Published by :
https://www.ijert.org/
An International Peer-Reviewed Journal

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 05 , May - 2026

Predicting User Purchases from Clickstream Data Using Machine Learning Models

D. Jaya Soniya
Assistant Professor

Amara Dathri Lakshmi Anuhya, Vinjamuri Pavan Mallesh, Pralayakaveri Anil, Gudivada Sai Rithwika
Student
Dept. of CSE-Data Science
St. Ann's College of Engineering & Technology, Chirala, Andhra Pradesh, India

Abstract — **Predicting user purchases from clickstream data is a challenging task owing to the dynamic and complex nature of online user behavior. Accurate prediction models are essential for e-commerce platforms to understand customer intent and support data-driven decision-making. This paper proposes a machine learning-based approach for efficient and reliable purchase prediction using clickstream data collected from a real-world multi-category e-commerce store. The system analyzes user interactions—including page views, product clicks, search queries, and cart activities—to capture behavioral patterns and identify purchase intent. Tree-based algorithms, namely Decision Tree, Random Forest, and LightGBM, are employed to handle high-dimensional and non-linear data effectively. A hybrid feature representation combining session-level statistics with recent user actions is utilized to enhance prediction performance. Experimental evaluation demonstrates improved classification performance in terms of accuracy, AUC-ROC, and G-Mean metrics. The proposed system provides a scalable and interpretable solution for purchase intent prediction, enabling businesses to optimize marketing strategies, enhance user experience, and increase conversion rates**.

Keywords — clickstream data; purchase prediction; machine learning; LightGBM; Random Forest; feature engineering; e-commerce

I. INTRODUCTION

Clickstream data refers to the sequence of actions performed by users while browsing a website or online application. In e-commerce platforms, every interaction—such as product clicks, page views, searches, cart additions, and purchases—is recorded as a digital event. These interactions collectively form clickstream data, which provides detailed insights into user behavior and browsing patterns. Researchers have established that such data plays a crucial role in understanding customer intent and predicting purchasing behavior by structuring user activities into meaningful analytical tasks [1].

Clickstream data is highly valuable for analyzing user navigation paths and identifying behavioral patterns across sessions. Each record typically includes attributes such as user ID, session ID, timestamp, product details, and type of action performed. Businesses utilize this information to improve website design, personalize user experiences, and optimize marketing strategies. However, due to the massive volume and complexity of clickstream data, manual analysis is not feasible. Therefore, advanced machine learning techniques are widely employed to efficiently process and extract meaningful insights from clickstream datasets [5], [7].

Traditional data analysis methods struggle to handle the large and complex patterns inherent in clickstream data. Machine learning provides an effective solution by analyzing user interactions and learning hidden patterns from historical data. Models can thus accurately predict whether a user is likely to complete a purchase during a given browsing session. The proposed system leverages LightGBM, Random Forest, and Decision Tree classifiers, combined with a hybrid feature representation strategy, to improve prediction accuracy and operational efficiency.

The remainder of this paper is organized as follows. Section II surveys related work. Section III describes the proposed methodology. Section IV presents the system design. Section V details the implementation. Section VI discusses testing. Section VII reports experimental results. Section VIII concludes the paper and outlines future directions.

II. LITERATURE SURVEY

Predicting user purchase intent using clickstream data has attracted considerable research attention. Cirqueira et al. [1] presented a comprehensive survey of customer purchase behavior prediction in e-commerce, highlighting the importance of session-based feature engineering. Moe and Fader [2] demonstrated that dynamic conversion behavior at e-commerce sites could be modeled through probabilistic frameworks, providing early evidence of the value of behavioral data for purchase prediction.

He and Garcia [3] addressed the class imbalance problem commonly encountered in purchase prediction datasets, where purchasing sessions constitute a minority class. Their work on learning from imbalanced data has influenced subsequent studies that employ oversampling and cost-

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 05 , May - 2026

sensitive learning techniques. Zavali et al. [4] utilized clickstream data to reveal distinct consumer segments, demonstrating the utility of unsupervised approaches as a precursor to supervised purchase prediction.

Liu et al. [5] conducted a comparative analysis of machine learning and deep learning models for customer behavior prediction, finding that gradient boosting methods consistently achieved competitive performance on large-scale datasets. Zolna et al. [6] proposed sequential user behaviour modelling for purchase prediction, emphasizing the importance of capturing temporal dependencies in clickstream sequences. Most recently, Tokuç and Dag [7] presented a systematic comparative study of clickstream data representations and machine learning models, concluding that hybrid representations combining session-level aggregates with recent action sequences yield superior results. Chen and Guestrin [8] introduced XGBoost, a scalable tree boosting system that established a benchmark for gradient boosting approaches in classification tasks.

Despite these advances, several gaps remain. Many existing models rely on basic features and ignore deeper behavioral signals such as navigation path diversity and session duration dynamics. Additionally, low prediction accuracy in conventional models such as decision trees and support vector machines on large-scale data motivates the adoption of advanced boosting frameworks. The present work addresses these gaps by employing a hybrid feature representation and applying LightGBM alongside ensemble methods to improve prediction reliability.

III. PROPOSED METHODOLOGY

A. System Overview

The proposed system aims to predict whether a user will make a purchase based on clickstream data collected from an e-commerce platform. The methodology encompasses data collection and preprocessing, feature extraction and engineering, machine learning model training, and performance evaluation. Important behavioral features—including session duration, number of page views, and product interactions—are used as predictors. Techniques such as data cleaning, feature selection, and hyperparameter tuning are systematically applied to optimize model performance.

B. Dataset

The dataset used in this project is the eCommerce Behavior Data from a Multi-Category Store, publicly available on Kaggle (REES46 dataset) [15]. It contains user interaction records collected from a real-world e-commerce platform and is widely employed for analyzing customer behavior and purchase prediction. Each record captures attributes including event type, session ID, user ID, product ID, category, brand, price, and timestamp.

C. Data Preprocessing

Raw clickstream data often contains noise, duplicate records, and incomplete sessions that can degrade model performance. The preprocessing pipeline addresses these issues through the following steps: (i) duplicate removal and elimination of invalid records; (ii) bot detection by filtering anomalous sessions with extreme event rates; (iii) session truncation after the first purchase event to prevent label leakage; (iv) real-time simulation by excluding the last few page views to mimic live prediction conditions; and (v) session filtering to remove sessions containing fewer than three actions. These measures collectively improve data quality and ensure reliable model training.

D. Feature Engineering and Selection

Feature engineering transforms raw clickstream records into structured representations that capture user behavioral patterns. Three complementary session representations are constructed. The Aggregated Representation summarizes session-level statistics such as total event count, session duration, product diversity, and pricing information. The Last-N Actions Representation encodes the most recent N user actions to capture short-term purchase intent. The Hybrid Representation combines aggregated session features with last-N action features to simultaneously capture global session patterns and recent browsing behavior. Following feature construction, correlation analysis, model-based feature importance ranking, and dimensionality reduction are applied to select the most informative variables. Table I lists the primary features employed in the system.

TABLE I. KEY FEATURES USED IN THE PROPOSED SYSTEM

Feature	Type	Description
Total Views	Behavioral	Total number of pages visited by the user
Product Clicks	Behavioral	Number of distinct products viewed
Session Duration	Behavioral	Total time spent during the browsing session
Add-to-Cart Actions	Behavioral	Indicates immediate purchase interest
Search Queries	Behavioral	Products actively searched by the user
Total Events	Behavioral	Total number of actions performed in the session

E. Data Splitting and Normalization

The dataset is partitioned using a 90/10 train-test split strategy, with cross-validation employed to ensure model reliability and mitigate overfitting. Numerical features are normalized using Min-Max Scaling to transform values to the [0, 1] range and Standardization to center and scale features using their mean and standard deviation. These transformations ensure that features are processed on a comparable scale by the machine learning models.

F. Machine Learning Models

Three tree-based classification algorithms are selected for purchase prediction. The Decision Tree classifier learns hierarchical decision rules from clickstream features using

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 05 , May - 2026

the Gini Index or Information Gain criteria, producing an easily interpretable model. The Random Forest algorithm constructs an ensemble of decision trees through bootstrap aggregation, combining predictions by majority voting to reduce overfitting and improve generalization. LightGBM (Light Gradient Boosting Machine) employs leaf-wise tree growth with gradient boosting, offering fast training speed, efficient handling of high-dimensional data, and superior classification accuracy on large-scale clickstream datasets.

G. Hyperparameter Tuning

Hyperparameter optimization is performed to maximize model performance. Table II summarizes the key hyperparameters tuned for each model.

TABLE II. HYPERPARAMETERS USED FOR MODEL OPTIMIZATION

Parameter	Description	Models Affected
Learning Rate	Controls the speed of model weight updates	LightGBM
Max Depth	Limits tree depth to reduce overfitting	Decision Tree, Random Forest
Number of Leaves	Determines tree structural complexity	LightGBM
Number of Estimators	Number of trees in the ensemble	Random Forest, LightGBM
Min. Samples Split	Minimum samples required to split a node	Decision Tree

After tuning, all three models demonstrate improved classification performance with higher accuracy, AUC-ROC, and G-Mean scores for predicting user purchase behavior.

IV. SYSTEM DESIGN

The system architecture is organized into five primary modules that interact to deliver end-to-end purchase prediction. The Data Collection Module gathers raw clickstream records from e-commerce platforms, capturing user interactions such as page views, product clicks, cart additions, and purchase events. The Data Preprocessing Module cleans the dataset by removing duplicates, handling missing values, and filtering invalid sessions. The Feature Engineering Module converts raw clickstream data into informative behavioral features representing user session patterns. The Model Training Module applies Decision Tree, Random Forest, and LightGBM classifiers to the processed feature set. The Prediction Module uses the trained model to classify each user session as a likely purchase or non-purchase event.

The database layer stores and manages clickstream records, session attributes, and model outputs in a structured relational format, ensuring data integrity, fast retrieval, and scalability. Security controls and access restrictions are implemented to protect user data. Model outputs and historical prediction logs are retained to support continuous performance monitoring and improvement.

V. IMPLEMENTATION

A. Software and Hardware Requirements

The system is implemented in Python 3.10 on Windows, macOS, or Linux. Core libraries include Scikit-learn for Decision Tree and Random Forest classifiers, LightGBM for gradient boosting, Pandas and NumPy for data processing, Hyperopt for hyperparameter tuning, and Matplotlib and Seaborn for visualization. A Streamlit-based web interface enables interactive exploration of predictions. Hardware requirements include a minimum of an Intel Core i5 processor and 8 GB RAM; a GPU is optional for accelerating training on large datasets.

B. Implementation Pipeline

The implementation follows a modular pipeline. Raw clickstream data is loaded from CSV files and subjected to deduplication, missing-value imputation, and session filtering. Session-level aggregated features are then computed, and the target variable (purchase: 1 / no purchase: 0) is derived from the presence of a purchase event within each session. The dataset is split into training (90%) and test (10%) subsets. LightGBM, Random Forest, and Logistic Regression classifiers are instantiated, trained, and evaluated using accuracy and AUC-ROC metrics. The best-performing model is serialized using Joblib for deployment. A representative excerpt of the core training and evaluation code is provided below:

```
features =
df.groupby('session_id').agg({'event_type':
'count',
    'price': ['mean','sum'], 'product_id':
'nunique'})
lgb_model = lgb.LGBMClassifier()
lgb_model.fit(X_train, y_train)
print('AUC:', roc_auc_score(y_test, y_prob))
```

VI. TESTING

The testing phase validates the correctness and reliability of each system module. Unit testing verifies individual components including data collection, preprocessing, feature engineering, and model training. Functional testing confirms that the system correctly predicts user purchases from clickstream inputs. Performance testing assesses model training speed and prediction latency, particularly under large dataset conditions. Evaluation testing validates that metrics such as accuracy, precision, recall, F1-score, and AUC-ROC are correctly computed. Table III summarizes the test cases executed and their outcomes.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 05 , May - 2026

TABLE III. TEST CASES FOR THE CLICKSTREAM PURCHASE PREDICTION SYSTEM

Test ID	Module	Description	Status
TC01	Data Collection	Clickstream dataset loads correctly without errors	Pass
TC02	Preprocessing	Missing values and duplicates are handled correctly	Pass
TC03	Feature Engineering	Behavioral features are generated accurately	Pass
TC04	Model Training	Decision Tree trains and produces predictions	Pass
TC05	Model Training	Random Forest trains and produces predictions	Pass
TC06	Model Training	LightGBM trains faster and yields accurate output	Pass
TC07	Prediction	Purchase / no-purchase output is generated correctly	Pass
TC08	Evaluation	Accuracy, Precision, Recall, F1-score computed	Pass

VII. RESULTS AND DISCUSSION

Experimental evaluation of the proposed system demonstrates that all three machine learning models—Decision Tree, Random Forest, and LightGBM—are successfully trained and produce meaningful purchase predictions from clickstream data. Among the three feature representation strategies, the Hybrid Representation achieves the highest prediction performance, outperforming both the Aggregated Representation and the Last-N Actions Representation. This result confirms that jointly encoding overall session statistics with recent user actions captures complementary behavioral signals that neither representation can provide independently [7].

LightGBM consistently demonstrates superior efficiency and predictive accuracy compared to the Decision Tree and Random Forest baselines, attributed to its leaf-wise growth strategy and built-in handling of high-dimensional sparse features. The system correctly identifies high-intent users likely to complete a purchase and provides probability estimates that can be used to prioritize marketing interventions. When no high-intent users are detected, the system returns a zero purchase likelihood output alongside strategic recommendations for user re-engagement.

Model evaluation metrics including accuracy, precision, recall, F1-score, AUC-ROC, and G-Mean are computed across all models and feature representations. The evaluation confirms that hyperparameter tuning further enhances classification performance, yielding a more balanced trade-off between precision and recall. The Streamlit-based interface allows interactive input of session features and real-time display of purchase probability, making the system accessible to non-technical business users.

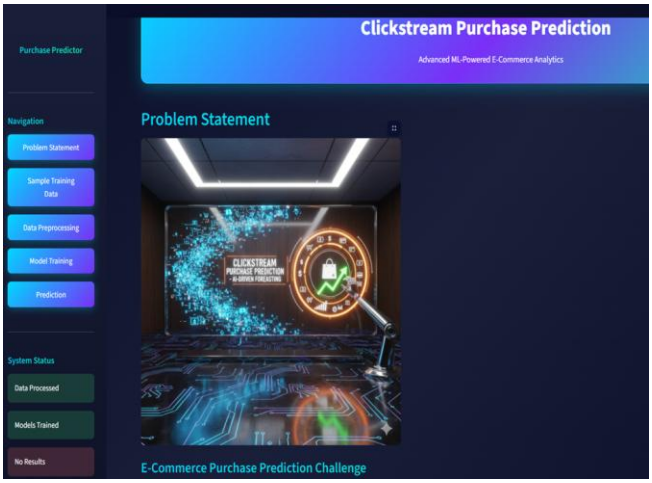

Figure 1. Figure showing Streamlit Application

VIII. CONCLUSION

This paper presented a machine learning-based system for predicting user purchases from clickstream data in e-commerce environments. The proposed approach employs Decision Tree, Random Forest, and LightGBM classifiers in conjunction with a hybrid feature representation that combines session-level aggregated statistics and recent user action sequences. Experimental results demonstrate that the hybrid representation achieves the highest prediction performance, and that LightGBM provides the best balance of efficiency and accuracy among the evaluated models [7], [8].

The work also highlights key challenges such as class imbalance and the necessity of meaningful feature selection for building reliable predictive models [3]. By leveraging a real-world multi-category e-commerce dataset and advanced machine learning techniques, the system provides a practical and scalable solution for purchase intent prediction. Integration of behavioral insights with robust predictive models enables businesses to optimize marketing strategies, deliver personalized recommendations, and increase conversion rates in modern e-commerce environments [1].

Future work will explore deep learning approaches—including recurrent neural networks and attention-based architectures—to model sequential dependencies in clickstream data more effectively. Real-time streaming integration, cross-platform behavioral data fusion, and coupling the prediction model with dynamic recommendation engines are also identified as promising directions for extending the proposed system.

REFERENCES

[1] D. Cirqueira, M. Hofer, D. Nedbal, M. Helfert, and M. Bezbradica, "Customer purchase behavior prediction in e-commerce," in Proc. Int. Workshop New Frontiers in Mining Complex Patterns, Springer, 2020.

[2] W. W. Moe and P. S. Fader, "Dynamic conversion behavior at e-commerce sites," Management Science, vol. 50, no. 3, pp. 326–335, 2004.

[3] H. He and E. A. Garcia, "Learning from imbalanced data," IEEE Trans. Knowl. Data Eng., 2009.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 05 , May - 2026

[4] M. Zavali, E. Lacka, and J. de Smedt, "Shopping hard or hardly shopping: Revealing consumer segments using clickstream data," IEEE Trans. Eng. Manag., vol. 70, no. 4, 2023.

[5] D. Liu, H. Huang, H. Zhang, X. Luo, and Z. Fan, "Enhancing customer behavior prediction in e-commerce: A comparative analysis of machine learning and deep learning models," Applied Computational Engineering, 2024.

[6] R. Zolna et al., "Sequential user behavior modeling for purchase prediction," Expert Systems with Applications, 2018.

[7] A. Aylin 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, 2025.

[8] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in Proc. ACM SIGKDD, 2016.

[9] H. Wang et al., "Tree-based models for customer behavior analysis in e-commerce," Electronics, 2023.

[10] J. Gan et al., "Gradient boosting approaches for purchase intent prediction," Computing and Informatics, 2021.

[11] M. Hendriksen et al., "Feature importance analysis for purchase intent prediction in e-commerce," Decision Support Systems, 2022.

[12] M. Saarela and S. Jauhiainen, "Comparison of feature importance measures as explanations for classification models," Social Network Analysis and Mining, 2021.

[13] D. Karl, "Forecasting e-commerce consumer returns: A systematic literature review," Management Review Quarterly, 2024.

[14] H. He and E. A. Garcia, "Learning from imbalanced data," IEEE Trans. Knowl. Data Eng., 2009.

[15] M. Kechinov, "E-commerce behavior data from a multi-category store," Kaggle Dataset, 2019.