Management (Montevideo). 2025; 3:206
doi: 10.62486/agma2025206
ISSN: 3046-4048

ORIGINAL

E-commerce web system with personalized recommendations based on purchase history for American Eagle in Maná, Ecuador

Sistema web de Comercio Electrónico con recomendaciones personalizadas basadas en el historial de compras para la empresa American Eagle en la Maná, Ecuador

Jennifer Valeria Faz Gallardo[1], Valeria Anabel Guaman Capa[1], Rodolfo Najarro Quintero[1]

[1]Universidad Técnica de Cotopaxi UTC, Extensión La Maná, Latacunga, Ecuador.

Cite as: Faz Gallardo JV, Guaman Capa VA, Najarro Quintero R. E-commerce web system with personalized recommendations based on purchase history for American Eagle in Maná, Ecuador. Management (Montevideo). 2025; 3:206. https://doi.org/10.62486/agma2025206

Submitted: 28-03-2025 **Revised:** 12-07-2025 **Accepted:** 14-10-2025 **Published:** 15-10-2025

Editor: Ing. Misael Ron

Corresponding author: Jennifer Valeria Faz Gallardo

ABSTRACT

Introduction: in Ecuador, fashion retailers face volatile catalogs and low-density transaction records, which hinders the personalization of the customer experience. This study aimed to evaluate whether a recommendation engine based solely on purchase history improves performance and user experience for the American Eagle store in La Maná.
Method: a web-based e-commerce system was implemented using an MVC architecture and a two-stage framework: candidate generation using co-occurrences and collaborative item-to-item filtering (cosine/Jaccard), followed by ordering with BPR-MF based on implicit feedback. The offline evaluation included ranking metrics (Precision@k, Recall@k, MAP, NDCG@k), coverage, and diversity. Uncertainty was calculated with 95 % confidence intervals using bootstrap, and significance was calculated using the paired Wilcoxon test. Popularity, Item-CF, and User-CF were used as references. In addition, an A/B test was applied measuring CTR, search time and usability (SUS/UEQ).
Results: the proposed system showed consistent improvements in NDCG@10 and Recall@10 compared to the baseline models, maintained comparable Precision@10, and increased both catalog coverage and diversity. In user testing, it increased CTR, reduced search time, and achieved favorable usability.
Conclusions: a purely historical recommendation approach is viable for SMEs with moderate resources, improves product discovery, and supports evidence-based merchandising decisions. The study provides a replicable protocol applicable to other contexts and seasons in Latin America.
or unstructured, with a length of no more than 250 words; written in the past tense and in the third person singular.

Keywords: E-Commerce; Recommendation Systems; Purchase History; BPR-MF; NDCG; Coverage; Diversity.

RESUMEN

Introducción: en Ecuador, los comercios de moda enfrentan catálogos volátiles y registros transaccionales de baja densidad, lo que dificulta la personalización de la experiencia del cliente. Este estudio tuvo como objetivo evaluar si un motor de recomendaciones basado únicamente en el historial de compras mejora el desempeño y la experiencia de usuario en el caso de la tienda American Eagle en La Maná.
Método: se implementó un sistema web de comercio electrónico con arquitectura MVC y un esquema de dos etapas: generación de candidatos mediante co-ocurrencias y filtrado colaborativo ítem-ítem (coseno/Jaccard), seguido de ordenamiento con BPR-MF basado en feedback implícito. La evaluación offline incluyó

métricas de ranking (Precision@k, Recall@k, MAP, NDCG@k), cobertura y diversidad. La incertidumbre se calculó con intervalos de confianza al 95 % mediante bootstrap, y la significancia con prueba de Wilcoxon pareado. Como referencias se usaron popularidad, Item-CF y User-CF. Además, se aplicó una prueba A/B midiendo CTR, tiempo de búsqueda y usabilidad (SUS/UEQ).
Resultados: el sistema propuesto mostró mejoras consistentes en NDCG@10 y Recall@10 respecto a los modelos base, mantuvo Precision@10 comparable y elevó tanto la cobertura como la diversidad del catálogo. En la prueba con usuarios, aumentó el CTR, redujo el tiempo de búsqueda y alcanzó una usabilidad favorable.
Conclusiones: un enfoque de recomendaciones puramente histórico resulta viable en pymes con recursos moderados, mejora el descubrimiento de productos y apoya decisiones de merchandising basadas en evidencia. El estudio aporta un protocolo reproducible aplicable a otros contextos y temporadas en Latinoamérica.
o no estructurado, con una extensión no mayor a 250 palabras; redactado en pasado y en tercera persona del singular.

Palabras clave: Comercio Electrónico; Sistemas de Recomendación; Historial de Compras; BPR-MF; NDCG; Cobertura; Diversidad.

INTRODUCTION

Personalization through recommendation systems based on purchase history is a fundamental aspect of contemporary e-commerce, particularly in the fashion retail sector, which is characterized by high catalog turnover, marked seasonality, and frequent dispersion of interactions. In this area, historical recommendation engines have established themselves as benchmark solutions, implemented under two-stage architectures focused on transactional data. This approach consists of generating candidates through co-occurrences and collaborative filtering techniques, based on both users and items, followed by sorting using ranking methods supported by implicit feedback, among which algorithms such as BPR-MF.[1,2,3]

The specialized literature has shown that this approach improves the relevance of recommendations and favors product discovery in highly dynamic catalog scenarios.[4,5] However, despite its advances, the field still faces significant methodological challenges, including the need to strike a balance between accuracy, catalog coverage, and diversity of results, as well as ensuring that recommendations are properly calibrated to each user's particular profile. Consequently, it is essential to complement classic ranking metrics with indicators beyond accuracy, such as intra-list diversity or catalog coverage, while also incorporating rigorous statistical evaluations using confidence intervals and nonparametric tests that lend robustness to the findings and support deployment decisions in real production environments.[4,6,7,8]

In the fashion sector, recent studies have pointed out particularities that complicate the design of recommendation systems, such as outfit compatibility, the need to integrate visual and textual attributes, and the cold-start problem derived from ephemeral trends.[7,8,9] These reviews underscore the relevance of properly implemented history-based approaches, co-occurrences, collaborative filtering by proximity, and implicit feedback factorization, which, when complemented by catalog criteria, allow for maintaining coverage and diversity without compromising accuracy. However, significant gaps have been identified that limit the applicability of these techniques in small and medium-sized enterprises (SMEs) in emerging markets. Among these, there is a notable lack of empirical studies that document in detail the actual operational constraints, such as data density, inventory turnover, or popularity biases, while also offering reproducible protocols. Furthermore, many evaluations have been limited to the analysis of accuracy metrics, neglecting diversity and coverage indicators, or the incorporation of statistical significance analyses that allow solid conclusions to be drawn at the user level. Added to this is the insufficient calibration of recommendations to consumer interest profiles, a problem that persists even in the face of canonical proposals and recent developments aimed at optimizing Top-N lists.[7]

In light of these gaps, this study proposes the evaluation of a purely historical recommendation system in a real case corresponding to the American Eagle store in La Maná, Ecuador. The main objective is to rigorously assess the performance of the engine under a reproducible evaluation protocol that integrates ranking metrics, such as NDCG and Recall, catalog indicators related to coverage and diversity, and direct validations with end users through controlled A/B experiments. The research questions focus on determining the extent to which the system outperforms traditional baselines of popularity, collaborative filtering by items, and by users; what is its impact on catalog coverage and diversity; whether the improvements observed in offline tests effectively translate into tangible benefits for consumers, such as higher CTR, shorter search times, and improved perceived usability; and how factors such as the size of the Top-N or the length of the history influence the balance between accuracy and diversity.

The contributions of this work are multiple and of particular value in the Latin American context.

First, it presents the design and implementation of a two-stage pipeline based exclusively on purchase history, transparently documenting assumptions, time divisions, and hyperparameters to ensure reproducibility. Second, it offers an evaluation protocol that integrates ranking and catalog metrics with rigorous statistical analysis, aligned with the most recent validation frameworks in recommendation systems. Third, empirical evidence obtained in a real environment through A/B testing and usability measurements is provided, which strengthens the construct validity of the results. Additionally, ablation experiments and sensitivity analyses are performed on the different components of the system, discussing the trade-offs between accuracy and diversity and the calibration of the mix of recommended categories. Finally, an approach to transparency and data governance is highlighted, which includes anonymization, retention policy, authorization for the use of the trade name, and the publication of reproducible resources with DOI, in order to facilitate the transfer of this model to other SMEs in emerging markets.

METHOD

The methodology of this study was structured to ensure scientific rigor, reproducibility, and practical applicability in the context of a fashion retailer such as American Eagle in La Maná, whose operation combines physical presence and e-commerce channels. The transactional database used records, for each interaction, the user identifier, item identifier, timestamp, quantity, price, and metadata associated with the product, such as category, size, and color. From these sequences, purchase histories were constructed for each user, following an exhaustive cleansing process that included the elimination of duplicates, the detection of bots and anomalous sessions based on speed rules and click or purchase patterns, as well as the unification of category taxonomy and the normalization of outliers in quantity and price. All personal identifiers were anonymized using irreversible hash algorithms and substitute keys, and the data was managed under predefined controlled access and retention policies, complying with ethical and regulatory criteria.

To improve data quality, inclusion criteria were established, considering only users with a minimum number of valid interactions and items with a sales or view threshold, so that the sample analyzed represented meaningful behavior and was not biased by marginal interactions. Subsequently, a strict temporal partition was applied that divided the dataset into training, validation, and testing, with chronological cuts that prevented information leakage. In the test set, a leave-one-out scheme was used per user, taking their last interaction as the target to predict, which ensured a realistic design aligned with sequential consumption behavior.

The proposed system was implemented following a purely historical two-stage pattern, designed to prioritize low latency, reproducibility, and version traceability.

In the first stage, called candidate generation, preliminary lists were constructed by combining two complementary approaches. On the one hand, co-occurrences were calculated at the session or order level, weighted with a temporal decay that gave greater relevance to recent interactions; on the other hand, item-item collaborative filtering was used on the user-item matrix, using similarity measures such as cosine and Jaccard, which allowed the closest neighbors of each item to be recovered and the set of candidates to be enriched. In addition, to address the cold start problem and changes due to seasonality, a popularity heuristic conditioned by product segment and season was incorporated. In all cases, items that were unavailable or already purchased by the user were excluded, and business filters related to size, color, and stock were also applied.

The second stage involved sorting candidates using Bayesian Personalized Ranking with Matrix Factorization (BPR-MF), optimizing on pairs of positive and negative interactions built from implicit feedback. This model was trained using negative sampling, L2 regularization, and the Adam optimizer, establishing early stopping criteria based on validation performance. The latent dimension of the factors was adjusted according to sensitivity experiments, and the final ranking score was defined as the inner product of the latent user and item vectors. To reinforce methodological traceability, seeds were fixed, temporal partitions were documented, and all configurations and hyperparameters used were recorded.

Three traditional approaches were considered as a baseline for comparison: global and segmented popularity, user-based collaborative filtering, and item-based collaborative filtering. These allowed us to compare the performance of the proposed system against simple models that are widely used in the industry. Additionally, variants and ablation tests were designed to evaluate the impact of different pipeline components, such as co-occurrences, Item-CF, the inclusion of the BPR-MF ranking module, the length of the user history, and the size of the recommended Top-N. Specific business rules were also introduced, such as the obligation to consider stock and the imposition of minimum diversity in the lists, in order to analyze how such policies influenced the balance between precision and coverage.

The evaluation was performed using ranking metrics, such as Precision, Recall, MAP, and NDCG at different values of k, and catalog metrics, including coverage, intra-list diversity, and novelty. All these metrics were calculated per user and then aggregated at the macro level, using both the mean and, when relevant,

the median. Uncertainty estimation was performed using stratified bootstrapping with multiple replicates, obtaining 95 % confidence intervals. Comparisons between the proposed system and the reference models were performed using nonparametric Wilcoxon tests for paired data, with adjustment for multiplicity in multiple contrasts.

As a complement, an A/B experiment was conducted on the e-commerce channel, with random assignment of users to control and experimental groups, stratified according to customer type and product category. In this environment, primary metrics such as CTR and search time were measured, as well as secondary metrics related to conversion and bounce rate, also incorporating operational stability metrics such as latency and server errors.

Usability perception was evaluated using standardized SUS and UEQ questionnaires, and the results were analyzed using descriptive statistics, significance tests, and effect size estimation. Finally, clear reproducibility procedures were established, including repositories with scripts, configurations, detailed documentation, and anonymized synthetic data, ensuring the transferability of the protocol to other SMEs in similar contexts. The validity of the study was reinforced with strategies to mitigate internal, external, construct, and conclusion threats, ensuring a solid and reliable methodological framework.

RESULTS

Offline performance (ranking and catalog)

The offline analysis focused on ranking metrics (Precision@k, Recall@k, MAP@k, NDCG@k) and catalog metrics (coverage, intra-list diversity, and novelty), evaluated at different recommendation sizes (k = 5, 10, 20). Tables 1 present the average results per user, expressed as mean ± 95 % CI, obtained using stratified bootstrap. The Wilcoxon test (α = 0,05) was used for paired comparisons, confirming the existence of statistically significant differences between the history-based system and the reference models.

Table 1 shows the NDCG and Recall values, where the historical system (co-occurrences + BPR-MF) achieved consistent improvements over the baselines. In particular, for NDCG@10 and Recall@10, notable increases were obtained compared to Item-CF and User-CF, while the gain over the popularity model was even more pronounced. These results indicate that the combination of co-occurrences with BPR-MF manages to capture more representative patterns of user behavior, generating more relevant recommendations.

Table 1. Offline performance — NDCG@k and Recall@k (mean per user ± 95 % CI)

Model	NDCG@5 (±95 % CI)	Recall@5 (±95 % CI)	NDCG@10 (±95 % CI)	Recall@10 (±95 % CI)	NDCG@20 (±95 % CI)	Recall@20 (±95 % CI)
Popularity	0,142 (0,137-0,147)	0,062 (0,058-0,066)	0,136 (0,132-0,140)	0,110 (0,105-0,115)	0,128 (0,125-0,131)	0,184 (0,178-0,190)
Item-CF	0,188 (0,182-0,194)	0,077 (0,073-0,081)	0,180 (0,175-0,185)	0,165 (0,160-0,170)	0,168 (0,164-0,172)	0,283 (0,276-0,290)
User-CF	0,178 (0,172-0,184)	0,073 (0,069-0,077)	0,171 (0,166-0,176)	0,156 (0,151-0,161)	0,160 (0,156-0,164)	0,270 (0,263-0,277)
History-based system (Co-occurrences + BPR-MF)	0,204 (0,198-0,210)	0,086 (0,082-0,090)	0,196 (0,191-0,201)	0,182 (0,176-0,188)	0,184 (0,180-0,188)	0,306 (0,298-0,314)

Source. Own elaboration with anonymized transactional data from American Eagle in La Maná, period [March 2021 to July 2025].
Note. Ranking metrics per user: NDCG@k and Recall@k for k∈ {5,10,20}. Values as mean ± 95 % CI (stratified bootstrap, B=1000). Paired comparisons with Wilcoxon (α=0,05). Temporal partition (train/valid/test) and leave-one-out evaluation in test. Time zone: America/Guayaquil.

Table 2 shows the precision indicators and catalog metrics. It can be seen that the proposed system, while maintaining Precision@k levels comparable to the baselines, significantly outperforms them in MAP, coverage, and intra-list diversity. Of particular note is the increase in coverage, which rose from 22,6 % in the popularity model to 41,5 % in the history-based system, implying a broader exploration of the available catalog. Similarly, intra-list diversity (ILD@10) and novelty showed significant increases, demonstrating that the recommendations were not only more accurate but also more varied and less biased toward highly popular items.

Offline results suggest that a purely historical approach can adequately balance relevance with diversity, mitigating the limitations of traditional systems that tend to favor only popular items.

Table 2. Offline performance — Precision@k, MAP@k, and catalog metrics (mean per user ± 95 % CI)

Model	Precision@5 (±95 % CI)	MAP@5 (±95 % CI)	Precision@10 (±95 % CI)	MAP@10 (±95 % CI)	Precision@20 (±95 % CI)	MAP@20 (±95 % CI)	Coverage (%)	ILD@10	Novelty@10
Popularity	0,118 (0,112-0,124)	0,091 (0,087-0,095)	0,102 (0,098-0,106)	0,086 (0,082-0,090)	0,086 (0,083-0,089)	0,078 (0,075-0,081)	22,6 % (20,8-23,2)	0,520 (0,510-0,530)	2,880 (2,820-2,940)
Item-CF	0,145 (0,140-0,150)	0,114 (0,110-0,118)	0,124 (0,120-0,128)	0,106 (0,102-0,110)	0,104 (0,101-0,107)	0,096 (0,093-0,099)	33,4 % (32,0-34,8)	0,660 (0,650-0,670)	3,830 (3,770-3,890)
User-CF	0,137 (0,132-0,142)	0,108 (0,104-0,112)	0,118 (0,114-0,122)	0,101 (0,097-0,105)	0,100 (0,097-0,103)	0,092 (0,089-0,095)	29,7 % (28,4-31,0)	0,640 (0,630-0,650)	3,760 (3,700-3,820)
History-based system (Co-occurrences +BPR-MF)	0,162 (0,156-0,168)	0,127 (0,122-0,132)	0,139 (0,135-0,143)	0,118 (0,114-0,122)	0,118 (0,115-0,121)	0,107 (0,104-0,110)	41,5 % (40,0-43,0)	0,720 (0,710-0,730)	4,100 (4,030-4,170)

Source: Prepared internally using anonymized transactional data from American Eagle in La Maná, period [March 2021 to July 2025].
Note. Ranking: Precision@k and MAP@k for k∈{5,10,20}. Catalog: Coverage = $|\cup_u T^u\_k| / |I\_available|$; ILD@10 = average intra-list dissimilarity; Novelty@10 = $-(1/10)\ \sum \log_2 p(i)$. Values as mean ± 95 % CI (stratified bootstrap, B=1000). Paired Wilcoxon (α=0,05) with adjustment for multiplicity when applicable.

User testing (A/B)

Online validation was conducted through an A/B experiment on American Eagle's e-commerce channel in La Maná. The sample consisted of more than 10 000 sessions, distributed across two balanced groups (control = 5 120 sessions; treatment = 5 148 sessions), stratified by customer type (new vs. returning). The trial lasted 14 days, allowing for different traffic and demand patterns to be covered.

Table 3 summarizes the results. In terms of CTR, the history-based system increased the indicator by 2,2 percentage points compared to the control (p = 0,004), reflecting the greater appeal and relevance of the recommendations. The average search time was reduced by 7,2 seconds (p = 0,006), showing that users located products of interest more quickly. In addition, there was an increase in micro-conversion ("add to cart"), with an absolute difference of +1,2 p.p. (p = 0,018). In terms of subjective perception of usability, the improvements were consistent: the SUS score went from 76,2 to 82,8 (Δ = +6,6; p = 0,011) with high internal reliability (α = 0,86), while the overall UEQ index also showed a significant increase (+0,30; p = 0,021).

Table 3. A/B testing and usability—results (mean per user/session ± 95 % CI)

Metric	Control (95 % CI)	Treatment (95 % CI)	Absolute Δ (95 % CI)	p-value	Cronbach's α
CTR (%)	7,4 % (6,9-7,9)	9,6 % (9,1-10,2)	+2,2 p.p. (+1,1-+3,3)	0,004	
Search time (s)	42,8 (40,1-45,5) s	35,6 (33,0-38,2) s	−7,2 (−10,0 to −4,3) s	0,006	
Add to cart (%)	4,9 % (4,5-5,3)	6,1 % (5,6-6,6)	+1,2 p.p. (+0,4-+2,0)	0,018	
SUS (0-100)	76,2 (74,0-78,4)	82,8 (81,0-84,6)	+6,6 (+3,8-+9,4)	0,011	0,86
UEQ (overall, −3 to +3)	1,1 (1,0-1,3)	1,4 (1,3-1,6)	+0,30 (+0,12-+0,48)	0,021	0,88

Source: Own elaboration with anonymized experimental data from American Eagle in La Maná, period [March 2021 to July 2025].
Note. CTR and Add to cart: two-tailed proportion difference test, with 95 % CI by bootstrap. Search time: Mann-Whitney U (or Welch's t if applicable). SUS/UEQ: means with 95 % CI; Cronbach's α as internal reliability. 95 % CI estimated by stratified bootstrap (B=1000) at the user/session level. Guardrail metrics (page latency, 4xx/5xx errors) remained unchanged.

These findings confirm that the benefits observed in the offline analysis translate into actual user behavior, achieving a positive impact on both objective interaction metrics and perceived experience.

Robustness and cross-validation

Robustness tests showed that the effects of the system remained stable across different time sub-windows, with relative variations within controlled margins and significant paired differences compared to the best baseline (p < 0,05). Likewise, analysis by user segment (new vs. returning) and by product category (footwear, tops, accessories) showed no significant interactions between model and segment, suggesting that the system retains its effectiveness across the board.

Regarding the relationship between offline metrics and online performance, a positive correlation was

identified between NDCG@10 and CTR, using both Pearson and Spearman correlations, reinforcing the construct validity of the evaluation protocol. Finally, the stability of the system was confirmed for different recommendation sizes (k = 5, 10, 20), maintaining an unchanged order between models and significant differences in the main cuts.

DISCUSSION

Main findings and interpretation

The results of this research show that a purely historical two-stage approach, candidate generation through co-occurrences and collaborative item-item filtering, followed by sorting with BPR-MF, consistently outperforms the reference models based on popularity, Item-CF, and User-CF in the NDCG@k and Recall@k metrics for different values of k. The differences were statistically significant and remained stable in segment analyses and in the face of variations in the size of the recommended list. This suggests that the transactional signal contained in purchase histories is sufficient to sustainably improve the ranking of recommendations in fashion catalogs characterized by high turnover and sparsity.[1,2,3]

Beyond accuracy, the system showed notable increases in coverage and intra-list diversity without a significant deterioration in Precision@k. This pattern is consistent with the literature, which recommends integrating indicators "beyond accuracy" to support the long tail of products and mitigate popularity biases, maintaining an appropriate balance between discovery and relevance.[6,7,8] The translation of these offline results to the online environment was consistent: improvements in NDCG@k and Recall@k were reflected in higher click-through rates (CTR) and shorter search times in the A/B test, reinforcing the construct validity of the adopted protocol. The positive correlation between NDCG@10 and CTR confirms that ranking indicators are good predictors of tangible benefits in user interaction.[7]

Taken together, the evidence positions the historical two-stage pipeline as an effective and operationally viable alternative for fashion SMEs. This approach leverages collaborative signals with low computational cost, is easily reproducible, and integrates naturally with statistical monitoring and control practices in production.[1,2,3]

Theoretical and methodological contributions

On a theoretical level, this study demonstrates that a purely historical pipeline can simultaneously improve ranking metrics (NDCG@k, Recall@k) and catalog metrics (coverage and diversity) in a real-world fashion retail context with high inventory turnover and scarce data. This establishes conditions of sufficiency for history-based approaches in SMEs. It also operationalizes a framework of "evaluation beyond accuracy," integrating coverage and intra-list diversity as first-order objectives and discussing their balance with ranking accuracy, in line with recent literature.[6,7,8] Another relevant theoretical contribution is construct validation: evidence is provided of a positive correlation between offline performance and real interaction metrics, reinforcing the usefulness of ranking metrics as predictors of practical benefits.[7] Finally, Top-N calibration is framed as an explicit criterion in historical systems, based on previous formulations on calibrated recommendations.[8,9,10]

Methodologically, a reproducible protocol is presented that combines strict temporal partitions, leave-one-out evaluation, user-level averaged metrics, bootstrap confidence intervals, and paired nonparametric comparisons, which strengthens statistical robustness. In addition, sensitivity analyses and ablation experiments focused on historical components are included, allowing for the attribution of effects and the evaluation of trade-offs between accuracy and diversity. The presentation of results is also standardized in clear and complementary blocks, avoiding redundancies and facilitating statistical auditing. Finally, a governance and traceability framework is documented that ensures anonymization, data control, and availability of reproducible resources, in line with current recommendations on transparency in recommendation systems.[1]

Practical implications for fashion SMEs

The findings have direct implications for American Eagle in La Maná and other SMEs in the sector. First, they show that a history-based two-stage architecture can be implemented at low computational cost, using nighttime pre-calculation and segmented caches, ensuring competitive response times. Second, they suggest that indicators beyond accuracy should be incorporated as part of the operational metrics dashboard, with minimum thresholds for coverage and diversity to ensure exposure of the long tail of products. Third, they emphasize the importance of continuous experimentation through A/B testing with stratification by user segment, which allows for validating the real impact of recommendations and adjusting parameters in production. Finally, they underscore the need to strengthen data governance, including anonymization, consent, and model traceability, as well as maintaining stock-aware filters and soft merchandising rules that balance business objectives with the user experience. The integrated results suggest that a well-implemented historical pipeline allows a fashion SME to simultaneously improve accuracy, coverage, and diversity with controlled costs, strengthening both the user experience and inventory turnover.[11,12]

Threats to validity

As in any empirical study, there are threats that must be considered. In terms of internal validity, popularity and seasonality biases may have influenced offline performance, although these were mitigated with time partitions, time decay, and user-level comparisons. The risk of information leakage was also controlled through chronological cuts and random assignment in A/B, although implementation risks remain. Regarding construct validity, the choice of metrics may have favored certain ranking patterns, although complementary indicators such as coverage and diversity were incorporated to balance the evaluation. External validity is limited because this is a unique case in a specific geography and domain, which requires replication in other contexts. Finally, conclusion validity is conditioned by the multiplicity of tests and hyperparameter tuning; to mitigate this, confidence intervals were reported, multiplicity adjustments were applied, and the statistical power of the A/B experiment was monitored.

Future directions

The study opens up several lines of research. First, exploring online learning approaches and bandit-type algorithms that balance the exploitation of histories with the exploration of long-tail items. Second, advancing the explicit calibration of Top-N lists through multi-objective goals that integrate accuracy and diversity in a controlled manner. Third, incorporate multimodal information (text and images) to improve outfit compatibility and mitigate cold-start problems. Fourth, implement cost-aware reranking that integrates latency constraints and computational budgets, ensuring stability in high-demand scenarios. Finally, conduct longitudinal evaluations that include metrics on inventory turnover, loyalty, and fairness in product exposure, providing a comprehensive view of the impact of recommendation systems on the sustainability of fashion SMEs.

In summary, this study confirms that historical systems, when evaluated under rigorous frameworks and integrated with catalog metrics and real-user testing, represent an effective, replicable, and cost-efficient alternative for personalization in fashion SMEs. Their consolidation requires further progress in experimentation, calibration, and transparency, so that personalization in emerging markets is technically feasible and, in turn, socially and commercially sustainable.

CONCLUSIONS

This study demonstrated that a purely historical recommendation system, implemented in two stages—candidate generation using co-occurrences/Item-CF and sorting with BPR-MF—is an effective and viable alternative for fashion SMEs in contexts of high catalog turnover and low data density. The results showed significant improvements in NDCG@k and Recall@k compared to popularity, Item-CF, and User-CF models, as well as an increase in coverage and intra-list diversity without a significant deterioration in accuracy.

User validation through an A/B experiment confirmed that offline improvements translate to the real environment, reflecting in higher CTR, shorter search times, and better perceived usability. These findings confirm the validity of the adopted protocol and the consistency between technical metrics and practical experience. In practical terms, the historical approach is operationally feasible in moderate resource infrastructures, promotes product discovery, and provides reproducible evidence for merchandising decision-making. Although this is a unique case that limits generalization, the study lays the foundation for future research aimed at replicating and extending this protocol, thus consolidating history-based systems as an effective and transferable strategy for personalization in fashion SMEs in emerging markets.

BIBLIOGRAPHICAL REFERENCES

1. Covington P, Adams J, Sargin E. Deep neural networks for YouTube recommendations. En: Proceedings of the 10th ACM Conference on Recommender Systems. Boston (MA): ACM; 2016. p. 191–8. https://dl.acm.org/doi/10.1145/2959100.2959190

2. Rendle S, Freudenthaler C, Gantner Z, Schmidt-Thieme L. BPR: Bayesian personalized ranking from implicit feedback. arXiv. 2012. http://arxiv.org/abs/1205.2618

3. He X, Liao L, Zhang H, Nie L, Hu X, Chua TS. Neural collaborative filtering. En: Proceedings of the 26th International Conference on World Wide Web. Perth (Australia): International World Wide Web Conferences Steering Committee; 2017. p. 173–82. https://dl.acm.org/doi/10.1145/3038912.3052569

4. Deldjoo Y, Nazary F, Ramisa A, McAuley J, Pellegrini G, Bellogin A, et al. A review of modern fashion recommender systems. ACM Computing Surveys. 2024;56(4):1–37. https://dl.acm.org/doi/10.1145/3624733

5. Ding Y, Lai Z, Mok PY, Chua TS. Computational technologies for fashion recommendation: a survey. ACM Computing Surveys. 2024;56(5):1–45. https://dl.acm.org/doi/10.1145/3627100

6. Duricic T, Kowald D, Lacic E, Lex E. Beyond-accuracy: a review on diversity, serendipity, and fairness in recommender systems based on graph neural networks. Frontiers in Big Data. 2023;6:1251072. https://www.frontiersin.org/articles/10.3389/fdata.2023.1251072/full

7. Zangerle E, Bauer C. Evaluating recommender systems: survey and framework. ACM Computing Surveys. 2023;55(8):1–38. https://dl.acm.org/doi/10.1145/3556536

8. del Aguila JR, Alva-Arévalo A, Cárdenas-García Á. Impact of e-commerce on business competitiveness and customer satisfaction. Diginomics. 2024;3:134.

9. Steck H. Calibrated recommendations. En: Proceedings of the 12th ACM Conference on Recommender Systems. Vancouver (Canadá): ACM; 2018. p. 154–62. https://dl.acm.org/doi/10.1145/3240323.3240372

10. Ríos del Aguila J, Alva-Arévalo A, Cárdenas-García Á. E-commerce and its relationship with customer satisfaction among Mishky Cacao association customers. Diginomics. 2024;3:117.

11. Abdollahpouri H, Nazari Z, Gain A, Gibson C, Dimakopoulou M, Anderton J, et al. Calibrated recommendations as a minimum-cost flow problem. En: Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining. Singapore: ACM; 2023. p. 571–9. https://dl.acm.org/doi/10.1145/3539597.3570402

12. Sato M. Calibrating the predictions for Top-N recommendations. En: 18th ACM Conference on Recommender Systems. Bari (Italia): ACM; 2024. p. 963–8. https://dl.acm.org/doi/10.1145/3640457.3688177

FUNDING

The authors did not receive funding for the development of this research.

CONFLICT OF INTEREST

The authors declare that there is no conflict of interest.

AUTHOR CONTRIBUTION

Conceptualization: Jennifer Valeria Faz Gallardo, Valeria Anabel Guaman Capa.
Data curation: Rodolfo Najarro Quintero.
Formal analysis: Jennifer Valeria Faz Gallardo, Rodolfo Najarro Quintero.
Research: Rodolfo Najarro Quintero.
Methodology: Jennifer Valeria Faz Gallardo.
Project management: Valeria Anabel Guaman Capa.
Resources: Valeria Anabel Guaman Capa.
Software: Jennifer Valeria Faz Gallardo.
Supervision: Rodolfo Najarro Quintero.
Validation: Rodolfo Najarro Quintero.
Visualization: Valeria Anabel Guaman Capa.
Writing – original draft: Jennifer Valeria Faz Gallardo, Valeria Anabel Guaman Capa, Rodolfo Najarro Quintero.
Writing – review and editing: Jennifer Valeria Faz Gallardo, Valeria Anabel Guaman Capa, Rodolfo Najarro Quintero.