IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 4 February 2025, accepted 19 February 2025, date of publication 5 March 2025, date of current version 14 March 2025.

*Digital Object Identifier 10.1109/ACCESS.2025.3548267*

RESEARCH ARTICLE

# Predicting User Purchases From Clickstream Data: A Comparative Analysis of Clickstream Data Representations and Machine Learning Models

**A. AYLİN TOKUÇ[1,2] AND TAMER DAG[3], (Member, IEEE)**
[1]Department of Computer Engineering, Kadir Has University, Fatih, 34083 Istanbul, Türkiye
[2]Valinor AI, SE9 4HA London, U.K.
[3]College of Engineering and Technology, American University of the Middle East, Egaila 54200, Kuwait

Corresponding author: A. Aylin Tokuç (20181102001@stu.khas.edu.tr)

**ABSTRACT** Predicting purchase events from e-commerce clickstream data is a critical challenge with significant implications for optimizing marketing strategies and enhancing customer experience. This study addresses this challenge by systematically evaluating and comparing multiple data representations – aggregated session attributes, recent user actions, and hybrid combinations – which bridges gaps in the existing literature and demonstrates the superiority of hybrid approaches. Unlike prior research, which typically focuses on single representations, our approach combines aggregated session-level summaries with granular, sequential user actions to capture both long-term and short-term behavioral patterns. Through comprehensive experimentation, we compared multiple machine learning models, including LightGBM, decision trees, gradient boosting, SVC, and logistic regression, using real-world e-commerce clickstream data. Notably, the hybrid representation with LightGBM achieved superior predictive performance, significantly outperforming alternative methods. Feature importance analysis revealed key factors influencing purchase likelihood, such as time since the last event, session duration, and product interactions. This study provides actionable insights into real-time marketing interventions by demonstrating the practical utility of hybrid data representations and efficient tree-based models. Our findings offer a scalable and interpretable framework for e-commerce platforms to enhance purchase predictions and optimize marketing strategies.

**INDEX TERMS** Clickstream data, customer behavior modeling, data representations, feature importance, gradient boosting, e-commerce, lightgbm, machine learning, model selection, purchase prediction.

## I. INTRODUCTION

The broad adoption of e-commerce has transformed the traditional shopping paradigm. This transition has prompted companies to invest heavily in predictive analytics, particularly in areas such as intent prediction (predicting what a customer is likely to do next), purchase prediction (predicting when a customer is likely to make a purchase), and churn prediction (predicting when a customer is likely to stop using a service), to enhance profitability and gain a competitive edge. Among these, purchase prediction is particularly notable for its potential to optimize marketing strategies such as personalized promotions and dynamic pricing. Accurately forecasting the likelihood of a purchase can enable e-commerce platforms to make data-driven decisions and increase conversion rates and customer satisfaction.

Clickstream data, capturing users' online interactions, is a valuable resource for predicting purchase behavior. However, its volume, high dimensionality, and inherent class imbalance – where purchase events represent only a small fraction of user sessions – pose significant challenges. Traditional approaches to purchase prediction typically rely

The associate editor coordinating the review of this manuscript and approving it for publication was Mu-Yen Chen.

on either aggregated session-level statistics or sequential representations of users' actions. Although the aggregated statistics summarize the overall session behavior, they may overlook the immediacy of recent user actions. Conversely, sequential representations capture short-term dynamics but lack the broader context of session-level attributes.

Although the use of clickstream data for intent prediction and conversion modeling has been widely studied, there remains a significant gap in understanding the impact of different data representations on the accuracy of purchase prediction. Despite their potential, the hybrid representations that integrate these complementary perspectives remain underexplored. Most studies focus exclusively on session aggregation or sequential modeling, limiting their ability to capture the nuanced interplay between long-term and short-term behaviors. This gap hinders the development of predictive models that can effectively balance the accuracy, interpretability, and scalability of real-world e-commerce applications. To address this limitation, our study systematically evaluated hybrid representations that combine aggregated session attributes with granular user actions. By leveraging the strengths of both approaches, we aim to provide a comprehensive understanding of how data representations affect predictive performance.

Some studies have applied Deep learning models such as LSTMs and Transformers, which require large computational resources and often sacrifice interpretability. By contrast, this study focuses on tree-based and classical models, which offer a better balance between predictive power, efficiency, and explainability. Furthermore, we rigorously evaluate the model using multiple metrics, including precision-recall (PR) curves, to ensure a comprehensive assessment of performance under class imbalance. In addition, we conducted a feature importance analysis to identify the critical factors. These insights contribute to a broader understanding of user behavior modeling and can inform the development of more effective real-time targeting strategies.

Our contributions are as follows:

- We introduce a novel framework for representing clickstream data that combines long-term session trends with immediate user actions, thus enabling a more comprehensive understanding of purchase behavior.
- We conducted a comprehensive comparison of several machine learning models, including LightGBM, decision trees, gradient boosting, random forests, linear SVC, and logistic regression, to predict purchase likelihood using three distinct data representations: aggregated, flattened recent actions, and hybrid.
- We provide actionable insights through feature importance analysis, highlighting the key behavioral and temporal factors that drive purchase decisions.

Experimental results on real-world e-commerce data demonstrate that hybrid representations significantly outperform single-representation approaches, with the LightGBM model achieving the highest predictive accuracy.

## II. RELATED WORK

Purchase prediction in e-commerce is a significant area of research because of its direct impact on marketing strategies and customer experience. Researchers have explored various data representations and machine learning algorithms to capture and predict user purchasing behavior. This section reviews existing approaches, highlights their limitations, and situates this study within the broader literature.

### A. DATA REPRESENTATION METHODS

Accurately representing the clickstream data is critical for reliable purchase predictions. Clickstream data encapsulate user interactions on websites, including clicks, page views, and search behaviors. Traditional models often rely on aggregated session statistics or sequential representations of user actions, whereas recent studies have explored hybrid approaches to leverage the strengths of both.

A conceptual framework by Cirqueira et al. [1] emphasizes the importance of decomposing customer behavior into comprehensive components to better understand purchase decisions. The framework identifies key tasks such as predicting buying sessions, purchase decisions, and customer intent, which inform the design of predictive models. Building on this, our study compares multiple data representations–aggregated session-level attributes, recent user actions, and hybrid combinations–to capture both long-term behavioral patterns and short-term decision-making cues.

#### 1) SESSION-LEVEL AGGREGATION METHODS

These methods aggregate user activity across all sessions, summarizing behaviors in terms of features such as total session duration, number of pages viewed, and average time per page. This approach captures high-level engagement patterns and is widely used in e-commerce analyses. For example, Moe and Fader [2] leveraged session attributes to model dynamic conversion behavior using variables such as views and time spent on site to understand purchase likelihood. Similarly, Wen et al. [3] derived click counts and session dwell time as well as temporal features to represent multi-behavioral trendiness. Zavali et al. [4] represented user sessions with counts (e.g., past visits, number of pages viewed) and categorical variables (e.g., device type, is cart opened) to segment their user base. However, the aggregation approach may oversimplify user behavior by overlooking fine-grained, action-level details.

#### 2) RECENT ACTIVITY METHODS

These methods emphasize the user's latest actions within a session, capturing the immediate behaviors that may precede a purchase. Sequential modeling methods, such as Long Short-Term Memory (LSTM) networks and recurrent neural networks (RNNs), have been adopted to effectively handle these temporal dynamics. Liu et al. [5] utilized clickstream

data over a 5-day sliding time window to predict the purchases on the next day with an LSTM. Koehn et al. [6] demonstrated the value of sequential modeling for real-time predictions. This approach enables real-time prediction before a session concludes but requires models capable of handling variable-length sequences.

#### 3) HYBRID METHODS

By combining session-level summaries with recent user actions, hybrid methods attempt to leverage both overarching behavioral patterns and immediate signals that are indicative of purchase intent. Bigon et al. [7] modeled various events, such as view, detail, add, remove, buy, and click, to classify user sessions, employing classifiers that capture both the sequence of actions and aggregated statistics. Similarly, Torkashvand et al. [8] introduced frameworks that merge content- and behavior-based features using deep learning architectures. Chakraborty et al. [9] combined aggregated user behavior data from a longer historical window with user activities over a short window in the current session. This dual approach provides a richer representation of user behavior. However, its reliance on neural networks makes interpretability and scalability challenging in real-time e-commerce applications.

Another approach is to represent clickstream sequences as vectors to capture the sequential and temporal aspects of the user behavior. By transforming user interactions into fixed-length vectors, these representations enable machine learning models to process complex behavioral patterns effectively. For example, Al Amoudi et al. [10] propose a vector-based encoding framework to improve the scalability and performance of models in high-dimensional clickstream datasets. Similarly, Olmezogullari and Aktas [11] highlighted the efficacy of vectorized representations in capturing user intent by encoding click sequences with positional and categorical attributes, which significantly enhance prediction accuracy in real-time systems. These studies underscore the versatility and effectiveness of fixed-length vector representations for addressing the challenges posed by high-dimensional and sequential data.

Our study advances this line of research by systematically comparing the performance of data representations. Unlike previous studies, we leverage more traditional and tree-based models, which offer both computational efficiency and interpretability.

### B. MACHINE LEARNING ALGORITHMS

A diverse array of machine learning models have been applied to purchase prediction tasks, ranging from traditional statistical techniques to advanced tree-based and deep learning architectures.

Traditional Models such as Logistic regression and decision trees, remain popular because of their simplicity and interpretability. Martinez et al. [12] conducted a comparative analysis of machine learning algorithms and observed that although less complex, they serve as effective baselines and are often computationally efficient for real-time predictions.

Sequential and probabilistic models, including Markov chains and Hidden Markov Models (HMMs), have also been used to capture user navigation and purchase behaviors in e-commerce settings. For instance, Bertsimas et al. [13] used Markov chains to estimate purchase probability based on user site navigation. Hidden Markov Models (HMMs) have also been adopted for sequential data modeling. Zucchini and MacDonald [14] discussed the effectiveness of HMMs in time-series analysis, which can be applied to clickstream data.

Ozyurt et al. [15] tailored a deep Markov Model and combined it with an attention network to learn long-term dependencies and capture different shopping phases in a user's journey.

Deep learning models such as Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) networks are commonly applied to sequence-heavy data representations such as clickstreams [16], [17]. Zolna et al. [18] introduced "user2vec," transforming user actions into input vectors without manual feature generation and then used LSTM-based models for prediction. Similarly, Jenkins [19] adopted an LSTM-based model to predict future user actions, given a user's clickstream history. However, these models are computationally intensive and require careful tuning to avoid overfitting on imbalanced datasets.

Verma [20] demonstrated the effectiveness of deep neural networks, including Multi-Layer Perceptron (MLP), Long Short-Term Memory (LSTM), and Temporal Convolutional Networks (TCN), for predicting purchases in e-commerce environments. These architectures outperformed traditional models such as XGBoost and Random Forest. However, their computational complexity and data requirements often pose challenges for real-time e-commerce applications.

Tree-based ensemble methods such as Random Forest, Gradient Boosting, and LightGBM are widely used in e-commerce owing to their ability to handle high-dimensional data and imbalanced class distributions. Wang et al. [21], Ulitzsch et al. [22] and Gan et al. [23] employed XGBoost to model session-level features and demonstrated its efficacy in capturing nonlinear relationships.

Recently, Tokuç and Dag [24] applied LightGBM to session-aggregated datasets for purchase prediction. Although this approach shows promise for computational efficiency, it does not address the potential added value of sequential or hybrid data representations, which are explored in this study.

This study builds on these works by evaluating a number of machine learning architectures across different data representations. We leverage more computationally efficient models, such as LightGBM and Gradient Boosting, while exploring the synergies between different data representations to enhance the predictive performance.

### C. FEATURE IMPORTANCE ANALYSIS

Feature importance analysis helps identify the key factors that influence purchase predictions. This is crucial for understanding user behavior and optimizing the predictive models.

Hendriksen et al. [25] demonstrated the value of feature importance analysis in enhancing the model performance for purchase intent prediction in e-commerce, particularly for anonymous and identified customers. Their work highlighted the impact of session data and historical customer behavior on predictive accuracy. Building on these insights, our study incorporates feature importance analysis across multiple machine learning models to identify the most influential predictors of purchase likelihood.

Tree-based models, such as XGBoost and LightGBM, are frequently used owing to their built-in feature importance capabilities. Wang et al. [21] and Gan et al. [23] employed tree-based models to highlight the importance of temporal features such as the recency of actions and user engagement metrics such as session frequency. Tokuç and Dag [24] investigated feature importance by training the LightGBM model. However, these studies have focused on a single data representation.

Zolna et al. [18] demonstrated the effectiveness of feature importance in sequential user actions but acknowledged the lack of interpretability in neural embeddings. Similarly, Saarela and Jauhiainen [26] compared different techniques for feature ranking, emphasizing the need for domain-specific metrics to improve the model transparency.

Building on previous work, there is a gap in further exploration to understand the interplay between session-level attributes and immediate actions in driving purchase predictions. Our work builds on these findings by employing multiple datasets to analyze the impact of different features. Moreover, this analysis not only enhances model interpretability but also provides actionable insights for e-commerce platforms to effectively tailor marketing interventions.

### D. GAPS IN LITERATURE

Although current purchase prediction methods have proven to be effective, they have notable limitations. Although powerful, deep learning models typically require extensive data and high computational resources, making them less accessible to all practitioners. These models can also act as "black boxes," offering limited interpretability, which is a drawback for understanding the factors influencing purchase decisions. In addition, their computational complexity, coupled with significant data requirements, limits their use in real-time applications.

A systematic literature review by Karl [27] highlights the diversity of machine learning applications in e-commerce, categorizing tasks such as purchase prediction, repurchase prediction, and product return prediction. Although this taxonomy shows a broad spectrum of approaches, it also underscores the need for frameworks that bridge session-level summaries with immediate user actions, a gap this study aims to address.

A recurring challenge in the literature is the underexploration of hybrid data representations that capture both short- and long-term user behaviors. Although aggregated session data offer a comprehensive view of user engagement, they often fail to account for the immediacy of actions leading up to a purchase. Conversely, sequential models focus on short-term behaviors but lack the broader context provided by the session summaries. This dichotomy leaves a gap in identifying the optimal representations that balance predictive performance, interpretability, and computational efficiency. Unlike previous studies, which focused solely on aggregated session statistics or recent user actions, our study bridges this gap by systematically evaluating hybrid data representations. This approach not only improves predictive performance but also offers interpretable insights for practical applications in e-commerce.

Class imbalance further compounded these challenges. Nonpurchase sessions vastly outnumber purchase sessions and skew model predictions towards the majority class [28]. Although techniques such as class weighting and resampling have been applied to address this issue, their effectiveness remains inconsistent, particularly for hybrid representations in which the interplay between minority and majority classes is complex.

To address these gaps, this study systematically evaluated multiple data representations, including aggregated session data, recent user actions, and hybrid combinations, across a range of machine learning models. By focusing on computationally efficient models, such as LightGBM, and employing feature importance analysis, this work contributes to the development of interpretable, scalable, and effective predictive frameworks for e-commerce. In doing so, it bridges the methodological gaps identified in previous studies and provides actionable insights into real-time applications.

## III. METHODOLOGY

This section outlines the dataset utilized for our experiments, the preprocessing techniques applied, the feature engineering methods used, the data representations examined, and the machine learning algorithms employed in the study. Fig. 1 illustrates the overall system schema for the proposed framework.

### A. DATASET DESCRIPTION

To assess the feasibility and effectiveness of the proposed modeling techniques, experiments were conducted using a clickstream e-commerce dataset. The dataset utilized in this research is the "eCommerce behavior data from a multi-category store," sourced from Kaggle [29] and published by REES46 [30]. This dataset encompasses clickstream data capturing user interactions on an online retail platform and documenting actions from October 2019 to April 2020. The data structure for an event is presented in Table 1.

**TABLE 1.** E-commerce clickstream dataset schema and description.

| Column Name | Data Type | Description |
|---|---|---|
| event_time | timestamp | A timestamp denoting the precise moment of the interaction (in UTC). |
| event_type | string | The type of user interaction, which may be a product view, add-to-cart action, or purchase. |
| product_id | integer | A unique identifier corresponding to the product involved in the event. |
| category_id | long | The identifier representing the category to which the product belongs. |
| category_code | string | A hierarchical classification code representing the product's category, when available. This field is generally populated for well-defined categories but may be omitted for miscellaneous or accessory items. |
| brand | string | A lowercase string indicating the brand name of the product, which may be absent in some instances. |
| price | float | The monetary value of the product, recorded as a floating-point number. |
| user_id | integer | A persistent identifier assigned to each unique user, facilitating cross-session tracking. |
| user_session | string | A temporary session identifier assigned to a user, which remains constant throughout a session but is updated each time the user returns to the platform after a prolonged interval. |

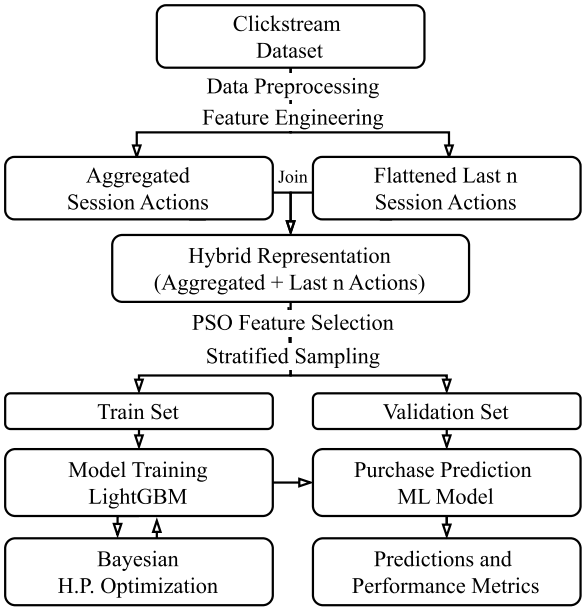


**FIGURE 1.** Clickstream data flow for purchase prediction modeling.

### B. CLICKSTREAM DATA REPRESENTATIONS

The way data are represented has an impact on model performance, particularly when dealing with high-dimensional and sequential data such as clickstreams. We explored three data representations to evaluate their predictive performances.

#### 1) FLATTENED LAST *N* ACTIONS PER SESSION

This dataset represents user sessions by focusing on the most recent $n$ actions taken by the users. This approach retains the last interactions, which are likely to influence purchasing decisions, allowing the model to leverage temporal patterns.

Each action is encoded with a set of features that describe user interaction, such as event type (view, cart, etc.), product details, and time elapsed since the previous action. Stationary session-level features, such as session duration and day of the week, were also included but were only stored once per session.

For each session, if the total number of actions exceeded $n$, only the last $n$ events were considered. Conversely, if the session had fewer than $n$ actions, data were padded to ensure consistent dimensions across all sessions.

This representation is particularly well-suited for machine learning models that perform better with fixed-length input arrays. In this study, we experimented with different values of $n$, to investigate the effect of the sequence length on the predictive performance.

The importance of a user's last actions within a session is tied to their potential to reveal the immediate purchase intent. Retaining the last five actions aims to balance capturing the recency of decisions, which maintains a broader context of the session's behavior. This mid-range sequence length was hypothesized to provide a sufficient context without overwhelming the model with redundant data. Conversely, using the last ten actions broadens the scope to include longer-term behaviors, capturing user exploration or changing interests within the session.

The decision to experiment with different temporal windows was driven by practical trade-offs, including computational complexity and model interpretability. As described in Section IV, a thorough comparative analysis determines the temporal representation that most effectively predicts purchase intent.

#### 2) AGGREGATED ACTIONS PER SESSION

Aggregated session statistics condense each session into a single row of summary statistics, capturing high-level behavioral patterns across the entire session. Attributes such as the total number of events (clicks, views, and carts), session duration, average product price, and temporal variables such as session start time and day of the week were generated to provide insight into overall engagement.

The main advantage of this approach is its simplicity and reduced computational cost because the entire session is reduced to a single row of features. However, this comes at the cost of losing granular event-level details, which may provide critical information for predicting user behavior. The aggregated approach aims to capture broad trends throughout a session, which could offer valuable insights into long-term behavioral tendencies, particularly for understanding user engagement or the overall flow of a session. In short, although this representation sacrifices detailed sequential information, it offers a computationally efficient summary of

**TABLE 2.** Event type distribution.

| Event Type | Number of Occurrences | Percentage (%) |
|---|---|---|
| Product View | 219,812,361 | 94.15 |
| Add to Cart | 9,991,458 | 4.28 |
| Purchase | 3,656,843 | 1.57 |

user activities that may correlate with purchasing tendencies, particularly for identifying broad engagement patterns.

#### 3) HYBRID REPRESENTATION

The aggregated representation aims to capture the overall characteristics of a session, whereas the flattened representation provides a more granular view of user behavior. The hybrid representation unions the features from both the flattened last $n$ actions and the aggregated action datasets. In other words, it is the concatenation of both prior datasets.

By combining the detailed view of recent actions from the flattened representation with the high-level summary provided by session aggregation, this approach seeks to capture both short- and long-term behavioral signals. This combined representation allows models to leverage both immediate user actions and overarching session characteristics, potentially offering a more comprehensive view of a user's intent.

By comparing these representations across various machine learning models, this study examines the potential of different clickstream data structures to improve purchase prediction accuracy. These comparisons are essential for identifying the optimal data representations for real-time e-commerce applications, where balancing predictive power with computational efficiency is critical.

### C. EXPLORATORY DATA ANALYSIS (EDA)

An initial Exploratory Data Analysis (EDA) was conducted to uncover patterns that could influence feature engineering and preprocessing decisions. Insights into key data distributions, such as session lengths, conversion rates, and the frequency of specific interactions, guided our handling of imbalanced classes and session length variations.

The distribution of event types (view, add-to-cart, and purchase) was analyzed to understand user interaction patterns. Table 2 shows the relative frequencies of each event type, revealing that product views dominate user activity at 94%, whereas purchases represent only a small fraction of all events. Similarly, only 3,063,117 sessions (5.85%) contained a purchase event, indicating an imbalanced dataset.

Further, the number of events per session and session duration were explored to understand user engagement. Table 3 presents the mean, standard deviation, minimum, and maximum values for several key session metrics, including the number of events, number of products, total session time in seconds, event frequency, view count, cart activity, and purchase behavior.

Of the 52,321,178 sessions considered, the average event count per session was 4.46 with a standard deviation of 7.34, while the average number of products interacted with was 2.84. Most sessions contained fewer than five events, with a long tail of sessions having significantly more interactions. Additionally, the average session duration is approximately 14,270 seconds, although highly variable (with a standard deviation of 279,292 seconds), reflecting a wide range of engagement levels. The average number of views per session was 4.20, whereas cart additions and purchases occurred less frequently, with averages of 0.19 and 0.07, respectively. Interestingly, the distribution of purchase value is highly skewed, with an average of 21.22 but with a maximum exceeding 120,000.

**TABLE 3.** Session statistics.

| Statistic | Median | Mean | Std Dev | Min | Max |
|---|---|---|---|---|---|
| Event Count | 2 | 4.46 | 7.33 | 1 | 5065 |
| Product Count | 1 | 2.84 | 4.43 | 1 | 5065 |
| Session Time (sec) | 41 | 14270.01 | 279292.01 | 0 | 10596947 |
| Event Per Second | 0.27 | 0.05 | 0.10 | 1.9e-7 | 102.00 |
| View Count | 2 | 4.20 | 7.12 | 0 | 5065 |
| Cart Count | 0 | 0.19 | 0.86 | 0 | 753 |
| Purchase Count | 0 | 0.07 | 0.32 | 0 | 252 |
| Purchase Value | 0.00 | 21.22 | 148.40 | 0.00 | 120503.56 |
| Event Per Product | 1 | 0.79 | 0.27 | 0.001 | 1.00 |
| Session Start Hour | 11 | 10.99 | 5.17 | 0 | 23 |
| Session Start Day | 4 | 3.99 | 2.00 | 1 | 7 |

**TABLE 4.** User statistics.

| Statistic | Median | Mean | Std Dev | Min | Max |
|---|---|---|---|---|---|
| Has Purchase (%) | 0 | 12.6 | 33.2 | 0 | 100 |
| Event Count | 23 | 23.30 | 77.97 | 1 | 61830 |
| Session Count | 5 | 5.23 | 38.51 | 1 | 60612 |
| Product Count | 12 | 11.72 | 29.87 | 1 | 10467 |
| Brand Count | 5 | 4.76 | 8.96 | 0 | 698 |
| Category Count | 4 | 4.06 | 7.18 | 1 | 455 |
| Event Per Session | 3.46 | 3.46 | 6.21 | 1.0 | 5065.0 |
| Purchase Count | 0 | 0.37 | 2.54 | 0 | 1456 |
| Cart Count | 0 | 0.96 | 4.35 | 0 | 1590 |
| View Count | 22 | 21.97 | 75.33 | 0 | 61821 |
| Total Spent | 0.00 | 111.01 | 1132.75 | 0.0 | 604082.01 |
| Avg Price Per Purchase | 0.00 | 35.03 | 139.61 | 0.0 | 2574.07 |

Similarly, events were analyzed on a user basis. The statistical summary is presented in Table 4. On average, 12.6% of the users make at least one purchase, indicating that a minority of them are converted to buyers. The average user makes 0.37 purchases overall, while the maximum purchase count for a single user is 1456. The high standard deviation in purchase counts (2.54) reflects considerable variability in user purchase behavior, with some users buying significantly more than others do.

Users have an average of 23.3 events over their sessions, though this number varies widely, with a standard deviation of 77.97. This indicates that while many users have minimal activity, some are highly engaged, as evidenced by the maximum of 61,830 events. Users interact with 5.23 sessions on average, although the maximum value shows that some users engage in over 60,612 sessions.

On average, users interact with 11.72 products and 4.76 brands, indicating a reasonable level of engagement with different products. However, the high maximum product

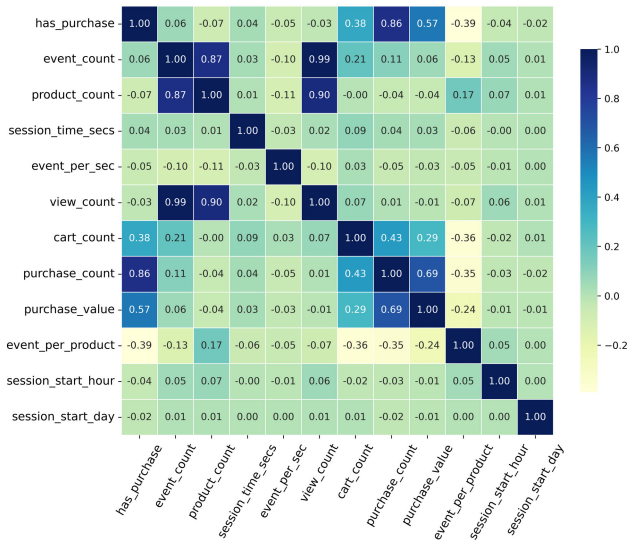


**FIGURE 2.** **Correlation matrix for session-based features, highlighting relationships between purchasing behavior and session metrics.**

count (10,467) and brand count (698) reveal that some users are engaging with a vast array of items.

Overall, cart activity was low, with an average cart count of 0.96. Most users do not add many items to their cart, although a few outliers engage heavily with this feature (a maximum of 1590 cart additions).

The average user spends approximately 111 on purchases, but there is a large variability in spending habits, with some users spending over 604,000. Similarly, the price per purchase varies significantly. The average price per purchase was 35.03; however, users made purchases as high as 2574.07.

When it comes to viewing events, users engage with 21.97 products on average, examining nearly 11.7 distinct products and 4.76 brands. This behavior reflects users' exploration before they make purchasing decisions.

The difference between the minimum and maximum event timestamps indicates how long a user has been active on the platform. This could reflect the level of engagement over time for different users, from short-term visitors to long-term customers.

To summarize, the user behavior analysis reveals a strong skew towards minimal engagement, with a small percentage of users contributing disproportionately to purchases and activity on the platform. The high variability in the number of events, purchases, and cart activities underscores the diversity of user interactions, from casual browsers to heavy spenders.

Finally, a correlation matrix is plotted in Fig. 2 to identify the relationships between key features such as session length, number of views, number of carts, and eventual purchase decisions.

The strongest positive correlation with `has_purchase` is observed for `purchase_count` (0.8628), indicating that sessions with a higher number of purchases almost certainly result in a purchase. Similarly, `purchase_value` has a strong positive correlation (0.5650), suggesting that higher purchase amounts are associated with purchase sessions. Interestingly, `cart_count` has a moderately strong positive correlation with purchase behavior (0.3750), indicating that sessions in which users add items to their cart are more likely to result in a purchase. This behavior aligns with typical user shopping patterns, where adding items to a cart is a precursor to completing a purchase.

The event count has a weak positive correlation with the label (0.0610), whereas the product count shows a small negative correlation (−0.0692). This suggests that a higher number of events during a session slightly increases the likelihood of a purchase; however, interacting with more distinct products may actually reduce the likelihood of a purchase. Similarly, the view count exhibits a very weak negative correlation with the label (−0.0268), implying that sessions with many product views are slightly less likely to result in a purchase. This may indicate browsing behavior without commitment.

The correlation between session time and `has_purchase` was weakly positive ($p = 0.0384$). This finding suggests that longer sessions may slightly increase the chance of a purchase, but the effect is minimal. Users may need more time to complete their purchase decisions, but the session duration alone is not a significant predictor. Both session start hours (−0.0359) and days (−0.0249) show very weak negative correlations with the label, indicating that the time of day or the day of the week has a negligible influence on purchasing behavior in the session.

The strongest negative correlation with `has_purchase` is seen in `event_per_product` (−0.3910), suggesting that sessions in which users spend more time interacting with each individual product are less likely to result in a purchase. This behavior may indicate indecision or exploratory browsing in which users examine products without intending to buy.

Unsurprisingly, there is a strong positive correlation between event count and product count (0.8748), indicating that sessions with more events typically involve interactions with more distinct products. Similarly, the event count is strongly correlated with the view count (0.9886), suggesting that many session events are related to product viewing. This further supports the idea that high-interaction sessions may represent exploratory browsing behavior.

On the other hand, the negative correlation between `event_per_product` and `has_purchase` (−0.3910) suggests that as users spend more time interacting with individual products, the likelihood of making a purchase decreases. This is indicative of extended browsing behavior or indecision rather than decisive purchasing. Events per second have a weak negative correlation with the label (−0.0521), indicating that sessions with a higher event frequency may not necessarily lead to purchases. This could be reflective of users quickly skimming products without significant engagement.

Overall, these insights suggest that predicting purchase outcomes is more closely linked to cart activity and

**TABLE 5.** Number of remaining rows in the dataset after preprocessing steps.

| Step | Number of Remaining Rows |
|---|---|
| All initial rows | 233,460,662 |
| Dropping Duplicates | 233,080,208 |
| Invalid Session Data | 233,080,156 |

**TABLE 6.** Number of empty cells by feature name.

| Feature Name | Number of Nulls |
|---|---|
| event_time | 0 |
| event_type | 0 |
| product_id | 0 |
| category_id | 0 |
| category_code | 47,496,567 |
| brand | 29,948,661 |
| price | 0 |
| user_id | 0 |
| user_session | 0 |
| date | 0 |

purchasing behavior within the session, while extensive browsing and indecision (high events per product) reduce the likelihood of making a purchase.

### D. DATA PREPROCESSING

The preprocessing phase entailed data cleaning, including handling missing values and filtering anomalous session lengths as described in the following subsections.

#### 1) DATA CLEANING AND VALIDATION

The dataset initially comprised 233,460,662 rows, including potential duplicates that could introduce bias or distort the analysis. A deduplication process was employed to remove redundant entries, thereby reducing the dataset to 233,080,208 rows. This step ensures the fidelity of the dataset and reduces the risk of overfitting owing to artificially inflated user behavior.

Further cleaning involved the removal of sessions with missing or invalid information. After filtering out sessions with invalid entries to maintain data integrity, the remaining number of rows was reduced to 233,080,156, as shown in Table 5.

The dataset was then examined for missing values, a detailed analysis is presented in Table 6. The only features with missing values were `category_code` and `brand`, which contained 47,496,567 and 29,948,661 null values, respectively. The `brand` feature was handled by encoding missing values as a distinct category labeled "*empty*." This approach preserved the interpretability of the model while avoiding potential biases introduced by imputation. Additionally, the `category_code` field was excluded from further analysis despite its hierarchical representation of product categories because it was both incomplete and redundant. On the other hand, the `category_id` feature fully captured the categorical information of products without missing values. Therefore, excluding the `category_code` feature did not result in loss of information. Apart from `category_code`, no other features were removed during preprocessing. Instead, all features underwent cleaning and transformation to ensure data integrity and consistency.

Outliers in the dataset, such as sessions exhibiting an abnormally high number of events per second or a disproportionate ratio of events to distinct products, were identified and removed. Such sessions are frequently indicative of automated interactions (*e.g.*, bot activity) rather than genuine user behavior. Specifically, sessions with more than one event per second or sessions with over 100 distinct products or a suspiciously high event-to-product ratio were removed. Further data imputation based on session statistics was not performed. This step ensures that the models are trained on genuine user interactions that reflect real-world behavior. Consequently, 2,990 anomalous sessions were excluded, thereby enhancing the integrity of the dataset.

#### 2) SESSION-LEVEL DATA REFINEMENT

Further cleaning was conducted to refine the dataset for purchase prediction tasks. This involved removing events that occurred within a session after a purchase. The predictive model aims to forecast the likelihood of a purchase, including post-purchase events, which would bias the analysis. By trimming sessions to include only pre-purchase events, the data align with the real-world context, where predictions must be made before purchase events. Hence, the model can better capture meaningful interactions that precede purchase decisions.

In addition, the last three page views of each session were excluded. This step is motivated by two key considerations.

- **Noise Reduction**: The final page views in a session often capture non-predictive behaviors, such as users navigating away from the purchase flow or lingering without decisive actions. Removing these events minimizes noise that can mislead the model.
- **Real-time Constraints**: In real-world applications, predictions must be made before users complete their interactions. Excluding the last few actions creates a buffer for simulating this real-time setting, allowing the model to predict using the data available up to the penultimate stages of the session [6].

Sessions with fewer than three actions were excluded because they lacked sufficient behavioral information to effectively inform the predictive models. Consequently, 29,155,161 sessions with fewer than three page views were eliminated, resulting in a dataset that focused on actionable sessions.

Conversely, sessions with more than 100 page views were truncated to this threshold to standardize data inputs and reduce computational overhead without compromising the quality of the predictive features.

When analyzing session times, substantial variability was observed, with a few sessions exhibiting exceedingly

**TABLE 7.** Session outcome vs number of sessions.

| Session Type | Number of Sessions | Percentage |
|---|---|---|
| Sessions with Purchase | 2,963,560 | 12.79 % |
| Sessions without Purchase | 20,199,466 | 87.21 % |

**TABLE 8.** Session outcome vs number of events.

| Session Type | Total Event Count | Percentage |
|---|---|---|
| Sessions with Purchase | 25,710,605 | 11.03 % |
| Sessions without Purchase | 207,369,551 | 88.97 % |

long durations, potentially skewing the average session time upward. These sessions were not removed from the dataset to better reflect unexpected variations in real-world datasets.

After completing all data preprocessing steps, the dataset was significantly refined to improve its reliability for training the predictive models. The final dataset comprises 23,163,026 user sessions, with 2,963,560 of these sessions leading to a purchase, representing a purchase ratio of 12.79%. The details of the session types and outcomes are presented in Table 7. The number of events in the sessions showed a similar distribution, as summarized in Table 8. Out of 233,080,156 actions in the dataset, 25,710,605 events belong to a session resulting in a purchase event, which is approximately 11% of the dataset.

### E. FEATURE ENGINEERING

Feature engineering plays a crucial role in transforming raw clickstream data into structured inputs for machine learning models. The feature engineering process introduces several additional attributes to enrich the dataset for predictive modeling. The enhanced clickstream dataset is described in Table 9.

The initial dataset included timestamps for each action in the `event_time` column. Temporal features such as `date`, `day_of_week`, and `is_weekend` were derived from these timestamps to capture the temporal dynamics of user behavior. These features provide granularity that helps to identify patterns in user activity across different days and peak periods.

Session-based features, including `session_time_so_far` and `session_last_event_time`, provide insight into user engagement within a session. These features measure the session duration and frequency, enabling the distinction between highly engaged users and casual browsers. Additionally, `time_since_last_event` captures the time between consecutive events, providing a measure of the user interaction frequency within a session. These session-based temporal engineered features were prioritized based on their strong correlations with purchase intent observed during EDA. These features effectively capture engagement dynamics and user decision-making processes, aligning well with the hybrid data representation.

**TABLE 9.** Enhanced clickstream dataset schema.

| Column Name | Description |
|---|---|
| `user_session` | Temporary session ID |
| `event_time` | Time when the event happened |
| `event_type` | Type of event: view, add to cart, or purchase |
| `product_id` | ID of the product |
| `category_id` | ID of the product's category |
| `brand` | Brand name of the product |
| `price` | Price of the product |
| `user_id` | Permanent user ID |
| `date` | Date of the event |
| `time_since_last_event` | Time since the previous event in session |
| `user_type` | Type of user (new or returning) |
| `session_time_so_far` | Duration of the session so far (seconds) |
| `day_of_week` | Day of the week when the event occurred |
| `is_weekend` | Binary indicator of if the event occurred on a weekend |
| `has_purchase_so_far` | Indicator of whether a purchase has occurred in the session so far |
| `session_count_so_far` | Number of sessions the user has had so far |
| `dist_product_count_so_far` | Number of distinct products viewed so far in the session |
| `dist_brand_count_so_far` | Number of distinct brands viewed so far in the session |
| `dist_category_count_so_far` | Number of distinct categories viewed so far in the session |
| `event_count_so_far` | Total number of events so far in the session |
| `session_last_event_time` | Time of the last event in the session |
| `label` | Binary indicator of whether a purchase event occurred |

User-specific features, such as `user_type` (first-time or returning user), `has_purchase_so_far`, and session based features `dist_product_count_so_far`, `dist_brand_count_so_far`, `dist_category_count_so_far`, and `event_count_so_far`, allow the tracking of user behavior over time. These features provide a high-level overview of user interactions; distinguish between different user types; and capture purchasing history, product preferences, and the diversity of interactions across products, brands, and categories.

Product-based features such as `product_id`, `price`, `category_id`, `category_code`, and `brand`, retain detailed information about the products involved in each session. These features allow for an analysis of how product attributes such as price and brand influence purchasing decisions.

The target variable, represented by the `label` feature, indicates whether a purchase event has occurred at the end of the session.

By incorporating these features, the dataset is enriched with a comprehensive set of temporal, session-based, user-specific, and product-related attributes. This approach ensures that the models leverage both granular and aggregated insights, addressing the gaps in previous studies that focused on narrower feature scopes.

As part of the feature engineering process, a correlation analysis was conducted for the aggregated session dataset to identify potential multicollinearity issues among features. The analysis revealed a moderate to high correlation between certain features, such as event count and product count ($r = 0.87$), as shown in the correlation matrix in Fig. 2. These features were retained because they capture complementary aspects of user engagement, which are critical for predicting purchase intent. For example, event count reflects overall session activity, while product count indicates the diversity of user interactions.

For sequential datasets, the correlation between occurrences of the same feature across different time steps is intrinsic to the data structure and reflective of temporal dependencies. Removing such features would lead to significant information loss and negatively impact the model's ability to leverage sequential patterns. By retaining these features, the model captures the evolving nature of user intent, which is particularly valuable for clickstream data analysis.

Additionally, tree-based models, such as LightGBM, are not negatively affected by multicollinearity, as they are non-parametric and can handle complex relationships among variables [31]. These models prioritize features based on their contribution to impurity reduction, minimizing redundancy, and ensuring stable predictions. This characteristic makes them well-suited for handling high-dimensional datasets with correlated features.

### F. FEATURE SELECTION

Feature selection aims to optimize the model's performance by focusing on the most relevant attributes in the dataset while reducing noise and computational complexity. Particle Swarm Optimization (PSO) is employed to identify high-impact features from an initial set of sessions and user behavior metrics.

PSO [32] is a nature-inspired optimization algorithm that uses a group of agents, called "particles," to search for the best solution collectively. Each particle iteratively moves closer to the best-known solution, balancing the exploration and convergence. Xue et al. [33] applied PSO for feature selection in classification tasks using a multi-objective approach, allowing for an explicit balance between the classification error and feature count.

To adapt PSO for feature selection in our feature space, we implemented the binary form of the algorithm using the `PySwarms` library in `Python` [34]. Each particle represents a unique binary mask, where each bit indicates whether a specific feature from the initial set should be included in the subset. As the particles move through the feature space, these selected features are used to train the LightGBM model. The AUC score, calculated for each particle, serves as the fitness function, allowing PSO to evaluate the predictive strength of each feature combination without penalizing the feature count.

The PSO algorithm was configured to balance the exploration and convergence. The number of particles (`n_particles`) was set to 10, ensuring a diverse search space while maintaining computational feasibility. The dimensionality of the search space (`dimensions`) was set to be equal to the number of features in the dataset. The algorithm was executed for a maximum of 50 iterations (`max_iters`), providing sufficient opportunity for convergence without excessive computational overhead.

The optimization process involves particles that adjust their positions in the binary search space based on the velocity update formula:

$$v_i^{(t+1)} = w \cdot v_i^{(t)} + c_1 \cdot r_1 \cdot \left(p_{\text{best}} - x_i^{(t)}\right) + c_2 \cdot r_2 \cdot \left(g_{\text{best}} - x_i^{(t)}\right) \quad (1)$$

In Equation 1; $v_i$ is the velocity of particle $i$ at time $t$, $w$ is the inertia weight, $c_1$ and $c_2$ are the acceleration coefficients, $r_1$ and $r_2$ are random numbers in [0, 1] and $p_{\text{best}}$ and $g_{\text{best}}$ denote the local and global best-known positions, respectively. The binary nature of the feature space is managed by applying a sigmoid transformation to the velocity update and using a probabilistic threshold to determine the binary state of the particle.

The key parameters for particle movement and optimization are defined in the `options` dictionary. The inertia weight ($w$) was set to 0.9 to balance exploration and exploitation; cognitive and social acceleration coefficients ($c_1$ and $c_2$) were both set to 0.5, allowing particles to consider both their personal best-known positions and the global best-known position in equal measure. To further enhance swarm diversity and mitigate premature convergence, the number of neighbors influencing each particle ($k$) was set to 5, whereas the number of informants ($p$) was set to 3.

For the aggregated dataset, which captures session-level summaries such as total events and average prices, PSO identified a subset of 12 key features that effectively encapsulate user behavior. Features such as `event_count` and `num_cart` emerged as critical indicators of engagement and intent, with the former representing the overall activity within a session and the latter directly tied to purchase behavior. Temporal features, including `session_start_hour` and `is_weekend`, were also selected, highlighting the role of shopping patterns influenced by the time of day and the day of the week. `average_price_per_event` further emphasized the importance of capturing user interactions with higher-value items, reflecting price sensitivity and potential purchasing power. On the other hand, features such as `user_session_count_so_far` and `category_count` were excluded during optimization because their contribution to the model performance was minimal.

In the flattened dataset, which captures sequential user actions, PSO reveals the importance of recent behaviors in predicting purchases. Features such as `time_since_last_event` and `price` consistently

**TABLE 10.** Train - test statistics.

| Statistic | Train Set | Test Set |
|---|---|---|
| Events | 167,242,845 | 18,602,787 |
| Sessions | 20,843,007 | 2,317,258 |

ranked high in importance, underscoring the relevance of immediate user interactions. The significance of features capturing the diversity of interactions, such as `dist_product_count_so_far` and `dist_brand_count_so_far`, further illustrated the role of user exploration in shaping purchase decisions. Conversely, features such as `category_id`, `event_count_so_far`, and `session _time_so_far` were eliminated because of their weak correlation with purchase outcomes.

### G. TRAIN - TEST SET DIVISION

We used a random stratified train-test split to train the models and evaluate their performance. Table 10 summarizes the resulting training and test set statistics.

The dataset was split into 90% training data and 10% testing data to ensure that the distribution of the target variable (purchase events) remained consistent across both sets. Stratified splitting is essential because of the class imbalance in the target variable, as purchase events are relatively rare compared with nonpurchase sessions. This method helps prevent any significant imbalance between the two sets and ensures that both positive and negative instances are adequately represented.

The 90-10 split ratio provides a reasonable data distribution for training and testing. A larger training set (90%) allows the model to learn from a substantial amount of data and capture patterns and relationships that can aid in predicting purchase behavior. The smaller test set (10%) was an independent evaluation set, providing an unbiased assessment of the model's performance on unseen data.

Additionally, to mitigate the effects of class imbalance, two main strategies were employed during the model training: Balanced Subsampling and Class Weight Adjustment.

Most models were trained using a balanced subsample of the training data, in which the number of purchase events (positive class) was artificially balanced by downsampling nonpurchase events (negative class). This prevents the models from overfitting to the majority class and ensures a more balanced decision boundary.

on the other hand, LightGBM can automatically adjust the weighting of classes in the loss function to account for the imbalance. This penalizes the misclassification of the minority class more heavily than that of the majority class, encouraging the model to focus on accurately predicting purchase events. This technique directly addresses the imbalance during training without subsampling or oversampling the dataset.

An imbalanced and larger training set is used for the LightGBM model. For the other models, a balanced and under-sampled dataset is utilized. However, to evaluate the true performance, all models were tested on the complete, unbalanced test set, offering a realistic assessment of how they would perform in production settings with imbalanced data.

### H. MODEL TRAINING

To predict purchase intent based on users' clickstream data, we employed a range of machine learning algorithms known for their effectiveness in classification tasks and ability to handle large-scale datasets. The models utilized included LightGBM, decision trees, gradient boosting, random forests, and logistic regression.

To further improve the model performance, hyperparameter optimization was performed for each algorithm using the Tree-structured Parzen Estimator (TPE) algorithm via `hyperopt` package [35].

TPE is a Bayesian optimization technique that adaptively searches the hyperparameter space by modeling the objective function as a mixture of Gaussian processes. Compared to grid or random search, TPE provides an efficient means of identifying optimal configurations for high-dimensional search spaces.

The objective function for optimization was defined as the negative Area Under the Receiver Operating Characteristic Curve (AUC-ROC) evaluated on a validation set. The best hyperparameters are detected over 50 trials, leading to the best AUC-ROC score.

The rest of the section defines the models trained and the hyperparameters tuned for each one. The search space of those hyperparameters and the best values are presented in Appendix A Table 11.

#### 1) LOGISTIC REGRESSION

Logistic regression [36] is a statistical model used for binary classification tasks. It models the probability that a given input point belongs to a certain class by using the logistic function. Despite its simplicity, logistic regression is a strong baseline and is widely used owing to its interpretability and efficiency on large datasets.

The primary hyperparameter optimized is the Regularization Parameter (`regParam`), which controls the strength of L2 regularization applied to the model to prevent overfitting. Smaller values correspond to stronger regularization, which helps reduce model complexity and the risk of overfitting.

#### 2) DECISION TREE

Decision trees [37] are non-parametric supervised learning models used for classification and regression tasks. They predict the value of a target variable by learning simple decision rules inferred from data features. The model splits the dataset into subsets based on the most significant attribute, creating a tree-like decision structure. Decision trees are intuitive and interpretable, making them useful for understanding the feature importance.

We searched for the optimal Maximum Depth and Minimum instances per Node values. Maximum Depth (`maxDepth`) limits the depth of the tree. A deeper tree can capture more complex patterns but may lead to overfitting. Minimum Instances per Node (`minInstancesPerNode`) specify the minimum number of samples required to split an internal node. Higher values make the model more conservative and help prevent overly complex trees.

#### 3) RANDOM FOREST

Random forests [38] are an ensemble learning method that constructs multiple decision trees during training and outputs the class that is the mode of the classes of individual trees. By averaging the results of multiple trees, random forests improve the predictive accuracy and control overfitting. They are effective in handling large datasets with higher dimensionality.

The key hyperparameters optimized were the number of decision trees in the forest (`numTrees`) and the maximum depth of each individual tree(`maxDepth`), ensuring that the forest was neither too shallow nor too complex. Increasing the number of trees typically improves performance, as more trees reduce the risk of overfitting. However, it also increases the computational cost. Maximum depth controls the complexity of the trees.

#### 4) GRADIENT BOOSTING

Gradient boosting [39] is an ensemble technique that builds models sequentially, with each new model attempting to correct the errors of previous models. It combines weak learners, typically decision trees, to form a strong predictive model. Gradient boosting is effective for capturing complex patterns in data and is known for its high predictive accuracy.

The main hyperparameters tuned for gradient boosting were the learning rate (`stepSize`), number of iterations (`maxIter`), and maximum depth(`maxDepth`). The learning rate and number of iterations work together to determine how quickly the model converges and how well it generalizes, whereas the maximum depth controls overfitting by limiting individual tree complexity.

#### 5) LIGHTGBM

LightGBM [40] is a gradient-boosting framework that uses tree-based learning algorithms. It is designed for efficiency and scalability and is particularly suitable for large-scale and high-dimensional data. LightGBM introduces techniques such as Gradient-based One-Side Sampling (GOSS) and Exclusive Feature Bundling (EFB) to reduce computational cost and improve accuracy. It handles categorical features naively and is robust against overfitting.

The search space for LightGBM included parameters that control the complexity of the model and its regularization, controlling the number of leaves (`numLeaves`), maximum depth of each tree (`maxDepth`) learning rate (`learningRate`), and data sampling during training (`baggingFreq`, `baggingFraction` and `featureFraction`).

### I. MODEL EVALUATION

The primary objective of this evaluation is to assess the performance of machine learning models across various data representations, with a particular focus on addressing the challenges posed by class imbalance in purchase prediction tasks. By leveraging a combination of metrics, this analysis provides a nuanced understanding of model efficacy and practical applicability in real-world e-commerce scenarios.

The models were evaluated by using a combination of general classification metrics, class imbalance-aware metrics, and diagnostic tools. Accuracy, precision, recall, and F1-score are used to measure the overall predictive quality of the models.

Given the skewed distribution of purchase versus nonpurchase events, metrics such as the Area Under the Receiver Operating Characteristic curve (AUC-ROC), Area Under the Precision-Recall curve (AUC-PR), geometric mean (gmean), and the Index Balanced Accuracy (IBA) are emphasized to provide a more nuanced understanding of model performance in handling imbalanced datasets. Another key metric considered in this study was the Validation Log Loss. Unlike metrics that focus on classification outcomes, the Log Loss evaluates the quality of probabilistic predictions. It penalizes incorrect predictions more heavily when the predicted probability is confidently wrong, thus providing a nuanced view of model calibration.

Confusion matrices were generated for each model to provide further insight into the distribution of the predicted versus actual classes.

In addition to standard classification metrics and visual diagnostics, a feature importance analysis was conducted for models that provide such insights. This analysis helps us understand which features play the most significant roles in predicting purchase events. The feature importances were calculated based on the contribution of each feature to the reduction in the impurity of the model.

## IV. EXPERIMENTAL RESULTS

This section presents the experimental setup, details the model training process across different data representations, analyzes the feature importance to identify the most influential factors in the prediction of purchase events, and compares the performance of various models using standard evaluation metrics.

### A. EXPERIMENTAL SETUP

The experiments were conducted on an x64-based PC equipped with an Intel 12th Gen i7-12700H processor running at 2.30 GHz with 14 cores and 16 gigabytes of physical RAM. The software environment consisted of Python 3.12.3 and Apache Spark version 3.3.2.

We implemented various machine learning algorithms using the Spark MLlib [41] library, which supports logistic

regression, SVC, decision tree, random forest, and gradient boosting models. For the LightGBM models, we utilized the SynapseML library [42], chosen for its optimized LightGBM integration with Spark.

The use of Apache Spark and its associated libraries allows for distributed computing, ensuring the efficient handling of the large-scale e-commerce clickstream dataset.

### B. MODEL TRAINING

To evaluate and compare the predictive accuracy of multiple machine learning models for purchase intent prediction in e-commerce, we trained each model using distinct representations of clickstream data. These representations include aggregated session-level attributes, flattened sequences capturing the most recent $n$ actions within a session, and hybrid datasets that combine both aggregated and sequential features.

Each model was trained using a 90-10 stratified train-test split, ensuring that the class distribution remained consistent across both sets. The 10% holdout set was reserved for final model evaluation, providing an unbiased estimate of generalization performance. This approach prevents data leakage and ensures that model performance is assessed on unseen data, reflecting real-world deployment scenarios.

Hyperparameter tuning was performed using the Tree-structured Parzen Estimator (TPE) via the `hyperopt` package, an adaptive Bayesian optimization method that efficiently searches the hyperparameter space. The objective function for optimization was the negative AUC-ROC score, ensuring that selected configurations improved model generalization.

Given the substantial class imbalance within the dataset, where nonpurchase events significantly outnumber purchase events, various strategies were implemented to mitigate potential bias in model training. For the LightGBM and decision tree models, which inherently manage class imbalance through parameter adjustments (*e.g.*, class weights), no additional resampling was necessary. For models without such intrinsic imbalance-handling capabilities, we applied random undersampling to the training data to produce a balanced dataset, ensuring that both purchase and nonpurchase events were equally represented. To comprehensively assess and differentiate model performance, we prioritized metrics designed to address class imbalance and provide nuanced insights into predictive accuracy.

### C. PERFORMANCE COMPARISON

Throughout this section, the performance of various models and data representations are evaluated. Model names and data representations are abbreviated to enhance table readability while retaining clarity. The model names are shortened as follows: LightGBM is represented by ''LGB,'' Decision Tree by ''DT,'' Gradient Boosting by ''GB,'' Random Forest by ''RF,'' Linear SVC by ''Lin SVC,'' and Logistic Regression by ''LR.'' Data representation types were also abbreviated: Aggregated Data is denoted as ''Agg,'' Recent 10 Actions as ''10A,'' Recent 5 Actions as ''5A,'' and Last Action Only as ''1A.'' Additionally, hybrid data representations combining recent actions with aggregated data were labeled as ''Hybrid-5A'' and ''Hybrid-1A'' for the last 5 actions and the most recent action unioned with aggregated dataset configurations, respectively.

Fig. 3 illustrates the precision-recall curves, showcasing each model's ability to identify the minority class (purchase events) at different thresholds. Additionally, the heatmap in Fig. 4 summarizes the performance metrics for the models trained on different datasets. The figures reveal clear trends in model performance across data representations. When comparing different configurations, it is evident that hybrid representations, which combine both aggregated session-level features and recent user actions, offer the best predictive performance across all metrics. They also demonstrate the best precision-recall trade-off. In contrast, using only aggregated features fails to capture the nuances of user behavior leading up to a purchase, while relying solely on recent user actions lacks a comprehensive overview provided by session aggregation. This synergy between recent user behavior and aggregate session information highlights the importance of incorporating multiple perspectives of user data to enhance predictive modeling in e-commerce settings.

Another critical factor is the number of recent user actions incorporated in hybrid and action-based representations. The inclusion of more recent actions, such as using the last 10 actions (10A) instead of the last 5 actions (5A), did not result in increased model performance. As the number of actions increases, the dimensionality of the input grows, requiring more computational resources and potentially leading to overfitting if not carefully managed. Interestingly, models trained on shorter sequences (1A, 5A) performed better, possibly due to increased noise or redundant information introduced with longer sequences (10A). In our experiments, the Hybrid-5A representation strikes a balance between computational efficiency and predictive accuracy, offering significant improvements without the excessive computational burden associated with larger action windows.

Of all the models evaluated, Logistic Regression (LR) performed the worst. Its simplicity limits its capacity to model the complex, nonlinear relationships inherent in user behavior data. In several configurations, LR recorded a Cohen's Kappa of 0, indicating that its predictions were no better than random guessing. Additionally, LR exhibited high rates of False Negatives (FN) and False Discovery Rates (FDR), reflecting its struggle to identify purchase events reliably. Similarly, Linear SVC failed to deliver competitive results, further underscoring the inadequacy of linear models for this task.

Random Forest (RF) and Decision Tree (DT) models demonstrate intermediate performance. While their interpretability and ability to model nonlinear patterns are advantageous, these models lacked the precision and balance of the more advanced gradient-boosting approaches. For

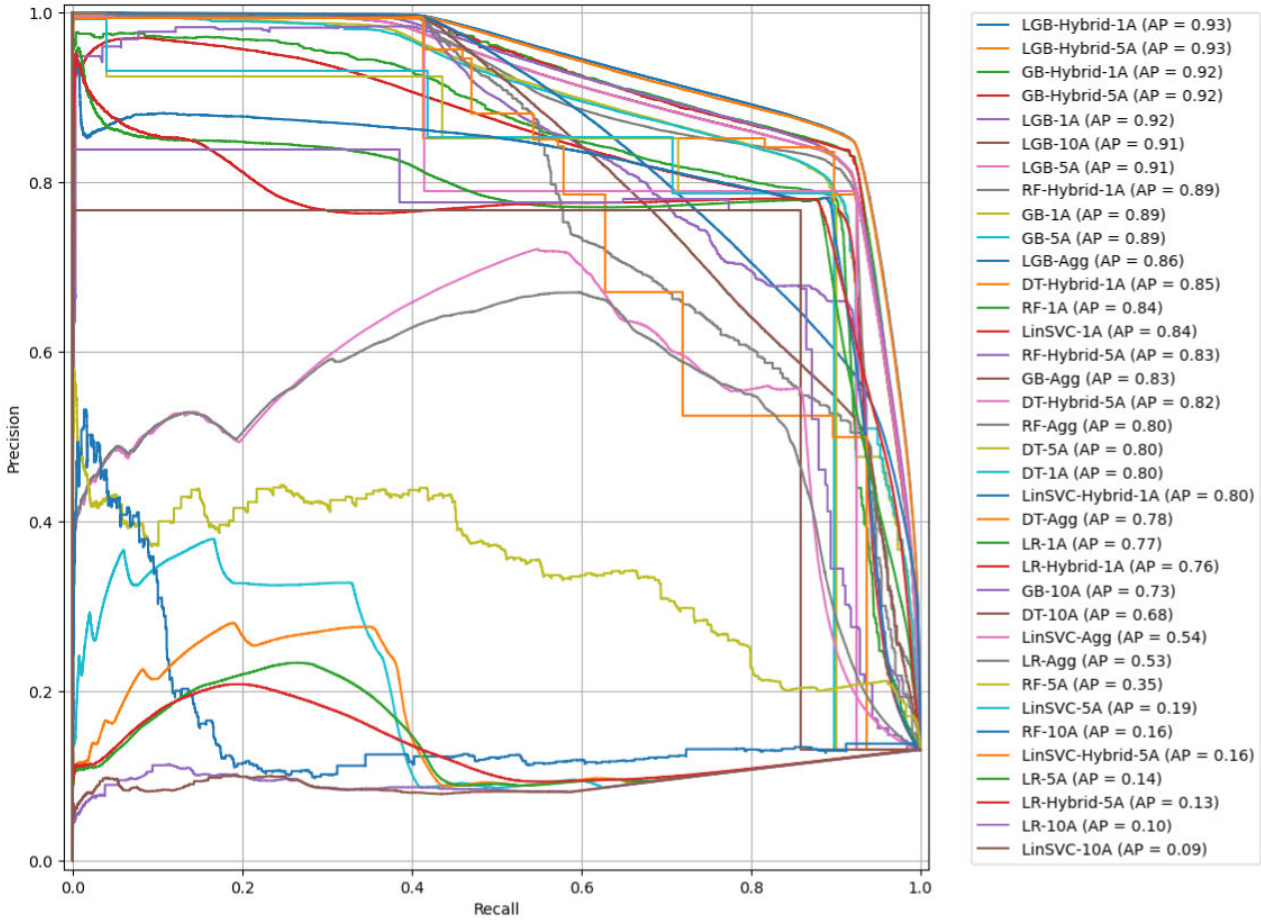


**FIGURE 3.** Precision-Recall curves for different models and data representations.

instance, RF models often showed higher variability across data representations, and DT configurations suffered from relatively poor calibration, as indicated by higher Validation Log Loss values.

The most promising results were achieved using Gradient Boosting (GB) and LightGBM (LGB) models, particularly when paired with hybrid data representations. Gradient Boosting models, such as GB-Hybrid-1A, achieved competitive F1 scores and geometric means, balancing sensitivity and specificity effectively. However, LightGBM, a highly optimized implementation of gradient boosting, consistently outperformed other models across all metrics.

The key performance metrics of the top-performing models are further discussed in Appendix B. In conclusion, LGB-Hybrid-1 emerged as the top-performing configuration, combining the predictive strength of hybrid data representations with the efficiency and accuracy of LightGBM. To provide a visual understanding of model performance, we present the confusion matrix and prediction distribution for this configuration.

The confusion matrix in Fig. 5 depicts the classification distribution for the purchase and nonpurchase predictions. This matrix indicates a high true positive rate (TPR) of 87%, illustrating the model's capability in correctly identifying purchase events. The false positive rate (FPR) is at a moderate level, which, combined with the high TPR, reflects the model's sensitivity to purchase intent without significantly compromising on nonpurchase accuracy. The relatively low rate of false negatives demonstrates the model's effectiveness in capturing minority class predictions, which is essential for applications in which missed purchases are costly. This performance underlines the model's suitability for real-world e-commerce applications, where identifying purchase events without excessive false positives or negatives is critical.

The prediction distribution in Fig. 6 offers insights into the confidence levels across the predicted classes. The top panel represents positive labels (purchases), with a significant concentration of probabilities close to 1, demonstrating the model's ability to confidently identify purchase events. This distribution indicates high precision for positive predictions, as the model effectively separates the true positive predictions from ambiguous cases. The bottom panel depicts negative labels (nonpurchases), which cluster near 0, reflecting the model's ability to differentiate nonpurchase events

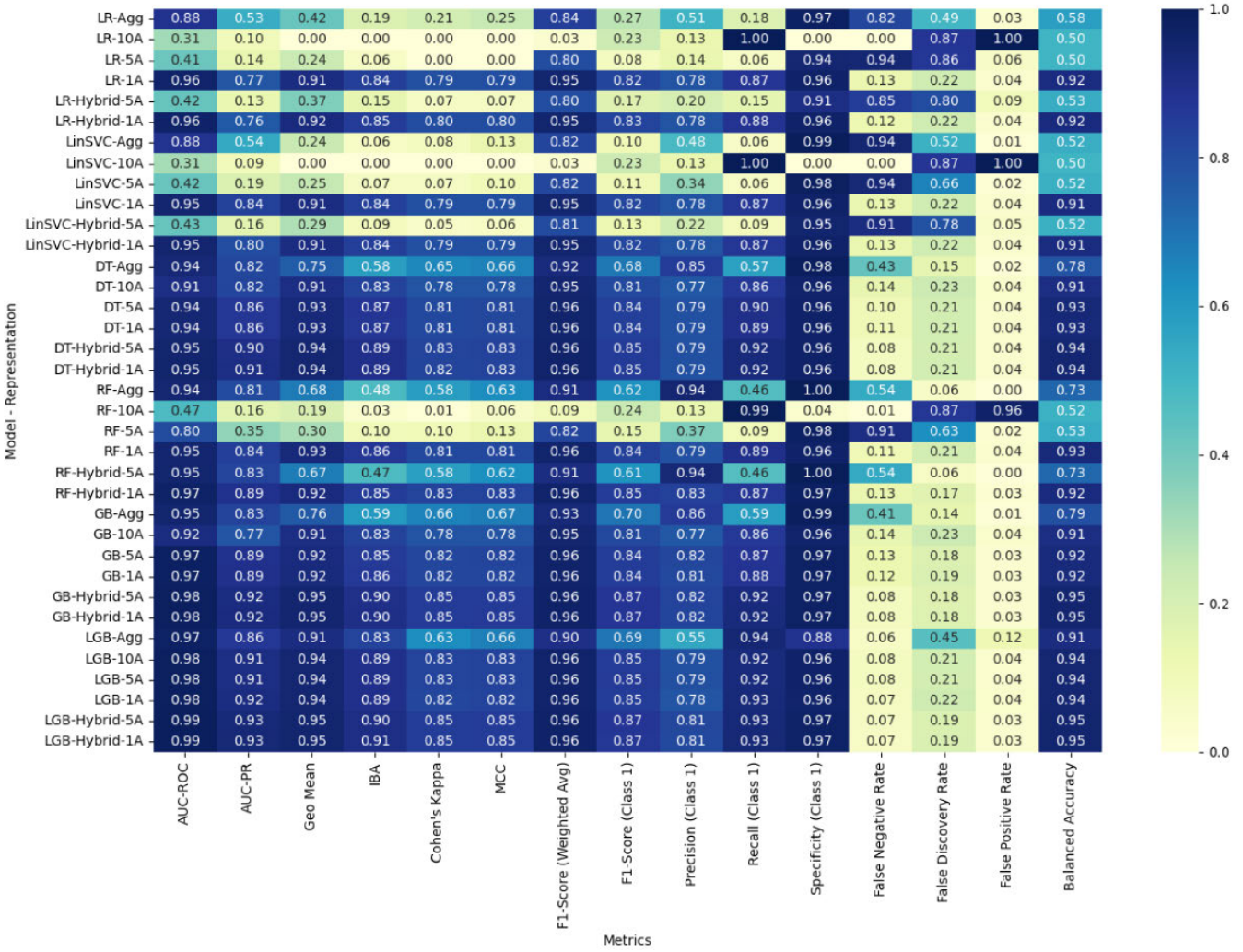


**FIGURE 4.** Performance metrics for different models and data representations.

accurately. Together, these distributions highlight the strong discriminative power of the model, as it minimizes the overlap between positive and negative probability scores, reinforcing its robustness under real-world conditions.

### D. FEATURE IMPORTANCE ANALYSIS

Feature importance analysis provides insights into the predictors driving purchase decisions, thus enhancing both the interpretability and practical utility of predictive models. In this study, feature importance was evaluated for tree-based models, including LightGBM, Gradient Boosting, Random Forest, and Decision Tree, using normalized importance scores.

Each model employs distinct techniques to compute feature importance. For LightGBM, importance scores are derived from both split-based and gain-based criteria, measuring how effectively each feature reduces the loss function during training. Similarly, Gradient Boosting and Random Forest models calculate importance by averaging impurity reductions across splits and trees, while Decision Trees use the total decrease in impurity attributed to each feature.

The heatmap in Fig. 7 visualizes the top 10 normalized feature importance scores across the models. Feature importance scores were normalized within each model as a percentage of the total importance, ensuring consistency and enabling direct comparisons. Temporal features such as `time_since_last_event` and session-level features such as `session_time_so_far` emerged as consistently significant predictors.

The Decision Tree models, while interpretable, identified fewer dominant features owing to their tendency to split on discrete variables. In the aggregated models, `num_view` and `num_cart` emerged as key contributors, reflecting the importance of user engagement. However, the feature importance values diminished for user actions, with only `event_type_int_1` showing consistent impact.

Random Forest models showed variability in feature importance, likely owing to their ensemble nature.

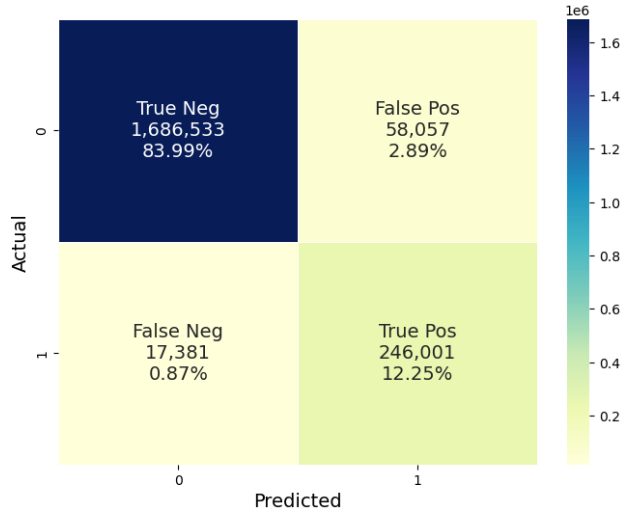


**FIGURE 5.** Confusion matrix for LightGBM Hybrid-1.

Aggregated data models strongly favored the number of events and the number of view events, with minimal contributions from demographic or static session features. For user actions data, features capturing recency and sequence (*e.g.* `event_type_int_1`, `session_time_so_far_1`) emerged as the most influential.

Gradient Boosting emphasizes the session duration, frequency of actions, and user engagement metrics. Sequential models highlight the importance of features related to recency, such as `time_since_last_event_1`, reaffirming the significance of the most recent behaviors in predicting purchases.

Across various configurations, LightGBM highlighted session-level features and recent user actions as the most influential predictors. For the aggregated session data, `session_start_time`, `session_time_sec`, and `event_count` consistently ranked high, thus underscoring the role of session dynamics in predicting user purchases. Notably, `average_price_per_event` and `product_count` further emphasize the importance of understanding user spending behavior within a session.

In LightGBM models utilizing sequential representations, features such as `time_since_last_event_1`, `price_1`, and `event_type_int_1` dominated, highlighting the predictive power of immediate user actions. The hybrid models combined the strengths of both approaches, with `session_start_time` and `event_count_so_far_1` frequently contributing significantly to model performance.

In aggregate data representations (Agg), no single feature dominated the importance rankings across the models. The prominence of `session_time_sec` suggests that longer user engagement positively correlates with purchase likelihood. However, features such as `brand_count` and `is_weekend` were consistently ranked as low-importance, indicating limited utility in distinguishing purchasing behavior from aggregated session summaries.

The last-*n*-actions representation (10A, 5A, and 1A) revealed the high predictive power of the most recent user actions. Features such as `time_since_last_event_1`, `event_type_int_1`, and `price_1` are among the most important, highlighting the role of the most recent user interactions in purchase decisions. Interestingly, the importance scores for deeper sequences (*e.g.*, `event_type_int_10`) tended to diminish, suggesting that older interactions contribute less to predictive accuracy.

Hybrid representations (Hybrid-1A and Hybrid-5A), which combine aggregated session features with recent user actions, provide a comprehensive view of user behavior. Features such as `session_start_time`, `event_count_so_far_1`, and `time_since_last_event_1` were highly ranked, indicating that combining short-term and long-term behavioral signals enhances model performance. This synergy between macro and micro-level user characteristics underscores the power of hybrid representations in capturing diverse predictors of purchase likelihood.

Across all models and representations, temporal features, such as `time_since_last_event` and session duration, consistently emerged as the most influential predictors. These attributes provide a clear indication of user engagement and immediacy in decision-making. Behavioral features, including the number of distinct products or brands interacted with during a session, further highlight the role of exploration in shaping purchase likelihood. Interestingly, features capturing long-term engagement ranked lower in importance, indicating that recent user actions had a more substantial impact on predicting purchases.

These findings underscore the importance of designing predictive frameworks that prioritize immediate user behavior without neglecting broader session-level context. For real-world e-commerce platforms, these insights translate into actionable strategies, such as tailoring marketing interventions based on time-sensitive features or dynamically adjusting promotions based on user engagement patterns.

## V. DISCUSSION

This study provides a detailed exploration of purchase prediction from e-commerce clickstream data, addressing the challenges of data representation and the predictive power of various machine learning models. The experimental results consistently demonstrate the superior performance of the LightGBM and Gradient Boosting models, particularly when applied to hybrid data representations that combine aggregated session attributes and recent user actions.

The comparative performance of the machine learning models in this study revealed several insights. Simpler models, such as Logistic Regression and Linear SVC, struggle to match the performance of others. These models lack the capacity to model complex, nonlinear relationships, which are critical for accurately predicting purchase intent

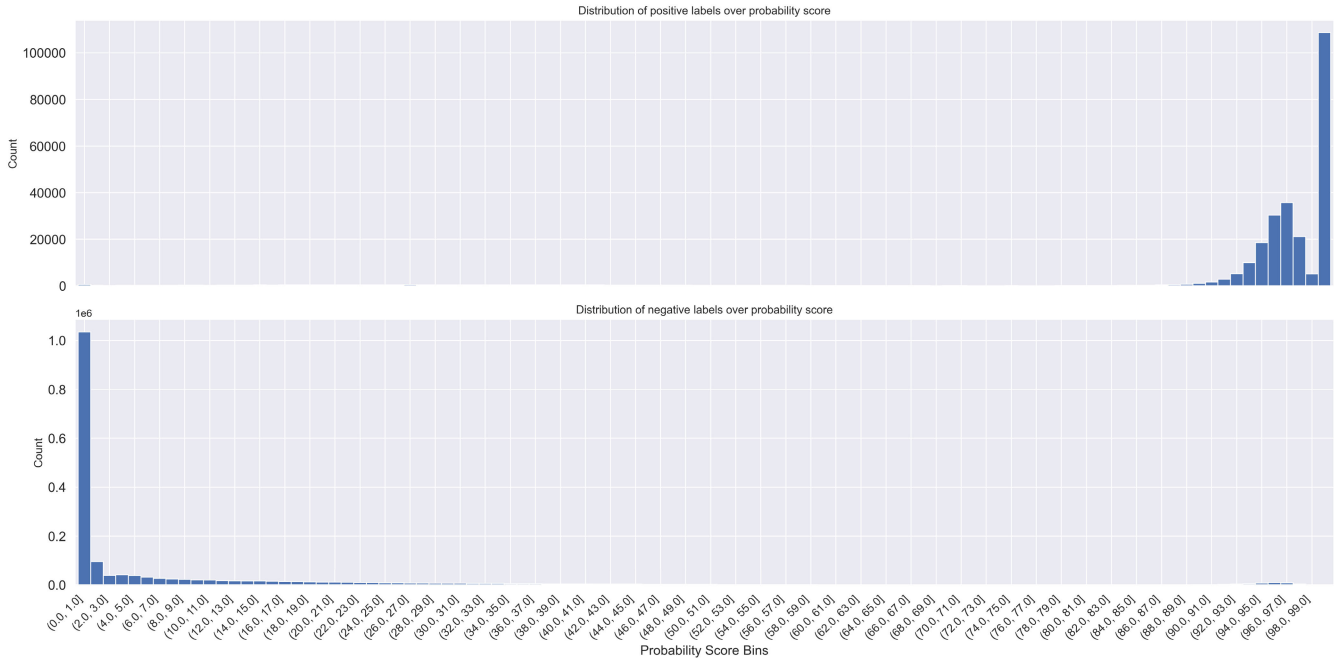


**FIGURE 6.** Distribution of positive and negative predictions vs. Actual labels for LightGBM Hybrid-1.

using clickstream data. Similarly, Decision Trees, although interpretable, often overfit in high-dimensional spaces and lack the robustness provided by ensemble approaches.

Random Forests, though an improvement over simpler models, exhibit higher variability in performance across data representations. Their inability to focus on minority classes effectively, coupled with higher computational demands for large datasets, limits their suitability for applications that require high predictive accuracy and scalability.

Gradient Boosting and LightGBM consistently delivered the best results across all data representations, demonstrating their ability to handle large-scale, high-dimensional datasets with nonlinear relationships. These models excel at capturing nonlinear relationships and complex interactions among features, which are prevalent in high-dimensional datasets such as the one used in this study. Additionally, these algorithms are particularly well-suited to address the class imbalance commonly observed in e-commerce datasets, where purchase events constitute only a small fraction of the data. LightGBM's use of class-weighted objective functions allows it to focus on the minority class without extensive preprocessing or sampling adjustments. Furthermore, its optimizations, such as Gradient-Based One-Side Sampling (GOSS) and Exclusive Feature Bundling (EFB), ensure computational efficiency even when processing large-scale datasets.

Additionally, this study uncovers that the choice of data representation significantly influences predictive performance. The comparison of data representations reveals that models based on recent user actions consistently outperform those relying solely on aggregated session-level attributes. This underscores the importance of temporal dynamics in capturing user intent. However, the hybrid approach–integrating recent actions with aggregated features–achieves the highest overall performance across all metrics. The combination of these perspectives allows models to leverage both temporal immediacy and broader session-level context, resulting in more accurate predictions.

Representations focused on the most recent user actions further highlight the importance of capturing temporal dependencies. These representations are particularly effective for models that can iteratively refine decision boundaries, such as Gradient Boosting. However, the reliance on sequential data alone may sacrifice context, as it excludes broader patterns captured by session-level aggregations. By contrast, hybrid representations strike a balance between these approaches, enabling models to harness the strengths of both granular and aggregated data.

The findings address a critical gap in existing research by demonstrating how hybrid representations effectively bridge the dichotomy between long-term behavioral patterns and immediate decision-making cues. Although hybrid representations offer improved predictive accuracy, they come at the cost of increased computational complexity. For real-time applications, platforms must balance the trade-off between model performance and processing overhead. The hybrid representation using the last 5 actions strikes a practical balance, delivering high accuracy without the excessive computational burden associated with longer sequences.

Feature importance analysis further highlights the predictive value of temporal features such as `time_since_last_event` and session-level attributes such as `session`

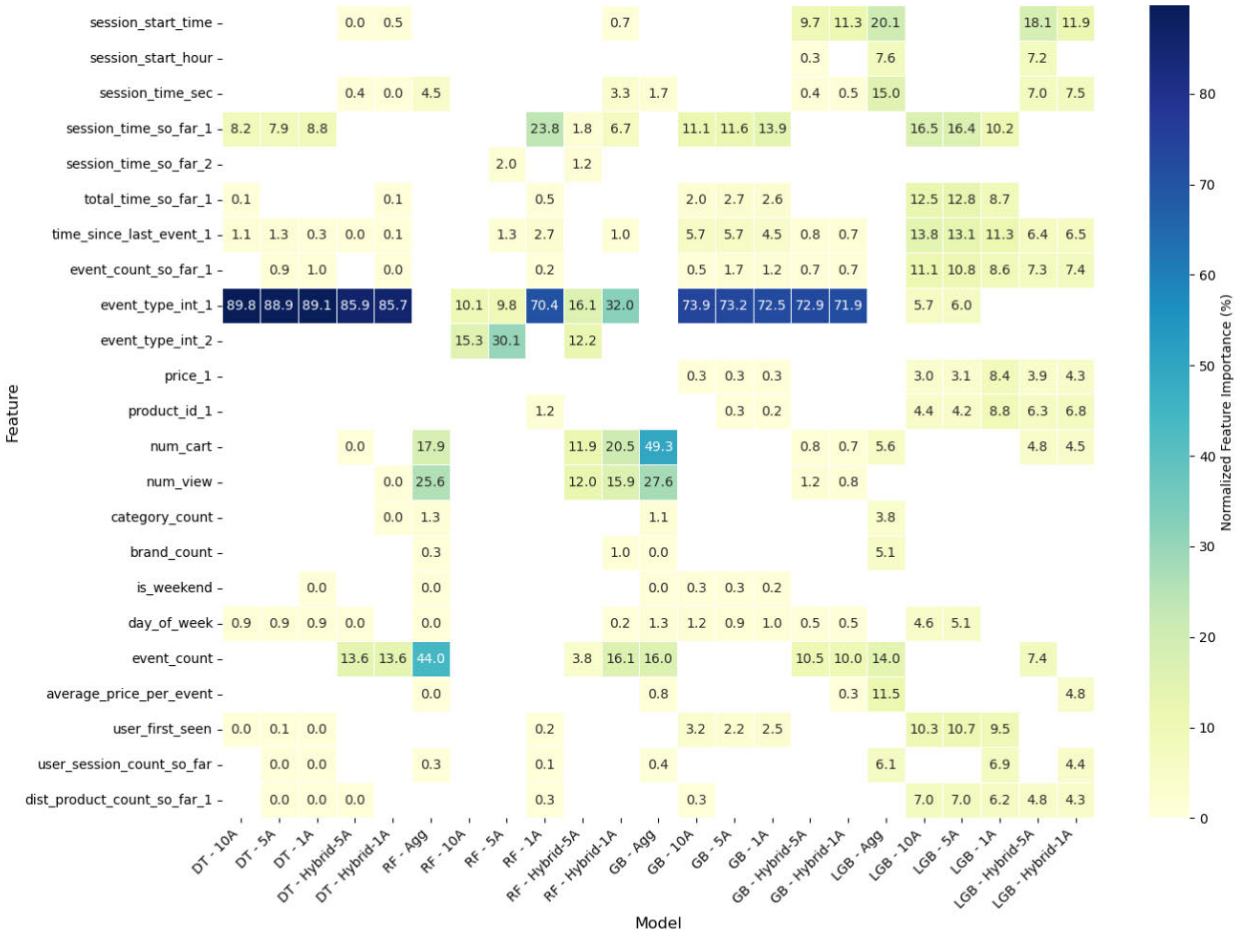


**FIGURE 7.** Normalized feature importance scores for the top 10 features across different machine learning models and data representations.

_time_sec denoting session length. These insights are particularly relevant for e-commerce platforms aiming to implement real-time targeting and personalized recommendations. For instance, the prominence of short-term behavioral signals suggests that real-time analytics systems should prioritize capturing and analyzing recent user actions to optimize conversion rates.

Despite these promising results, this study had several limitations. The use of a single dataset may limit the generalizability of findings to other e-commerce domains with distinct user behaviors or product offerings. Future research should validate these insights across diverse datasets and industries to ensure their broader applicability. Although this study provides strong empirical validation using a realistic e-commerce dataset, future work will validate these findings by incorporating additional datasets. Evaluating the model performance across multiple datasets will further assess the generalizability of the proposed representations in different domains. However, ensuring fair cross-dataset comparisons requires careful alignment of schema, session structures, and event granularity, which remains an ongoing challenge. These considerations will be a focus of future research. Nevertheless, the dataset used in this study reflects typical e-commerce behaviors, characterized by a predominance of browsing events, a smaller subset of cart additions, and an even smaller proportion of purchase events, which is consistent with other datasets in the domain. This similarity supports the generalizability of our findings across diverse e-commerce platforms.

While this study demonstrates the effectiveness of tree-based models in mitigating multicollinearity and achieving high predictive performance, certain trade-offs exist. The retention of highly correlated features, while essential for capturing patterns in the dataset, may complicate the interpretability of the model, particularly in contexts where actionable insights are required for decision-making. Similarly, although tree-based models such as LightGBM and Gradient Boosting achieve superior predictive performance, their complexity presents interpretability challenges. These models rely on ensembles of decision trees, which, despite offering feature importance scores, do not produce straightforward, human-interpretable decision rules. In contrast,

simpler models such as Logistic Regression and Decision Trees provide greater transparency but struggle to capture complex feature interactions, leading to reduced predictive accuracy. This trade-off is especially relevant for e-commerce applications where both explainability (e.g., for marketing strategies) and predictive power (e.g., for real-time targeting) are required. Future work could explore methods for improving interpretability, such as SHAP values or feature importance visualization, to better explain how correlated features contribute to model predictions.

In addition, the class imbalance in purchase prediction tasks poses challenges for model training and evaluation. Although strategies such as class weighting and balanced sampling mitigated these issues, they may not fully replicate real-world scenarios in which nonpurchase events dominate.

In analyzing session times, we observed substantial variability, with a few sessions exhibiting exceedingly long durations, potentially skewing the average session time upward. the mean session time was significantly higher than the median, indicating a right-skewed distribution heavily influenced by outliers. Supplementary analyses showed that removing outliers had minimal impact on the model performance, underscoring the robustness of the results.

Deploying purchase prediction models in live e-commerce systems presents several challenges, particularly in ensuring scalability, latency, and robust monitoring. The proposed system is implemented using PySpark, a production-friendly framework that is well-suited for handling large-scale data processing and distributed computing. Tree-based models such as LightGBM are well-suited for production because of their computational efficiency; however, for real-time predictions on platforms with high user traffic, optimizations such as quantization and model pruning are essential to minimize latency.

Monitoring the deployed system requires tracking both performance and operational metrics, such as latency and resource utilization. In addition, incorporating feedback loops, such as user interactions with predictions (*e.g.*, clicks or purchases), enables iterative model refinement. A/B testing or multivariate experiments are critical for validating the effectiveness of the model compared to existing systems, allowing businesses to quantify gains in metrics such as conversion rate or revenue per session. These practices not only ensure the reliability of the model but also pave the way for scalable, adaptive e-commerce applications capable of maintaining high performance in dynamic environments.

Future work could explore the integration of deep learning models, such as transformers or recurrent neural networks, to capture sequential dependencies in user behavior more effectively. Although computationally intensive, these models may further enhance performance by learning complex temporal patterns.

Moreover, expanding the feature set to include contextual factors, such as user demographics or external events, could enrich the model's predictive capacity. Incorporating such features may help capture broader behavioral trends and improve predictions in evolving e-commerce landscapes.

## VI. CONCLUSION

This study addresses a critical challenge in e-commerce: predicting purchase events using clickstream data. By systematically evaluating the impact of different data representations –aggregated session attributes, recent user actions, and hybrid combinations– this study provides a comprehensive understanding of how data structures influence predictive performance. Through detailed experimentation with multiple machine learning models, including LightGBM, Gradient Boosting, Random Forest, and logistic regression, we demonstrate the efficacy of hybrid representations in combining long-term session trends with immediate user behavior.

The findings underscore the superiority of hybrid representations, particularly when paired with efficient tree-based models such as LightGBM. The LightGBM model trained on Hybrid-1A representation consistently outperformed all other configurations, achieving the highest predictive accuracy and interpretability. These results validate the hypothesis that integrating short and long-term behavioral signals captures user intent more effectively than single-representation approaches.

The findings have practical implications for e-commerce platforms aiming to optimize user targeting and marketing strategies. The predictive power of recent user actions suggests that real-time analytics systems should prioritize capturing and analyzing the latest clickstream data to enhance recommendation engines and retargeting campaigns.

One of the most consistently influential features in our models, `time_since_last_event`, captures the time elapsed between consecutive user actions within a session. A lower value for this feature typically indicates rapid interactions, often signaling heightened engagement or purchase intent. Conversely, longer gaps between actions may suggest hesitation or disengagement, thereby reducing the likelihood of conversion. Integrating such influential features into dynamic pricing models or personalized recommendations can drive higher conversion rates by leveraging immediate user intent. The hybrid representation framework is not only applicable to purchase prediction but can also be adapted for related tasks such as churn prediction or personalized recommendations, offering a flexible and powerful tool for various domains.

The relative underperformance of purely aggregated data underscores the limitations of relying solely on session-level summaries. Although aggregated features are computationally efficient, their inability to capture nuanced, real-time user behavior restricts their effectiveness. Platforms with limited computational resources may benefit from hybrid models, which balance the granularity of recent actions with the broader context provided by the aggregated data.

**TABLE 11.** Hyperparameter search results for models.

| Model / Hyperparameter | Search Range | Agg | 10A | 5A | 1A | Hybrid-5A | Hybrid-1A |
|---|---|---|---|---|---|---|---|
| **LightGBM** | | | | | | | |
| numLeaves | [20, 150] | 53.000000 | 28.000000 | 73.000000 | 134.000000 | 145.000000 | 104.000000 |
| maxDepth | [3, 10] | 6.000000 | 5.000000 | 8.000000 | 7.000000 | 8.000000 | 6.000000 |
| learningRate | [0.01, 0.35] | 0.320439 | 0.250357 | 0.260237 | 0.248018 | 0.103169 | 0.223268 |
| featureFraction | [0.4, 1.0] | 0.609906 | 0.608363 | 0.640018 | 0.657464 | 0.462789 | 0.799158 |
| minDataInLeaf | [20, 150] | 56.000000 | 65.000000 | 42.000000 | 255.000000 | 80.000000 | 111.000000 |
| baggingFraction | [0.5, 1.0] | 0.726721 | 0.623935 | 0.704498 | 0.623935 | 0.789714 | 0.663305 |
| baggingFreq | [0, 10] | 6.000000 | 5.000000 | 8.000000 | 1.000000 | 3.000000 | 7.000000 |
| **GBM** | | | | | | | |
| stepSize | [0.01, 0.3] | 0.274528 | 0.203451 | 0.198140 | 0.210170 | 0.129266 | 0.155013 |
| maxIter | [50, 200] | 126 | 157 | 192 | 74 | 51 | 86 |
| maxDepth | [3, 10] | 5 | 7 | 7 | 6 | 7 | 6 |
| **DT** | | | | | | | |
| maxDepth | [3, 10] | 6 | 5 | 9 | 8 | 8 | 7 |
| minInstancesPerNode | [1, 50] | 16 | 17 | 5 | 13 | 11 | 10 |
| **RF** | | | | | | | |
| numTrees | [50, 200] | 150 | 100 | 200 | 120 | 180 | 140 |
| maxDepth | [3, 10] | 6 | 5 | 9 | 8 | 8 | 7 |
| **LR** | | | | | | | |
| regParam | [1e-5, ..., 1e-1, 1] | 1e-2 | 1e-3 | 1e-2 | 1e-1 | 1e-3 | 1e-3 |

**TABLE 12.** Key performance metrics for top performing models and data representations.

| Model - Representation | AUC-ROC | AUC-PR | Geo Mean (Class 1) | F1-Score Weighted | F1-Score (Class 1) | Precision (Class 1) | Recall (Class 1) | Log Loss Validation |
|---|---|---|---|---|---|---|---|---|
| LGB-Hybrid-1A | 0.986521 | 0.933718 | 0.950224 | 0.963556 | 0.867056 | 0.809059 | 0.934008 | 0.140536 |
| LGB-Hybrid-5A | 0.986135 | 0.931873 | 0.949879 | 0.964152 | 0.868918 | 0.813574 | 0.932342 | 0.142097 |
| GB-Hybrid-1A | 0.982178 | 0.919223 | 0.946622 | 0.964825 | 0.870519 | 0.823055 | 0.923791 | 0.172797 |
| GB-Hybrid-5A | 0.981866 | 0.916986 | 0.945634 | 0.964828 | 0.870352 | 0.824588 | 0.921494 | 0.173628 |
| LGB-1A | 0.981993 | 0.916715 | 0.943765 | 0.957180 | 0.845580 | 0.776522 | 0.928120 | 0.160986 |
| LGB-5A | 0.980684 | 0.90976 | 0.943536 | 0.959268 | 0.852069 | 0.790137 | 0.924535 | 0.164752 |
| LGB-10A | 0.980674 | 0.909979 | 0.943240 | 0.958934 | 0.850954 | 0.788372 | 0.924327 | 0.165226 |
| GB-1A | 0.973415 | 0.891054 | 0.922886 | 0.958117 | 0.844052 | 0.812032 | 0.878701 | 0.143460 |
| GB-5A | 0.973672 | 0.889585 | 0.920845 | 0.958298 | 0.844203 | 0.816486 | 0.873868 | 0.143466 |

This research advances the field by bridging a significant gap in the literature: the underexploration of hybrid data representations. Although previous studies have predominantly focused on either aggregated or sequential data, this study systematically evaluates their integration, highlighting the potential for scalable, interpretable, and high-performing predictive frameworks.

This study contributes to the broader understanding of user behavior modeling by offering actionable insights for e-commerce platforms. Real-time prediction systems can benefit from prioritizing immediate user actions, whereas hybrid approaches provide a comprehensive framework for capturing diverse behavioral signals. These findings also highlight the potential to improve marketing strategies, personalized recommendations, and demand forecasting through advanced predictive modeling.

Future research could explore more sophisticated sequential models, such as transformers, to capture deeper temporal dependencies in clickstream data. Another possible direction is to incorporate external factors such as promotions and seasonal trends. Additionally, applying this framework across different domains and datasets would validate its generalizability and further refine its applicability in diverse e-commerce settings.

In summary, the main findings of our study are as follows:

- Models trained on hybrid user actions outperform those trained on session-level attributes or recent user actions.
- LightGBM achieved the highest predictive accuracy, whereas the linear models underperformed.
- Temporal and behavioral features, such as time since the last event, session duration, and product interactions, were identified as the most influential factors in determining purchase likelihood.

In conclusion, this study not only enhances the understanding of data representation strategies for purchase prediction but also provides a practical and scalable framework for leveraging hybrid representations in real-world applications, paving the way for more sophisticated and effective e-commerce solutions.

## APPENDIX A
## HYPERPARAMETER SEARCH SPACE AND PARAMETERS

In Section III-H, we introduced the machine learning models used to predict whether a session will result in a purchase action.

The hyperparameter tuning process involved a TPE-based Bayesian search over predefined parameter ranges for each model to identify the best-performing configurations across

different dataset representations. Table 11 summarizes the hyperparameter search space for each model, as well as the optimal combination for each model-dataset variant. The table includes the results for LightGBM, Gradient Boosting Machines (GBM), Decision Trees (DT), Random Forests (RF), and Logistic Regression (LR) models. The columns represent the hyperparameter search range for the hyperparameter and the data representation.

## APPENDIX B
## KEY PERFORMANCE METRICS FOR THE TOP PERFORMING MODELS AND DATA REPRESENTATIONS

Table 12 presents the key performance metrics of the top-performing models. Among the evaluated models, the hybrid representations of LightGBM (LGB) demonstrated superior performance across the metrics. Notably, LightGBM with the Hybrid-1A representation stands out as the best-performing configuration, achieving the highest Validation AUC-ROC (0.9865), AUC-PR (0.9337) and Weighted F1-Score (0.9636). This indicates its superior capability to discriminate between purchase and nonpurchase sessions. The Geometric Mean (0.9502) and Validation Log Loss (0.1405) further underscore its robustness, particularly in managing imbalanced data. The combination of immediate user actions with aggregated session features in the Hybrid-1A representation captures both short and long-term behavioral signals, leading to better predictive performance.

## REFERENCES

[1] D. Cirqueira, M. Hofer, D. Nedbal, M. Helfert, and M. Bezbradica, "Customer purchase behavior prediction in e-commerce: A conceptual framework and research agenda," in *Proc. Int. Workshop New Frontiers Mining Complex Patterns*. Cham, Switzerland: Springer, May 2020, pp. 119–136.

[2] W. W. Moe and P. S. Fader, "Dynamic conversion behavior at e-commerce sites," *Manage. Sci.*, vol. 50, no. 3, pp. 326–335, Mar. 2004, doi: 10.1287/mnsc.1040.0153.

[3] Z. Wen, W. Lin, and H. Liu, "Machine-learning-based approach for anonymous online customer purchase intentions using clickstream data," *Systems*, vol. 11, no. 5, p. 255, May 2023, doi: 10.3390/systems11050255.

[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, pp. 1353–1364, Apr. 2023, doi: 10.1109/TEM.2021.3070069.

[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," *Appl. Comput. Eng.*, vol. 55, no. 1, pp. 181–195, Jul. 2024, doi: 10.54254/2755-2721/55/20241475.

[6] D. Koehn, S. Lessmann, and M. Schaal, "Predicting online shopping behaviour from clickstream data using deep learning," *Expert Syst. Appl.*, vol. 150, Jul. 2020, Art. no. 113342, doi: 10.1016/j.eswa.2020.113342.

[7] L. Bigon, G. Cassani, C. Greco, L. Lacasa, M. Pavoni, A. Polonioli, and J. Tagliabue, "Prediction is very hard, especially about conversion. Predicting user purchases from clickstream data in fashion e-commerce," 2019, *arXiv:1907.00400*.

[8] A. Ghadami and T. Tran, "TriDeepRec: A hybrid deep learning approach to content- and behavior-based recommendation systems," *User Model. User-Adapted Interact.*, vol. 34, no. 5, pp. 2085–2114, Nov. 2024, doi: 10.1007/s11257-024-09418-w.

[9] A. Chakraborty, V. Raturi, and S. Harsola, "BBE-LSWCM: A bootstrapped ensemble of long and short window clickstream models," in *Proc. 7th Joint Int. Conf. Data Sci. Manage. Data (11th ACM IKDD CODS 29th COMAD)*, Jan. 2024, pp. 350–358, doi: 10.1145/3632410.3632452.

[10] S. A. Amoudi, A. Alhothali, R. Mirza, H. Assalahi, and T. Aldosemani, "Click-based representation learning framework of Student navigational behavior in MOOCs," *IEEE Access*, vol. 12, pp. 121480–121494, 2024, doi: 10.1109/ACCESS.2024.3450514.

[11] E. Olmezogullari and M. S. Aktas, "Pattern2Vec: Representation of clickstream data sequences for learning user navigational behavior," *Concurrency Comput., Pract. Exper.*, vol. 34, no. 9, p. 6546, Apr. 2022, doi: 10.1002/cpe.6546.

[12] A. Martínez, C. Schmuck, S. Pereverzyev, C. Pirker, and M. Haltmeier, "A machine learning framework for customer purchase prediction in the non-contractual setting," *Eur. J. Oper. Res.*, vol. 281, no. 3, pp. 588–596, Mar. 2020, doi: 10.1016/j.ejor.2018.04.034.

[13] D. J. Bertsimas, A. J. Mersereau, and N. R. Patel, "Dynamic classification of online customers," in *Proc. SIAM Int. Conf. Data Mining*, May 2003, pp. 107–118, doi: 10.1137/1.9781611972733.10.

[14] W. Zucchini and I. L. MacDonald, *Hidden Markov Models for Time Series: An Introduction Using R*. London, U.K.: Chapman & Hall, 2009, doi: 10.1201/9781420010893.

[15] Y. Ozyurt, T. Hatt, C. Zhang, and S. Feuerriegel, "A deep Markov model for clickstream analytics in online shopping," in *Proc. ACM Web Conf.*, Apr. 2022, pp. 3071–3081, doi: 10.1145/3485447.3512027.

[16] Y. Zhang, H. Dai, C. Xu, J. Feng, T. Wang, and J. Bian, "Sequential click prediction for sponsored search with recurrent neural networks," in *Proc. Conf. AAAI Artif. Intell.*, 2014, vol. 28, no. 1, pp. 1369–1375, doi: 10.1609/aaai.v28i1.8917.

[17] 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 Comput. Appl.*, vol. 31, no. 10, pp. 6893–6908, Oct. 2019, doi: 10.1007/s00521-018-3523-0.

[18] K. Żołna and B. Romański, "User modeling using LSTM networks," in *Proc. 31st AAAI Conf. Artif. Intell.*, 2017, pp. 1–8.

[19] P. Jenkins, "ClickGraph: Web page embedding using clickstream data for multitask learning," in *Companion Proc. World Wide Web Conf.*, May 2019, pp. 37–41, doi: 10.1145/3308560.3314198.

[20] A. Verma, "Consumer behaviour in retail: Next logical purchase using deep neural network," 2020, *arXiv:2010.06952*.

[21] W. Wang, W. Xiong, J. Wang, L. Tao, S. Li, Y. Yi, X. Zou, and C. Li, "A user purchase behavior prediction method based on XGBoost," *Electronics*, vol. 12, no. 9, p. 2047, Apr. 2023, doi: 10.3390/electronics12092047.

[22] E. Ulitzsch, V. Ulitzsch, Q. He, and O. Lüdtke, "A machine learning-based procedure for leveraging clickstream data to investigate early predictability of failure on interactive tasks," *Behav. Res. Methods*, vol. 55, no. 3, pp. 1392–1412, Jun. 2022, doi: 10.3758/s13428-022-01844-1.

[23] L. Gan, "XGBoost-based e-commerce customer loss prediction," *Comput. Intell. Neurosci.*, vol. 2022, no. 1, Jul. 2022, Art. no. 1858300.

[24] A. A. Tokuç and T. Dağ, "Customer purchase intent prediction using feature aggregation on e-commerce clickstream data," in *Proc. 8th Int. Artif. Intell. Data Process. Symp. (IDAP)*, Sep. 2024, pp. 1–5.

[25] M. Hendriksen, E. Kuiper, P. Nauts, S. Schelter, and M. de Rijke, "Analyzing and predicting purchase intent in e-commerce: Anonymous vs. identified customers," 2020, *arXiv:2012.08777*.

[26] M. Saarela and S. Jauhiainen, "Comparison of feature importance measures as explanations for classification models," *Social Netw. Appl. Sci.*, vol. 3, no. 2, p. 272, Feb. 2021, doi: 10.1007/s42452-021-04148-9.

[27] D. Karl, "Forecasting e-commerce consumer returns: A systematic literature review," *Manage. Rev. Quart.*, May 2024, doi: 10.1007/s11301-024-00436-x.

[28] H. He and E. A. Garcia, "Learning from imbalanced data," *IEEE Trans. Knowl. Data Eng.*, vol. 21, no. 9, pp. 1263–1284, Sep. 2009, doi: 10.1109/TKDE.2008.239.

[29] M. Kechinov. (2019). *Ecommerce Behavior Data From Multi Category Store*. Accessed: May 5, 2022. [Online]. Available: https://www.kaggle.com/datasets/mkechinov/ecommerce-behavior-data-from-multi-category-store

[30] REES46 Inc. (2025). *REES46 for Ecommerce*. Accessed: Jan. 25, 2025. [Online]. Available: https://rees46.com/

[31] S. Chowdhury, Y. Lin, B. Liaw, and L. Kerby, "Evaluation of tree based regression over multiple linear regression for non-normally distributed data in battery performance," in *Proc. Int. Conf. Intell. Data Sci. Technol. Appl. (IDSTA)*, Sep. 2022, pp. 17–25, doi: 10.1109/IDSTA55301.2022.9923169.
[32] J. Kennedy and R. Eberhart, "Particle swarm optimization," in *Proc. Int. Conf. Neural Netw. (ICNN)*, vol. 4, 1995, pp. 1942–1948, doi: 10.1109/ICNN.1995.488968.
[33] B. Xue, M. Zhang, and W. N. Browne, "Particle swarm optimization for feature selection in classification: A multi-objective approach," *IEEE Trans. Cybern.*, vol. 43, no. 6, pp. 1656–1671, Dec. 2013, doi: 10.1109/TSMCB.2012.2227469.
[34] L. J. V. Miranda. (2017). *PySwarms: A Python-Based Particle Swarm Optimization (PSO) Library*. Accessed: Jan. 25, 2025. [Online]. Available: https://pyswarms.readthedocs.io/en/latest/
[35] J. Bergstra, B. Komer, C. Eliasmith, D. Yamins, and D. D. Cox, "Hyperopt: A Python library for model selection and hyperparameter optimization," *Comput. Sci. Discovery*, vol. 8, no. 1, Jul. 2015, Art. no. 014008, doi: 10.1088/1749-4699/8/1/014008.
[36] D. R. Cox, "The regression analysis of binary sequences," *J. Roy. Stat. Soc. B, Stat. Methodol.*, vol. 20, no. 2, pp. 215–232, Jul. 1958, doi: 10.1111/j.2517-6161.1958.tb00292.x.
[37] J. R. Quinlan, "Induction of decision trees," *Mach. Learn.*, vol. 1, no. 1, pp. 81–106, Mar. 1986, doi: 10.1007/bf00116251.
[38] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001, doi: 10.1023/a:1010933404324.
[39] J. H. Friedman, "Greedy function approximation: A gradient boosting machine," *Ann. Statist.*, vol. 29, no. 5, pp. 1189–1232, Oct. 2001, doi: 10.1214/aos/1013203451.
[40] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, and W. Ma, "LightGBM: A highly efficient gradient boosting decision tree," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, and S. Vishwanathan, Eds., 2017, pp. 1–9.
[41] Apache Softw. Found. (2023). *MLlib: Machine Learning Library—Spark 3.5.0 Documentation*. Accessed: Sep. 30, 2024. [Online]. Available: https://spark.apache.org/docs/3.5.0/ml-guide.html
[42] Microsoft. (2025). *SynapseML: Scalable and Distributed Machine Learning With Apache Spark*. Accessed: Sep. 30, 2024. [Online]. Available: https://microsoft.github.io/SynapseML/

**A. AYLİN TOKUÇ** received the B.S. degree in mathematics and the M.S. degree in computer science from Bilkent University, Ankara, Türkiye, in 2006 and 2008, respectively. She is currently pursuing the Ph.D. degree in computer engineering with Kadir Has University.

She has held various roles in software development and management across multiple industries. She is currently a Data Science Consultant in London, U.K. Her work primarily focuses on machine learning model development, predictive analytics, and AI-driven decision-making systems. She has authored research on data science applications on real-life business problems. Her research interests include data science, machine learning, artificial intelligence, and information retrieval.

**TAMER DAG** (Member, IEEE) received the M.S. and Ph.D. degrees in electrical and computer engineering from Northeastern University, Boston, MA, USA.

He has industry experience as a Senior Software Engineer with Lucent Technologies, where he worked on the design and implementation of high-performance network protocols and telecommunications systems. Currently, he is a full-time Associate Professor with the Department of Computer Engineering, American University of the Middle East (AUM), Kuwait.

Prof. Dag has served on the technical program and organizing committees of several international conferences and has published over 50 peer-reviewed research articles in international journals and conference proceedings. His research interests include wireless sensor networks, image processing, and machine learning, with a particular focus on autonomous network optimization, intelligent data analysis, and deep learning-based image classification.