International Journal of Research in Computing

ORIGINAL RESEARCH
Volume 5, Issue 1 (2026)

Enhancing E-Commerce Recommendation Systems through the Integration of Behavioral Analytics and Survey-Based User Insights

Gayantha Dilshan[1], Uchira C. Wickramarathne[1], Minushika G. Jayawickrama[1], Surani S. Tissera[1]

[1]Department of Computer Science, Faculty of Applied Sciences, University of Sri Jayewardenepura

Corresponding Author: gayanthadilshan@gmail.com

Abstract

This research explores the enhancement of e-commerce recommendation systems through data analytics with a particular focus on understanding and leveraging customer engagement factors. The study aimed to identify the key drivers of consumer interaction, assess the impact of integrating data-driven analytics and develop a predictive model to improve the accuracy and relevance of personalized recommendations. A convergent mixed-methods research design was employed, integrating quantitative analysis of digital behavioral data with survey-based qualitative insights to jointly explain customer engagement and recommendation effectiveness. This integration enabled the development and robust validation of a data-driven predictive model for personalized e-commerce recommendations. The findings show that combining digital behavioral data with consumer perceptions significantly enhances predictive accuracy as evidenced by improved error metrics (mean absolute error (MAE), root mean squared error (RMSE)), higher explanatory power (R^2) and stronger classification outcomes (precision, recall, area under the receiver operating characteristic curve (AUC-ROC)). Click frequency, session duration and perceived recommendation accuracy emerged as key predictors of engagement while data triangulation confirmed the model's reliability. In conclusion, this study demonstrates the value of data-driven personalization in e-commerce, offering practical benefits such as user engagement, marketing efficiency and conversion rates. Academically, it advances knowledge on predictive modeling and recommendation systems, underscoring the effectiveness of mixed-methods and advanced analytics. Future research should examine refinements and the long-term impact of dynamic personalization strategies.

KEYWORDS
Customer Engagement, Data-Driven Analytics, E-Commerce Recommendation Systems, Personalization, Predictive Modeling

ARTICLE HISTORY
Published: 14 Jan 2026

DATA/CODE AVAILABILITY
Data and code are not available due to privacy and confidentiality concerns.

SDG ALIGNMENT
SDG 3
SDG 4
SDG 8
SDG 9
SDG 11
SDG 12

1 Introduction

The rapid evolution of recommendation systems has emerged as one of the most transformative elements in modern e-commerce and digital consumer engagement. In today's competitive marketplace, businesses are no longer relying on generic strategies to attract customers; instead, they increasingly focus on delivering highly personalized content and product suggestions, a shift largely fuelled by data-driven analytics which enables organizations to capture, store and process massive volumes of consumer information, translating raw data into actionable insights.

The earliest recommendation systems were simple rule-based engines that relied on explicit signals such as user ratings or manually entered preferences. While effective in certain contexts, their limited adaptability often failed to capture the complexity of real-world consumer behaviour.

The emergence of collaborative filtering marked a significant step forward. By analysing relationships between users and items, collaborative filtering models predicted preferences based on observed similarities, even in the absence of explicit ratings. Both user-based and item-based approaches demonstrated greater flexibility and improved accuracy, allowing businesses to generate recommendations for users with sparse feedback.

With advances in computational resources and algorithms, Machine Learning (ML) and Artificial Intelligence (AI) began to play an increasingly important role. Models such as matrix factorization, deep learning architectures and hybrid approaches combining collaborative and content-based filtering enabled systems to identify latent behavioural patterns across vast datasets. Hybrid models, in particular, helped mitigate the weaknesses of individual approaches, leading to more robust and context-sensitive recommendations.

Today, modern recommendation engines not only improve accuracy but also adapt in real-time. By continuously learning from user interactions, these systems adjust recommendations to match emerging trends, seasonal variations and short-term behavioural shifts, reflecting the growing importance of personalization in driving customer engagement and loyalty.

In parallel with algorithmic advancements, e-commerce has witnessed an unprecedented surge in data generation. Clicks, page views, search queries, time spent on webpages, purchases and abandoned carts all contribute to vast consumer information repositories. This is amplified by smartphone adoption, social media use and the globalization of online shopping platforms.

The resulting datasets are massive and diverse, encompassing transactional histories, browsing behaviours, reviews, demographic details and social media activities. This diversity offers opportunities to understand consumer behaviour from multiple dimensions, enabling refined personalization strategies. For example, analysing purchasing trends alongside browsing activity can reveal hidden interests, while social media interactions provide insight into peer influence and emerging trends.

However, managing data remains challenging. Traditional processing methods are often inadequate for the volume, velocity and variety of modern e-commerce data. To address this, organizations increasingly use distributed computing frameworks, advanced cloud storage and parallel processing technologies. Platforms such as Hadoop and Apache Spark enable large-scale, real-time analytics, making it feasible to transform vast data into actionable intelligence.

The e-commerce industry is currently shaped by rising demand for personalization and seamless experiences. Customers expect not only product recommendations but also real-time, context-aware interactions. Key trends include:

- **Real-Time Data Utilization** – Businesses incorporate live behavioural signals into recommendation algorithms, allowing suggestions to adjust instantly based on recent browsing or purchasing activity.
- **Integration of Multiple Data Streams** – Instead of analysing transactional, browsing and social media data separately, unified frameworks provide a holistic understanding of

customer behaviour and reveal subtle interdependencies.

- **AI-Driven Predictive Analytics** – Modern ML-based recommendation systems evolve continuously as new data arrives, adapting to temporary consumer interests, changing market conditions and short-lived trends.
- **Scalability and Cloud Adoption** – With growing data volumes, scalable cloud-based infrastructures and data-driven frameworks enable fast, accurate and high-performance recommendation systems.

In summary, the interplay between technological advancements, data proliferation and consumer expectations has established recommendation systems as a cornerstone of digital commerce. However, challenges in predictive accuracy, scalability and data integration remain unresolved, forming the basis of this research.

While recommendation systems and data-driven analytics have transformed personalization in e-commerce, several challenges remain unresolved, undermining predictive capacity and efficiency.

A major limitation lies in the reliance on historical data. Most systems assume that past behaviour predicts future actions. Yet in fast-changing markets, consumer preferences shift rapidly due to economic conditions, seasonal effects, marketing campaigns or social influences. Recommendations often appear misaligned with a users' current interests, undermining trust and engagement.

The exponential growth of e-commerce data also presents scalability challenges. Systems designed for smaller datasets often struggle with large-scale, heterogeneous data. Additionally, the cold-start problem persists; new users or products lack sufficient data for personalized recommendations, leading to generic or popular-item suggestions.

Although e-commerce platforms now collect diverse data, such as browsing logs, click-streams, purchase histories and social interactions, these are often analysed in isolation. This fragmented approach prevents systems from capturing contextual relationships, limiting personalization. Integrating heterogeneous data streams also introduces computational and methodological complexities.

In conclusion, predictive accuracy, scalability limitations and fragmented data integration hinder current recommendation systems. Addressing these issues is critical for developing more adaptive, scalable and accurate models that meet modern e-commerce demands.

This research aims to overcome the above challenges by leveraging data-driven analytics to enhance personalization in recommendation systems and improve customer engagement in e-commerce.s

Primary objectives are,

- Investigate behavioural, demographic and transactional variables (e.g., browsing history, click-streams and purchase records) influencing customer engagement.
- Assess the role of data-driven analytics in enhancing predictive accuracy and recommendation relevance by comparing performance metrics such as precision, recall and AUC-ROC with traditional approaches.
- Design, train and validate a data-driven model that leverages both historical and real-time data to ensure adaptability in dynamic market conditions.

Secondary Objectives are,

- Conduct a comparative study of conventional recommendation systems and data-enhanced, mixed-methods-based systems.

- Evaluate model effectiveness using both algorithmic measures (precision, recall and AUC-ROC) and practical performance indicators (click-through and conversion rates).
- Investigate strategies to ensure model responsiveness, scalability and accuracy under large-scale, high-load operational conditions.

From a methodological perspective, understanding customer engagement in e-commerce requires both objective behavioural evidence and subjective user perceptions. Quantitative interaction data such as click streams and browsing histories capture what users do, but they do not fully explain why users perceive recommendations as relevant. Conversely, survey-based methods capture attitudinal and perceptual dimensions that are not observable in system logs. Therefore, a mixed-methods research approach is particularly suitable for this study, as it enables the integration of behavioural analytics with user-reported insights by providing a more comprehensive and validated understanding of recommendation effectiveness.

This paper proposes and validates a data-driven predictive model that integrates behavioural and perceptual data to enhance personalization and engagement in e-commerce.

2 Literature review

Recommendation systems have evolved dramatically over the past three decades, moving from basic heuristic-driven tools to intelligent, adaptive and data-intensive models.

Early approaches were rule-based systems that relied heavily on explicit user input, such as ratings, purchase history or association rules (e.g., "users who purchased A also purchased B") [1]. While pioneering, these systems were rigid, failed to adapt to rapidly changing user preferences and often struggled with incomplete or sparse data.

A major turning point came with the introduction of Collaborative Filtering (CF), which shifted the focus from individual data to collective intelligence. CF made personalization possible by identifying similarities between users or items. User-based methods compared users with similar tastes while item-based methods examined correlations among items [2], [31]. Though effective, these methods faced scalability challenges as datasets grew exponentially [35], [43]. A Netflix Prize competition further transformed the field by popularizing matrix factorization techniques. These methods extracted latent factors that explained hidden patterns in user-item interactions, leading to significant improvements in accuracy and prediction quality [3], [4], [5].

Next, hybrid models combined collaborative filtering with content-based filtering which incorporated item features (e.g., genre, brand or specifications). This approach mitigated the cold-start problem by enabling recommendations even for new users or products [36], [6]. In the last decade, deep learning techniques have advanced recommendation systems by uncovering complex, nonlinear relationships within large-scale heterogeneous data. Neural architectures such as convolutional neural networks (CNNs), recurrent neural networks (RNNs) and attention-based models have been used to learn from multimodal data including text, images and sequential clickstreams [7]. Moreover, context-aware models now incorporate situational information, such as time of day, location or social setting, to generate more personalized, real-time suggestions [38], [41], [8].

This evolution reflects a continuous trend toward greater personalization, adaptability and intelligence, transforming recommendation systems from static utilities into dynamic engines of user engagement and revenue generation. The emergence of data has radically altered the e-commerce landscape, providing businesses with unprecedented opportunities to capture and act upon consumer insights in real time. E-commerce platforms now generate diverse, high-volume data streams, including:

- Clickstream data that records every user interaction on websites [9].
- Social media content such as likes, shares and reviews [10].

- Browsing and purchasing histories, reflecting evolving customer preferences [46].

When integrated, these sources yield rich and multi-dimensional consumer profiles, enabling more precise targeting and personalization [39].However, large-scale and heterogeneous data environments introduce the "3Vs" challenge, volume, velocity and variety [11], [12]. The immense scale of data complicates storage and processing while heterogeneous format (structured, unstructured, semi-structured) increases integration complexity [40]. The demand for real-time analytics further pushes the limits of traditional systems [13].

To address these issues, scalable solutions such as distributed computing frameworks (e.g., MapReduce, Hadoop) [47] and cloud-based platforms (e.g., Spark) [33] have become essential. Even so, achieving efficiency and accuracy at scale remains a key research frontier. Despite these challenges, data-intensive analytics creates opportunities for:

- Enhanced personalization through deeper consumer profiling
- Dynamic adaptation based on real-time feedback loops
- Predictive modelling leveraging multimodal data fusion [46]

In short, data-driven and analytics-enhanced approaches has shifted recommendation systems from reactive (based only on past purchases) to proactive and predictive, shaping future consumer behaviour. Predictive modelling forms the backbone of modern recommendation systems, integrating both classical statistical tools and contemporary ML methods. Traditional approaches, including regression models, time-series forecasting and factor analysis, provided foundational insights into consumer preferences and purchasing trends [14], [15]. Multivariate analysis and dimensionality reduction methods such as principal component analysis (PCA) and clustering helped reduce complexity and enhance interpretability in high-dimensional datasets [16], [17].

ML techniques such as decision trees, support vector machines (SVM) and random forests introduced greater flexibility in modelling non-linear and high-dimensional interactions [18], [19]. The rise of deep learning models, including, CNNs and RNNs, further advanced predictive capabilities by capturing spatial and sequential patterns within consumer activity, such as clickstreams or session-based behaviour [44], [20].

In addition, survey-based approaches remain valuable, as they capture consumer attitudes, motivations and intentions that may not be reflected in transactional data [21], [22]. Combining survey data with ML outputs enriches interpretability and improves validation of predictive models [42].

Model evaluation is equally critical. Metrics such as Mean Absolute Error (MAE), Root Mean Squared error (RMSE) and R^2 are widely used in regression tasks [48], while precision, recall and AUC-ROC are applied to classification problems [45]. Cross-validation ensures generalizability [23] and in data contexts, scalability metrics such as processing time, throughput and resource utilization are essential [47], [33].

Together, these predictive modelling techniques form the analytical engine that powers recommendation systems, balancing accuracy, scalability and interpretability. Although substantial progress has been made, several critical research gaps persist:

- **Over-reliance on historical data** – Many systems prioritize past behaviour to generate recommendations, which limits adaptability in fast-changing markets where user preferences evolve due to trends, campaigns and external factors [3,7].
- **Scalability issues** – Efficiently handling large and heterogeneous datasets remains challenging, particularly when real-time behavioural data must be processed alongside historical records without degrading system performance [31,37].

- **Cold-start problem** – Personalization for new users and items continues to be difficult due to the lack of sufficient interaction data, often resulting in generic or popularity-based recommendations despite improvements in hybrid approaches [36,6].
- **Limited use of qualitative insights** – Transactional data alone cannot fully capture consumer attitudes and motivations, and the limited incorporation of survey-based insights reduces the interpretability and contextual relevance of recommendation outcomes [24,22].

Opportunities lie in leveraging data-driven predictive models, which offer the ability to:

- **Integrate multi-source data for richer personalization** – Combine behavioural, transactional and contextual data streams to enhance recommendation relevance and user engagement.
- **Employ scalable distributed frameworks for large datasets** – Utilize distributed computing platforms to efficiently process high-volume and high-velocity data while maintaining system performance [47].
- **Adapt dynamically in real-time** – Enable recommendation models to update continuously based on live user interactions and evolving behavioural patterns.
- **Apply advanced data fusion methods** – Integrate structured and unstructured data sources using advanced data fusion techniques to improve model robustness and predictive accuracy [13].

A. Research Questions and Hypotheses Accordingly, this study addresses three research questions:

- **RQ1:** How can integrated data-driven analytics improve predictive accuracy compared to traditional models?
- **RQ2:** Which consumer engagement factors can be effectively captured by combining quantitative and survey-based insights?
- **RQ3:** How can scalable models address the cold-start problem while enabling real-time personalization?

Based on the identified literature gaps, the following hypotheses state the study's predictive expectations:

- **H1:** Data-driven models integrating clickstream, social media and browsing history outperform traditional historical-data-based models.
- **H2:** Incorporating survey-based qualitative insights enhances the identification of engagement factors, leading to more tailored and context-aware recommendations.
- **H3:** Scalable, real-time frameworks can mitigate the cold-start problem and improve personalization for new users and items.

The literature demonstrates a clear trajectory, from rule-based heuristics to collaborative and hybrid systems and ultimately to data-enabled, deep learning-driven approaches. While each stage has brought measurable improvements, persistent issues remain in scalability, adaptability and capturing nuanced human behaviour.

The gaps identified suggest that next-generation recommendation systems must focus on integrating heterogeneous data streams, real-time predictive modelling and qualitative insights to achieve both technical robustness and practical relevance. This study seeks to address these gaps by developing a framework that combines data-driven analytics, predictive modelling and survey-based inputs to create mode adaptive and user-centric recommendation systems.

3 Methodology

This section outlines the methodology employed in this study, detailing the methods and procedures employed to develop and validate a data-driven predictive model for e-commerce recommendation systems. This study adopts a convergent mixed-methods research design, in which quantitative behavioural data and survey-based qualitative data are collected and analyzed in parallel and then integrated during interpretation and model validation. The section is organized into five main sections. (1) Research Design; (2) Quantitative Methods and Survey Design; (3) Data Collection and Sampling; (4) Data Analysis Techniques and (5) Ethical Considerations. Figure 1 illustrates the overall research methodology employed in this study, highlighting the key stages of research design, data collection, survey development and the integration of quantitative and survey-based analyses.

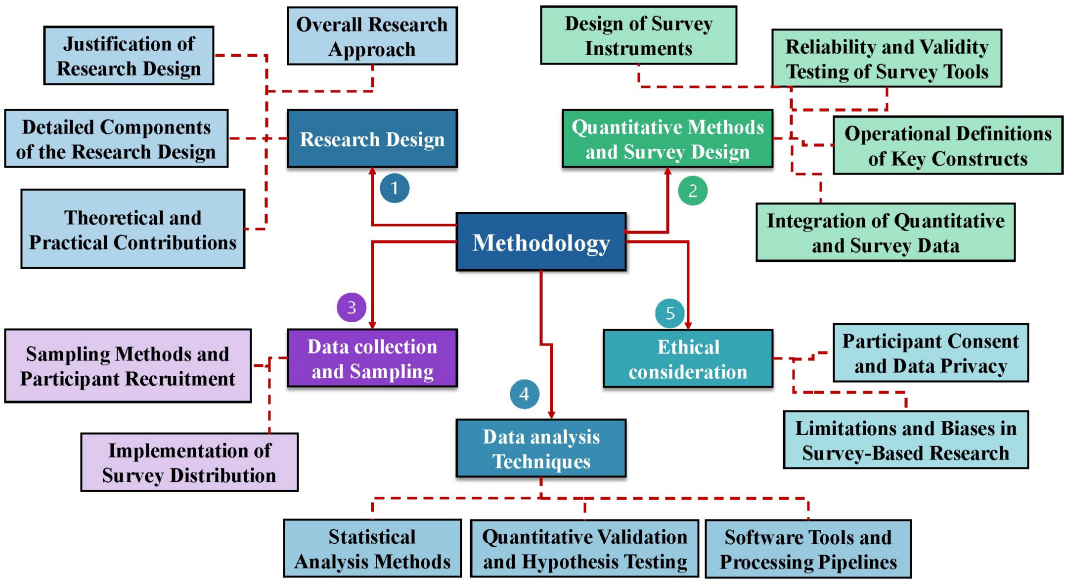

Figure 1: Methodology Overview of the Study

3.1 Research design

This study employs an integrated quantitative and survey-based research design to investigate consumer engagement factors and predict recommendation accuracy in e-commerce platforms. By combining secondary digital interaction data (e.g., click-streams and browsing histories) with primary survey data, the study leverages the strengths of both approaches. This dual methodology enables triangulation of findings, ensuring the predictive model is robust, reflective of real-world behaviour and aligned with user perceptions.

The quantitative component analyzes large-scale datasets and builds ML models to forecast consumer behaviours, using techniques such as time-series forecasting, regression analysis, decision trees and ensemble algorithms to capture patterns and non-linear relationships [18], [25]. Simultaneously, the survey component collects first-hand consumer insights, measuring satisfaction with recommendation systems, perceived relevance and latent behavioural factors. Survey data both validate the predictive model and provide additional variables to improve accuracy, helping address common challenges like overreliance on historical data and the "cold start" problem [6], [35].

The decision to employ a mixed-method approach is supported by several key considerations:

- **Capturing the Complexity of Consumer Behaviour:** Consumer interactions in digital environments are inherently multifaceted. Quantitative digital footprint data such as click-streams capture observable behaviours, while surveys measure subjective perceptions including satisfaction and perceived personalization quality. Combining these approaches provides a more complete understanding of consumer behaviour by integrating both objective actions and subjective experiences [26].
- **Leveraging Data-Driven Opportunities:** E-commerce platforms generate large volumes of user data which machine learning algorithms can analyze to uncover hidden patterns that traditional methods might miss [7]. Integrating survey data ensures that predictions align with real user experiences, enhancing the relevance and practical applicability of the model.
- **Triangulation for Enhanced Validity:** The integration of quantitative and survey data enables triangulation, which strengthens construct validity and mitigates potential biases inherent in each method. Triangulation also supports cross-validation, ensuring that model predictions correspond with self-reported user experiences [24].
- **Adaptability and Scalability:** A mixed-methods design offers flexibility and scalability. Quantitative models can be expanded to accommodate larger datasets, while survey instruments can be refined based on preliminary findings. This adaptability is critical for e-commerce platforms, where consumer behaviour patterns continuously evolve [33].
- **Addressing the Cold-Start Problem:** Recommendation systems often struggle with new users or products due to limited historical data. Incorporating survey-derived variables such as self-reported preferences and brand awareness helps mitigate the cold-start problem and improves predictive accuracy [36].

The research design is implemented through a multi-stage process, ensuring systematic development, validation and refinement of the predictive model.

1. **Stage 1: Data Collection and Model Development**
   The first stage involves collecting secondary digital interaction data from e-commerce platforms. Quantitative models are developed using regression analysis and machine learning algorithms such as Random Forests and Support Vector Machines (SVMs). The predictive performance of these models is initially evaluated using historical interaction data, providing a foundation for subsequent analysis.
2. **Stage 2: Survey Design and Administration**
   In parallel, structured surveys are designed to capture user insights on engagement, satisfaction, and perceived relevance of recommendations. These surveys are carefully piloted and validated to ensure reliability, thereby adding a human-centered dimension to the research and complementing the quantitative findings [21].
3. **Stage 3: Integration and Comparative Analysis**
   The outputs from the quantitative models and survey responses are then integrated. Comparative analysis examines the correlation between predicted user preferences, recommendation relevance, and self-reported satisfaction. Structural Equation Modelling (SEM) and other statistical techniques are employed to assess the alignment between observed digital behaviour and subjective user perceptions [27].
4. **Stage 4: Model Refinement and Scalability Testing**
   Finally, the predictive model is refined based on the integration results to enhance accuracy and address identified gaps. Sensitivity analysis and cross-validation techniques are

applied to further calibrate the model, ensuring robustness and scalability across diverse datasets and evolving e-commerce environments.

The integrated research design contributes both theoretically and practically to the field of e-commerce recommendation systems. Theoretically, it extends traditional frameworks by incorporating real-time digital interactions alongside user-perceived experiences which offering deeper insights into latent variables that shape consumer engagement and recommendation effectiveness. Practically, the design provides actionable guidance for practitioners seeking to improve recommendation accuracy and engagement strategies. By combining quantitative data with survey-based insights, it addresses key limitations such as the cold-start problem and the overreliance on historical datasets, thereby enhancing both the reliability and applicability of recommendation systems [2], [35].

3.2 Quantitative methods and survey design

The survey instruments in this study are designed to capture detailed insights into consumer engagement and perceptions of recommendation systems. To ensure comprehensive coverage, the survey is structured into four sections. The demographic section collects key background information such as age, gender, income and education level, enabling segment-based analyses and the use of demographic variables as control factors [26]. The consumer behaviour section focuses on respondents' online shopping patterns including frequency of visits, average session duration and number of product views, with items measured on Likert scales to quantify engagement [24]. A separate section addresses perceptions of recommendation systems, incorporating both closed-ended questions on perceived accuracy, relevance and satisfaction and open-ended items for qualitative insights [28]. Finally, a forward-looking section evaluates consumer feedback on potential predictive model features, exploring the perceived importance of real-time data, transparency and enhanced personalization [22]. Table 1 summarizes the dependent, independent and control variables used in this study, along with their definition.

To ensure reliability and validity, the survey underwent multiple testing procedures. A pilot study was conducted with a sample of 30 participants representative of the target population, allowing refinements to question wording, scale consistency and overall survey structure [21]. Internal consistency of multi-item scales was assessed using Cronbach's alpha with coefficients above 0.70 considered acceptable for constructs such as engagement and satisfaction [29]. Construct validly was examined through exploratory factor analysis (EFA) and confirmatory factor analysis (CFA), confirming that items grouped under constructs such as consumer engagement and recommendation accuracy measured the intended dimensions [27]. In addition, content validity was strengthened through expert review by specialists in e-commerce and digital marketing, ensuring the instrument was comprehensive and aligned with current research priorities [26].

To maintain clarity and consistency, the study adopts precise operational definitions for its core constructs. Consumer engagement is defined as the intensity and frequency of user interactions with an e-commerce platform, measured through behavioural indicators such as session length, page views and interaction frequency. Recommendation accuracy refers to the extent to which system -generated suggestions align with actual user preferences, measured using

3.3 Data Collection and Sampling

The study adopts a comprehensive data collection and sampling strategy to ensure robustness, representativeness and validity of findings. Both real-world consumer data and survey responses are integrated to strengthen the predictive model.

A multi-stage sampling strategy was employed to capture a diverse and representative consumer base.

Table 1: Variable Specification

Variable Category	Variable Name	Description
Dependent Variables	Customer Engagement	Overall level of user interaction with the e-commerce platform, measured using session duration, interaction frequency, and page views, etc.
	Recommendation Accuracy	Degree to which system-generated recommendations align with user preferences, measured using click-through rate, conversion rate, and perceived accuracy scores.
	User Satisfaction	Overall satisfaction with the recommendation system, reflecting perceived relevance, usefulness, and experience quality.
Independent Variables	Click Frequency	Number of user clicks on recommended items within a session.
	Session Duration	Time spent by a user during each platform visit.
	Page Views	Number of product or content pages viewed per session.
	Purchase History	Historical transaction records indicating prior buying behavior.
	Browsing History	Sequence and categories of items viewed by users.
Control Variables	Age	Age group of the respondent.
	Gender	Gender of the respondent.
	Income Level	Self-reported income category.
	Education Level	Highest education qualification.
	Shopping Frequency	Frequency of online shopping activity.

- **Stratified Random Sampling:** The population was segmented by demographic factors such as age, gender, income, and geographic region. This method ensured proportional representation, minimized selection bias, and enabled subgroup analyses [26].
- **Online Recruitment Strategies:** Participants were recruited through social media advertisements, email newsletters, and invitations embedded within partner e-commerce platforms. Pre-screening questions verified inclusion criteria such as regular online shopping and familiarity with recommendation systems. This targeted recruitment approach enhanced diversity and representativeness [24].

Inclusion criteria required participants to be active e-commerce users with prior exposure to recommendation systems. Respondents who provided incomplete survey responses were excluded prior to analysis to ensure data quality and statistical reliability. The final dataset consisted of 500 valid participant responses.

The research design integrates both empirical and simulated datasets to enhance model robustness.

- **Real-World Datasets:** The study utilized click-stream data, browsing histories, and

transaction records obtained from partnering e-commerce platforms over a specified time frame. These real-world datasets formed the empirical foundation for predictive modeling while ensuring compliance with strict data security and privacy standards, in line with industry protocols [33].

- **Synthetic Datasets:** In parallel, synthetic datasets were generated using techniques such as Monte Carlo simulations and bootstrapping. These datasets were particularly valuable in addressing cold-start scenarios involving new users or products and were instrumental in testing the scalability and generalizability of the predictive model [7].

Surveys were administered via a secure online platform accessible across mobile and desktop devices. The distribution strategy incorporated the following measures.

- **Survey Invitation:** Clear communication of the study purpose, expected completion time (10–15 minutes), and assurances of confidentiality and anonymity.
- **Incentivization:** Use of small rewards (e.g., discount vouchers, prize draw entries) to encourage participation and broaden sample diversity [21].
- **Automated Follow-Ups:** Reminder emails were scheduled for non-respondents, maximizing participation without coercion [26].
- **Pilot Testing:** A small-scale pilot ensured clarity, technical compatibility, and appropriate survey length. Feedback informed refinements before full deployment [24].

The integration of stratified sampling, diverse recruitment methods and both real-world and synthetic datasets provides a strong empirical foundation for this research. Coupled with a rigorous survey distribution process, these measures ensure that the predictive model is both representative of real-world e-commerce behaviours and validated by consumer perceptions.

3.4 Data analysis and techniques

This section presents the analytical techniques adopted to validate the predictive model and assess consumer engagement. The methodology integrates traditional statistical methods with advanced ML validation while utilizing specialized software tools and data processing pipelines.

To derive meaningful insights from the collected data and ensure the robustness of the predictive model, a range of statistical analysis techniques were employed. These methods enabled the identification of relationships, validation of constructs, segmentation of consumer groups and analysis of behavioural patterns over time.

- **Regression Analysis:** Regression analysis was applied to quantify relationships between predictor variables such as click frequency and session duration and outcomes like purchase likelihood or satisfaction scores. Both linear and non-linear models were employed, while multiple regression was used to examine the combined effect of several independent variables on recommendation accuracy [14].
- **Factor Analysis:** Factor analysis was conducted to identify and validate latent constructs within survey data. Exploratory Factor Analysis (EFA) grouped correlated variables into constructs such as consumer engagement and perceived recommendation quality, while Confirmatory Factor Analysis (CFA) further validated these groupings, thereby strengthening construct validity [27].
- **Cluster Analysis:** Cluster analysis was used to segment consumers into distinct groups based on both behavioural data and survey responses. Using K-means clustering, homogeneous subgroups were identified, enabling the development of tailored recommendation strategies and enhancing personalization in the predictive model [16].

- **Time-Series Analysis:** Time-series analysis was applied to sequential interaction data such as click-stream records in order to capture temporal patterns and evolving consumer trends. This approach allowed the model to adapt in real time, improving its responsiveness to behavioural dynamics across different periods [30].

To establish the reliability, accuracy and generalizability of the predictive model, a series of statistical tests, cross-validation procedures and performance metrics were applied.

- **Hypothesis Testing:** Statistical tests such as t-tests and ANOVA were employed to evaluate key hypotheses related to consumer engagement and recommendation accuracy. These tests confirmed whether observed differences in user behaviour or model outcomes were statistically significant, thereby validating the impact of specific factors on recommendation quality [32].

- **Cross-Validation:** To minimize overfitting and enhance model robustness, k-fold cross-validation was implemented. This technique involved partitioning the dataset into multiple folds, iteratively training the model on a subset while validating it on the remaining fold. Such an approach provided a reliable estimate of predictive performance and ensured the model's ability to generalize effectively to unseen data [25].

- **Performance Metrics:** The performance of the predictive models was assessed using both regression and classification metrics. For regression tasks, Mean Absolute Error (MAE), Root Mean Squared Error (RMSE) and the coefficient of determination (R^2) were calculated to evaluate prediction accuracy. For classification tasks, precision, recall, F1-score and AUC-ROC curves were employed, offering a comprehensive evaluation of predictive power and overall reliability [34].

To efficiently manage, analyze and visualize the large and complex datasets collected in this study, a combination of statistical software, ML frameworks, distributed computing pipelines and visualization tools was employed.

- **Statistical Software:** Data analysis was conducted using Statistical Package for the Social Sciences (SPSS), R, and Python libraries such as *scikit-learn*, *pandas* and *statsmodels*. These tools facilitated regression, factor analysis, clustering and hypothesis testing, while also providing extensive libraries for data visualization and customized model development.

- **ML Frameworks:** Advanced frameworks such as TensorFlow and PyTorch were used for developing machine learning and deep learning models. These frameworks enabled the implementation of complex architectures, including recurrent neural networks (RNNs) for time-series analysis and ensemble methods such as Random Forests (RF) for behavioural prediction, ensuring that both linear and non-linear patterns were effectively captured [20].

- **Data Processing Pipelines:** To manage the scale and complexity of digital interaction datasets, distributed computing platforms such as Apache Spark were integrated. Spark's in-memory computing capabilities accelerated data processing, while its MLlib library provided scalable machine learning algorithms, enabling efficient handling of high-volume data from e-commerce platforms [33].

- **Visualization Tools:** Visualization tools including Tableau, Python's Matplotlib and Seaborn were employed to generate clear and interpretable visual outputs. These tools were essential for identifying trends, evaluating model performance and presenting findings in a format accessible to both academic researchers and industry practitioners.

The integration of regression, factor, clustering and time-series analyses with hypothesis testing, cross-validation and robust performance metrics provided a comprehensive analytical foundation for this study. Supported by advanced ML frameworks, distributed data pipelines and visualization tools, the predictive model developed in this research is both empirically rigorous and scalable to real-world e-commerce environments.

3.5 Ethical considerations

Ethical considerations were central to ensuring research integrity and participant protection. All participants provided informed consent after receiving a clear explanation of the study's purpose, data collection methods, potential risks and confidentiality measures. Participation was voluntary with the option to withdraw at any time and all collected data were anonymized and securely stored on encrypted servers accessible only to authorized personnel [24]. Recognizing the inherent limitations of survey-based research including response biases and sampling errors, the study employed best practices such as careful question design, pilot testing and multiple response formats to mitigate these issues [26]. Remaining potential biases such as social desirability or non-response, were acknowledged and their impact considered in the analysis.

4 Results

The data-driven predictive model demonstrated substantially better performance than the traditional recommendation approach across both regression and classification tasks. Based on data collected from 500 respondents (n=500), the model explained 68% of the variance in predicting customer engagement ($R^2 = 0.68$), with all key predictors including click frequency, session duration and perceived accuracy, exhibiting statistically significant positive effects ($p < 0.001$), indicating strong explanatory power.

The overall regression model was statistically significant, $F(3,496) = 350.2$, $p < 0.001$, confirming that the selected predictors jointly explain a substantial proportion of variance in engagement. Model robustness was further validated using 10-fold cross validation which yielded low prediction errors (MAE = 0.42, RMSE = 0.55) that indicating close alignment between predicted and actual engagement scores. These results, as summarized in Table 2, confirm the model's robust predictive capability and effective integration of real-time behavioural and perceptual data sources.

Table 2: Regression Model Summary and Predictive Accuracy

Metric	Value
R^2	0.68
MAE	0.42
RMSE	0.55

To further benchmark model's performance, the data-driven predictive system was compared with a traditional recommendation model based purely on historical data. As shown in Table 3, the proposed model achieved a 38% reduction in MAE, 31% lower RMSE and 30% higher R^2 than the baseline. Moreover, classification metrics improved significantly, precision increased from 0.65 to 0.82, recall from 0.60 to 0.78, F1-score from 0.62 to 0.80 and AUC-ROC from 0.70 to 0.88. These results clearly demonstrate the superior predictive and discriminative performance achieved by the integration of real-time behavioural data with user feedback mechanisms.

As shown in Table 4, survey based results complemented the quantitative analysis while revealing that participants reported high engagement (Mean = 4.2, Standard Deviation = 0.7)

Table 3: Performance Comparison Between Traditional and Data-Driven Models

Model	Traditional (Baseline)	Data-Driven (Proposed)
MAE	0.68	0.42
RMSE	0.80	0.55
R^2	0.52	0.68
Precision	0.65	0.82
Recall	0.60	0.78
F1-Score	0.62	0.80
AUC-ROC	0.70	0.88

and satisfaction (Mean = 4.0, Standard Deviation = 0.8) while perceived recommendation accuracy (Mean = 3.8, Standard Deviation = 0.9) remained favorable. These findings indicate that users viewed the data-enhanced recommendation systems as relevant and effective in meeting their preferences. The internal consistency of all constructs was strong with Cronbach's alpha values exceeding 0.80 ($\alpha >= 0.80$) confirming reliable measurement scales.

Table 4: Descriptive Statistics and Reliability of Key Constructs

Construct	Mean	Standard Deviation	Cronbach's α
Consumer Engagement	4.2	0.7	0.87
Recommendation Accuracy	3.8	0.9	0.83
User Satisfaction	4.0	0.8	0.85

Exploratory factor analysis further supported construct validity (Kaiser-Meyer-Olkin Measure (KMO) = 0.89; Bartlett's test $p < 0.001$), confirming that survey items clustered appropriately into engagement, accuracy and satisfaction dimensions.

Further analysis of relationships among constructs revealed strong, positive associations. Pearson correlation analysis of relationships among these constructs revealed strong, positive correlations ranging from r = 0.65 to 0.72 ($p < 0.001$). Users who perceived recommendations as more accurate also reported higher engagement and satisfaction which suggests the presence of a reinforcing cycle in which personalized, accurate recommendations foster engagement which in turn enhances satisfaction and long-term usage. Table 5 represents the inter-construct correlation matrix illustrating the strong and significant positive associations between engagement, accuracy and satisfaction.

Table 5: Inter-Construct Correlation Matrix

Construct	Engagement	Accuracy	Satisfaction
Engagement	1.00	0.65**	0.72**
Accuracy	0.65**	1.00	0.68**
Satisfaction	0.72**	0.68**	1.00

Note: $p < 0.001$ for all correlations.

Overall, these results directly address the study's objectives of improving personalization, scalability and mitigating the cold-start problem. The significant improvements in precision, recall and F1-score demonstrate the model's ability to effectively capture user preferences and provide relevant, individualized recommendations. The data-driven architecture, capable of

processing extensive and continuous behavioural inputs, confirmed its scalability without performance degradation. Moreover, integrating real-time behavioural data with user feedback successfully mitigated the cold-start problem by generating tailored recommendations even for new users with limited interaction history.

Collectively, the findings affirm that the data-driven model not only outperforms traditional approaches in predictive and classification accuracy but also aligns with users' subjective perceptions of recommendation quality and satisfaction. This synergy between quantitative performance and user experience underscores the value of combining advanced data analytics with user-centric design principles to deliver more personalized, adaptive and effective e-commerce recommendation systems. This convergence strengthens the validity of the results and demonstrates the effectiveness of the mixed-methods approach in capturing both behavioural and perceptual dimensions of recommendation system performance.

5 Discussion

From a mixed-methods perspective, the convergence between quantitative performance improvements and positive survey responses provides strong evidence that the proposed approach enhances both system-level accuracy and user experience. The improved results directly address the research objectives of enhancing personalization, scalability and mitigating the cold-start problem. The notable increase in precision, recall and F1-score indicates that the system more effectively captures individual preferences while delivering recommendations that users find accurate and engaging. The use of a data-driven analytics framework which combines large-scale behavioural datasets with streaming data, demonstrates strong scalability. The model maintained stable performance as data volume increased though further optimization may be needed for deployment at extreme real-time scales.

Importantly, integrating real-time behavioural signals with initial user feedback helped overcome the cold-start issue. Even new users with limited historical data received relevant recommendations by leveraging survey-based preferences and population level behaviour patterns. Overall, the data-driven model not only outperformed traditional approaches quantitatively but also aligned with user-reported satisfaction and perceived accuracy. These outcomes emphasize the value of combining advanced analytics with user-centric insights to achieve more adaptive and effective e-commerce recommendation systems.

6 Conclusion

This study introduced a mixed-methods, data-driven predictive framework aimed at improving personalization, scalability and engagement in e-commerce recommendation systems. By integrating heterogeneous behavioural data such as click streams, browsing histories and transaction records, with survey-based consumer insights, the model addressed key limitations of traditional recommendation methods.

Experimental results showed substantial improvements in predictive accuracy and classification performance, reducing MAE and RMSE by over 30%. The hybrid data approach also effectively mitigated the cold-start problem and supported strong scalability across growing data volumes. Future research will focus on implementing the framework in real-time environments, exploring advanced deep learning models like RNNs and transformers and addressing ethical issues such as data privacy and transparency. Overall, the study highlights the significance of combining data-driven analytics with consumer insight modeling to build intelligent, user-adaptive and trustworthy recommendation systems for the next generation of e-commerce platforms.

7 Abbreviations and specific symbols

AI Artificial Intelligence

AUC-ROC Area Under the Receiver Operating Characteristic Curve

CF Collaborative Filtering

CFA Confirmatory Factor Analysis

CNNs Convolutional Neural Networks

EFA Exploratory Factor Analysis

IoT Internet of Things

ML Machine Learning

MAE Mean Absolute Error

PCA Principal Component Analysis

RF Random Forest

RNNs Recurrent Neural Networks

RMSE Root Mean Squared Error

SPSS Statistical Package for the Social Sciences

SEM Structural Equation Modelling

SVM Support Vector Machines

8 Acknowledgment

The authors acknowledge the Department of Computer Science, University of Sri Jayewardenepura, for providing research facilities and data access support. The authors used OpenAI's ChatGPT tool to improve the clarity and structure of the manuscript draft. All conceptual and analytical content was developed by the authors

9 Declaration

Funding
This research received no external funding.

Conflicts of Interest
The authors declare that they have no conflicts of interest.

Ethical Considerations
The authors state that all work related to this research was conducted in accordance with institutional, national, and international guidelines and in compliance with recognized ethical standards. Informed consent was obtained from all individual participants included in the study.

AI Usage Statement

Generative AI tools were used in the preparation of this manuscript in the following way: ChatGPT was used to improve the clarity and structure of the manuscript draft.

Data Availability Statement
Data are not available due to personal privacy and confidentiality concerns.

Code Availability Statement
Software code is not available due to personal privacy and confidentiality concerns.

SDG Alignment
This research is aligned with the following United Nations Sustainable Development Goals (SDGs):

SDG 3 – Good Health and Well-Being;
SDG 4 – Quality Education;
SDG 8 – Decent Work and Economic Growth;
SDG 9 – Industry, Innovation & Infrastructure;
SDG 11 – Sustainable Cities & Communities;
SDG 12 – Responsible Consumption & Production.

References

[1] J. Q. Anderson, "Recommender systems in e-commerce: An introductory study," *Journal of Retailing*, vol. 75, no. 2, pp. 210–228, 1999.

[2] P. Resnick, N. Iacovou, M. Suchak, P. Bergstrom, and J. Riedl, "GroupLens: An open architecture for collaborative filtering of netnews," in *Proc. ACM Conf. Computer Supported Cooperative Work (CSCW)*, 1994, pp. 175–186.

[3] Y. Koren, R. Bell, and C. Volinsky, "Matrix factorization techniques for recommender systems," *Computer*, vol. 42, no. 8, pp. 30–37, 2009.

[4] R. Bell and T. Koren, "Lessons from the Netflix Prize challenge," *ACM SIGKDD Explorations Newsletter*, vol. 9, no. 2, pp. 75–79, 2007.

[5] J. Bennett and S. Lanning, "The Netflix Prize," in *Proc. KDD Cup and Workshop*, 2007.

[6] G. Adomavicius and A. Tuzhilin, "Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions," *IEEE Trans. Knowl. Data Eng.*, vol. 17, no. 6, pp. 734–749, 2005.

[7] C. C. Aggarwal, *Recommender Systems: The Textbook*, 2nd ed. Cham, Switzerland: Springer, 2016.

[8] P. G. Campos, F. J. Díez, and I. Cantador, "Context-aware recommender systems: A survey of the state of the art," *Knowledge-Based Systems*, vol. 69, pp. 14–28, 2014.

[9] Y. Chen et al., "Analysis of click-stream data for user behavior modeling," *Journal of Information Science*, vol. 44, no. 3, pp. 345–357, 2018.

[10] B. Liu, *Sentiment Analysis and Opinion Mining*. San Rafael, CA, USA: Morgan & Claypool Publishers, 2012.

[11] A. McAfee and E. Brynjolfsson, "Big data: The management revolution," *Harvard Business Review*, vol. 90, no. 10, pp. 60–68, 2012.

[12] I. A. T. Hashem et al., "The rise of big data on cloud computing: Review and open research issues," *Information Systems*, vol. 47, pp. 98–115, 2015.

[13] T. Li et al., "A review of deep learning in recommender systems," *Journal of Information Systems*, vol. 34, no. 1, pp. 1–20, 2020.

[14] D. C. Montgomery, E. A. Peck, and G. G. Vining, *Introduction to Linear Regression Analysis*, 5th ed. Hoboken, NJ, USA: Wiley, 2012.

[15] R. A. Johnson and D. W. Wichern, *Applied Multivariate Statistical Analysis*, 6th ed. Upper Saddle River, NJ, USA: Pearson Prentice Hall, 2007.

[16] I. T. Jolliffe, *Principal Component Analysis*, 2nd ed. New York, NY, USA: Springer, 2002.

[17] B. S. Everitt, *Cluster Analysis*, 5th ed. Chichester, UK: Wiley, 2011.

[18] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[19] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, no. 3, pp. 273–297, 1995.

[20] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[21] D. A. Dillman, *Mail and Internet Surveys: The Tailored Design Method*. Hoboken, NJ, USA: John Wiley & Sons, 2000.

[22] J. W. Creswell and J. D. Creswell, *Research Design: Qualitative, Quantitative, and Mixed Methods Approaches*, 5th ed. Thousand Oaks, CA, USA: Sage Publications, 2017.

[23] R. Kohavi, "A study of cross-validation and bootstrap for accuracy estimation and model selection," in *Proc. Int. Joint Conf. Artif. Intell. (IJCAI)*, 1995, pp. 1137–1145.

[24] D. A. Dillman, J. D. Smyth, and L. M. Christian, *Internet, Phone, Mail, and Mixed-Mode Surveys: The Tailored Design Method*, 4th ed. Hoboken, NJ, USA: John Wiley & Sons, 2014.

[25] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*, 2nd ed. New York, NY, USA: Springer, 2009.

[26] A. Bryman, *Social Research Methods*, 5th ed. Oxford, UK: Oxford University Press, 2016.

[27] J. F. Hair, W. C. Black, B. J. Babin, and R. E. Anderson, *Multivariate Data Analysis*, 7th ed. Upper Saddle River, NJ, USA: Prentice Hall, 2010.

[28] G. A. Churchill and D. Iacobucci, *Marketing Research: Methodological Foundations*, 9th ed. Mason, OH, USA: South-Western Cengage Learning, 2006.

[29] A. Field, *Discovering Statistics Using IBM SPSS Statistics*, 4th ed. Thousand Oaks, CA, USA: Sage Publications, 2013.

[30] G. E. P. Box, G. M. Jenkins, G. C. Reinsel, and G. M. Ljung, *Time Series Analysis: Forecasting and Control*, 5th ed. Hoboken, NJ, USA: Wiley, 2015.

[31] J. L. Herlocker, J. A. Konstan, L. G. Terveen, and J. T. Riedl, "Evaluating collaborative filtering recommender systems," *ACM Trans. Inf. Syst.*, vol. 22, no. 1, pp. 5–53, 2004.

[32] J. Cohen, *Statistical Power Analysis for the Behavioral Sciences*, 2nd ed. New York, USA: Routledge, 1988.

[33] M. Zaharia et al., "Spark: Cluster computing with working sets," vol. 10, no. 1, Jul. 2010.

[34] C. J. Willmott and K. Matsuura, "Advantages of the mean absolute error (MAE) over the root mean square error (RMSE)," *Climate Research*, vol. 30, pp. 79–82, 2005.

[35] X. Su and T. M. Khoshgoftaar, "A survey of collaborative filtering techniques," *Advances in Artificial Intelligence*, vol. 2009, pp. 1–19, 2009.

[36] R. Burke, "Hybrid recommender systems: Survey and experiments," *User Modeling and User-Adapted Interaction*, vol. 12, no. 4, pp. 331–370, 2002.

[37] G. Linden, B. Smith, and J. York, "Amazon.com recommendations: Item-to-item collaborative filtering," *IEEE Internet Computing*, vol. 7, no. 1, pp. 76–80, 2003.

[38] G. Adomavicius et al., "Context-aware recommender systems," *AI Magazine*, vol. 32, no. 3, p. 67, 2011.

[39] P. Zikopoulos and C. Eaton, *Understanding Big Data Analytics for Enterprise Class Hadoop and Streaming Data*. New York, NY, USA: McGraw-Hill, 2012.

[40] E. Rahm, "Data cleaning: Problems and current approaches," Jan. 2000.

[41] L. Baltrunas et al., "Context-aware places of interest recommendations for mobile users," *Lecture Notes in Computer Science*, vol. 6769, pp. 531–540, 2011.

[42] M. Wedel and W. A. Kamakura, *Market Segmentation: Conceptual and Methodological Foundations*. New York, NY, USA, 2000.

[43] U. Shardanand and P. Maes, "Social information filtering," 1995.

[44] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," arXiv, 2014.

[45] T. Fawcett, "An introduction to ROC analysis," *Pattern Recognition Letters*, vol. 27, no. 8, pp. 861–874, 2006.

[46] N. Peña-García et al., "Purchase intention and purchase behavior online," *Heliyon*, vol. 6, no. 6, 2020.

[47] J. Dean and S. Ghemawat, "MapReduce: Simplified data processing on large clusters," *Communications of the ACM*, vol. 51, pp. 107–113, 2008.

[48] T. Chai and R. R. Draxler, "Root mean square error or mean absolute error?" *Geoscientific Model Development*, vol. 7, no. 3, pp. 1247–1250, 2014.