlighting recent advancements and their applications in e-commerce. Section III describes the data mining and analysis process, detailing the methods used for data collection, preprocessing, transformation, feature engineering, and modeling. Section IV presents the results

doi: 10.18178/ijml.2025.15.3.1180

and interpretation, including trend analysis, model coefficients, and model performance. Finally, Section V concludes the paper, discussing the implications of the findings and suggesting future research directions.

II. State of the Art

Sales forecasting is an essential task for e-commerce platforms, providing crucial impacts on business processes. Accurate sales predictions enable better understanding of financial status, workforce management, and improvement of supply chain systems. The continuous advancements in machine learning and statistical methods provide more robust and accurate tools for e-commerce sales forecasting. These techniques are crucial for businesses to stay competitive and meet the growing demands of the global e-commerce market. Multiple studies highlight the importance of sales predictions in inventory planning, competitive pricing, and timely promotional strategies, thereby driving the growth and stability of e-commerce platforms.

A notable study presents a sales prediction model that leverages deep learning to handle complex and unstructured data. The model integrates consumer behavior analysis, selecting input variables such as product images, prices, discounts, and historical sales. It utilizes three types of neural networks: fully connected neural networks for structured data, convolutional neural networks for image data, and recurrent neural networks for sales sequence data. These models are combined to form a deep neural network for feature representation. A final fully connected neural network is used to train the prediction model, which outperforms traditional methods like exponential regression and shallow neural networks in prediction accuracy. [4].

Building on this, Li et al. presents a cascaded hybrid neural network model for predicting commodity demand on e-commerce platforms using multimodal data. By integrating historical order information and product sentiment data, the model combines bi-directional long short-term memory (LSTM) and gated recurrent unit (GRU) networks to improve prediction accuracy. The model achieved an average absolute error of 0.1682 and a root mean square error (RMSE) of 0.4537 for weekly forecasts, and 0.8611 with an RMSE of 8.1938 for long-term demand predictions. These results demonstrate the model's effectiveness in enhancing forecast accuracy and supporting inventory management [5].

Recent advancements have continued to refine these techniques. A study introduces a CNN-LSTM hybrid model, optimized using a Genetic Algorithm (GA), to predict stock prices more accurately. By combining CNN's feature extraction and LSTM's handling of long-term dependencies, the model predicts the next day's closing price using 20 days of stock and technical data. The model, evaluated on Korea Stock Index (KOSPI) data, showed better accuracy than standalone CNN, LSTM, and CNN-LSTM models. This approach can aid investors and policymakers in making more informed decisions [6]. On the other hand, Datsko conducted a complex comparison of statistical and econometric methods, showing that advanced statistical techniques can complement machine learning approaches for enhanced forecasting accuracy [7].

Rasappan *et al.* proposes an optimized machine learning algorithm, EGJO-LSTM, for sentiment analysis of e-commerce product reviews. The approach involves data collection, pre-processing, feature selection, and classification. Using techniques like LF-MICF and IGWO for feature selection, EGJO-LSTM classifies sentiments into negative, positive, or neutral. Tested on Amazon.com data, it significantly outperforms traditional and hybrid methods, with improvements of 25% in precision and 32% in accuracy [8].

Moreover, the global growth of e-commerce sales underscores the importance of sophisticated forecasting models. A 2024 report from Shopify projects that global e-commerce sales will reach $6.33 trillion, reflecting an 8.8% annual increase. This growth highlights the increasing reliance on advanced forecasting models to manage large-scale operations effectively [9]. On the other hand, Asana discusses various sales forecasting methods, including pipeline forecasting and multivariable analysis, which help businesses predict sales more accurately by considering multiple factors such as market trends and economic conditions [10].

In 2024, the focus has also shifted towards integrating AI-driven analytics for e-commerce demand forecasting. Companies are leveraging AI to handle large datasets and provide dynamic pricing strategies, ensuring they can respond quickly to market changes and consumer behavior. This approach minimizes errors and optimizes business performance, underscoring the critical role of advanced analytics in modern e-commerce operations [11].

Recent studies have highlighted the profound impact of machine learning (ML) on enhancing marketing performance, particularly in e-commerce. For instance, [12] demonstrates how ML-driven technological advancements have transformed various aspects of e-commerce, including supply chain management, value chain optimization, and customer relationship management. The paper emphasizes the role of ML in increasing operational efficiency and driving more informed decision-making. By analyzing diverse e-commerce models and showcasing practical use cases, it underscores the critical benefits of integrating ML into e-commerce strategies, leading to improved outcomes and competitiveness.

Additionally, an innovative approach introduces an ensemble deep learning framework designed to analyze customer behavior for improved personalized recommendations and customer satisfaction in e-commerce. This framework integrates three powerful algorithms—CNN, GAN, and LSTM-RNN—into two distinct modules: Purchasing Intention (PI) and Abandonment Analysis (AA). It effectively captures customer interactions, refining recommendations with greater precision. The model demonstrates superior accuracy and generalization, outperforming existing models in delivering highly tailored recommendations [13].

Similarly, another study explores the application of machine learning techniques in demand forecasting for the manufacturing sector, evaluating methods like regression, time series fore- casting, neural networks, and ensemble models. It concludes by emphasizing the need for improving data quality, enhancing model interpretability, and considering ethical implications in ML-based demand forecasting. [14].

III. Data Mining and Analysis Process

The Knowledge Discovery in Databases (KDD) technique was employed in this paper due to its systematic approach to discovering valuable insights from large datasets. The KDD process provides a structured methodology for transforming raw data into useful knowledge through a series of well-defined steps, ensuring the integrity and quality of the data throughout the process. By applying the KDD technique, this study was able to effectively uncover patterns and trends in the e-commerce data.

A. Data Collection and Selection

The initial step in the study involved collecting and selecting relevant data. A comprehensive dataset from a leading household items retailer in Saudi Arabia was utilized, which includes various transactional and customer-related information. The dataset spans from September 2020 to September 2023. However, due to the insufficient volume of data in the early records from 2020, the analysis focused on purchases made from January 2022 to the end of September 2023. This adjustment ensured a more robust dataset for the analysis, as illustrated in Fig. 1.

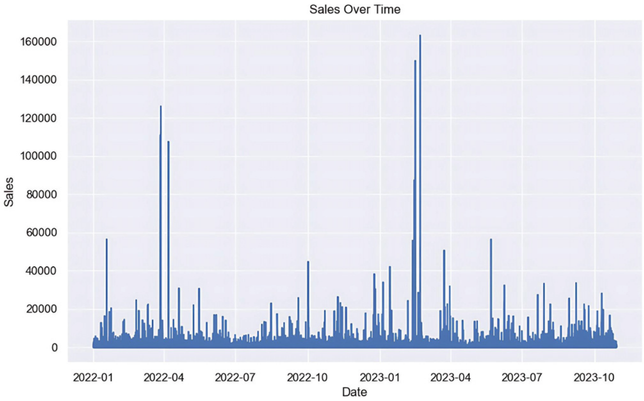

Fig. 1. Sales over time.

Table 1. Description of dataset attributes

Attribute	Description
ORDER NUMBER	Unique identifier for each order.
BUSINESS_DATE	The date when the business transaction was recorded.
TRAN_SEQ_NO	Transaction sequence number for tracking the order process.
ORDER QTY	The quantity of items ordered.
ORDER CREATE DATE	The date and time when the order was created.
CUST NAME	The name of the customer who placed the order.
CUST_MOBILE	The mobile number of the customer, serving as the primary and unique key.
CUST ADD	The address of the customer.
CUST CITY	The city where the customer is located.
CUST DISTRICT	The district where the customer is located.
SHIP DATE	The date and time when the order was shipped.
TENDER_TYPE	The type of payment method used (e.g., COD - Cash on Delivery).
LATEST ORDER STATUS	The most recent status of the order (e.g., Delivered, Pending).
DELIVERY_DATE	The date when the order was delivered to the customer.
ORDER STATUS	The overall status of the order.
ITEMS	The price of the same item(s) in the order.
COD Charges	Charges applicable for Cash on Delivery.
Delivery Charges	Charges for delivering the order.
VAT	Value Added Tax applied to the order.
TENDER	Detailed description of the payment method.
total_amount	The total cost of the order, calculated as ITEMS + COD Charges + VAT.

Eventually, the dataset encompasses order numbers, transaction dates, order quantities, customer details, financial charges (including VAT and delivery charges), item prices, and more as showing in Table 1. This rich dataset provides a solid foundation for analyzing sales patterns and developing predictive models. The primary key for this study is the customer's phone number, as it is the unique identifier across all records. The dataset contains 19198606 rows, with some customers making multiple orders and some orders comprising different items separated into different rows. Nevertheless, the addition of a new column, total_amount was mandatory to encapsulate the total cost associated with each order by summing the item price (ITEMS), Cash on Delivery charges (COD_Charges), and Value Added Tax (VAT).

Additionally, the ORDER_NUMBER attribute requires careful handling to accurately reflect transactions where multiple items are purchased in a single order. For example, if a customer orders three different items—such as cups, plates, and a shower curtain—in one transaction, each item will be recorded in a separate row, resulting in the ORDER _NUMBER being duplicated across these rows.

B. Data Preprocessing

Data preprocessing is a crucial phase that ensures the quality and consistency of the dataset. This step involves cleaning the data by handling missing values and converting data types to suitable formats. Ensuring data integrity at this stage is essential to avoid any disruptions in subsequent analyses. Non-numeric values were converted to numeric types, and any resulting missing values were filled with zeros where appropriate. This preparation step is vital for effective feature engineering and modeling, ensuring that algorithms can process the data without errors.

Given that purchases can be influenced by the date and specific periods of the year, it was essential to consider factors such as yearly discounts and traditional buying patterns in Saudi Arabia. Significant household item purchases typically occur before the two Eids, Ramadan and Adha, whose dates vary annually. Studying these trends is crucial in e-commerce sales forecasting because it allows businesses to anticipate demand spikes and optimize inventory management, marketing strategies, and customer service. For example, during Ramadan, there is a marked increase in the purchase of food items, clothing, and gifts as families prepare for the holy month and the subsequent celebrations. Similarly, the period leading up to Eid al-Adha often sees a surge in the sale of household items, gifts, and festive decorations. By understanding and forecasting these patterns, e-commerce platforms can ensure they have adequate stock, launch timely promotions, and tailor their marketing efforts to maximize sales.

To better understand the patterns and trends in the sales data, new columns for the month, day of the week, year, and date were created. This transformation allows for more granular analysis, enabling the identification of specific time-based trends and seasonal effects on sales performance. For instance, analyzing sales performance by the day of the week can reveal if certain days have higher sales volumes, which can inform marketing and inventory decisions. Similarly, breaking down sales by month (as indicated in Fig. 2) can help identify seasonal trends and monthly growth. After creating these time-based columns, all other columns in the dataset referring to the order date were omitted to avoid redundancy and reduce complexity. This simplification ensures that the dataset is streamlined, making it easier to analyze and interpret the time-related patterns without the confusion of multiple date-related columns. Additionally, the shipping date was excluded as it falls outside the scope of this analysis.

Furthermore, all attributes were validated to match specific formats, such as email addresses, phone numbers, and dates. Rows with missing purchase amounts were removed since they do not contribute to the analysis. However, rows with missing attributes such as customer name, address, or district were replaced with ’Unknown’ to preserve the purchase amount data, which is crucial for understanding the overall pattern. This approach ensures that the dataset remains comprehensive and that critical sales data is not lost.

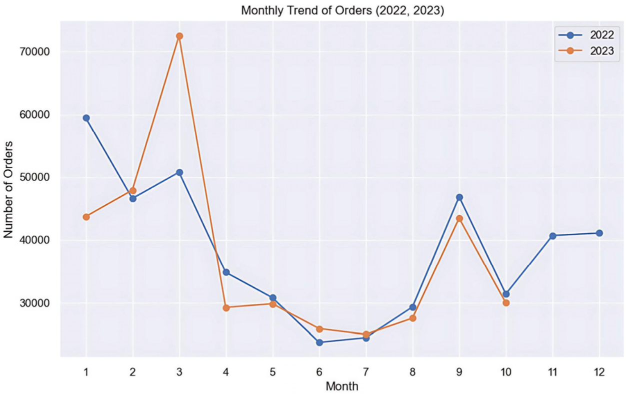

Fig. 2. Monthly sales trend.

Last, breakpoint detection was implemented as illustrated in Fig. 3 to identify significant shifts or changes within the dataset over time. This technique helps in pinpointing the highest sales overall the dataset helping in understanding periods of significant change, which can be linked to promotional campaigns, seasonal effects, or external factors impacting sales.

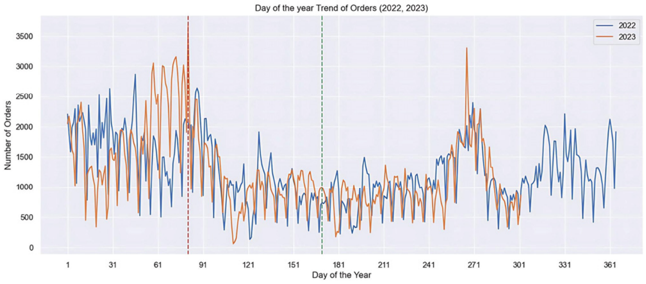

Fig. 3. Order volume with breakpoints for 2022 and 2023.

C. Data Transformation and Feature Engineering

Encoding categorical variables was an essential step in this process. All categorical data must be converted into a numerical format suitable for machine learning algorithms applied in this study. The method employed included identifying non- numeric columns and applying different encoding techniques based on the cardinality of the columns: For columns with fewer unique values (low cardinality), label encoding was used, assigning a unique integer to each category. For columns with a high number of unique values (high cardinality), frequency encoding was applied, replacing each category with the frequency of its occurrence in the dataset.

On the other hand, feature engineering is a key process that involves creating new features to improve the predictive power of the model. In this study, several important features were engineered:

1) Recency: Days since the last purchase
2) Frequency: Number of orders by each customer
3) Monetary: Total amount spent by each customer

Including these attributes is important as they help in understanding customer behavior. Recency indicates how recently a customer made a purchase, which can suggest their engagement level. Frequency shows how often a customer buys, reflecting their loyalty. Monetary value represents the total spending of a customer, highlighting their economic contribution.

Afterwards, linear regression analysis was conducted on ’Recency’ and ’Monetary’ to gain more insights into purchase frequency. The coefficients of the linear regression model indicate the relationship between each predictor (independent variable) and the target variable (purchase frequency). This analysis helps in identifying the impact of each factor on customer purchasing behavior, enabling more accurate predictions and targeted strategies.

D. Data Modeling

Various machine learning techniques were systematically applied to identify patterns and relationships within the data, with the goal of making sales predictions and deriving insights. This phase transforms the preprocessed data into actionable knowledge through model selection, hyperparameter tuning, and evaluation.

The dataset, consisting of nearly 19 million rows and 24 attributes, is initially divided into training and testing subsets to ensure models are evaluated on unseen data, which helps in assessing their generalizability. A sample size of 100000 rows is used for initial testing and hyperparameter tuning to expedite the modeling process and mitigate computational constraints. Sampling ensures that the models are evaluated quickly without compromising the integrity of the insights derived from the data.

Several machine learning algorithms are employed, each with distinct strengths and suitable use cases. Linear Regression is used for its simplicity and efficiency in modeling linear relationships between the dependent and independent variables [15]. Although it provides an excellent baseline, its assumptions of linearity and independence of errors can limit its applicability in more complex scenarios. Decision Trees offer a non-linear approach by recursively partitioning the data space, capturing interactions between features that linear models might miss [16]. They are easy to interpret but can suffer from overfitting, as indicated by their tendency to perform exceptionally well on training data but less so on test data.

To mitigate overfitting, ensemble methods such as Random Forest and Gradient Boosting are utilized. Random Forest combines multiple decision trees through bagging to enhance predictive accuracy and robustness by averaging their pre- dictions [17]. This technique significantly reduces overfitting and improves generalization performance. Gradient Boosting builds trees sequentially, with each tree correcting errors made by its predecessor, often leading to superior predictive performance but at the cost of increased computational complexity [18].

Hyperparameter tuning is performed using GridSearchCV, a methodical approach to identify the optimal parameter values for each model by evaluating all possible combinations over a specified parameter grid [19]. This ensures that each model operates under the best conditions for the given dataset, enhancing their predictive performance and robustness.

A summary of the pros and cons of each model is provided in Table 2 for quick reference.

Table 2. Summary of model pros and cons

Model	Pros	Cons
Linear Regression	Simple to implement, efficient, interpretable	Assumes linearity, independence of errors, limited in capturing complex relationships
Decision Tree	Non-linear, easy to interpret, captures feature interactions	Prone to overfitting, can be unstable with small variations in data
Random Forest	Reduces overfitting, robust, handles high dimensional data well	Computationally intensive, less interpretable than single trees
Gradient Boosting	High predictive performance, handles complex relationships	Computationally complex, risk of overfitting, requires careful tuning

E. Model Evaluation

Evaluating models on both training and testing datasets is crucial for understanding their performance and generalizability. Here is the key evaluation metrics used:

1) Mean Absolute Error (MAE): Average magnitude of errors in predictions.
2) Root Mean Squared Error (RMSE): Average prediction error in the same units as the original data.
3) R-squared (R^2): Proportion of variance in the dependent variable predictable from the independent variables. It ranges from 0 to 1, where 1 indicates perfect prediction.
4) Accuracy (Train and Test): Measures the proportion of correct predictions over the total predictions for both training and testing datasets. Displaying the accuracy for both training and testing datasets is crucial for detecting overfitting and underfitting.

IV. Results and Interpretation

A. Data Trend

Monthly trend, as showing in Fig. 2 reveals significant seasonal variations. The peak time of purchase in March is indeed attributed to the Eid of Ramadan. Conversely, a drop in the summer months, particularly June and July, occurs as most families travel, leading to a decrease in purchase flow. Both years almost follow the same trend, indicating consistent seasonal patterns suggesting that consumer behavior and sales during these periods are predictable, which is crucial for planning inventory and marketing strategies.

Similarly, based on Fig. 1 and Fig. 3 the sales trends observed highlight fluctuations during key events and periods of heightened consumer activity. The first significant sales spike occurs around early 2022, coinciding with the lead-up to the Eid al-Adha holiday, a time known for increased purchasing as consumers buy gifts, clothing, and food. Another major peak is seen in March 2023, corresponding with the beginning of Ramadan, when consumers typically engage in extensive shopping for food and gifts in preparation for the holy month. This is reflected in the sharp increase in sales during this period.

Throughout the rest of the year, there are several smaller peaks and troughs that reflect ongoing promotional activities, seasonal sales, and other events influencing consumer behavior. These fluctuations underscore the impact of cultural and religious events on shopping patterns, as well as the effectiveness of sales and marketing strategies in driving consumer demand.

B. Model Coefficients

The coefficients derived from the linear regression model provide a detailed understanding of how various predictors influence purchase frequency. Each predictor's coefficient indicates the nature and strength of its relationship with the target variable, purchase frequency.

1) Recency (0.010727): This positive coefficient indicates that as the days since the last purchase increase, purchase frequency also increases. This counterintuitive result may suggest that customers who haven't purchased recently are starting to buy more frequently, possibly due to re- engagement strategies or changing needs.
2) Monetary (0.001549): The positive coefficient for monetary value shows that higher spending correlates with increased purchase frequency. This is expected, as high- spending customers are more engaged and likely to make frequent purchases, highlighting their economic importance.
3) Order Quantity (-0.491885): The negative coefficient for order quantity implies that customers who buy in larger quantities tend to purchase less frequently. This likely reflects bulk buying behavior, where customers purchase large amounts less often, reducing the need for frequent orders.

Briefly, these insights reveal that high-spending customers and those recently re-engaged tend to purchase more frequently, while bulk buyers purchase less often.

C. Data Modeling

The performance of each model is evaluated based on the various metrics as detailed in Table 3.

Table 3. Model performance metrics

Model	RMSE	MAE	R^2	Train Accuracy	Test Accuracy
Linear Regression	4.6137	2.6805	0.9999	0.9999	0.9999
Decision Tree	186.2625	2.7390	0.8119	1.0000	0.8119
Random Forest	113.7388	4.5349	0.9299	0.9906	0.9299
Gradient Boosting	91.0188	4.5969	0.9551	0.9996	0.9551

Linear Regression shows extremely high accuracy on both training (Train Accuracy = 0.9999) and testing sets (Test Accuracy = 0.9999), with very low error metrics (RMSE = 4.6137, MAE = 2.6805, R^2 = 0.9999). This suggests that the linear relationship assumptions are highly suitable for the dataset, resulting in nearly perfect predictions. However, the simplicity of the model means it might not capture more complex patterns in the data.

The Decision Tree model exhibits perfect accuracy on the training set (Train Accuracy = 1.0), indicating overfitting, as evidenced by the significantly lower accuracy (Test Accuracy = 0.8119) and higher RMSE (186.2625) on the testing set. The MAE for the testing set is 2.7390, and R^2 is 0.8119. This model captures complex patterns in the training data but fails to generalize well to unseen data, making it less reliable for predicting future sales accurately.

Random Forest achieves better generalization compared to Decision Tree, with improved accuracy (Train Accuracy = 0.9906, Test Accuracy = 0.9299) and lower error metrics on the testing set (RMSE = 113.7388, MAE = 4.5349, R^2 = 0.9299). This ensemble method effectively reduces overfitting and enhances predictive performance. The Random Forest model balances complexity and generalization, making it a strong candidate for practical applications.

Gradient Boosting demonstrates the best performance among all models, with the lowest RMSE (91.0188) and highest R^2 (0.9551) on the testing set. The model also shows high accuracy on both training (Train Accuracy = 0.9996) and testing sets (Test Accuracy = 0.9551), and an MAE of 4.5969. This model's sequential learning approach results in superior predictive accuracy, making it highly effective for the sales forecasting task. The Gradient Boosting model's ability to correct its errors iteratively leads to enhanced prediction capabilities, outperforming other models in this study.

In summary, while Linear Regression provides an excellent baseline with nearly perfect predictions, Gradient Boosting offers the best overall performance for generalization to new data. Decision Tree suffers from overfitting, and Random Forest effectively mitigates this issue, offering robust predictions. Among all models, Gradient Boosting is identified as the best model due to its superior performance in terms of RMSE and R^2, indicating its strong predictive power and ability to generalize well to unseen data. This makes it the most suitable model for enhancing sales forecasting accuracy, aiding strategic decision-making in the e-commerce sector.

V. Conclusion

The findings of this study underscore the importance of advanced feature engineering and machine learning algorithms in improving sales forecasting accuracy in the e-commerce sector. By leveraging a comprehensive dataset from a leading household items retailer in Saudi Arabia, key patterns and trends influencing sales performance were identified. The integration of domain-specific features, such as recency, frequency, and monetary value, provided valuable insights into customer behavior and purchasing patterns. The comparison of various machine learning models, including Linear Regression, Decision Tree, Random Forest, and Gradient Boosting, revealed that Gradient Boosting offers the best predictive performance. Incorporating advanced machine learning techniques into e-commerce operations can significantly enhance sales forecasting accuracy, crucial for inventory management, customer satisfaction, and strategic decision-making. The study highlights the potential of combining traditional machine learning models with state-of-the-art algorithms to achieve better forecasting outcomes.

Future work should emphasize the critical role of time series forecasting techniques, given that sales trends are influenced by specific periods of the year, such as holidays and promotional events. Additionally, exploring other applications like customer segmentation and sentiment analysis can offer deeper insights into customer preferences and behaviors, enabling more targeted marketing efforts.

Moreover, future research should consider exploring additional machine learning models, such as neural networks, and integrating other data sources, such as social media trends and economic indicators, to enhance predictive power. These additional data sources can provide a more holistic view of the factors affecting sales, leading to more accurate and robust forecasts.

In conclusion, this study provides a comprehensive method- ology for sales forecasting in the e-commerce sector. The integration of domain-specific features and advanced machine learning algorithms has proven effective in improving prediction accuracy. The findings offer valuable insights to optimize e-commerce operations and drive business growth. By incorporating time series forecasting, exploring new machine learning models, and integrating additional data sources, future research can further advance the field, providing more robust and actionable insights for the e-commerce industry.

Conflict of Interest

The authors declare no conflict of interest.

Author Contributions

Dr. Roxane Mallouhy is the sole author of this paper. She conducted the research, performed the data analysis, and authored the manuscript. The author confirms that all aspects of the work have been carried out independently and approves the final version of the paper.

Acknowledgments

This work was supported by Al Yamama University. I would like to thank the college of engineering and architecture (COEA) for their assistance.

References

[1] Worldwide retail e-commerce sales from 2014 to 2027. (2024). [Online]. Available: https://www.statista.com/statistics/379046/worldwide-retail-e-commerce-sales

[2] eMarketer. (2023). Worldwide ecommerce forecast 2023. [Online]. Available: https://www.emarketer.com/content/worldwide-ecommerce-forecast-2023

[3] D. Herhausen, S. F. Bernritter, E. W. Ngai, A. Kumar, and D. Delen, "Machine learning in marketing: Recent progress and future research directions," *Journal of Business Research*, vol. 170, 114254, 2024.

[4] W. Qian and Y. Wang, "Analyzing e-commerce market data using deep learning techniques to predict industry trends," *Journal of Organizational and End User Computing (JOEUC)*, vol. 36, no. 1, pp. 1–22, 2024.

[5] C. Li, "Commodity demand forecasting based on multimodal data and recurrent neural networks for e-commerce platforms," *Intelligent Systems with Applications*, vol. 22, 200364, 2024.

[6] H. Baek, "A CNN-LSTM stock prediction model based on genetic algorithm optimization," *Asia-Pacific Financial Markets*, vol. 31, no. 2, pp. 205–220, 2024.

[7] O. Kosovan and M. Datsko, “Complex comparison of statistical and econometrics methods for sales forecasting,” in *Proc. the Computational Methods in Systems and Software*, Springer, 2023, pp. 340–355.

[8] P. Rasappan, M. Premkumar, G. Sinha, and K. Chandrasekaran, “Transforming sentiment analysis for e-commerce product reviews: Hybrid deep learning model with an innovative term weighting and feature selection,” *Information Processing & Management*, vol. 61, no. 3, 103654, 2024.

[9] Global ecommerce sales growth report. (2024). [Online]. Available: https://www.shopify.com, Shopify, 2024.

[10] Asana. Sales operations planning. (2024). [Online]. Available: https://asana.com/resources/sales-operations-planning

[11] Retalon. E-commerce demand forecasting. (2024). [Online]. Available: https://retalon.com/blog/e-commerce-demand-forecasting

[12] M. Sharma, V. Sharma, and R. Kapoor, “Study of e-commerce and impact of machine learning in e-commerce,” in *Empirical Research for Futuristic E-Commerce Systems: Foundations and Applications*. IGI Global, 2022, pp. 1–22.

[13] K. Mamta and S. Sangwan, “Aapidl: An ensemble deep learning-based predictive framework for analyzing customer behaviour and enhancing sales in e-commerce systems,” *International Journal of Information Technology*, vol. 16, no. 5, pp. 3019–3025, 2024.

[14] M. R. Hasan, “Addressing seasonality and trend detection in predictive sales forecasting: A machine learning perspective,” *Journal of Business and Management Studies*, vol. 6, no. 2, pp. 100–109, 2024.

[15] F. Galton, “Regression towards mediocrity in hereditary stature,” *The Journal of the Anthropological Institute of Great Britain and Ireland*, vol. 15, pp. 246-263, 1886.

[16] J. R. Quinlan, “Induction of decision trees,” *Machine Learning*, vol. 1, pp. 81–106, 1986.

[17] L. Breiman, “Random forests,” *Machine Learning*, vol. 45, pp. 5–32, 2001.

[18] J. H. Friedman, “Greedy function approximation: A gradient boosting machine,” *Annals of Statistics*, vol. 29, no. 5, pp. 1189–1232, 2001.

[19] T. Bartz-Beielstein, “Hyperparameter tuning,” in *Online Machine Learning: A Practical Guide with Examples in Python*. Springer, 2024, pp. 125–140.