Journal of Innovations in Computer Science and Trends in IT, (April 2026), Vol 3, Issue1, ISSN 3048-4707

Intelligent Framework for analyzing Customer Purchasing Behavior using Machine Learning

Harsh Arora[1] and Meenu[2]
[1]Associate Professor, Department of Computer Application,
Banarsidas Chandiwala Institute of Information Technology, New Delhi
harsh@bciit.ac.in
[2]Assistant Professor, Department of Computer Application,
Institute of Innovation in Technology and Management, New Delhi
drmeenu.knsl@gmail.com

***Abstract*: The rapid expansion of electronic commerce platforms has produced massive volumes of customer interaction data, creating significant opportunities for understanding purchasing behaviour. Accurate analysis of customer activities helps organizations improve recommendation systems, customer retention, personalization, and sales planning. Conventional analytical approaches often struggle to process large and complex datasets efficiently. This study proposes an intelligent machine learning framework for analysing customer behaviour in e-commerce environments. Supervised learning techniques are applied to transactional and behavioural datasets to identify meaningful consumer patterns. Experimental observations demonstrate that machine learning approaches provide higher predictive accuracy for customer segmentation, purchase prediction, and churn identification when compared with traditional statistical approaches.**

Keywords: E-commerce platforms, machine learning methods, classification models, clustering approaches, recommendation engines, and customer behaviour analytics.

I. Introduction

The development of e-commerce technology has transformed the relationship between organizations and consumers by supporting instant communication, personalized services, and seamless online purchasing experiences. In each of their interactions, which include product browsing, placing orders, review of items and product feedback customers generate what is in fact very valuable digital info. We look at this info very closely to identify what customers prefer and to also predict what they will buy in the future. We look at how consumers interact with online platforms, what factors play a role in their decisions, and how their interests change over time. Machine learning we see as very efficient for the task of finding out hidden patterns in large complex data sets without the need for explicit programming. Also, from past research we see that

Journal of Innovations in Computer Science and Trends in IT, (April 2026), Vol 3, Issue1, ISSN 3048-4707

machine learning plays a very key role in e commerce analytics. in depth behavioral analysis. Clustering methods like K-Means are a mainstay for customer segmentation and also see large use of Algorithms such as Decision Trees and Support Vector Machines for purchase prediction [1]. Also we see growth in the use of deep learning in recommendation systems and sentiment analysis. That said many current studies use narrow data sets or report results from a single learning technique. In this study we aim to improve on that which we do by looking at many machine learning models through the use of practical ecommerce variables for better overall behavioral analysis.

II. Research Objectives

This research focuses on understanding customer behaviour in e-commerce systems by applying machine learning techniques to data analysis. The study is designed to recognize shopping trends and anticipate customer decisions using different classification methods. It also aims to categorize users according to their purchasing preferences and online interaction patterns. Customer segmentation is further utilized to assist businesses in creating customized marketing campaigns and promotional activities. Moreover, the performance of several machine learning models is evaluated to identify the algorithm that provides the highest accuracy and reliability in predicting e-commerce customer behaviour.

III. Methodology

Data preprocessing plays a significant role in improving the quality and reliability of datasets used for machine learning analysis.

Missing data in the dataset is reduced by way of appropriate imputation techniques which in turn helps to reduce bias and also to prevent information loss. We use mean imputation for numerical attributes and mode imputation for categorical attributes which in turn helps to maintain statistical consistency. The dataset we use in this work is of customer information which we have anonymized from e commerce platforms [2].

These data include customer IDs, demographic info, which we also look at browser behavior, purchase frequency, what products were bought, which product categories they fall in, and also the time spent by the user on the platform. Before we get into the modeling stage we do in depth

Journal of Innovations in Computer Science and Trends in IT, (April 2026), Vol 3, Issue1, ISSN 3048-4707

preprocessing which includes getting rid of inconsistent data, normalizing numbers and encoding of categories. We start the machine learning process by collecting data from related sources which are customer databases, transaction logs, and user interaction records. We clean and transform the collected info during pre-processing to make sure it is in a format that machines can work with. Then we choose which algorithms to use based on the type of problem we are trying to solve and our research goals. The dataset we worked with was split into two parts: training and validation. This helps our models learn from the data without getting too good at fitting the training data. We check how our models do using things, like accuracy, precision, recall and F1-score.

We keep adjusting the model settings and how we prepare the data until we get results. Once we have a model that works well we use it for tasks like predicting what customers will buy and suggesting products [3].

To get our data ready we use a technique called Min-Max normalization. This makes all the numerical data have the scale, which helps our models learn better. We also change data into numbers that a computer can understand using one-hot encoding.

All these steps help reduce mistakes in the data make it more consistent and make our predictions better.

Journal of Innovations in Computer Science and Trends in IT, (April 2026), Vol 3, Issue1, ISSN 3048-4707

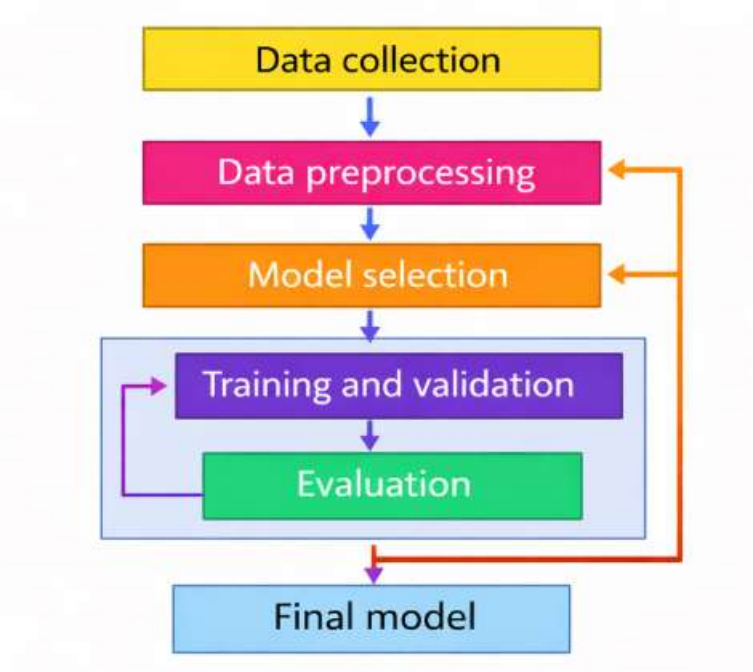

Figure 1: Workflow of the Machine Learning Model Development Process

Figure 1 shows how a machine learning system works from start to finish, emphasizing the iterative and sequential processes that go into creating a successful predictive model.

Various sources, such as customer databases, e-commerce platforms (such as eBay and Shopify), transaction records, and user interaction logs, are utilized to obtain the raw data. The data is prepared for further processing using preprocessing. This stage manages missing values, removes irrelevant or orthogonal data, normalizes the numerical and categorical information, and encodes it in a suitable format. Enhanced dataset accuracy and reliability are achieved through these operations [4]. Thus, the data that has been processed becomes organized and suitable for application in machine learning algorithms. Machine learning models are selected based on the nature of their problem, research objectives and characteristics after they have been processed through preprocessing.

For proper model assessment, the dataset is first sorted into training and validation sets before training begins. The model learns through the available data during the training process, and validation is used to fine-tune parameters and reduce overfitting. After the training, performance

Journal of Innovations in Computer Science and Trends in IT, (April 2026), Vol 3, Issue1, ISSN 3048-4707

metrics such as accuracy, precision, recall, F1-score, and error rate are used to evaluate models' performance. Several evaluation metrics are also available. If these obtained results don't work, the preprocessing techniques used or algorithms chosen can be modified to improve overall performance. This procedure is carried out repeatedly until the desired level of accuracy and reliability is attained, and then the most effective model is utilized in real-world scenarios, such as anticipating customer behavior in e-commerce settings.

Min-Max normalization is employed in feature scaling, which involves converting numerical values to a fixed range for the first time. This technique helps machine learning algorithms improve their performance on distance and gradient-based learning models [5] by automatically encoding categorical features into numerical form. Collectively, these preprocessing methods improve the data quality, reduce noise, and increase the predictive accuracy and effectiveness.

IV. Machine Learning Techniques

This research applies various machine learning techniques to analyze and forecast customer behavior in e-commerce systems. The use of logistic Regression as a baseline method for determining whether he is likely to purchase elicits the explanation that buying outcomes depend on the relationship between independent variables. Decision Tree algorithms are utilized to identify the characteristics that have a significant impact on consumer choices. Feature values are used to create hierarchical structures in the model, which offers an unambiguous approach to providing information and explanations for conducting behavioral analysis [6]. Rules-based decision support and feature significance can be achieved through the use of Decision Trees. The use of Random Forest, an ensemble learning technique, is employed to enhance the accuracy and stability of predictions. By training multiple decision trees on randomly chosen subsets of data, this technique combines them. By combining the features, overfitting is reduced and generalization performance is improved, allowing the model to capture complex behavioral patterns more effectively [7]. Support Vector Machine is utilized as a classification technique in this study. SVM establishes a superior decision boundary that permits the greatest degree of separation between customer types.

Journal of Innovations in Computer Science and Trends in IT, (April 2026), Vol 3, Issue1, ISSN 3048-4707

High-dimensional datasets are a powerful tool for classifying customer behaviour in online shopping environments, making this method highly efficient.

V. Results and Discussion

The proposed machine learning models are experimentally evaluated using standard performance indicators including accuracy, precision, recall, and F1-score. The algorithms with the highest accuracy rates in terms of prediction are Random Forest (91.2%), Support Vector Machine (88.6%), Decision Tree (85.3%), and Logistic Regression (82.5%). Moreover, the high level of performance of Random Forest shows that it can capture both complex and non-linear relationships within customer datasets. K-Means clustering, along with other classification techniques, can be used to categorize consumers into groups such as high-value customers, price-sensitive buyers, low-engagement users, and occasional shoppers. These consumer groups provide valuable information for individualized marketing, advertising and recommendation systems. These findings demonstrate that ensemble learning techniques are superior to individual classification methods, owing to their reliability, reduced overfitting, and improved generalization capabilities. By utilizing these data, companies can enhance customer experience, optimize pricing strategies, and design more targeted advertising campaigns. The proposed framework has practical implications for demand forecasting, churn prediction, personalized recommendations, and digital marketing optimization. While these results are encouraging, the study does have some limitations.... The use of high-quality datasets is crucial for effective prediction, and processing large-scale data requires a considerable amount of computational power. Ethical management of customer information and data privacy are also significant challenges in ecommerce analytics.

The Scikit-learn library and Python programming language are employed for preprocessing, training, and evaluation procedures. Random Forest is deemed the most balanced operator in terms of accuracy, precision, recall and F1-score. Correlation analysis shows strong positive associations between various behavioural variables, including platform usage, browsing activity, time spent, and purchase frequency/transaction value [8]. However, demographic factors such as age have weaker correlations with purchasing behaviour. The survey also highlights the importance of

Journal of Innovations in Computer Science and Trends in IT, (April 2026), Vol 3, Issue1, ISSN 3048-4707

maintaining privacy and social networks. These observations suggest that customer decisions are more likely to be influenced by behavioural attributes than demographic variables alone.

Table 1: Comparative Analysis of Machine Learning Models Used

Algorithm	Accuracy	Precision	Recall	F1-Score
Logistic Regression	82.5 %	0.81	0.80	0.80
Decision Tree	85.3 %	0.84	0.83	0.83
SVM	88.6 %	0.87	0.86	0.86
Random Forest	**91.2 %**	**0.90**	**0.91**	**0.90**

Table 1 shows the comparative analysis of machine learning models which is depicted above. The linear relationships between the datasets numerical features are found and measured using correlation analysis that is shown in figure 2. With a correlation coefficient of 0.71, the results show a strong positive relationship between the number of pages viewed and the amount of time spent on the platform, suggesting that users who browse more pages typically stay on the e-commerce site longer. In a similar way, there is a strong positive correlation of 0.68 between purchase frequency and transaction value, indicating that regular buyers are more likely to generate larger transaction amounts. However, with a correlation value of 0.21, the relationship between age and purchase frequency is comparatively weak, suggesting that demographic factors have little effect on consumer behaviour. All things considered, these findings demonstrate that behavioural characteristics are more significant predictors of customer behaviour in e-commerce analytics than demographic characteristics, showing stronger interrelationships [9].

Journal of Innovations in Computer Science and Trends in IT, (April 2026), Vol 3, Issue1, ISSN 3048-4707

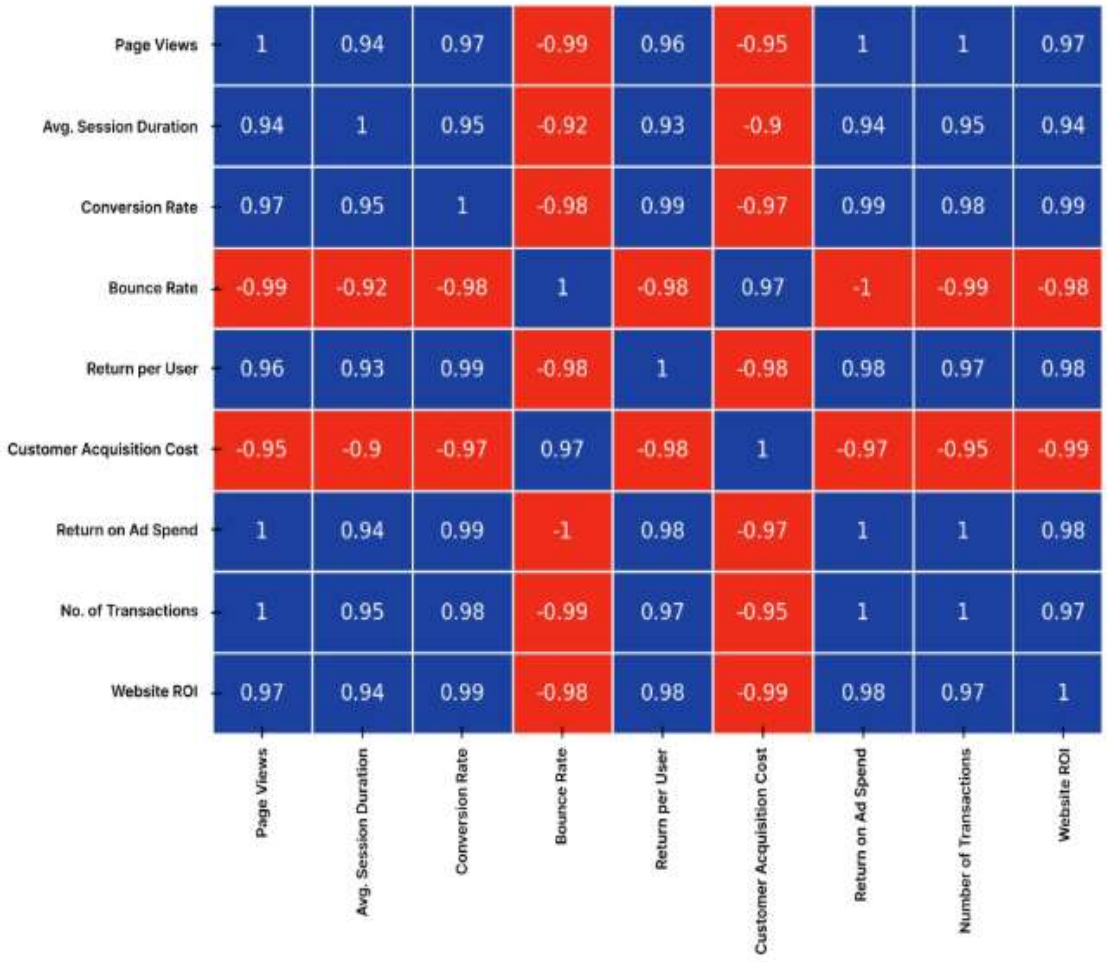

Figure 2: Correlation Heatmap of e-commerce Store Metrics

Journal of Innovations in Computer Science and Trends in IT, (April 2026), Vol 3, Issue1, ISSN 3048-4707



Figure 3: Word Cloud of Customer Reviews

Figure 3 is a word cloud visualization created from customer reviews of a dataset pertaining to restaurants or e-commerce, where the size of each word indicates how frequently it appears in the reviews [10]. Larger font sizes are used for prominent words like “food,” “service,” “great,” “friendly,” “nice,” “staff,” and “atmosphere,” suggesting that patrons often highlight these features when discussing their experiences. This implies that the main elements affecting customer satisfaction are the caliber of the food, the effectiveness of the service, the conduct of the employees, and the general atmosphere. Customer sentiment is generally positive, as evidenced by terms like "excellent," "recommended," "delicious," "lovely," and "good," which show positive opinions of the offerings and services.

VI. Conclusion and Future Scope

The study concludes that machine learning techniques provide a scalable and efficient framework for analysing customer behaviour in e-commerce systems. Classification and clustering algorithms generate valuable insights into purchasing patterns, customer preferences, and behavioural trends.

Journal of Innovations in Computer Science and Trends in IT, (April 2026), Vol 3, Issue1, ISSN 3048-4707

The proposed framework improves predictive accuracy and supports data-driven decision-making for enhancing customer satisfaction and business growth.

Future research may focus on integrating sentiment analysis from customer reviews, enabling real-time behavioural prediction, and applying reinforcement learning approaches for intelligent recommendation systems. Advanced deep learning models such as Long Short-Term Memory networks can also be explored for analysing sequential customer interactions and improving predictive performance in dynamic e-commerce environments.

References

1. B. Lebichot, T. Verhelst, Y. -A. Le Borgne, L. He-Guelton, F. Oblé and G. Bontempi, "Transfer Learning Strategies for Credit Card Fraud Detection," in IEEE Access, vol. 9, pp. 114754-114766, 2021, doi: 10.1109/ACCESS.2021.3104472.
2. E. Manohar, P. Jenifer, M. S. Nisha, and B. Benita, "A Collective Data Mining Approach to Predict Customer Behaviour," 2021 Third International Conference on Intelligent Communication Technologies and Virtual Mobile Networks (ICICV), 2021, pp. 1310-1316, doi: 10.1109/ICICV50876.2021.9388558.
3. V. Shrirame, J. Sabade, H. Soneta and M. Vijayalakshmi, "Consumer Behaviour Conference on Electronics, Computing and Communication Technologies (CONECCT), Bangalore, India, 2020, pp. 1-6, doi: 10.1109/CONECCT50063.2020.9198562.
4. B. Lebichot, T. Verhelst, Y. -A. Le Borgne, L. HeGuelton, F. Oblé and G. Bontempi, "Transfer Learning Strategies for Credit Card Fraud Detection," in IEEE Access, vol. 9, pp. 114754-114766, 2021, doi: 10.1109/ACCESS.2021.3104472.
5. S. Shahriar, A. R. Al-Ali, A. H. Osman, S. Dhou and M. Nijim, "Machine Learning Approaches for EV Charging Behaviour: A Review," in IEEE Access, vol. 8, pp. 168980-168993, 2020, doi: 10.1109/ACCESS.2020.3023388.
6. S. Shahriar, A. R. Al-Ali, A. H. Osman, S. Dhou and M. Nijim, "Machine Learning Approaches for EV Charging Behaviour: A Review," in IEEE Access, vol. 8, pp. 168980-168993, 2020, doi: 10.1109/ACCESS.2020.3023388.
7. D. Damkevala, R. Lunavara, M. Kosamkar and S. Jayachandran, "Behaviour Analysis using Serverless Machine Learning," 2019 6th International Conference on Computing for Sustainable Global Development (INDIA Com), New Delhi, India, 2019, pp. 1068-1072.
8. F. D. Pereira, E. H. T. Oliveira, D. Fernandes, and A. Cristea, "Early Performance Prediction for CS1 Course Students using a Combination of Machine Learning and an Evolutionary Algorithm," 2019 IEEE 19th International Conference on Advanced Learning Technologies (ICALT), Macie, Brazil, 2019, pp. 183-184, doi: 10.1109/ICALT.2019.00066.
9. Z. Zhao, H. Tang, X. Zhang, X. Qu, X. Hu, and J. Lu, “Classification of children with autism and typical development using eye-tracking data from face-to-face conversations: Machine learning

Journal of Innovations in Computer Science and Trends in IT, (April 2026), Vol 3, Issue1, ISSN 3048-4707

model development and performance evaluation," Journal of Medical Internet Research, vol. 23, no. 8, p. e29328, 2021.

10. J. Pfeiffer, T. Pfeiffer, M. Meisner, and E. Weis, "Eye-tracking-based classification of information search behaviour using machine learning: evidence from experiments in physical shops and virtual reality shopping environments," Information Systems Research, vol. 31, no. 3, pp. 675–691, 2020.