IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Machine Learning-Based Customer Behavior Analysis and Purchase Prediction in E-Commerce: A Comparative Framework

[1]Ms. Shreya Santosh Garte, [2]Dr. Priti V. Pancholi

[1]Computer Science and Engineering, M. S. Bidve Engineering College Latur, India
[2]Computer Science and Engineering, M. S. Bidve Engineering College Latur, India

Abstract: The growth of e-commerce has created large volumes of behavioral and transactional data that can be used to anticipate customer purchasing decisions. This paper presents a practical machine learning framework for customer behavior analysis and purchase prediction using demographic attributes, browsing activity, previous purchase information, and product-related features. The study compares Logistic Regression, Decision Tree, Support Vector Machine, and Random Forest classifiers within a structured pipeline consisting of data cleaning, categorical encoding, normalization, feature selection, model training, and evaluation. The documented experimental implementation uses an 80:20 training–testing split and evaluates models through accuracy, precision, recall, and F1-score. Comparative observations indicate that Random Forest provides the strongest overall balance of predictive capability, robustness, and resistance to overfitting, while Decision Tree offers high interpretability, Logistic Regression provides a computationally efficient baseline, and Support Vector Machine remains competitive for complex decision boundaries. The findings are consistent with published e-commerce studies in which ensemble models achieve strong purchase-intention prediction. The proposed framework can support customer segmentation, personalized marketing, product recommendation, sales forecasting, and more efficient allocation of marketing resources. The paper also identifies data quality, class imbalance, changing customer preferences, and model interpretability as important considerations for deployment.

Index Terms **-** customer behavior analysis, purchase prediction, machine learning, e-commerce, Random Forest, predictive analytics, customer intelligence

I. Introduction

Digital businesses continuously collect customer information from online transactions, e-commerce platforms, mobile applications, browsing sessions, and customer relationship management systems. These data streams contain signals about preferences, engagement, past transactions, and product interests. Converting such signals into reliable purchase-intention estimates is important because firms can use them to prioritize high-potential customers, personalize communication, optimize recommendation systems, and allocate marketing resources more efficiently. The supplied project identifies customer behavior analysis as the process of examining customer actions and decision patterns to understand purchasing behavior and emphasizes that conventional surveys and descriptive statistical methods are increasingly limited by the scale and complexity of digital data.

Machine learning provides a practical alternative because classification algorithms can learn relationships between customer attributes and historical purchase outcomes. Purchase prediction is commonly formulated as a binary classification problem in which a model estimates whether a customer or session is likely to result in a purchase. Prior research has shown that clickstream, session, and behavioral features can provide useful predictive information. Sakar et al. [1] demonstrated that online shopping intention can be predicted from session and page-view information, while more recent work has emphasized feature engineering, class balancing, and ensemble learning for stronger predictive performance [2].

This paper converts the project framework into a concise journal-style study focused on four widely used supervised classifiers: Logistic Regression (LR), Decision Tree (DT), Support Vector Machine (SVM), and Random Forest (RF). The objective is not to introduce a highly complex architecture, but to establish an understandable and reusable machine learning pipeline for customer analytics. The main contribution is a structured comparison of these models, with emphasis on predictive usefulness, interpretability, computational practicality, and business relevance.

II. Related Work

Purchase-intention modelling has evolved from descriptive customer segmentation toward real-time and session-level prediction. Moe [3] showed that navigational clickstream behavior can distinguish browsing and buying patterns, establishing the value of behavioral traces in online retail. Sakar et al. [1] later proposed a real-time system that combined page-view and session features with Random Forest, SVM, and neural models. Their findings showed that combining clickstream and session information improves predictive capability, highlighting the importance of behavioral context rather than relying on a single customer attribute. Ensemble learning has received particular attention because purchase decisions often depend on nonlinear interactions between many variables. Random Forest aggregates predictions from multiple decision trees and is designed to reduce variance and overfitting [4]. Baati and Mohsil [5] applied Random Forest to online shopper intention and reported significantly higher accuracy and F1-score than the compared Naïve Bayes and C4.5 techniques. Satu and Islam [2] evaluated extensive preprocessing, feature transformation, balancing, and feature-selection strategies and found Random Forest to be a stable classifier, with a reported best accuracy of 92.39% and F-score of 0.924 on their engineered benchmark subset. More recent studies continue to investigate ensembles and richer behavioral representations, indicating that the research direction remains active [6], [7]. Despite these advances, a practical gap remains between sophisticated research systems and models that are easy to implement in typical academic or small-business settings. Many implementations emphasize predictive accuracy but provide limited attention to the balance among data preparation, model interpretability, scalability, and actionable business use. The present framework addresses this gap by comparing classical models using a transparent end-to-end process that can be reproduced with standard Python machine learning libraries.

Table 1. Selected literature on online purchase-intention prediction

Study	Data / Approach	Models	Main finding
Sakar et al. [1]	Session and clickstream information	RF, SVM, MLP, LSTM	Behavioral and session features support real-time intention prediction.
Baati & Mohsil [5]	Online shopper behavior	NB, C4.5, RF	RF produced significantly stronger accuracy and F1-score.
Satu & Islam [2]	UCI online shopper dataset with feature engineering	Multiple classifiers	RF was stable; best reported accuracy 92.39%.
Tokuç & Dag [7]	Real-world clickstream representations	LR, SVC, tree and boosting models	Hybrid behavioral representations improved purchase prediction.
Cao-Van [6]	Online Shoppers dataset	RF, XGBoost, LightGBM, GB, SVM	Modern ensembles and imbalance handling remain effective directions.

III. Problem Formulation

Let each customer or browsing instance be represented by a feature vector x = [x1, x2, …, xn], where the attributes may include demographic information, browsing duration, pages visited, previous purchases, product category, website interactions, and related customer signals. The target variable y is binary, where y = 1 denotes a purchase and y = 0 denotes no purchase. The learning objective is to estimate a classification function f(x) that predicts the purchase outcome for unseen instances while maintaining sufficient generalization for changing customer behavior.

The framework considers three practical requirements. First, the model should achieve reliable discrimination between likely buyers and non-buyers. Second, the preprocessing pipeline should transform heterogeneous customer attributes into a consistent modelling representation. Third, the final predictions should remain useful to business users, which means that model performance must be considered together with interpretability and computational cost.

IV. Proposed Methodology

The proposed methodology follows a conventional supervised learning workflow. Customer data are first collected and inspected for missing values, duplicate records, inconsistent formats, and obvious outliers. Categorical attributes are encoded into numerical form, and continuous variables are normalized or standardized where required. Feature selection is then used to retain variables with meaningful predictive information and to reduce unnecessary model complexity. The processed data are divided into training and testing subsets, after which the four classifiers are trained and compared using a common evaluation protocol.

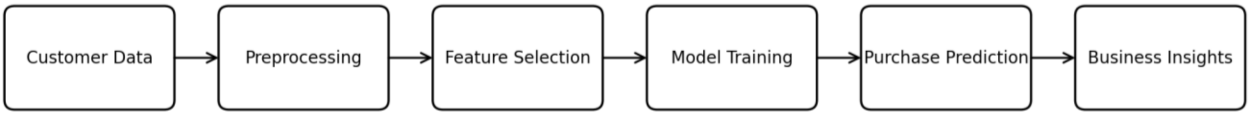

Figure 1. Machine learning workflow for customer behavior analysis and purchase prediction.

4.1 Data Representation and Preprocessing

The project dataset is structured around customer demographic details, browsing behavior, and purchase history. Representative fields include Customer ID, Age, Gender, Annual Income, Browsing Time, Pages Visited, Previous Purchases, Product Category, and Purchase Status. Customer ID is treated as an identifier rather than a predictive feature. Numerical attributes are checked for scale differences, while categorical variables are encoded before model fitting. Missing or duplicate records should be removed or imputed according to the amount and pattern of missingness. These steps are important because unreliable or inconsistent input data directly affect prediction quality.

Table 2. Representative customer attributes used by the framework

Attribute	Type	Analytical role
Age	Numeric	Demographic purchasing pattern
Gender	Categorical	Customer profile variable
Annual Income	Numeric	Purchasing capacity indicator
Browsing Time	Numeric	Engagement intensity
Pages Visited	Numeric	Depth of session interaction
Previous Purchases	Numeric	Historical loyalty / repeat-buying signal
Product Category	Categorical	Product-interest context
Purchase Status	Binary target	0 = no purchase, 1 = purchase

4.2 Classification Algorithms

Logistic Regression is used as an interpretable probabilistic baseline. It estimates purchase probability through the logistic function and is computationally efficient when relationships are approximately linear. Decision Tree constructs hierarchical decision rules and is particularly useful when business users need to understand how variables lead to a classification. However, a single tree can overfit training data when its depth is not controlled.

Support Vector Machine identifies a separating hyperplane that maximizes the margin between classes. With nonlinear kernels, SVM can represent more complex customer decision boundaries, but it may require careful tuning and becomes computationally demanding on large datasets. Random Forest is an ensemble of decision trees trained with bootstrap sampling and random feature selection. By aggregating many decorrelated trees, RF generally reduces the instability and overfitting associated with a single tree [4]. This makes it well suited to customer behavior data containing nonlinear interactions and mixed predictive effects.

4.3 Evaluation Measures

The project evaluates predictive performance using Accuracy, Precision, Recall, and F1-score. Accuracy measures the proportion of correctly classified instances. Precision measures the proportion of predicted buyers who actually purchase, whereas Recall measures the proportion of actual buyers correctly identified by the model. F1-score is the harmonic mean of Precision and Recall and is particularly useful when the positive purchase class is smaller than the non-purchase class. For highly imbalanced datasets, future implementations should also report AUROC or precision–recall AUC because overall accuracy alone can be misleading.

Table 3. Evaluation metrics

Metric	Formula	Interpretation
Accuracy	(TP + TN) / (TP + TN + FP + FN)	Overall correctness
Precision	TP / (TP + FP)	Reliability of positive purchase predictions
Recall	TP / (TP + FN)	Ability to identify actual buyers
F1-score	2PR / (P + R)	Balance between precision and recall

V. Experimental Setup

The documented implementation uses Python in a Jupyter Notebook environment with common data-science libraries such as Pandas, NumPy, Scikit-learn, and Matplotlib. The processed customer dataset is divided into 80% training data and 20% testing data. Logistic Regression, Decision Tree, SVM, and Random Forest are trained on the same modelling dataset and evaluated using the four metrics described above. The use of a consistent split allows direct model comparison under the same data conditions.

For a stronger future implementation, stratified cross-validation should be preferred to a single split, especially when purchases form a minority class. Class weighting or SMOTE can also be considered to reduce bias toward the majority class; SMOTE is a widely used synthetic oversampling method for imbalanced classification [8]. The current framework remains deliberately simple so that it can be implemented and understood without advanced computational infrastructure.

VI. Results and Discussion

The experimental observations documented in the project show that the four classifiers provide different trade-offs. Random Forest achieved the best overall prediction performance and was selected as the primary model because it handled complex customer patterns effectively while reducing overfitting. SVM produced competitive classification results but required greater computational effort. Decision Tree provided transparent and understandable decision rules, while Logistic Regression offered a fast baseline with moderate predictive capability. These observations are summarized in Table 4.

Table 4. Comparative outcome of the documented implementation

Model	Predictive outcome	Interpretability	Computational demand	Overall role
Logistic Regression	Moderate / baseline	High	Low	Efficient benchmark
Decision Tree	Moderate	Very high	Low	Explainable rules
SVM	Competitive	Medium–low	Medium–high	Complex boundaries
Random Forest	Highest overall	Medium	Medium	Preferred model

The observed importance of browsing time, page visits, annual income, and previous purchases is logically consistent with the structure of purchase-intention prediction. High browsing activity indicates customer engagement, while past purchases capture repeat-buying behavior and customer loyalty. Income can provide an approximate indication of purchasing capacity, although it should be interpreted carefully because demographic variables may introduce fairness and privacy concerns. In operational use, feature importance should therefore be reviewed not only for predictive power but also for ethical and regulatory suitability.

The preference for Random Forest is also consistent with published evidence. Baati and Mohsil [5] found Random Forest significantly stronger than their compared classifiers for online shopper intention. Satu and Islam [2] reported a best Random Forest accuracy of 92.39%, F-score of 0.924, and AUROC of 0.975 under different feature-engineering configurations. These external values should be treated as literature benchmarks rather than as numerical results of the present project, because the supplied project documentation records the comparative ranking of models but does not preserve per-model metric values. The agreement between the documented ranking and published evidence nevertheless strengthens the practical case for Random Forest as a strong baseline for this application.

From a business perspective, prediction scores can be converted into actionable customer segments. Customers with high purchase probability may receive personalized recommendations or timely offers, while customers with moderate probability may be targeted with engagement-oriented content. Aggregate predictions can also support demand forecasting and inventory planning. However, prediction should not be treated as a substitute for customer understanding; models can degrade as preferences, product ranges, website design, and market conditions change. Periodic retraining and monitoring are therefore necessary.

VII. Practical Implications

The proposed framework is appropriate for e-commerce firms seeking a low-complexity predictive analytics solution. A retailer can integrate purchase probabilities into customer relationship management dashboards, recommendation engines, or campaign-management systems. Because the framework uses common classification algorithms, implementation costs can remain relatively low and models can be retrained as new data arrive. Random Forest is particularly useful when prediction quality is the priority, whereas Decision Tree or Logistic Regression may be preferable when business stakeholders require maximum transparency.

The framework also supports marketing efficiency. Rather than exposing all customers to identical promotions, firms can rank customers according to predicted purchase probability and focus resources on segments with the greatest expected response. This approach can reduce unnecessary advertising expenditure and improve conversion opportunities. At the same time, organizations should apply privacy-by-design practices, minimize unnecessary personal data, and assess whether demographic variables create unfair targeting effects.

VIII. Limitations and Future Work

The main limitation of the present paper is that the original project record provides comparative conclusions but does not contain the exact numerical score of each classifier. For this reason, unsupported values have not been reconstructed. Future work should preserve the complete experiment log, including the dataset version, model hyperparameters, random seed, confusion matrices, per-class metrics, and cross-validation results. This would improve reproducibility and support statistical comparison between models.

Future research can also examine Gradient Boosting, XGBoost, LightGBM, neural networks, and sequence models for richer clickstream data. Deep learning models such as LSTM networks can capture temporal relationships across browsing actions [1], while recent work suggests that hybrid session and event representations can improve purchase prediction [7]. Explainable AI methods such as SHAP may be added to make ensemble predictions more transparent. Finally, real-time deployment should study concept drift so that the model can adapt when customer behavior changes over time.

IX. Conclusion

This paper presented a practical machine learning framework for customer behavior analysis and purchase prediction in e-commerce. The framework integrates data preprocessing, feature selection, classification, and performance evaluation using Logistic Regression, Decision Tree, SVM, and Random Forest. The documented comparison identifies Random Forest as the strongest overall model because of its robustness, ability to capture nonlinear customer patterns, and reduced tendency to overfit compared with an individual decision tree. The findings are consistent with established and recent purchase-intention studies. By translating behavioral data into purchase probabilities, the framework can support personalized marketing, recommendation, customer segmentation, and data-driven business decisions. Future work should focus on reproducible numerical evaluation, cross-validation, imbalance handling, explainability, and real-time adaptation.

References

[1] C. O. Sakar, S. O. Polat, M. Katircioglu, and Y. Kastro, "Real-time prediction of online shoppers' purchasing intention using multilayer perceptron and LSTM recurrent neural networks," Neural Computing and Applications, vol. 31, no. 10, pp. 6893–6908, 2019. https://doi.org/10.1007/s00521-018-3523-0

[2] M. S. Satu and S. F. Islam, "Modeling online customer purchase intention behavior applying different feature engineering and classification techniques," Discover Artificial Intelligence, vol. 3, art. 36, 2023. https://doi.org/10.1007/s44163-023-00086-0

[3] W. W. Moe, "Buying, searching, or browsing: Differentiating between online shoppers using in-store navigational clickstream," Journal of Consumer Psychology, vol. 13, nos. 1–2, pp. 29–39, 2003.

[4] L. Breiman, "Random forests," Machine Learning, vol. 45, no. 1, pp. 5–32, 2001. https://doi.org/10.1023/A:1010933404324

[5] K. Baati and M. Mohsil, "Real-Time Prediction of Online Shoppers' Purchasing Intention Using Random Forest," in Artificial Intelligence Applications and Innovations, IFIP AICT, vol. 583, pp. 43–51, Springer, 2020. https://doi.org/10.1007/978-3-030-49161-1_4

[6] K. Cao-Van, "Forecasting Online Shoppers' Purchase Intentions Using Machine Learning and Ensemble Modeling," in Context-Aware Systems and Applications, LNICST, vol. 678, pp. 210–227, Springer, 2026. https://doi.org/10.1007/978-3-032-24399-7_15

[7] A. A. Tokuç and T. Dag, "Predicting User Purchases From Clickstream Data: A Comparative Analysis of Clickstream Data Representations and Machine Learning Models," IEEE Access, vol. 13, pp. 43796–43817, 2025.

[8] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique," Journal of Artificial Intelligence Research, vol. 16, pp. 321–357, 2002. https://doi.org/10.1613/jair.953

[9] C. Cortes and V. Vapnik, “Support-vector networks,” Machine Learning, vol. 20, pp. 273–297, 1995. https://doi.org/10.1007/BF00994018

[10] F. Pedregosa et al., “Scikit-learn: Machine Learning in Python,” Journal of Machine Learning Research, vol. 12, pp. 2825–2830, 2011.

[11] G. Suchacka and G. Chodak, “Using association rules to assess purchase probability in online stores,” Information Systems and e-Business Management, vol. 15, no. 3, pp. 751–780, 2017. https://doi.org/10.1007/s10257-016-0329-4

[12] University of California, Irvine, “Online Shoppers Purchasing Intention Dataset,” UCI Machine Learning Repository, 2018. https://doi.org/10.24432/C5F88Q