A Review on E-commerce Sales Forecasting

Dr. U.M.Fernandes Dimlo
HOD, Dept of CSE, Sreyas Institute of Engineering and Technology,
Telangana,India.

Bachu Suma Sri
Dept of CSE, Sreyas Institute of Engineering and Technology,
Telangana,India.

Guntupalli Mokshitha
Dept of CSE, Sreyas Institute of Engineering and Technology,
Telangana,India.

Shiva Prasad Reddy B
Dept of CSE, Sreyas Institute of Engineering and Technology,
Telangana,India.

Masampally Pranav
Dept of CSE, Sreyas Institute of Engineering and Technology,
Telangana,India.

***Abstract*— Accurate sales forecasting is a critical component of e-commerce operations, enabling businesses to make informed decisions about inventory, pricing, and customer engagement. This paper presents an e-commerce sales forecasting system built using Python and Streamlit, focusing on a simplified dataset devoid of external influences such as seasonality or holidays. The dataset consists of attributes containing information like order details, product information and transaction records. The proposed system employs a Random Forest model to predict sales for the next six months and provides analytical insights into total revenue, units sold, sales trends, and product purchase patterns. The system also includes user authentication for secure access, dataset uploading for training, and a report generation feature for actionable insights. Designed with simplicity and functionality in mind, this work demonstrates the potential of accessible tools for small and medium-sized e-commerce businesses to enhance strategic planning and operational efficiency. Future developments aim to incorporate external factors for more robust predictions.**

Keywords—Random Forest, Streamlit, Predictive Analytics, Simplified Dataset, Revenue Analysis, Sales Trends, User Authentication, Report Generation.

I. INTRODUCTION

The rapid expansion of the e-commerce industry has reshaped the global retail environment, making accurate sales forecasting a critical tool for effective decision-making. Forecasting enables businesses to optimize inventory levels, design strategic marketing campaigns, and manage demand fluctuations. However, traditional forecasting approaches, such as moving averages and ARIMA models, often fail to address the intricacies of modern e-commerce systems due to limitations in adaptability and their inability to incorporate complex consumer behavior patterns. This work introduces a simplified yet robust e-commerce sales forecasting system designed to address these gaps. Built on a Python-based platform with a Streamlit interface, the system leverages a straightforward dataset that excludes external factors like seasonality and holidays. The dataset consists of attributes containing information like order details, product information and transaction records. By focusing on intrinsic attributes, the model offers businesses a streamlined solution to predict sales trends without the complexity of external data integration. The proposed system is built around a Random Forest algorithm, selected for its reliability in handling structured datasets and its ability to provide accurate predictive analytics. The system's architecture includes secure user authentication, dataset upload functionality, and automated model training. Key outputs include comprehensive sales insights such as revenue trends, total units sold, monthly and weekly sales patterns, product purchase distributions, and payment method preferences. The model forecasts sales for the next six months, enabling businesses to make informed decisions. The system demonstrates the potential for accessible tools to empower small and medium-sized e-commerce enterprises. Future work includes incorporating real-time data, external factors, and advanced algorithms to enhance forecasting precision and scalability.

OBJECTIVES AND METHODOLOGY

The primary objective of this work is to develop an efficient e-commerce sales forecasting system that provides actionable insights for small and medium-sized businesses. This system employs a simplified dataset comprising attributes such as Order ID, Product ID, Product Category, Quantity, Unit Price, Total Price, Order Date, Customer ID, Payment Type, and Order Status, intentionally excluding external factors such as seasonality or holidays to maintain a focused scope. Using a Python-based Streamlit application, the system integrates user authentication for secure access, dataset uploading for training, and a machine learning model based on Random Forest to predict sales for the next six months. The methodology begins with preprocessing the dataset by cleaning, handling missing values, and normalizing it to improve data quality. The Random Forest model is then trained to identify patterns in historical data, with performance evaluated using metrics like Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE). Once trained, the model generates insights such as total revenue, units sold, monthly and weekly sales patterns, product purchase percentages, and payment method distributions, all presented through interactive visualizations. The system also features a report generation module that allows users to download detailed summaries of analyses and forecasts. Designed for accessibility, the interface ensures an intuitive experience for users, making the system both functional and scalable for future enhancements, including real-time data integration and support for external factors.

II. LITERATURE SURVEY

Sales forecasting is a critical task in the e-commerce industry, where businesses are constantly striving to predict future demand in order to optimize inventory, marketing, and customer engagement strategies. Over the years, various

statistical methods, such as ARIMA (Auto-Regressive Integrated Moving Average), have been traditionally used to forecast sales based on historical data (Box & Jenkins, 1976). However, these methods have limitations when dealing with large, complex datasets that are typical in modern e-commerce environments. With the advent of machine learning, more advanced techniques, including regression models, decision trees, and neural networks, have emerged as powerful tools for sales prediction. For instance, Random Forest, an ensemble learning method, has been successfully applied to capture complex, non-linear relationships in e-commerce sales data, outperforming traditional linear models in several cases (Breiman, 2001). Moreover, recent studies have incorporated additional variables such as advertisement spending, customer demographics, and external market factors, further enhancing prediction accuracy (Chong et al., 2017). Despite the promising results, challenges remain in handling dynamic and noisy data, as well as ensuring model interpretability and scalability for large-scale applications. Thus, ongoing research continues to explore hybrid approaches that combine statistical techniques with machine learning algorithms to address these issues and improve forecasting precision.

Fig. 1: E-commerce Sales Forecasting Logo

Fig. 2: Importance of Forecasting

III. PROPOSED SYSTEM

The proposed system is designed to predict future e-commerce sales based on historical sales data. This system aims to provide an efficient, user-friendly tool for e-commerce businesses to forecast their future sales performance, assisting in inventory planning, marketing strategies, and financial planning. By leveraging machine learning algorithms, particularly the Random Forest model, the system forecasts sales over the next six months, offering detailed insights into revenue, sales trends, and product purchase patterns. The system is simple yet powerful, providing a comprehensive solution through a seamless user interface developed with Streamlit, enabling users to upload their datasets, train the model, and receive predictions and insights.

A. User-Interference Design:

The system's interface is crafted to be intuitive and straightforward, ensuring ease of use for individuals with basic technical knowledge. Upon accessing the system, users are first prompted to sign up or log in. This authentication process is vital for secure access and ensuring that user data remains confidential. After logging in, users are directed to the data upload section, where they can upload their sales data in CSV format. Once the dataset is uploaded, the system preprocesses the data, displaying a preview for confirmation. The user interface features a clean dashboard that presents the processed data along with insights such as total revenue, total units sold, and payment type analysis. These results are presented in visually appealing graphs and charts, offering a clear and understandable view of the sales data.

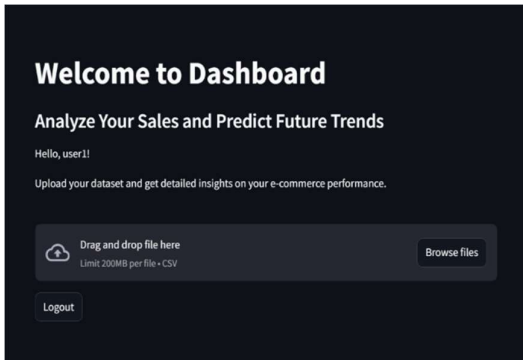


Fig. 3: HomePage

B. Data Preprocessing:

Upon receiving the uploaded dataset, the system begins by performing necessary preprocessing tasks. This includes cleaning the data by handling missing values, removing duplicates, and converting categorical variables, such as Product Category and Payment Type, into numerical values using encoding techniques. Additionally, continuous variables such as Unit Price and Total Price are normalized to maintain consistency across the dataset. This step ensures that the data is in an appropriate format for training the model. Once the data is preprocessed, it is divided into training and testing sets, allowing the system to validate the accuracy of the model after training.

C. Model Training and Sales Prediction:

The heart of the proposed system lies in its sales prediction capability, powered by a Random Forest model. This machine learning model is well-suited for the task, as it can handle complex relationships in the dataset and generate accurate predictions for future sales. The system trains the model using the historical data, focusing on factors such as product category, quantity sold, and total price. After training, the model is used to predict sales for the next six months. These predictions include key metrics like the total revenue for the upcoming months. The model's predictions are presented on the dashboard, offering users valuable insights into future sales trends.

D. Data Analysis and Insights Generation:

In addition to sales predictions, the system provides a range of data analysis and insights to enhance decision-making. These insights include detailed reports on total revenue, total units sold, and a breakdown of sales by product category. The system also generates payment type analysis, showing which payment methods are most commonly used. One of the key features of the system is its ability to display predicted future sales, which can assist businesses in making informed decisions about inventory management, marketing campaigns, and resource allocation. Visualizations of these insights, such as pie charts, bar graphs, and line graphs, make it easy for users to interpret and act on the data.

Fig.4 : Total Revenue

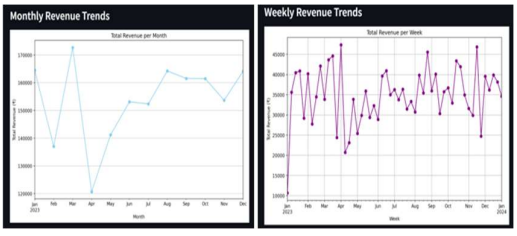


Fig.5: Monthly, Weekly Revenue

E. Evaluation Metrics and Report Generation:

After generating the sales predictions, the system evaluates the performance of the Random Forest model by using standard machine learning metrics such as R-squared, Mean Absolute Error (MAE), and Root Mean Squared Error (RMSE). These metrics give users a clear understanding of how well the model is performing and help identify areas where improvements can be made. Furthermore, the system includes a report generation feature that compiles the analysis, predictions, and evaluation metrics into a downloadable report. This report provides users with a comprehensive summary of their sales forecasting results, which can be saved or shared for further analysis or decision-making.

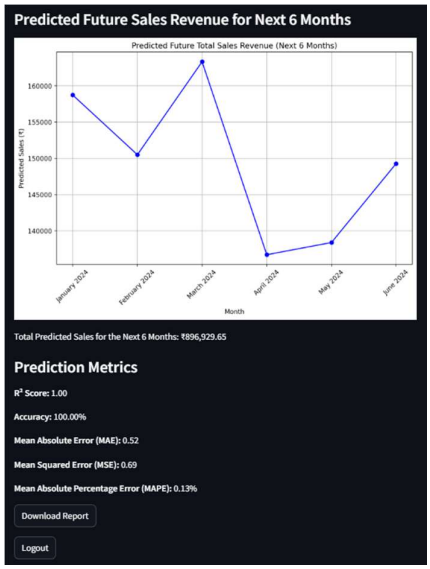


Fig.6 : Prediction for Future Sales

F. System Workflow:

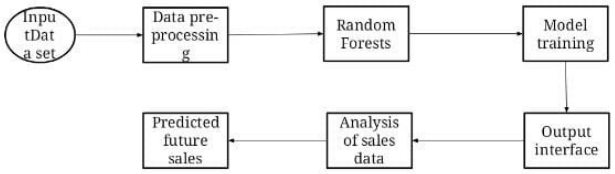
The system operates in a clear, logical flow:

1. **Initialization**: The user logs in and uploads their dataset through the user interface.
2. **Data Preprocessing**: The system cleans and preprocesses the data, making it ready for model training.
3. **Model Training and Prediction**: The Random Forest model is trained and used to predict future sales.
4. **Results Display**: The predicted sales and other insights are displayed on the dashboard.
5. **Evaluation and Report Generation**: The system calculates evaluation metrics and generates a downloadable report for the user.
6. **Logout**: The user can securely log out of the system after completing their tasks.

IV. IMPLEMENTATION

The e-commerce sales forecasting system integrates machine learning, data preprocessing, and a simple graphical user interface to deliver accurate sales predictions. The system begins by allowing users to authenticate via a sign-up and sign-in process, with user credentials stored in CSV files for simplicity. Once logged in, users can upload their e-commerce sales dataset. The dataset is then preprocessed, ensuring that all missing values are handled appropriately and that categorical variables are encoded numerically for compatibility with machine learning algorithms. Following this preprocessing step, a Random Forest model is trained using the processed dataset, enabling the system to predict sales for the next six months. The system provides outputs such as total revenue, units sold, and monthly and weekly sales, along with additional insights like payment type distribution and product purchase percentages. These insights help businesses understand trends and make data-driven decisions based on historical sales data and predictive analysis.

A. Architecture Diagram:



This design processes the uploaded dataset using Python's pandas library for data manipulation. After data preprocessing, the system proceeds with model training. The Random Forest model, known for its ability to handle complex datasets and produce reliable predictions, is chosen for forecasting sales. The model is trained on various features, such as product categories, quantities, unit prices, and order dates. Once trained, the model is used to predict sales for the next six months. The system's interface leverages Streamlit, allowing for seamless interaction between the user and the predictive model. The interface provides real-time visualizations of sales data, such as total revenue, units sold, and predicted sales, helping users interpret trends and future predictions. The system not only calculates and displays these insights but also generates a downloadable report containing the analysis and forecasts, enabling users to make informed business decisions based on the predictions.

B. Test Cases:

Test Case Id	**Description**	**Input**	**Expected Output**	**Remarks**
TC1	User authentication (sign up and login)	Valid user credentials	User is signed up or logged in successfully	Ensure user details are stored and login is validated correctly
TC2	Dataset upload functionality	Valid CSV file with e-commerce sales data	Data uploaded successfully without errors	Ensure the file is correctly formatted and contains no missing values
TC3	Handling missing data in uploaded dataset	CSV file with missing values in certain columns	System handles missing data (e.g.,by filling with default values or asking for input)	Test with incomplete data to ensure proper handling of null values
TC4	Model training process	Preprocessed data with feature selection	Model is trained without errors and accuracy metrics improve over epochs	Verify correct feature engineering and that the training completes
TC5	Monthly sales prediction for next 6 months	Trained model and sales data for the last 6 months	Predicted sales for the next 6 months are output as expected	Ensure the prediction is based on trained model with correct feature inputs

TC 6	Payment type prediction analysis	Data with various payment types (credit card, PayPal, etc.)	Payment type prediction percentages are displayed correctly	Validate prediction logic and output percentages for payment types
TC 7	Sales visualizati -on (monthly and weekly)	Data for multiple months, includin -g daily sales data	Visualized graphs for total sales per month and per week	Ensure the correct format of visualizatio -n (bar charts/line graphs)
TC 8	Model prediction for new data	New CSV file with e-commer -ce sales data	Predicted future sales based on the new data	Test with new data to verify model prediction accuracy for unseen data
TC 9	Report generation and download	Data after model training and predictio -n	A downloadab -le report file with all insights and predictions	Ensure the report is generated and downloadab -le with correct details

V. DISCUSSION

A. Usability and User Experience:

The usability and user experience (UX) of an e-commerce sales forecasting system are pivotal for ensuring that users can easily understand and interpret the insights presented. The system's user interface has been designed to be intuitive and accessible, with clear navigation and easy-to-read visualizations. The dashboard enables users to interact with their dataset and gain valuable insights promptly. The ease of use of this forecasting tool lies in the simple input method: uploading a CSV file with sales data and receiving predictions based on the model's analysis. The results, such as predicted future sales, revenue, and insights on various factors like advertising spend, are displayed in visually appealing charts and graphs.

For users unfamiliar with machine learning, the system provides clear explanations of the predictions and model evaluation metrics, ensuring that they understand the reasoning behind the forecasts. However, challenges may arise when handling large datasets or when the data isn't well-preprocessed, leading to errors or inaccurate results. To enhance usability, future versions of the system could incorporate features like automatic data cleaning or error detection. This would make the system more user-friendly for non-technical users, further improving accessibility.
Moreover, feedback mechanisms, such as tooltips or pop-up guides, could assist users in better understanding how to use the system and interpret the results. By providing clear visual feedback and ensuring quick response times, the system fosters an engaging and efficient user experience, ultimately promoting adoption in a business or academic setting.

B. Data Preprocessing and Model Training:

Data preprocessing is an essential step in any machine learning system, especially when dealing with large e-commerce datasets. In this system, preprocessing ensures that the sales data is clean, consistent, and suitable for model training. It involves tasks like handling missing data, encoding categorical variables, and normalizing numerical features to improve the model's performance. For instance, the training process uses feature engineering to extract meaningful patterns from the raw dataset, including aspects such as monthly sales, and product categories.

Once the data is preprocessed, the model undergoes training using regression techniques that is random forest, which is well-suited for predicting continuous values like sales. The model's accuracy improves over time as it learns from the data and fine-tunes the prediction process. Moreover, the normalization of features ensures that all inputs have a similar scale, preventing the model from being biased toward certain variables.
Additionally, data augmentation techniques, such as rotating or transforming features, can be employed to improve the robustness of the model. This step ensures that the model can generalize well to various market conditions and handle real-world variations in the data. A key component of the model's success is its ability to process historical sales data efficiently, predicting future sales with high accuracy based on past trends and external factors like advertising spend.

C. Applications in Real World Scenarios:

The applications of e-commerce sales forecasting extend across various industries and provide invaluable insights to businesses seeking to improve their sales strategies and customer satisfaction. In retail and e-commerce, this system can predict future sales based on historical data and advertising efforts, helping businesses plan their inventory and marketing budgets effectively. The model's ability to forecast sales revenue helps companies optimize their supply chain, ensuring that they maintain sufficient stock without overstocking.

In marketing, the system's analysis of the relationship between advertising spend (e.g., TV, social media,

newspapers) and sales can help businesses allocate their marketing budgets more effectively. By predicting which channels will yield the highest return on investment (ROI), companies can tailor their campaigns for maximum impact. Furthermore, in sectors like logistics and warehousing, accurate sales forecasts allow companies to anticipate demand and streamline their distribution strategies, leading to cost savings and improved efficiency. The e-commerce sales forecasting model can also be integrated into business intelligence tools, offering real-time insights that inform decision-making processes across departments. For instance, customer service teams can use these predictions to anticipate high-demand periods and optimize staffing levels. Financial analysts can leverage these forecasts to adjust revenue projections and set realistic financial targets.

Moreover, the ability to predict future sales provides businesses with an edge in customer satisfaction. By ensuring that popular products are in stock and that marketing campaigns align with consumer interests, companies can better meet customer expectations and enhance the overall shopping experience. As the system continues to evolve, it can integrate with AI-powered recommendation engines to predict customer preferences and drive personalized marketing strategies.

These real-world applications demonstrate how e-commerce sales forecasting can transform business operations, enabling companies to make data-driven decisions, optimize resources, and enhance customer experiences.

VI. CONCLUSION AND FUTURE SCOPE

The development of this e-commerce sales forecasting system marks a significant step towards empowering small and medium-sized businesses with accessible, data-driven tools for strategic planning. By utilizing a simplified dataset that excludes external factors like seasonality and holidays, the system focuses on delivering reliable predictions while minimizing computational complexity. It employs a Random Forest algorithm to forecast sales for the next six months, supported by features such as user authentication, dataset upload, and report generation, all integrated into a user-friendly Python-based Streamlit application. Key insights, including total revenue, units sold, monthly and weekly sales trends, product purchase percentages, and payment type distributions, provide comprehensive analytics to enhance decision-making processes.

The work successfully demonstrates the effectiveness of machine learning techniques in e-commerce sales forecasting by offering scalability, adaptability, and actionable insights. Its simplicity ensures that businesses with limited technical expertise can leverage predictive analytics to optimize inventory, streamline operations, and improve customer satisfaction. Moreover, the modular design of the system allows for seamless integration of additional features in future iterations.

However, like any initial implementation, the system has its limitations. The exclusion of external factors such as economic conditions, competitor activities, and seasonal trends restricts the scope of its predictions, particularly in dynamic or highly competitive markets. Additionally, the current reliance on manual data uploads prevents real-time forecasting capabilities, which are critical for responding to sudden market changes, flash sales, or demand surges. Addressing these challenges opens avenues for substantial future enhancements.

Future developments could include the integration of external variables such as economic indicators, social media trends, and promotional activities to improve the robustness and applicability of the model. Real-time data processing capabilities could enable businesses to make instantaneous decisions during high-demand periods, while the incorporation of advanced algorithms, including ensemble models and explainable AI techniques, could further refine prediction accuracy. Enhancing the user interface with interactive dashboards and real-time visualizations would improve the system's usability and appeal. Furthermore, extending the system to handle multi-channel sales data, regional segmentation, and SKU-level forecasting could broaden its utility for diverse business needs.

In conclusion, this work highlights the transformative potential of predictive analytics in the e-commerce industry. By offering an efficient and user-centric solution, the system provides a foundation for businesses to optimize operations and stay competitive in a rapidly evolving marketplace. As future enhancements are implemented, this tool has the potential to evolve into a comprehensive decision-support system, driving growth, innovation, and efficiency across the e-commerce sector.

VII. REFERENCES

[1] F. Karimova, A survey of e-commerce recommender systems, Eur. Sci. J., vol. 12, no. 34, pp. 75-89, 2016.
[2] Z. Huo, Sales prediction based on machine learning, in 2021 2nd Int. Conf. E-Commerce Internet Technol. (ECIT), IEEE, 2021.
[3] M.V. Rajesh and S. Rao Chintalapudi, A review on applications of machine learning in e-commerce, Adv. Appl. Math. Sci., vol. 20, no. 11, pp. 2831-2841, 2021.
[4] B.M. Pavlyshenko, Machine-learning models for sales time series forecasting, Data, vol. 4, no. 1, pp. 15, 2019.
[5] B. Lakshmanan, S.N. Palaniappan, R. Vivek, and K. Viswanathan, Sales demand forecasting using LSTM network, in Artif. Intell. Evol. Comput. Eng. Syst., Springer Singapore, 2020.
[6] X. Qi, et al., From known to unknown: Knowledge-guided transformer for time-series sales forecasting in Alibaba, arXiv preprint arXiv:2109.08381, 2021.
[7] Y. Ensafi, et al., Time-series forecasting of seasonal items sales using machine learning–A comparative analysis, Int. J. Inf. Manag. Data Insights, vol. 2, no. 1, pp. 100058, 2022.
[8] Z. Zhang and C. Nuangjamnong, The impact factors toward online repurchase intention: A case study of Taobao e-commerce platform in China, Int. Res. E-J. Bus. Econ., vol. 7, no. 2, pp. 35-56, 2022.
[9] H.-F. Chen and S.-H. Chen, How website quality, service quality, perceived risk and customer satisfaction affects repurchase intention? A case of Taobao online shopping, in Proc. 10th Int. Conf. E-Education, E-Business, E-Management and E-Learning, 2019.
[10] C. Chengjie and Q. Wei, Taobao user purchase behavior prediction and feature analysis based on ensemble learning, in Proc. 2023 IEEE Int. Conf. e-Business Eng. (ICEBE), IEEE, 2023.
[11] Y. Dai and J. Huang, A sales prediction method based on LSTM with hyper-parameter search, J. Phys.: Conf. Ser., vol. 1756, no. 1, IOP Publishing, 2021.
[12] D. Pliszczuk, et al., Forecasting sales in the supply chain based on the LSTM network: The case of furniture industry,

2021.
[13] A.B. Wardak and J. Rasheed, Bitcoin cryptocurrency price prediction using long short-term memory recurrent neural network, Eur. J. Sci. Technol., vol. 38, pp. 47-53, 2022.
[14] Y. Qi, et al., A deep neural framework for sales forecasting in e-commerce, in Proc. 28th ACM Int. Conf. Inf. Knowl. Manag., 2019.