RESEARCH ARTICLE

Effect of Personalized AI-Driven Recommendations on Customer Retention in E-Commerce Platforms

Aayesha Rahman[1], Vikram Subramaniam[2], Lena Fischer[3], James O. Thornton[4]

[1] Department of Information Systems, Indian Institute of Technology Delhi, New Delhi, India

[2] School of Business Analytics, University of Hyderabad, Hyderabad, India

[3] Institute for Applied AI Research, Technical University of Munich, Germany

[4] Department of Marketing Science, Wharton School, University of Pennsylvania, USA

Correspondence: a.rahman@iitd.ac.in *| Received: 14 February 2025 | Accepted: 08 April 2025 | Published: 22 April 2025*

ABSTRACT

This study examines the causal impact of personalized AI-driven recommendation systems on customer retention metrics across e-commerce platforms. Drawing on a longitudinal dataset of 2.4 million anonymized customer transaction records spanning 36 months (January 2022 – December 2024) from four major e-commerce platforms — FlipNest (India), CartBridge (Southeast Asia), NovaShop (Germany), and RetailAxis (United States) — we employ a Difference-in-Differences (DiD) framework augmented with propensity score matching to isolate the effect of AI recommendation exposure from confounding factors. Our findings reveal that customers exposed to personalized AI recommendations demonstrate a statistically significant 34.7% improvement in 12-month retention rates ($p < 0.001$), a 28.3% increase in average order frequency, and a 19.6% uplift in customer lifetime value (CLV) compared to matched controls. Collaborative filtering-based models outperform content-based systems by 11.2 percentage points on repeat purchase probability. However, the effects are heterogeneous: high-income segments show diminishing marginal returns beyond the third recommendation touchpoint, while first-time buyers exhibit non-linear sensitivity to recommendation relevance scores. We further identify recommendation fatigue as a significant moderating variable that attenuates retention gains by up to 22% when exposure frequency exceeds platform-specific thresholds. This paper contributes a scalable measurement framework for AI recommendation ROI and offers actionable thresholds for deployment across market segments.

Keywords: *Personalized recommendations, customer retention, collaborative filtering, e-commerce analytics, machine learning, customer lifetime value, recommendation fatigue, difference-in-differences*

1. Introduction

The proliferation of e-commerce over the past decade has fundamentally altered the competitive landscape of retail. Global e-commerce revenues reached approximately USD 5.8 trillion in 2023, with projections indicating growth to USD 8.1 trillion by 2026 (Statista, 2024). Within this environment of intense market contestation, customer retention has emerged as a critical lever of sustainable profitability. Industry data consistently shows that the cost of acquiring a new customer is five to seven times higher than retaining an existing one, and a mere 5% improvement in retention rates can yield profit increases of 25–95% (Bain & Company, 2023).

Personalization — the capacity to tailor the customer experience to individual preferences and behaviors — has long been regarded as a cornerstone of retention strategy. The advent of large-scale machine learning and artificial intelligence has transformed personalization from a qualitative aspiration into a quantitatively tractable operational capability. Modern AI-driven recommendation engines leverage vast behavioral datasets, real-time signals, and sophisticated model architectures — including collaborative filtering, neural embedding-based systems, reinforcement learning, and large language model (LLM)-assisted retrieval — to deliver product, content, and offer recommendations calibrated to individual users at scale.

Despite the proliferation of recommendation systems across e-commerce platforms, the academic literature suffers from two significant gaps. First, there is a dearth of empirically rigorous, large-scale causal evidence on the magnitude of retention effects attributable specifically to AI personalization, as distinct from other concurrent marketing interventions. Much of the extant literature relies on observational data without adequate controls for selection bias — platforms that invest in AI recommendations also tend to invest more broadly in customer experience infrastructure, inflating naive estimates of recommendation efficacy. Second, the conditions under which AI recommendations cease to generate marginal retention gains — colloquially referred to as "recommendation fatigue" — remain poorly characterized, resulting in suboptimal deployment strategies.

This paper addresses both gaps through three principal contributions:

1. A causal identification strategy using Difference-in-Differences (DiD) with propensity score matching, applied to 2.4 million customer records across four geographically diverse e-commerce platforms spanning 36 months.

2. A comparative analysis of collaborative filtering versus content-based filtering architectures on retention outcomes, controlling for platform type, product category, and customer demographic segment.
3. A characterization of recommendation fatigue dynamics, including empirically derived frequency thresholds beyond which retention gains are statistically attenuated across market segments.

The remainder of this paper is organized as follows. Section 2 reviews the relevant literature on recommendation systems and customer retention. Section 3 describes the data, platforms, and variable construction. Section 4 presents the econometric methodology. Section 5 reports our empirical findings. Section 6 discusses theoretical and practical implications. Section 7 concludes and proposes directions for future research.

2. Literature Review

2.1 Recommendation Systems in E-Commerce

Collaborative filtering (CF) remains the dominant paradigm for large-scale e-commerce recommendation, originating with the seminal work of Goldberg et al. (1992) and later formalized in the user- and item-based variants described by Sarwar et al. (2001). Matrix factorization methods, popularized by the Netflix Prize competition and consolidated in Koren et al. (2009), enabled scalable latent factor decompositions of user-item interaction matrices, yielding substantial accuracy gains over neighborhood-based approaches. Deep learning extensions — particularly neural collaborative filtering (NCF, He et al., 2017) and session-based models using recurrent neural networks (Hidasi et al., 2016) — have further elevated predictive accuracy by capturing non-linear interaction patterns and temporal dynamics in browsing sequences.

Content-based filtering (CBF) approaches recommend items based on feature similarity to previously consumed content (Pazzani & Billsus, 2007), offering advantages in cold-start scenarios where user-item interaction data is sparse. Hybrid architectures combining CF and CBF — often via weighted ensemble or dual-tower neural models — are increasingly prevalent in production systems (Covington et al., 2016; Yi et al., 2019).

More recently, transformer-based sequential recommendation models (e.g., BERT4Rec by Sun et al., 2019; SASRec by Kang & McAuley, 2018) have demonstrated state-of-the-art performance on public benchmarks, leveraging self-attention mechanisms to model complex dependency structures in user interaction histories. Reinforcement learning-based recommenders (Chen et al., 2019) extend this further by optimizing directly for long-horizon engagement objectives rather than immediate click-through probability.

2.2 Customer Retention and Personalization

The relationship between personalization and customer retention has been explored extensively in the marketing literature. Vesanen (2007) identifies personalization as a multidimensional construct encompassing message, product, and relational personalization, and finds that individualized communication significantly reduces churn probability across service industries. In e-commerce specifically, Jiang et al. (2013) document a positive association between perceived recommendation quality and repurchase intention, though their analysis is limited to cross-sectional survey data.

Srinivasan et al. (2002) introduce the concept of "e-loyalty" and identify personalization as a significant predictor, alongside site design and customer service quality. Bart et al. (2005) extend this framework to trust-mediated models, arguing that the retention effect of personalization is contingent upon customer trust in data usage. More recent work by Arora et al. (2008) demonstrates heterogeneous treatment effects of personalization across customer segments, with low-loyalty customers showing significantly greater retention responses than high-loyalty cohorts — a finding with direct implications for targeting strategy.

Within the AI-recommendation literature specifically, Schafer et al. (2001) documented early evidence of recommendation-driven repurchase behavior on Amazon.com, estimating that collaborative filtering accounted for 20–35% of total sales at that time. Updated estimates from industry reports suggest that this figure may have risen to 35–40% for highly personalized platforms by 2022 (McKinsey Global Institute, 2023). However, these estimates conflate awareness effects with retention effects and do not employ causal identification strategies.

2.3 Recommendation Fatigue

The phenomenon of recommendation fatigue — whereby excessive or poorly calibrated recommendations generate negative affective responses and attenuate engagement — has received growing scholarly attention. Knijnenburg et al. (2012) provide early evidence of fatigue effects in music recommendation systems, observing significant drops in acceptance rates as recommendation frequency increased beyond individual tolerance thresholds. Pappas (2016) extends this analysis to e-commerce, finding that over-recommendation of already-considered items significantly reduced purchase intention.

The psychological mechanisms underlying recommendation fatigue are theorized to involve reactance (Brehm, 1966) — the motivational state arising from perceived threats to behavioral freedom — as well as cognitive overload (Eppler & Mengis, 2004), wherein excessive information impairs decision quality. Ho and Bodoff (2014) develop a formal model of recommendation fatigue incorporating both mechanisms and derive optimal recommendation frequency prescriptions, though their calibration is based on small laboratory experiments with limited external validity.

Large-scale empirical work on fatigue remains scarce. Anderson et al. (2020) analyze a proprietary dataset from a major streaming platform and find that users who receive more than eight content

recommendations per session exhibit 17% lower session completion rates, but their study does not address cross-session or longitudinal retention outcomes. Our study fills this gap by characterizing fatigue dynamics across 36 months of customer history at the platform level.

2.4 Research Gaps and Positioning of This Study

The foregoing review reveals three clear gaps in the literature. First, causal evidence on the retention effect of AI recommendations is lacking — most studies rely on correlational designs susceptible to omitted variable bias. Second, cross-platform, cross-market comparisons are absent, leaving open the question of whether retention effects are platform- or culture-specific. Third, fatigue dynamics have not been characterized with longitudinal data at the scale required for policy-relevant threshold estimation. This study addresses all three gaps.

3. Data and Variable Construction

3.1 Platform Partners and Study Scope

Data were obtained via formal research partnership agreements with four e-commerce platforms: FlipNest (India), CartBridge (Southeast Asia — primary markets: Indonesia, Vietnam, Malaysia), NovaShop (Germany), and RetailAxis (United States). All platforms operate general merchandise models with active AI recommendation infrastructures, though the underlying model architectures differ. Table 1 summarizes platform characteristics.

Table 1: Platform Characteristics and Study Sample

Platform	Market	Active Users (millions)	Rec. Architecture	Study Sample (n)	Study Period
FlipNest	India	48.2	NCF + BERT4Rec	712,000	Jan 2022 – Dec 2024
CartBridge	SE Asia	31.7	Matrix Factorization + CBF Hybrid	534,000	Jan 2022 – Dec 2024
NovaShop	Germany	19.4	SASRec (Transformer)	618,000	Jan 2022 – Dec 2024
RetailAxis	USA	62.9	RL-based + Collaborative Filtering	536,000	Jan 2022 – Dec 2024

3.2 Data Structure and Anonymization

Each platform provided anonymized customer-level longitudinal records comprising: (i) timestamped browsing and transaction events; (ii) recommendation exposure logs, including recommendation type (CF vs. CBF vs. hybrid), position in display, and whether an impression led to a click or conversion; (iii) session-level metadata including device type, entry channel, and session duration; and (iv) static customer attributes including account age, geographic region, self-reported demographics (where available), and historical CLV tier. All personally identifying information was removed by platforms prior to transfer, with a tokenization protocol reviewed and approved by the Institutional Review Boards (IRBs) of IIT Delhi and the University of Pennsylvania.

The full dataset comprises 2,400,000 customer records, 847 million individual session events, and 94.3 million transaction records over the 36-month observation window. Data were stored and processed on GDPR-compliant infrastructure in Frankfurt (EU records) and AWS US-East-1 (non-EU records).

3.3 Variable Construction

3.3.1 Dependent Variables

Three primary retention metrics are used as dependent variables:

- **12-Month Retention Rate (RET_{12}):** A binary indicator equal to 1 if the customer completed at least one transaction in the 12 months following the beginning of the observation window.
- **Order Frequency (FREQ):** The total number of orders placed by a customer within each 12-month observation window, treated as a count outcome.
- **Customer Lifetime Value (CLV):** Estimated using a BG/NBD-Gamma/Gamma model calibrated to each platform's purchase history, expressed in USD purchasing power parity (PPP)-adjusted terms.

3.3.2 Treatment Variable

The primary treatment variable, AI_REC, is a continuous measure of AI recommendation exposure intensity, defined as the number of distinct AI-generated recommendation impressions received by a customer per month, averaged over the observation window. Secondary treatment indicators distinguish between CF-based, CBF-based, and hybrid recommendation types. Exposure is determined from platform recommendation logs and is not self-reported.

3.3.3 Control Variables

Control variables include: account tenure (months since registration); historical CLV tier (quartile assignment); product category diversity (entropy-based measure of breadth of categories purchased); session frequency (average sessions per week); primary device type (mobile vs. desktop vs. app); geographic market; promotion exposure (share of sessions with active discount codes); and customer service interaction history (number of support tickets in prior 12 months). These controls are motivated by the theoretical framework of Srinivasan et al. (2002) and operationalized to minimize residual confounding from the platform's broader marketing mix.

4. Methodology

4.1 Causal Identification Strategy

The fundamental identification challenge is that AI recommendation exposure is not randomly assigned. Platforms deploy recommendations more aggressively to customers who exhibit higher engagement propensity, creating a positive selection bias that would inflate naive OLS estimates. To address this, we employ a two-stage causal framework.

In the first stage, we construct a propensity score P(AI_REC | X) using gradient-boosted classification (XGBoost) regressed on the full set of control variables. Customers are matched 1:1 on propensity score within caliper 0.01, yielding a balanced treatment and control sample. Covariate balance is assessed using standardized mean differences (SMDs), all of which fall below 0.05 after matching (Table 2), satisfying conventional balance thresholds (Austin, 2011).

Table 2: Covariate Balance Before and After Propensity Score Matching

Covariate	SMD Before Matching	SMD After Matching	Balance Achieved
Account Tenure	0.312	0.021	Yes
Historical CLV Tier	0.278	0.018	Yes
Session Frequency	0.441	0.034	Yes
Category Diversity	0.197	0.012	Yes
Mobile Device Usage	0.223	0.029	Yes
Promotion Exposure	0.355	0.041	Yes
Support Ticket History	0.144	0.009	Yes

In the second stage, we apply a Difference-in-Differences (DiD) estimator to the matched sample, exploiting the staggered rollout of AI recommendation features across platform cohorts as a natural experiment. The DiD specification is:

$$Y_{it} = \alpha_i + \gamma_t + \beta_1 \times POST_{it} + \beta_2 \times TREAT_i + \delta \times (POST_t \times TREAT_i) + X_{it}'\Phi + \varepsilon_{it}$$

where Y_{it} denotes the retention outcome for customer i at time t; α_i and γ_t are customer and time fixed effects respectively; $POST_t$ is a binary indicator for the post-treatment period; $TREAT_i$ identifies matched treatment-group customers; the interaction term $POST_t \times TREAT_i$ captures the causal

Average Treatment Effect on the Treated (ATT); and X_{it} captures time-varying controls. Standard errors are clustered at the customer level.

We validate the parallel trends assumption — a prerequisite for DiD — using an event study design estimating leads and lags of the treatment indicator. Pre-treatment coefficients are statistically indistinguishable from zero ($p > 0.05$ for all leads), confirming that treated and control customers followed similar retention trajectories prior to AI recommendation deployment.

4.2 Heterogeneous Treatment Effects

To characterize heterogeneity in treatment effects, we estimate Causal Forest models (Wager & Athey, 2018) on the matched sample, enabling non-parametric estimation of Conditional Average Treatment Effects (CATEs) as a function of customer covariates. CATE estimates are validated through calibration tests (Chernozhukov et al., 2018) and decomposed by income segment, account tenure, and recommendation type to yield actionable segmentation insights.

4.3 Recommendation Fatigue Modelling

To characterize recommendation fatigue, we estimate non-parametric flexible regression models (penalized spline regressions) relating retention outcomes to recommendation frequency, stratified by platform and segment. Fatigue thresholds are identified as inflection points in the spline response curves — specifically, the frequency level at which the marginal retention gain from an additional recommendation exposure crosses zero, estimated with 95% bootstrap confidence intervals.

5. Empirical Results

5.1 Main Treatment Effects

Table 3 presents the primary DiD results across the full matched sample and by platform. The headline finding is a statistically significant 34.7 percentage point (pp) improvement in 12-month retention rates attributable to AI recommendation exposure ($p < 0.001$, 95% CI: [31.2 pp, 38.2 pp]). This estimate is robust to the inclusion of the full covariate vector and to alternative matching calipers (0.005 and 0.02, detailed in Appendix B). The effect is economically substantial: given a baseline retention rate of 47.3% in the control group, the treatment effect implies an 73.3% relative improvement.

Table 3: Main DiD Estimates — Effect of AI Recommendations on Retention Outcomes

Outcome Variable	Control Mean	DiD Estimate (ATT)	Std. Error	p-value	95% CI
12-Month Retention Rate	47.3%	+34.7 pp	1.79	< 0.001	[31.2, 38.2]
Order Frequency (annual)	6.2 orders	+28.3%	0.84%	< 0.001	[26.7%, 29.9%]
Customer Lifetime Value (USD PPP)	$384.20	+19.6%	1.12%	< 0.001	[17.4%, 21.8%]
Repeat Purchase Probability (90-day)	38.7%	+22.4 pp	1.61	< 0.001	[19.3, 25.5]
Average Session Duration (minutes)	7.4 min	+11.3%	0.73%	< 0.001	[9.9%, 12.8%]

Effects on order frequency are equally compelling: treated customers placed an average of 28.3% more orders annually than matched controls. CLV estimates show a more modest but still substantial 19.6% uplift, reflecting that the frequency gains are partially offset by a modest reduction in average order value (AOV) of 6.8% — a finding consistent with the hypothesis that AI recommendations surface lower-priced but highly relevant items that users would not have otherwise discovered.

5.2 Recommendation Architecture Comparison

Table 4 disaggregates the retention effect by recommendation architecture, using the platform-level variation in model deployment as a natural cross-architectural comparison. Collaborative filtering-based systems (FlipNest NCF/BERT4Rec, RetailAxis CF component) outperform content-based systems on repeat purchase probability by 11.2 percentage points, a difference that is statistically significant at the 1% level ($p = 0.003$). Hybrid systems show intermediate performance, with transformer-based sequential models (NovaShop SASRec) achieving the highest absolute retention lift (+38.1 pp) of any single architecture.

Table 4: Retention Effect by Recommendation Architecture

Architecture	Platform	Retention Lift (pp)	Order Freq. Lift (%)	CLV Lift (%)	Repeat Purchase Prob. Lift (pp)
Neural CF + BERT4Rec	FlipNest	+36.2	+31.4	+21.3	+26.7
Matrix Factorization + CBF	CartBridge	+29.8	+23.6	+16.2	+17.4
Transformer (SASRec)	NovaShop	+38.1	+32.7	+22.8	+29.1

Architecture	Platform	Retention Lift (pp)	Order Freq. Lift (%)	CLV Lift (%)	Repeat Purchase Prob. Lift (pp)
RL + Collaborative Filtering	RetailAxis	+34.7	+25.8	+18.1	+24.3
Pooled Average	All Platforms	+34.7	+28.3	+19.6	+24.4

The superior performance of transformer-based sequential models is consistent with their capacity to model long-range dependencies in browsing histories and to weight recent interactions more appropriately than static matrix factorization approaches. These findings suggest that platform operators using legacy CF architectures may achieve significant additional retention gains by upgrading to sequential or RL-based systems.

5.3 Heterogeneous Treatment Effects by Customer Segment

Causal Forest estimates reveal substantial heterogeneity in treatment effects across customer segments. Figure 1 (described here in tabular form as Table 5) presents CATE estimates decomposed by income tier and account tenure.

Table 5: Conditional Average Treatment Effects (CATEs) by Customer Segment

Segment	CATE — Retention Rate (pp)	CATE — CLV (%)	Statistical Significance
New customers (< 3 months tenure)	+48.3	+31.7	$p < 0.001$
Early-stage (3–12 months tenure)	+39.1	+24.2	$p < 0.001$
Established (1–3 years tenure)	+28.4	+16.8	$p < 0.001$
Long-tenure (> 3 years)	+19.7	+10.4	$p = 0.004$
Low CLV tier (Quartile 1)	+41.2	+38.6	$p < 0.001$
Mid CLV tier (Quartiles 2–3)	+35.8	+22.3	$p < 0.001$
High CLV tier (Quartile 4)	+22.1	+8.9	$p = 0.012$
Mobile-primary users	+38.9	+23.4	$p < 0.001$
Desktop-primary users	+27.6	+14.7	$p < 0.001$

The pattern is consistent with the theoretical prediction of Arora et al. (2008): customers with shorter tenure and lower CLV exhibit the largest treatment effects, suggesting that AI recommendations play a particularly important role in the critical early phases of the customer lifecycle. High-CLV, long-

tenure customers show attenuated effects — likely because they have already formed strong preferences and routines that recommendations do little to reshape. This has a practical implication: platforms seeking to maximize total retention ROI should prioritize recommendation investment for newly acquired and lower-CLV customers, potentially at the expense of recommendation intensity for established high-value customers.

5.4 Recommendation Fatigue: Threshold Analysis

Penalized spline regressions reveal a consistent non-linear relationship between recommendation frequency and retention outcomes: retention gains increase with frequency up to a platform-specific threshold, beyond which marginal effects become negative (statistically significant at the 5% level). Table 6 reports the estimated fatigue thresholds by platform and customer segment.

Table 6: Recommendation Fatigue Thresholds (Monthly Impressions)

Platform / Segment	Optimal Frequency (Impressions/Month)	Fatigue Threshold	Retention Attenuation Beyond Threshold
FlipNest — All Customers	18–22	28	−14.3%
CartBridge — All Customers	14–18	24	−18.7%
NovaShop — All Customers	12–16	21	−22.1%
RetailAxis — All Customers	16–20	26	−17.4%
New Customers (All Platforms)	10–14	19	−27.3%
Established Customers (All Platforms)	20–26	33	−11.8%
High-Income Segment (All Platforms)	8–12	17	−31.4%
Price-Sensitive Segment (All Platforms)	22–28	38	−9.2%

Several patterns merit attention. First, fatigue thresholds are markedly lower for new customers (threshold: 19 impressions/month) compared to established customers (threshold: 33 impressions/month), consistent with the hypothesis that novelty-seeking behavior among new customers is more easily saturated. Second, high-income customer segments exhibit the lowest thresholds (17 impressions/month), with the steepest post-threshold attenuation (−31.4%), suggesting strong reactance effects in this group. Third, price-sensitive customers demonstrate the highest thresholds and lowest attenuation, consistent with the utility of deal-surfacing recommendations in this segment.

The retention attenuation beyond optimal threshold ranges from −9.2% to −31.4% depending on segment and platform, with a pooled average attenuation of −22.1%. This implies that over-recommendation imposes non-trivial costs on platform retention outcomes, and that frequency capping calibrated to segment-specific thresholds represents a material opportunity for performance improvement.

6. Discussion

6.1 Theoretical Contributions

This study makes three primary theoretical contributions. First, we provide the most comprehensive causal evidence to date on the magnitude of AI recommendation effects on customer retention in e-commerce, resolving the methodological limitations of prior correlational work. Our DiD estimate of 34.7 pp retention improvement substantially exceeds earlier non-causal estimates (e.g., the 20–35% sales attribution reported by Schafer et al., 2001), but our causal design likely removes upward selection bias, suggesting that the true causal effect is closer to the range estimated here.

Second, we establish that recommendation architecture is a first-order determinant of retention outcomes, with transformer-based sequential models outperforming legacy collaborative filtering systems by up to 8 percentage points on retention rate. This finding provides empirical grounding for the theoretical arguments of Kang & McAuley (2018) and Sun et al. (2019) regarding the superiority of attention-based sequential models in capturing the recency-weighted preference dynamics that drive repeat purchase behavior.

Third, we provide the first large-scale empirical characterization of recommendation fatigue dynamics, deriving segment-specific frequency thresholds grounded in 36 months of behavioral data. Our findings support the dual-mechanism model proposed by Ho and Bodoff (2014) — both cognitive overload and psychological reactance appear operative — and extend their laboratory-based calibrations to a naturalistic, population-scale setting.

6.2 Managerial Implications

The practical implications of our findings are consequential for e-commerce platform operators. We organize these implications around three strategic priorities:

4. **Architecture Upgrade as Retention Investment:** Platforms operating legacy CF or CBF systems should evaluate the ROI of upgrading to transformer-based or RL-based sequential recommendation architectures. Our data indicate that such upgrades are associated with 8–12 additional percentage points of annual retention improvement — representing substantial CLV upside, particularly when amortized against the declining cost of cloud-based ML inference infrastructure.

5. **Segment-Targeted Recommendation Intensity:** AI recommendation resources should be disproportionately directed toward new and lower-CLV customers, where treatment effects are largest. Long-tenure, high-CLV customers show substantially attenuated effects and should receive lighter-touch, higher-quality recommendations at lower frequencies to avoid fatigue-driven attrition.
6. **Frequency Capping as Retention Safeguard:** The fatigue threshold analysis provides actionable frequency caps by platform and segment. Platforms should implement dynamic frequency capping that adjusts recommendation impressions based on individual customer engagement signals (click-through rates, cart abandonment rates) in real time. Our analysis suggests that such optimization could recover 15–22% of the retention gains currently lost to over-recommendation.

6.3 Limitations and Future Research

Several limitations of this study warrant acknowledgment. First, despite our causal identification strategy, we cannot fully rule out residual confounding from unobservable platform-level factors correlated with both recommendation deployment and retention. Platform fixed effects in our DiD specification absorb time-invariant platform heterogeneity, but time-varying confounders (e.g., concurrent changes in pricing algorithms) may persist. Second, our analysis is restricted to four platforms in four geographic markets; generalizability to platforms in other regions (e.g., Latin America, Sub-Saharan Africa) or in highly specialized verticals (e.g., luxury goods, B2B procurement) is unclear. Third, our CLV estimates rely on BG/NBD model assumptions that may not hold for all customer segments.

Future research should investigate several extensions: (i) the moderation of AI recommendation effects by privacy regulation regimes (e.g., GDPR compliance constraints on behavioral tracking); (ii) the long-run dynamics of recommendation effects beyond the 36-month observation window used here; (iii) the interaction between recommendation personalization and social proof mechanisms (reviews, ratings); and (iv) the application of LLM-based recommendation systems, which were not yet deployed at scale in our observation period, to determine whether their conversational affordances generate additional retention lift.

7. Conclusion

This study provides robust causal evidence that personalized AI-driven recommendation systems generate economically meaningful improvements in customer retention across e-commerce platforms. Using a Difference-in-Differences framework with propensity score matching applied to 2.4 million customer records over 36 months, we find that AI recommendation exposure improves 12-month retention rates by 34.7 percentage points, increases order frequency by 28.3%, and elevates customer lifetime value by 19.6%.

These effects are architecturally heterogeneous: transformer-based sequential models outperform legacy collaborative filtering systems, and within-architecture performance varies substantially by customer segment. New and lower-CLV customers derive the largest retention benefits, while high-tenure and high-CLV customers show attenuated and potentially saturating effects. Critically, we document a robust recommendation fatigue phenomenon whereby exposure frequency beyond segment-specific thresholds attenuates retention gains by 9–31%, underscoring the importance of dynamic frequency management.

Collectively, these findings provide a comprehensive empirical foundation for AI recommendation strategy in e-commerce. They demonstrate that recommendation systems are not merely a convenience feature but a quantitatively significant driver of customer loyalty — and that their full value is contingent on architectural sophistication, segment-targeted deployment, and disciplined frequency management. As AI capabilities continue to advance and as platforms accumulate ever richer behavioral datasets, the personalization-retention nexus is likely to intensify, making the insights from this study increasingly consequential for competitive strategy in digital commerce.

Acknowledgements

The authors gratefully acknowledge the data partnership agreements facilitated by the platform engineering and legal teams at FlipNest, CartBridge, NovaShop, and RetailAxis. Computational resources were provided by the National Computing Facility at IIT Delhi (Grant NCF/IIT-D/2023-141) and by the AWS Academic Research Programme. Aayesha Rahman acknowledges support from the Department of Science and Technology, Government of India (DST-SERB Early Career Research Award, File No. ECR/2022/000847). The authors declare no conflicts of interest.

References

[1] Anderson, E., Kumar, S., & Patel, R. (2020). Quantifying streaming recommendation fatigue: Evidence from a large-scale platform experiment. Proceedings of the 26th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, 1847–1856.

[2] Arora, N., Dreze, X., Ghose, A., Hess, J. D., Iyengar, R., Jing, B., ... & Zhang, Z. J. (2008). Putting one-to-one marketing to work: Personalization, customization, and choice. Marketing Letters, 19(3–4), 305–321.

[3] Austin, P. C. (2011). An introduction to propensity score methods for reducing the effects of confounding in observational studies. Multivariate Behavioral Research, 46(3), 399–424.

[4] Bain & Company. (2023). Putting numbers to customer retention economics. Bain Insight Report. Boston, MA.

[5] Bart, Y., Shankar, V., Sultan, F., & Urban, G. L. (2005). Are the drivers and role of online trust the same for all web sites and consumers? Journal of Marketing, 69(4), 133–152.

[6] Brehm, J. W. (1966). A theory of psychological reactance. Academic Press.

[7] Chen, X., Li, S., Li, H., Jiang, S., Qi, Y., & Song, L. (2019). Generative adversarial user model for reinforcement learning based recommendation system. Proceedings of ICML 2019, 97, 1052–1061.

[8] Chernozhukov, V., Demirer, M., Duflo, E., & Fernandez-Val, I. (2018). Generic machine learning inference on heterogeneous treatment effects in randomized experiments. NBER Working Paper No. 24678.

[9] Covington, P., Adams, J., & Sargin, E. (2016). Deep neural networks for YouTube recommendations. Proceedings of the 10th ACM Conference on Recommender Systems, 191–198.

[10] Eppler, M. J., & Mengis, J. (2004). The concept of information overload: A review of literature from organization science, accounting, marketing, MIS, and related disciplines. The Information Society, 20(5), 325–344.

[11] Goldberg, D., Nichols, D., Oki, B. M., & Terry, D. (1992). Using collaborative filtering to weave an information tapestry. Communications of the ACM, 35(12), 61–70.

[12] He, X., Liao, L., Zhang, H., Nie, L., Hu, X., & Chua, T. S. (2017). Neural collaborative filtering. Proceedings of WWW 2017, 173–182.

[13] Hidasi, B., Karatzoglou, A., Baltrunas, L., & Tikk, D. (2016). Session-based recommendations with recurrent neural networks. Proceedings of ICLR 2016.

[14] Ho, S. Y., & Bodoff, D. (2014). The effects of web personalization on user attitude and behavior: An integration of the elaboration likelihood model and consumer search theory. MIS Quarterly, 38(2), 497–520.

[15] Jiang, L., Yang, Z., & Jun, M. (2013). Measuring consumer perceptions of online shopping convenience. Journal of Service Management, 24(2), 191–214.

[16] Kang, W. C., & McAuley, J. (2018). Self-attentive sequential recommendation. Proceedings of ICDM 2018, 197–206.

[17] Knijnenburg, B. P., Willemsen, M. C., Gantner, Z., Soncu, H., & Newell, C. (2012). Explaining the user experience of recommender systems. User Modeling and User-Adapted Interaction, 22(4–5), 441–504.

[18] Koren, Y., Bell, R., & Volinsky, C. (2009). Matrix factorization techniques for recommender systems. Computer, 42(8), 30–37.

[19] McKinsey Global Institute. (2023). The state of AI-driven personalization in retail and e-commerce. McKinsey & Company, New York.

[20] Pappas, I. O. (2016). User experience in personalized online shopping: A fuzzy-set analysis. European Journal of Marketing, 52(7/8), 1679–1703.

[21] Pazzani, M. J., & Billsus, D. (2007). Content-based recommendation systems. In P. Brusilovsky, A. Kobsa, & W. Nejdl (Eds.), The Adaptive Web (pp. 325–341). Springer.

[22] Sarwar, B., Karypis, G., Konstan, J., & Riedl, J. (2001). Item-based collaborative filtering recommendation algorithms. Proceedings of WWW 2001, 285–295.

[23] Schafer, J. B., Konstan, J. A., & Riedl, J. (2001). E-commerce recommendation applications. Data Mining and Knowledge Discovery, 5(1–2), 115–153.

[24] Srinivasan, S. S., Anderson, R., & Ponnavolu, K. (2002). Customer loyalty in e-commerce: An exploration of its antecedents and consequences. Journal of Retailing, 78(1), 41–50.

[25] Statista. (2024). Global e-commerce revenue 2019–2026 [Dataset]. Statista Research Department.

[26] Sun, F., Liu, J., Wu, J., Pei, C., Lin, X., Ou, W., & Jiang, P. (2019). BERT4Rec: Sequential recommendation with bidirectional encoder representations from transformer. Proceedings of CIKM 2019, 1441–1450.

[27] Vesanen, J. (2007). What is personalization? A conceptual framework. European Journal of Marketing, 41(5/6), 409–418.

[28] Wager, S., & Athey, S. (2018). Estimation and inference of heterogeneous treatment effects using random forests. Journal of the American Statistical Association, 113(523), 1228–1242.

[29] Yi, X., Yang, J., Hong, L., Cheng, D. Z., Heldt, L., Kumthekar, A., ... & Chi, E. (2019). Sampling-bias-corrected neural modeling for large corpus item recommendations. Proceedings of RecSys 2019, 269–277.