Explainable Next-Purchase Recommendations: A Multistakeholder Framework

Maciej Mozolewski[1,2,*], Honorata Zych[1], Sabri Manai[1], Krzysztof Kutt[1] and Grzegorz J. Nalepa[1]

[1]*Department of Human-Centered Artificial Intelligence, Institute of Applied Computer Science, Jagiellonian University, Krakow, Poland*
[2]*Jagiellonian Human-Centered AI Lab, Mark Kac Center for Complex Systems Research, Jagiellonian University, Krakow, Poland*

Abstract

We present a next-purchase recommendation system that combines advanced algorithms with explainable AI (XAI) to learn individual customer preferences from purchase histories and deliver personalized recommendations that enhance user engagement and inform marketing strategy. Our approach provides dual-layer, multistakeholder explanations: targeted communications that promote personalized marketing messages for customers and strategic insights for business stakeholders (e.g., marketing departments), reducing cognitive load and fostering trust. The system also addresses cold-start scenarios and leverages implicit feedback. Experiments on the MovieLens dataset demonstrate a balanced trade-off between accuracy, novelty, and explainability, potentially lowering users' decision-making effort.

Keywords

Explainable AI, Recommender systems, Decision support, Cold-start mitigation, Implicit feedback

1. Introduction

In professional contexts, AI supports managers by meeting information demands during decision making and reducing cognitive load [1]. Our focus is on recommendation systems, studied for decades and popularized by platforms like Amazon and Netflix, yet requiring more than mere suggestion generation. Modern recommendations should not only match user preferences but also surface unexpected and novel items—so-called *serendipitous* recommendations [2, 3]. Trustworthiness is essential in the increasingly popular *multistakeholder* environments, where end-users and business stakeholders interact [4], so recommendation systems must explain their predictions [5] and address user-facing transparency alongside stakeholder goals [6]. Thus, we focus on two stakeholder groups: (1) *end-users* (e.g., students on an e-learning platform, readers on a news site, or shoppers using a retail app, who directly receive and act on recommendations), and (2) *business stakeholders* (e.g., service providers, system owners, marketers, system administrators, or curriculum designers in an educational context).

This article is part of the *PEER – The Hyper-Expert Collaborative AI Assistant* project[1], an EU-funded Horizon Europe initiative redefining human–AI collaboration for complex decision making through user-centered design, dynamic engagement, and transparent reasoning. PEER develops AI solutions for manufacturing, warehouse management, and smart inclusive cities, with recommendation and preference modeling among its core research areas. Although we propose an explainable recommendation system for a retail use case, limited pilot data were not used in this study. Our work also aligns with the 2025 Workshop on *"AI for understanding human behavior in professional settings"* (BEHAIV)[2],

BEHAIV-2025: AI for understanding human behavior in professional settings, 25 October 2025, Bologna, Italy.
*Corresponding author.
m.mozolewski@uj.edu.pl (M. Mozolewski); honorata.zych@doctoral.uj.edu.pl (H. Zych); sabri.manai@doctoral.uj.edu.pl (S. Manai); krzysztof.kutt@uj.edu.pl (K. Kutt); grzegorz.j.nalepa@uj.edu.pl (Grzegorz J. Nalepa)
0000-0003-4227-3894 (M. Mozolewski); 0009-0002-2026-3177 (H. Zych); 0009-0009-2242-7391 (S. Manai); 0000-0001-5453-9763 (K. Kutt); 0000-0002-8182-4225 (Grzegorz J. Nalepa)

[1]https://peer-ai.eu/en/
[2]https://slawomir-nowaczyk.github.io/BEHAIV-2025

which emphasizes understanding experts' information demands during decision making and reducing cognitive load via explanatory AI to enhance safety and satisfaction at work.

Our contributions are:

- An end-to-end framework uniting diverse ML models, hybrid serendipity mechanisms and XAI to support collaborative decision making.
- Tailored *multistakeholder* explanations for end-users (e.g., customers) and business stakeholders (e.g., system owners).
- A retail recommendation tool that learns individual preferences to provide clear, context-aware product suggestions for in-store fulfillment, with simple explanations and built-in support for novelty and cold-start cases.

The rest of the paper is organized as follows: Section 2 covers background; Section 3 describes system architecture, data preprocessing steps, and the models; Section 4 details the dataset, evaluation protocol, performance metrics and reports both quantitative and qualitative findings, including model comparisons, cold-start analysis, and explanation examples; Section 5 analyzes trade-offs, limitations, and practical implications; Section 6 concludes and outlines future work; finally, Section 6 acknowledges support.

2. Related Work and Background

The first traces of the recommendation problem in scientific literature appeared in 1978 with the paper [7], which proposed a rule-based system. Subsequnetly, several other classical methods were introduced including collaborative filtering [8], content-based filtering[9] and association rule-based approaches [10]. A common feature of this classical recommendation algorithms is their highly interpretable nature. Explanations can be easily derived from rules themselves [7], listing what similar users have purchased [11, 8] or providing confidence ranking scores for similar items [9].

As deep learning models began to gain significant attention, they were also adopted for the recommendation problem [12]. A variety of architectures have since been explored, including autoencoders [13, 14], graph neural networks[15], and transformer-based models [16, 17]. These models excel at capturing complex and higher-order relationships. However, their increased complexity makes them opaque "black boxes", so researchers use popular model-agnostic techniques like SHAP [18, 14] and LIME [14] for explanation. In addition, counterfactual explanations have emerged as a promising direction, offering insights by showing how slight changes in input could alter recommendation outcomes [19, 20, 21].

Explanations enhance transparency, build trust, and support acceptance in recommender systems, aligning with the XAI Manifesto's call for transparency, accountability, and understandability [22]. Classical approaches provided straightforward justifications: rule-based systems explained outcomes through explicit logic[7]; collaborative filtering offered transparency by referencing the actions of similar users - for example, explaining a recommendation with "users similar to you liked this item" [8, 11]. Content-based methods highlighted shared item attributes - "this item is recommended because it shares features with items you liked" or confidence scores [11, 9]. Social explanations, such as neighbor ratings, their similarity to the user, and temporal dynamics, further enriched user trust [8, 11]. To demystify the decision-making of black-box models, model-agnostic tools such as SHAP and LIME are often employed, typically using plot-based visualizations to show feature contributions. Recent advances also incorporate large language models (LLM) to translate these explanations into more accessible natural language summaries [18]. Counterfactual explanations, often presented in natural language, offer another powerful approach, illustrating how slight changes in user behavior or preferences might lead to different outcomes [11, 9, 18]. Beyond textual or statistical formats, visual explanation methods have gained attention. These include word clouds that emphasize relevant terms [9, 23], *"tagsplanations"*

that interpret recommendations via users' sentiment toward descriptive tags and tag relevance [24], and graph-based visualizations that depict items and preferences as interconnected nodes and edges, thereby tracing the semantic or behavioral logic behind recommendations [11, 23]. Together, these diverse strategies reflect a shift from merely generating recommendations to constructing rich, multimodal explanations that are transparent and user-centered.

Recommender systems often operate in multi-sided environments where users, product providers, and platform owners have distinct - sometimes conflicting - goals. Traditional algorithms typically prioritize user utility, overlooking broader stakeholder objectives [25].

To address this, multistakeholder recommendation systems [4] explicitly model and balance the interests of users, providers, and platforms. This is especially important in platforms like eBay, Etsy, or Airbnb, where sustained engagement from all parties is critical to long-term success [26].

Despite growing interest in explainability, multistakeholder explanations remain underexplored, with few studies addressing how to tailor them to different stakeholder needs. [6] in their study on job recommender systems, draw from the literature the idea that explanations should either be individually tailored to each stakeholder, or that a single explanation may be adapted in presentation depending on the stakeholder's level of expertise. Building on this, they explored several explanation modalities suited to different roles, including graph-based visualizations (showing weighted paths in a knowledge graph), LLM-generated textual summaries, and feature attribution bar charts. Their results show clear preferences among stakeholder types: candidates favored short textual explanations for quick judgment, hiring managers preferred graph-based views for a more technical overview, and recruiters benefited most from detailed textual narratives. In a related contribution, [27] examined explanation strategies in enterprise decision-making and identified counterfactual explanations as particularly effective in multistakeholder contexts, as they enhance transparency while safeguarding stakeholder privacy and preference sensitivity. Together, these findings point toward the need for adaptive, role-sensitive, and privacy-aware explanation frameworks, a topic still in its early stages of research.

3. Methodology

Recommendation systems personalize experiences using historical and interaction data, yet often optimize only one goal. Our proposal introduces a transparent, multistakeholder workflow: it employs user profiles to train models, generates dual XAI explanations for consumers and managers, and applies business-rule filtering to deliver trusted recommendations. We used the Cornac framework [28] for its support of multiple recommendation approaches and explainability. The selected models represent distinct algorithmic paradigms, enabling comparison in terms of ranking quality and runtime efficiency.

The Figure 1 illustrates recommendation process as proposed by us: first, the model is trained using historical user profiles; next, new user data is fed into the trained model to generate a set of candidate recommendations. These recommendations are then passed through an explanation module, which produces human-readable rationales for each suggested item. After explanations are generated, business rules (defined by business stakeholders; but also end-user needs and preferences) are applied to filter out ineligible products. Finally, the system delivers the filtered items and their explanations directly to the end-user, while simultaneously presenting a higher-level system explanation to business stakeholders.

3.1. Model Training and Hyper-parameter Tuning

All experiments were executed in Google Colab notebooks with the *Cornac framework* [28], which supplies unified routines for data loading, model optimisation, and evaluation. [3].

The data source was the *MovieLens 100K implicit-feedback matrix*. This dataset resembles the proprietary data collected in the *PEER project* (shopping baskets), but those could not be used due to confidentiality restrictions imposed by the *PEER project* use-case owner at the time of writing. It was

[3]Colab notebook with code and instructions on how to run it is available at:
https://github.com/sabri-manai/Explainable-Next-Purchase-Recommendations-A-Multistakeholder-Framework

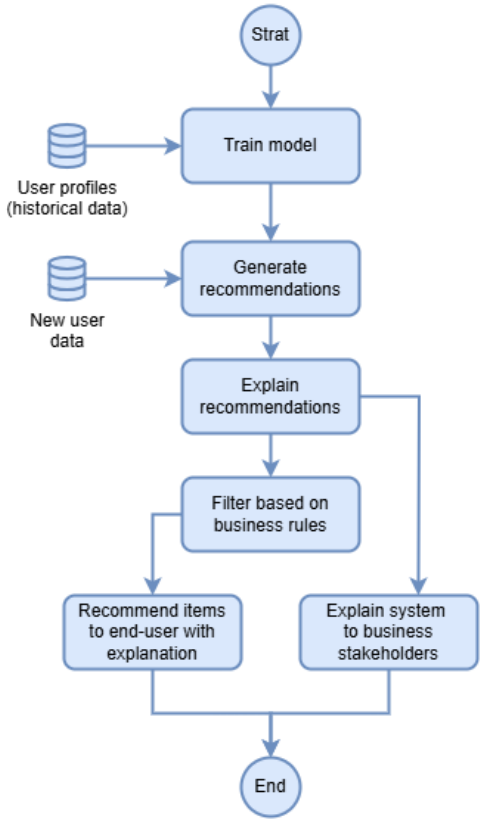

Figure 1: High-level architecture of the transparent, *multistakeholder* recommendation system.

partitioned into training, validation, and test subsets in a 70 : 20 : 10 ratio by the framework's *RatioSplit* utility, and ratings of four or higher were treated as positive interactions.

Model definitions. To ensure robust performance on sparse implicit-feedback data, we selected four state-of-the-art recommenders spanning the main paradigms. These include Bayesian Personalized Ranking (BPR) [29], Matrix Factorization (MF) [30], Hierarchical Poisson Factorisation (HPF) [31], and a Variational Auto-Encoder for Recommendations (RecVAE) [32], all implemented via their corresponding Cornac model classes.

Evaluation metrics. Five ranking metrics and two runtime measures are used throughout the study. *AUC* is the probability that a randomly chosen positive interaction (rating ≥ 4) is ranked ahead of a randomly chosen negative one. *MAP* averages, across users, the precision observed at each relevant item, rewarding early hits. *NDCG@10* normalises discounted cumulative gain at rank ten by the ideal DCG, so higher values mean that relevant items appear nearer the top. *Precision@10* is the fraction of relevant items in the first ten positions, whereas *Recall@10* is the fraction of each user's relevant catalogue retrieved within that cut-off. Runtimes are also logged: a single *Time (s)* during validation, and separate *Train (s)* and *Test (s)* columns for the final evaluation.

Tuning strategy. Only the latent dimension was varied, running a grid search for each model, and HPF achieved its highest validation AUC at k=19. This single-parameter sweep served the dual purpose of capacity control and of testing the hypothesis that a latent space of 19 factors would mirror the 19 MovieLens genre indicators examined later in Section 3.2. All remaining parameters were left at their Cornac defaults, as preliminary runs showed negligible sensitivity outside the capacity dimension.

3.2. Categories Alignment of Latent Factors

Although Hierarchical Poisson Factorization produces purely numerical item factors, an explicit link to human-readable categories (movie genres) was required for explanatory purposes. The alignment proceeded in three steps.

Correlation matrix. First, the factor loadings for every movie were merged with the 19 binary genre indicators supplied by the MovieLens metadata. For each latent dimension and each genre, the Pearson correlation coefficient was calculated, yielding a $k \times 19$ matrix whose entries quantify how strongly a factor is expressed by titles in a given genre.

One–to–one matching. Each latent factor correlates with several genres, yet a single genre label is required for interpretation. The absolute correlation values were therefore negated. This conversion turns the task of maximising correlations into the minimisation form expected by the Hungarian algorithm. Applying the algorithm to the negated matrix produced a one-to-one assignment that links every factor to the genre with which it shares its strongest absolute correlation.

Interpretation and use. The resulting factor–genre map was visualized as a heat-map (Fig.4), allowing latent themes to be read off at a glance. The same mapping was later used to aggregate SHAP attributions to genre level for user-facing explanations and to support business stakeholders-oriented factor steering. No model weights were modified in this procedure, and the alignment step was applied entirely post-hoc.

3.3. Filtering with user preferences

To enhance the personalization of recommendations and tailor them to individual user preferences, we incorporate a lightweight user profiling mechanism. To maintain system robustness and avoid excessive storage requirements, we do not retain the complete user interaction history. Instead, we store only a minimal subset of preference information - specifically, categories explicitly marked as disliked by the user during interactions with the system. This information is stored in the user profile. Filtering based on user preferences is performed in two stages. First, a preliminary filtering step removes items the user has already seen from the set of model-predicted recommendations (user history is obtained on the fly from the dataset). Second, we further refine the candidate set by excluding any movies that belong to genres identified as disliked in the user's profile. After each recommendation round, the user is prompted to update their preferences by specifying any additional disliked genres. This process is sequential and adaptive: each time a recommendation is provided, the user has the opportunity to revise their preferences, and these updates are immediately incorporated into subsequent filtering steps.

3.4. Serendipity and Business Rules

In order to combine serendipity with business goals, we propose a hybrid approach that blends business-driven content promotion (via unpopular items) with user-centric novelty (through category diversity) 2, ensuring mutual benefit for both the user and the platform. [4].

To enrich the recommendation system with diversity and surprise, we implemented a serendipitous recommendation module that selects two types of product suggestions for each user: a random unpopular pick and a category-novelty-based pick. These recommendations are intentionally designed to highlight content that lies outside the user's usual viewing patterns and the platform's typical popularity trends.

A key business rule guiding this module is the promotion of underexposed or lesser-known products. Increasing their visibility and sales is strategically important for broadening market appeal, optimizing inventory turnover, and enhancing overall profitability. This goal is achieved through the following

[4]The code for serendipity and business rules integration along with user profiles can be found at: https://github.com/hzych/Recommendations.

steps. First, we identify unpopular content by selecting products that fall within the lowest quartile of overall rating frequency - specifically, those with a number of user ratings below the 25th percentile of the distribution of rating counts. These are the items we aim to elevate. Next, we apply a series of filters: (1) the product must not have been rated by the user, (2) it must not belong to any category the user has explicitly disliked, and (3) it must have an average rating of at least 3.0, ensuring a minimum level of quality.

From the resulting candidate set, two types of recommendations are generated:

- *Random Pick*: An item is selected at random, introducing an element of surprise and unpredictability.
- *Category-Novelty Pick*: An item is selected based on the novelty of its category profile relative to the user's historical preferences. This process includes an additional step of computing a category novelty score, which assigns higher values to products containing categories the user has interacted with less frequently, thereby encouraging exploration into unfamiliar content areas.

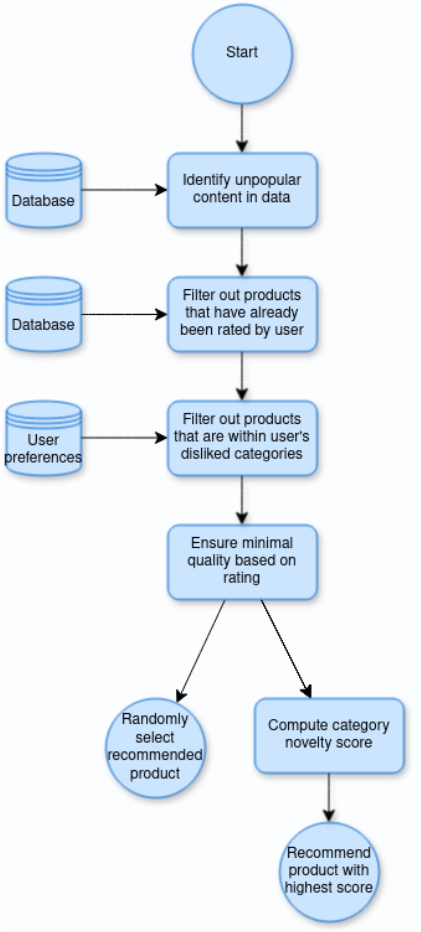

Figure 2: Hybrid approach algorithm for serendipity and business rules integration.

Other common business rules, such as filtering out-of-stock items, prioritizing high-margin or sponsored products, enforcing regional availability, and capping repeat recommendations, can be incorporated via simple post-filtering or by adjusting recommendation scores.

Validation	AUC	MAP	NDCG@10	Time (s)
BPR	0.90	0.09	0.12	1.54
MF	0.73	0.04	0.05	1.60
HPF	0.92	0.12	0.15	1.79
RecVAE	0.91	0.08	0.09	3.25

Table 1
Validation performance on MovieLens 100K.

Test	AUC	MAP	NDCG@10	Train (s)	Test (s)
BPR	0.90	0.14	0.19	2.06	2.22
MF	0.73	0.05	0.08	0.23	1.80
HPF	0.93	0.19	0.26	30.42	3.94
RecVAE	0.92	0.12	0.14	286.38	4.44

Table 2
Test performance on MovieLens 100K.

3.5. Explanation module

To explain recommendations to end-users, we apply SHAP at the genre level on each top suggestion, revealing how individual genre preferences influence the score. We chose SHAP for its community validation, model-agnostic applicability (it works with any recommendation model), and axiomatic guarantees: local accuracy (attributions sum to the prediction), consistency (higher-impact features receive larger values), and missingness (absent features score zero).

For business stakeholders, we employ a *latent-factor loading matrix* from our matrix factorization model: rows correspond to latent dimensions and columns to genres, showing how factors map to domain concepts. While this model-specific analysis validates the global structure of factorization-based recommenders, analogous loading or correlation analyses could be devised for other factor-driven architectures; pure black-box models without explicit factors would require alternative techniques such as embedding-based concept extraction.

4. Results

In this section we present both quantitative and qualitative evidence for the effectiveness of the proposed pipeline. We begin by benchmarking four representative recommendation models on the MovieLens 100K dataset, then analyse their robustness in simulated cold-start scenarios. We further evaluate the clarity of user- and stakeholder-oriented explanations and, finally, discuss how post-filtering rules aimed at serendipity and business constraints affect overall performance.

4.1. Model Comparison

Across both validation and test splits (Tables 1 and 2), the Hierarchical Poisson Factorisation (HPF) model delivers the strongest ranking quality - topping AUC, MAP, and NDCG@10 - while remaining reasonably fast to evaluate. Bayesian Personalised Ranking (BPR) follows closely on accuracy and is far cheaper to train, making it a pragmatic second choice. RecVAE reaches near-HPF AUC scores but incurs the heaviest training cost, and Matrix Factorisation (MF) trades accuracy for speed, achieving the lowest metrics yet the quickest training time.

4.2. Cold-Start Analysis

To assess robustness under sparse histories, a synthetic cohort of ten *brand-new* users was created, each seeded with only three to six randomly selected past movies. Despite the limited input, the

Title	Genres	Score
Killing Fields, The (1984)	Drama, War	12.82
Afterglow (1997)	Drama, Romance	0.14
Waiting for Guffman (1996)	Comedy	2.51
Cat People (1982)	Horror	2.28
Band Wagon, The (1953)	Comedy, Musical	0.19

Table 3
Interaction history for user `new-u-4`.

Title	Genres	Score
Star Wars (1977)	Action, Adventure, Romance, Sci-Fi, War	47.95
Godfather, The (1972)	Action, Crime, Drama	35.73
Silence of the Lambs, The (1991)	Drama, Thriller	34.10
One Flew Over the Cuckoo's Nest (1975)	Drama	32.99
Raiders of the Lost Ark (1981)	Action, Adventure	32.23
Schindler's List (1993)	Drama, War	31.37
Fargo (1996)	Crime, Drama, Thriller	29.44
Casablanca (1942)	Drama, Romance, War	27.59
Return of the Jedi (1983)	Action, Adventure, Romance, Sci-Fi, War	26.94
Shawshank Redemption, The (1994)	Drama	26.73

Table 4
Top-10 recommendations for user `new-u-4`.

HPF model typically returned lists that were still genre-coherent: on average, *roughly seven of the ten* recommended titles shared at least one genre with the user's seeds, while the remaining two to three items introduced new genres and thus encouraged discovery. Genre-level SHAP visualisations indicated that each suggestion was usually driven by one or two strongly positive genres (e.g., *Mystery* $\approx$ +72%) and tempered by mildly negative ones (e.g., *Drama* $\approx$ -8%). These observations suggest that the pipeline can preserve relevance and provide intuitive explanations even when only a handful of interactions are available.

4.3. User-Focused Explanations

To illustrate explanation quality for cold-start users, and to simulate diverse preferences and sparse histories, ten synthetic user profiles were created, enabling controlled evaluation of the explanation module. As the focus was on validating the explanation mechanism, no real users were involved. We showcase the results for user labeled `new-u-4`. Table 3 gives this user's sparse history: five titles dominated by *Drama/War* with a touch of *Comedy* and *Horror*. Table 4 shows the Top-10 recommendations, headed by *Star Wars (1977)*.

Visual explanation. Figure 3 aggregates user- and item-factors into genre-level SHAP values. **War** (+52.1%), **Animation** (+37.4%), and **Crime** (+8.9%) provide the strongest positive signals; **Fantasy** (−0.3%) exerts a mild negative influence.

Textual explanation. Explanations enable the generation of a textual recommendation prompt by populating a predefined template:

Based on your viewing history (<HistoryMovies>), we're excited to recommend <RecommendedMovie> as a perfect fit for you. This choice is driven by your strong preference for <Genre1>, <Genre2>, and <Genre3> genres. Because you've indicated you don't enjoy <DislikedGenre>, we've omitted it entirely, and since you're neutral on <Genre4> and <Genre5>, those genres played virtually no role in this pick. We think you're going to love it!

In case of user `new-u-4`:

Based on your viewing history (Killing Fields; Afterglow; Waiting for Guffman; Cat People; Band Wagon), we're excited to recommend Star Wars as a perfect fit for you. This choice is driven by your strong preference

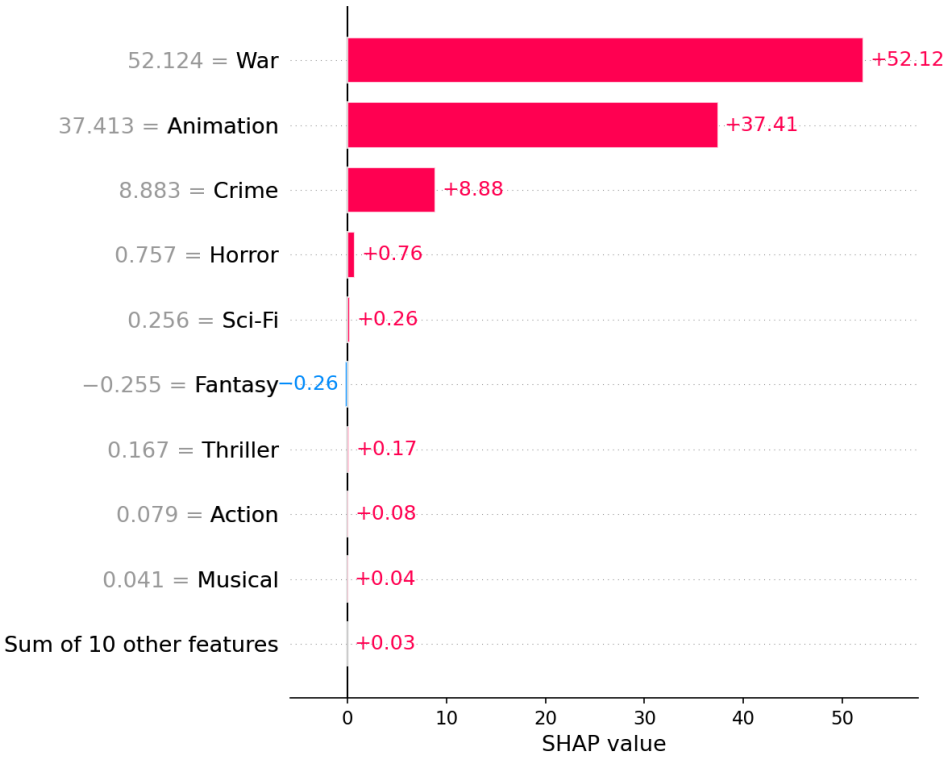

Figure 3: Genre-level SHAP explanation for the top recommendation.

for War, Animation, and Crime genres. Because you've indicated you don't enjoy Fantasy, we've omitted it entirely, and since you're neutral on Adventure and Comedy, those genres played virtually no role in this pick. We think you're going to love it!

4.4. Stakeholder-Focused Explanations

Figure 4 translates the abstract HPF embedding into a genre–factor matrix that is easy for non-technical stakeholders to interpret. Each row represents one of the model's latent dimensions, while each column corresponds to a movie genre. Darker shades indicate that a given factor is strongly expressed by items in that genre. Several pronounced patterns emerge: one factor activates almost exclusively for *Western* titles, another peaks for *Horror*, and a third clearly tracks *Animation/Children's* content, while neighboring factors jointly capture the *Action–Adventure* spectrum.

These visual cues let marketing teams map otherwise opaque latent variables to recognizable content themes. By amplifying or suppressing specific factors in a campaign, they can steer the recommender toward inventory that best matches a target segment. For example, pushing Factor 9 to feature Halloween releases, or tuning down the Western-specific factor in regions where that genre under-performs.

4.5. Preference modeling

To illustrate how filtering according to user preferences works, we create a profile for user `new-u-4`, specifying *Romance* as a disliked genre. This allows us to simulate a personalized content screening process based on explicit user preferences. Table 5 shows the recommendations that remain after applying this filter, effectively removing all titles associated with the unwanted genre. This step demonstrates how simple preference-based filtering can help tailor recommendations to better align with a user's tastes and avoid suggesting content they are likely to reject.

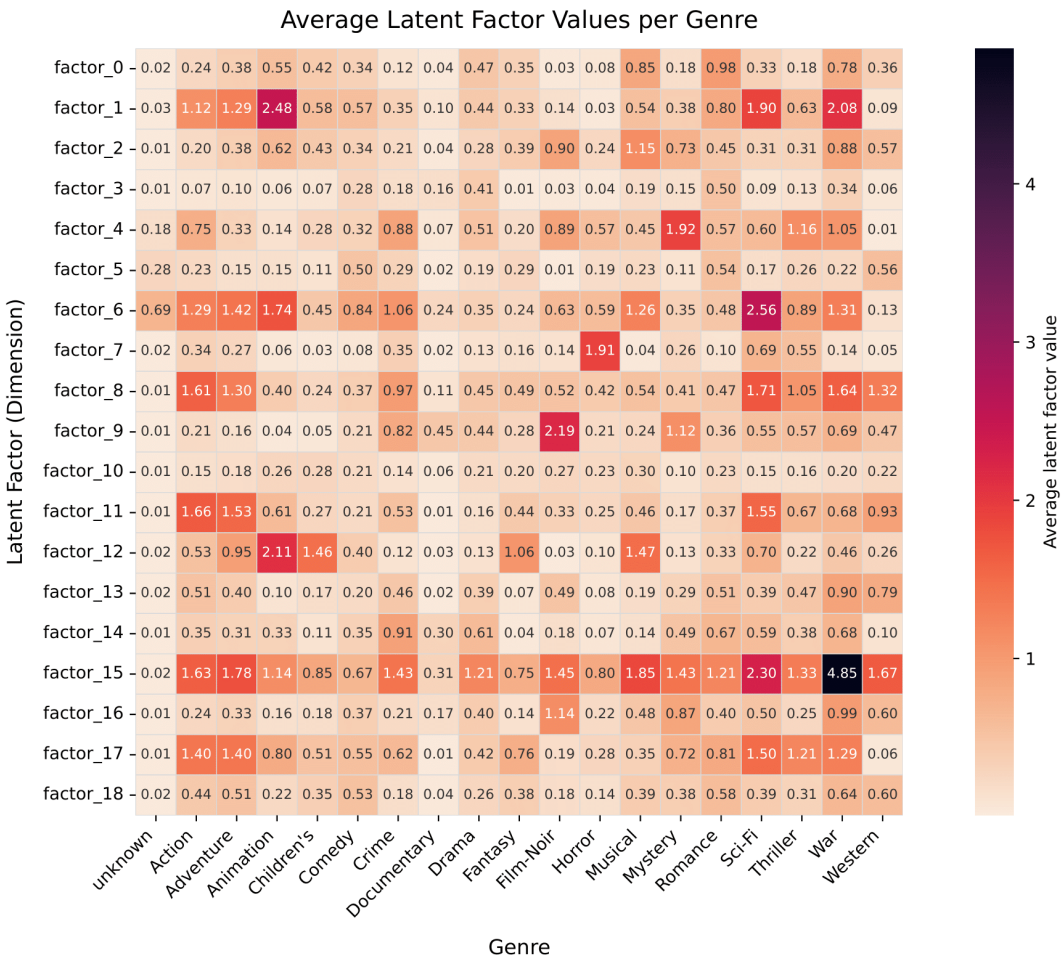

Figure 4: Average latent-factor strength per genre. Darker cells mark stronger associations.

Title	Genres
Shawshank Redemption, The (1994)	Drama
Silence of the Lambs, The (1991)	Drama, Thriller
Fargo (1996)	Crime, Drama, Thriller
Godfather, The (1972)	Action, Crime, Drama
Raiders of the Lost Ark (1981)	Action, Adventure
Schindler's List (1993)	Drama, War
One Flew Over the Cuckoo's Nest (1975)	Drama

Table 5
Recommendations after filtering out disliked genres for user `new-u-4`.

4.6. Serendipity and Business Rules

To illustrate the serendipitous recommendations, table 6 shows both the random pick and the novel genre pick with their respective attributes, such as title, genres, average rating (*Avg Rating*), and number of ratings (*Num Ratings*).

We can observe that the random pick adheres well to the business rules of serendipitous recommendation. Although the genre, *Drama*, is familiar to the user, the selected film has received only a single rating. Despite its limited exposure, the rating is the maximum possible - 5.0 - indicating a potentially high-quality item. This makes it a strong candidate for serendipity: an overlooked film that may align with the user's preferences, offering the possibility of surprising satisfaction.

Type	Title & Genres	Avg Rating	Num Ratings
Random Pick	*Aiqing wansui (1994)* Genre: Drama	5.00	1
Novel Pick	*Best Men (1997)* Genre: Action, Comedy, Crime, Drama	3.40	5

Table 6
Serendipitous recommendations for `new-u-4`: a random pick and a genre-novelty-based pick.

In contrast, the novelty-based pick excels in fulfilling both core goals of serendipity-novelty and unexpected relevance. The recommended title falls within the genres *Action* and *Crime*, which both, according to the user's history 3, represent a new area of interest. Additionally, with just five ratings, it remains relatively undiscovered by the broader user base. This not only increases the chance of offering the user something fresh, but also supports business objectives such as content discovery and catalog diversification.

Taken together, these two recommendations exemplify complementary approaches to serendipity: one driven by quality and underexposure, the other by genre novelty and user exploration. This demonstrates the effectiveness of using both randomization and personalized novelty scoring in surfacing engaging, lesser-known content.

5. Discussion

In this work, we address a clear literature gap: despite their state-of-the-art accuracy, modern next-purchase recommendation pipelines remain opaque "black boxes" to both end-users and organizational stakeholders. To bridge this gap in a *multistakeholder* context, we integrate explainable AI techniques that surface the drivers of each suggestion, reduce cognitive load during decision making, and build trust. We also embed business rules: serendipity and novelty picks driven by user profiles to introduce under-exposed yet relevant items, balancing discovery with precision. This demonstrates that transparency, exploratory novelty, and high-performance recommendation can coexist while supporting professional information needs and decision processes.

Our *multistakeholder* approach delivers tailored explanations for both end-users and business stakeholders. For end-users, we apply SHAP to generate model-agnostic, genre-level attributions that clarify which past interactions or categories influenced each recommendation, fostering trust, engagement, and enabling targeted marketing messages. For business stakeholders, we visualize model embeddings via a latent-factor loading matrix, mapping each latent dimension to movie genres, to link model structure with domain concepts, support strategy refinement, business-rule tuning, model-selection decisions, and thus potentially reduce cognitive load. This approach generalizes to any model: when explicit factors are absent, we would project user/item vectors (e.g., via PCA or UMAP) for equivalent interpretability.

To ensure that new users receive meaningful suggestions while still enabling exploration of unexpected content, we leveraged lightweight user profiles that store minimal preference information. By embedding business rules: novelty and serendipity filters, promotion of under-exposed items, and category-based diversity, we balanced personalized recommendations with overarching organizational goals. To address the cold-start problem, we simulated ten new users with only a few interactions each and evaluated model performance using *NDCG@10*, complemented by SHAP explanations to highlight the most influential genre contributions, demonstrating robust recommendation quality even in sparse-data scenarios.

Despite promising results, this study has limitations. It relies solely on the MovieLens dataset: pilot data from the *PEER project* were not yet available at the time of writing, and the system has not been tested in a live environment. Moreover, new users and profiles were simulated, so real-world dynamics may differ.

6. Conclusion and Future Work

This work shows that a state-of-the-art next-purchase recommendation pipeline - enriched with explainable AI, business-rule–driven serendipity, and user-profile–driven novelty - can deliver both high accuracy and transparency, reduce users' cognitive load, and satisfy the needs of end-users and organizational stakeholders.

Expanding beyond the movie domain, we plan to test the methodology on diverse digital content platforms - such as the proprietary data from the *PEER project*, music streaming services, e-learning platforms, and news portals - validating its effectiveness in live environments and refining business-rule logic under real-world user behaviors.

Acknowledgments

This paper is part of a project that has received funding from the European Union's Horizon Europe Research and Innovation Programme, under Grant Agreement number 101120406. The paper reflects only the authors' view and the EC is not responsible for any use that may be made of the information it contains.

The research has been supported by a grant from the Priority Research Area (DigiWorld) under the Strategic Programme Excellence Initiative at Jagiellonian University.

Contribution of Maciej Mozolewski for the research for this publication has been supported by a grant from the Priority Research Area (DigiWorld) under the Mark Kac Center for Complex Systems Research Strategic Programme Excellence Initiative at Jagiellonian University.

Declaration on Generative AI

During the preparation of this work, the authors used OpenAI *o4-mini-high* and *GPT-5*. The tools assisted with drafting and revising text, paraphrasing and style polishing, abstract drafting, formatting guidance, and programming support (code drafting and refactoring). All AI-assisted content and code were reviewed, tested, and edited by the authors, who accept full responsibility for the final manuscript.

References

[1] M. Westphal, M. Vössing, G. Satzger, G. B. Yom-Tov, A. Rafaeli, Decision control and explanations in human-ai collaboration: Improving user perceptions and compliance, Computers in Human Behavior 144 (2023) 107714. URL: https://www.sciencedirect.com/science/article/pii/S0747563223000651. doi:10.1016/j.chb.2023.107714.

[2] Y. Kim, S. Oh, C. Noh, E. Hong, S. Park, Design of a serendipity-incorporated recommender system, Electronics 14 (2025). URL: https://www.mdpi.com/2079-9292/14/4/821. doi:10.3390/electronics14040821.

[3] F. Abbas, Serendipity in recommender system: A holistic overview, in: 2018 IEEE/ACS 15th International Conference on Computer Systems and Applications (AICCSA), 2018, pp. 1–2. doi:10.1109/AICCSA.2018.8612895.

[4] R. Burke, G. Adomavicius, T. Bogers, T. D. Noia, D. Kowald, J. Neidhardt, Özlem Özgöbek, M. S. Pera, N. Tintarev, J. Ziegler, De-centering the (traditional) user: Multistakeholder evaluation of recommender systems, 2025. URL: https://arxiv.org/abs/2501.05170. arXiv:2501.05170.

[5] R. Confalonieri, L. Coba, B. Wagner, T. R. Besold, A historical perspective of explainable artificial intelligence, WIREs Data Mining and Knowledge Discovery 11 (2021) e1391. URL: https://wires.onlinelibrary.wiley.com/doi/abs/10.1002/widm.1391. doi:10.1002/widm.1391.

[6] R. Schellingerhout, F. Barile, N. Tintarev, A co-design study for multi-stakeholder job recommender system explanations, in: L. Longo (Ed.), Explainable Artificial Intelligence, Springer Nature Switzerland, Cham, 2023, pp. 597–620.

[7] W. van Melle, Mycin: a knowledge-based consultation program for infectious disease diagnosis, International Journal of Man-Machine Studies 10 (1978) 313–322. URL: https://www.sciencedirect.com/science/article/pii/S0020737378800492. doi:https://doi.org/10.1016/S0020-7373(78)80049-2.

[8] J. L. Herlocker, J. A. Konstan, J. Riedl, Explaining collaborative filtering recommendations, in: Proceedings of the 2000 ACM Conference on Computer Supported Cooperative Work, CSCW '00, Association for Computing Machinery, New York, NY, USA, 2000, p. 241–250. URL: https://dl.acm.org/doi/10.1145/358916.358995. doi:10.1145/358916.358995.

[9] L. M. de Campos, J. M. Fernández-Luna, J. F. Huete, An explainable content-based approach for recommender systems: a case study in journal recommendation for paper submission, User Modeling and User-Adapted Interaction 34 (2024) 1431–1465. URL: https://link.springer.com/article/10.1007/s11257-024-09400-6. doi:10.1007/s11257-024-09400-6.

[10] E. Kannout, H. S. Nguyen, M. Grzegorowski, Speeding up recommender systems using association rules, in: N. T. Nguyen, T. K. Tran, U. Tukayev, T.-P. Hong, B. Trawiński, E. Szczerbicki (Eds.), Intelligent Information and Database Systems, Springer Nature Switzerland, Cham, 2022, pp. 167–179.

[11] M. A. Chatti, M. Guesmi, A. Muslim, Visualization for recommendation explainability: A survey and new perspectives, ACM Trans. Interact. Intell. Syst. 14 (2024). URL: https://dl.acm.org/doi/10.1145/3672276. doi:10.1145/3672276.

[12] S. Zhang, L. Yao, A. Sun, Y. Tay, Deep learning based recommender system: A survey and new perspectives, ACM Comput. Surv. 52 (2019). URL: https://dl.acm.org/doi/10.1145/3285029. doi:10.1145/3285029.

[13] S. Sedhain, A. K. Menon, S. Sanner, L. Xie, Autorec: Autoencoders meet collaborative filtering, in: Proceedings of the 24th International Conference on World Wide Web, WWW '15 Companion, Association for Computing Machinery, New York, NY, USA, 2015, p. 111–112. URL: https://dl.acm.org/doi/10.1145/2740908.2742726. doi:10.1145/2740908.2742726.

[14] C. V. Roberts, E. Elahi, A. Chandrashekar, On the bias-variance characteristics of lime and shap in high sparsity movie recommendation explanation tasks, 2022. URL: https://arxiv.org/abs/2206.04784. arXiv:2206.04784.

[15] C. Gao, Y. Zheng, N. Li, Y. Li, Y. Qin, J. Piao, Y. Quan, J. Chang, D. Jin, X. He, Y. Li, A survey of graph neural networks for recommender systems: Challenges, methods, and directions, ACM Trans. Recomm. Syst. 1 (2023). URL: https://dl.acm.org/doi/10.1145/3568022. doi:10.1145/3568022.

[16] P. Cao, P. Liò, Genrec: Generative sequential recommendation with large language models, 2024. URL: https://arxiv.org/abs/2407.21191. arXiv:2407.21191.

[17] J. Ji, Z. Li, S. Xu, W. Hua, Y. Ge, J. Tan, Y. Zhang, Genrec: Large language model for generative recommendation, in: Advances in Information Retrieval: 46th European Conference on Information Retrieval, ECIR 2024, Glasgow, UK, March 24–28, 2024, Proceedings, Part III, Springer-Verlag, Berlin, Heidelberg, 2024, p. 494–502. URL: https://doi.org/10.1007/978-3-031-56063-7_42. doi:10.1007/978-3-031-56063-7_42.

[18] M. Narvekar, K. Bharucha, V. Vishwanath, N. Gabani, S. Fernandes, Enhancing interpretability in diverse recommendation systems through explainable ai techniques, Journal of Computational Analysis and Applications (JoCAAA) 32 (2024) 447–456. URL: https://eudoxuspress.com/index.php/pub/article/view/1427.

[19] J. Zhong, E. Negre, Shap-enhanced counterfactual explanations for recommendations, in: Proceedings of the 37th ACM/SIGAPP Symposium on Applied Computing, SAC '22, Association for Computing Machinery, New York, NY, USA, 2022, p. 1365–1372. URL: https://dl.acm.org/doi/10.1145/3477314.3507029. doi:10.1145/3477314.3507029.

[20] J. Tan, S. Xu, Y. Ge, Y. Li, X. Chen, Y. Zhang, Counterfactual explainable recommendation, in: Proceedings of the 30th ACM International Conference on Information & Knowledge Management, CIKM '21, Association for Computing Machinery, New York, NY, USA, 2021, p. 1784–1793. URL: https://dl.acm.org/doi/10.1145/3459637.3482420. doi:10.1145/3459637.3482420.

[21] O. Barkan, V. Bogina, L. Gurevitch, Y. Asher, N. Koenigstein, A counterfactual framework for

learning and evaluating explanations for recommender systems, in: Proceedings of the ACM Web Conference 2024, WWW '24, Association for Computing Machinery, New York, NY, USA, 2024, p. 3723–3733. URL: https://dl.acm.org/doi/10.1145/3589334.3645560. doi:10.1145/3589334.3645560.

[22] L. Longo, M. Brcic, F. Cabitza, J. Choi, R. Confalonieri, J. D. Ser, R. Guidotti, Y. Hayashi, F. Herrera, A. Holzinger, R. Jiang, H. Khosravi, F. Lecue, G. Malgieri, A. Páez, W. Samek, J. Schneider, T. Speith, S. Stumpf, Explainable artificial intelligence (xai) 2.0: A manifesto of open challenges and interdisciplinary research directions, Information Fusion 106 (2024) 102301. URL: https://www.sciencedirect.com/science/article/pii/S1566253524000794. doi:10.1016/j.inffus.2024.102301.

[23] I. Al-Hazwani, T. Luo, O. Inel, F. Ricci, M. El-Assady, J. Bernard, Scrollypoi: A narrative-driven interactive recommender system for points-of-interest exploration and explainability, in: Adjunct Proceedings of the 32nd ACM Conference on User Modeling, Adaptation and Personalization, UMAP Adjunct '24, Association for Computing Machinery, New York, NY, USA, 2024, p. 292–304. URL: https://dl.acm.org/doi/10.1145/3631700.3665183. doi:10.1145/3631700.3665183.

[24] J. Vig, S. Sen, J. Riedl, Tagsplanations: explaining recommendations using tags, in: Proceedings of the 14th International Conference on Intelligent User Interfaces, IUI '09, Association for Computing Machinery, New York, NY, USA, 2009, p. 47–56. URL: https://dl.acm.org/doi/10.1145/1502650.1502661. doi:10.1145/1502650.1502661.

[25] H. Abdollahpouri, R. Burke, Multistakeholder recommender systems, in: F. Ricci, L. Rokach, B. Shapira (Eds.), Recommender Systems Handbook, Springer US, New York, NY, 2022, pp. 647–677. URL: 10.1007/978-1-0716-2197-4_17. doi:10.1007/978-1-0716-2197-4_17.

[26] H. Abdollahpouri, G. Adomavicius, R. Burke, I. Guy, D. Jannach, T. Kamishima, J. Krasnodebski, L. Pizzato, Multistakeholder recommendation: Survey and research directions, User Modeling and User-Adapted Interaction 30 (2020) 127–158. URL: https://link.springer.com/article/10.1007/s11257-019-09256-1. doi:10.1007/s11257-019-09256-1.

[27] G. Cornacchia, F. M. Donini, F. Narducci, C. Pomo, A. Ragone, Explanation in multi-stakeholder recommendation for enterprise decision support systems, in: A. Polyvyanyy, S. Rinderle-Ma (Eds.), Advanced Information Systems Engineering Workshops, Springer International Publishing, Cham, 2021, pp. 39–47.

[28] A. Salah, Q.-T. Truong, H. W. Lauw, Cornac: A comparative framework for multimodal recommender systems, Journal of Machine Learning Research 21 (2020) 1–5. URL: http://jmlr.org/papers/v21/19-805.html.

[29] S. Rendle, C. Freudenthaler, Z. Gantner, L. Schmidt-Thieme, Bpr: Bayesian personalized ranking from implicit feedback, in: Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence, UAI '09, AUAI Press, Arlington, Virginia, USA, 2009, p. 452–461.

[30] Y. Koren, R. Bell, C. Volinsky, Matrix factorization techniques for recommender systems, Computer 42 (2009) 30–37. URL: https://ieeexplore.ieee.org/document/5197422. doi:10.1109/MC.2009.263.

[31] P. Gopalan, J. M. Hofman, D. M. Blei, Scalable recommendation with hierarchical poisson factorization, in: Proceedings of the Thirty-First Conference on Uncertainty in Artificial Intelligence, UAI'15, AUAI Press, Arlington, Virginia, USA, 2015, p. 326–335.

[32] I. Shenbin, A. Alekseev, E. Tutubalina, V. Malykh, S. I. Nikolenko, Recvae: A new variational autoencoder for top-n recommendations with implicit feedback, in: Proceedings of the 13th International Conference on Web Search and Data Mining, WSDM '20, Association for Computing Machinery, New York, NY, USA, 2020, p. 528–536. URL: https://dl.acm.org/doi/10.1145/3336191.3371831. doi:10.1145/3336191.3371831.