SRYAHWA PUBLICATIONS

Journal of Advertising and Public Relations
ISSN: 2639-1953 | Volume 5 Issue 2, 2025
https://doi.org/10.22259/2639-1953.0502003

RESEARCH ARTICLE

Advertising Wisely: A Comprehensive Study of State-of-the-Art Recommendation Models for Cold-Start Scenarios in Online Magazine Advertising

Yu DU[1], Erwann Lavarec[2]

[1,2]*Department of Research and Development, CloudIs Mine, Montpellier, France.*

Received: 26 August 2025 Accepted: 09 September 2025 Published: 19 September 2025

Corresponding Author: Yu DU, Department of Research and Development, CloudIs Mine, Montpellier, France.

Abstract

Recommender systems are crucial for traffic driven online media platforms, aligning relevant advertisements with target audiences to enhance engagement and revenue. However, selecting effective methods remains challenging, especially under cold-start conditions where user interaction data is sparse. This study evaluates 13 state-of-the-art recommendation models—including rule-based, machine learning, and deep learning methods—using eight standard metrics, providing actionable insights for online advertising practitioners. Additionally, we propose as impleyet effective strategy to mitigate the cold-start issue by reformulating interaction data to increase its density. Our results show that: (1) deep learning models typically outperform classical machine learning approaches in cold-start scenarios; (2) the proposed data reformulation significantly improves accuracy across all tested models without reducing recommendation diversity; and (3) when data sparsity decreases, light weight machine learning methods can outperform complex deep learning models, offering practical and efficient solutions for real-world deployment.

Keywords: Online Advertising, Recommendation System, Cold-Start Problem, User Interaction Data, Machine Learning.

1. Introduction

In traffic-driven online media and magazine platforms, advertising and recommender systems are essential for delivering personalized content and maximizing user engagement (Binns,2016; Malthouse et al., 2019; Zhao et al.,2020). The primary objective of these web publishers is to deliver advertisers' brand information to the most relevant group of users, who are likely to take a desired action (e.g., clicking on an ad) after being exposed to specific types of advertisement impressions. A commonpricing modelis Cost-Per-Click (CPC), where web publishers (e.g.,online magazine platforms) earn revenue each time a user clicks on a displayed advertisement (Mohan,2020; Najafi-As adolahi and Fridgeirsdottir,2014). Consequently, implementing an effective recommender system is vital for online media platforms,with the aim of suggesting them ost relevant advertisements to users and thereby maximizing both click-through rates (CTR) and advertising revenue.

Since the early 1990s, the field of recommender systems has evolved significantly, producing a wide array of models and technologies, such as collaborative filtering (CF) and content-based approaches (Dongetal.,2022; Javedetal.,2021; Sarwaretal.,2001). Recent models use advanced deep learning techniques, enhancing the accuracy of recommendations (Zhang et al.,2019). For web publishers, choosing the most suitable recommender system from the myriad of options for online advertising is particularly challenging. The efficacy of recommendation approaches canvary considerably, especially in cases involving cold-start users with minimal or no prior interactions— which is a

Citation: Yu DU, Erwann Lavarec. Advertising Wisely: A Comprehensive Study of State-of-the-Art Recommendation Models for Cold-Start Scenarios in Online Magazine Advertising. Journal of Advertising and Public Relations. 2025; 5(2): 23-33.

prevalent challenge in there alm of online advertising (Panetal.,2019). This issue is typically manifested as extreme sparsity in the user–item interaction matrix, where the vast majority of entries are unobserved. In a multitude of real-world scenarios, particularly in traffic-driven platforms, the number of unique individuals (often anonymous visitors) can far exceed the number of items (e.g., advertisement candidates), further exacerbating the sparsity problem. This high user-to-item ratio presents challenges for learning reliable user preferences and training effective recommendation models (Bobadilla et al.,2013).

In light of these challenges, the present paper sets out a comprehensive study that rigorously evaluates a wide variety of thir teen different recommender systems using a production-level, real- world tracking dataset in the field of online media and magazine platforms. The objective of this study is two fold: first, to provide an empirical analysis; and second, to offer insights that help practitioners make more informed decisions about which recommendation approach to prioritize for initial consideration. Further more, a simpleyet effective strategy is proposed to mitigate the cold- start problem by reducing data sparsity, and Its impact on the accuracy of the tested recommender systems is evaluated. The contributions of our study are as follows:

- We conduct arigorous,real-world performance bench marking of 13 state-of-the-art recommender models using production user interaction data from an online magazine and advertising platform, offering actionable insights for practitioners aiming to optimize advertisement targeting in cold-start conditions.
- We introduce a simple yet effective sparsity-reduction strategy to mitigate the cold-start problem, enhancing recommendation accuracy without compromising diversity or coverage.

The remainder of the paper is structured as follows. Section 2 briefly introduces the 13 recommender approaches evaluated in this study. Section 3 discusses the cold-start challenge in online advertising, outlines our proposed sparsity-reduction strategy, and formulates the research questions. Section 4 describes the experimental protocol, including dataset characteristics, evaluation metrics, and model implementation.Section 5 presents the results and key findings, followed by a discussion of the study's limitations and implications. Finally, Section 6 concludes the paper.

2. Considered Recommendation Approaches

As outlined in Section 1, this study examines 13 state-of-the-art recommendation models drawn from leading conferences and journals with in there commender systems community. These models Represent a broad spectrum of algorithmic paradigms, which we group into three main categories: rule-based approaches, machine learning (ML)-based approaches, and deep learning (DL)-based models (cf.Table 1). The following points provide a concise overview of each model's underlying principles.

Table1. *Summary of considered recommendation models.*

Recommendation Model	Reference	Type
EASE	(Steck,2019)	ML-based
FM	(Rendle,2010)	ML-based
Item2vec	(BarkanandKoenigstein,2016)	ML-based
ItemKNN	(Sarwaretal.,2001)	Rule-based
LightGCN	(Heetal.,2020)	DL-based
MF	(Korenetal.,2009)	ML-based
MostPop	(Jietal.,2020)	Rule-based
Multi-VAE	(Liangetal.,2018)	DL-based
NeuMF	(Heetal.,2017)	DL-based
NFM	(HeandChua,2017)	DL-based
NGCF	(Wangetal.,2019)	DL-based
PureSVD	(Cremonesietal.,2010)	ML-based
SLIM	(NingandKarypis,2011)	ML-based

- *Embarrassingly Shallow Auto Encoder (EASE).* A simple linear recommender tailored for sparse implicit feedback.Instead of deep autoencoders, it learns a closed-form item–item weight matrix

to reconstruct user interactions and often rivals or surpasses complex neural methods, including Multi-VAE (Liang et al.,2018; Steck,2019; Zhang et al.,2020).

- *Factorization Machines (FM).* A generalized factorization model that captures pairwise interactions among arbitrary features (users,items,andsidedata). It keeps MF-like efficiency while flexibly integrating context in sparse settings (Koren et al.,2009; Rendle,2010).
- *Neural Item Embedding (Item2vec).* An embedding-based recommendation model inspired by Word2vec. It treats user interaction sequences like sentences and co-occurring items like words, learning vector representations that capture item similarity. These embeddings can then be used for collaborative filtering, often out performing classic MF on benchmark datasets (Barkan and Koenigstein,2016; Koren et al.,2009; Mikolov et al.,2013).
- *Item-based Collaborative Filtering (Item KNN).* A memory-based method that recommends items similar to those a user already consumed. It computes item–item similarity from historical interactions and aggregate esneighbors to produce rankings; no model training is required (Sarwar et al.,2001).
- *Simplified Graph Convolution Network (Light GCN).* A graph-based approach that propagates user and item embeddings over the interaction graph while deliberately removing feature transformations and nonlinearities. This stream lined design improves generalization and accuracy over earlier GNN recommenders like NGCF (Heetal.,2020; Wuetal.,2021).
- *Matrix Factorization (MF).* A foundational latent factor model that represents users and items as vectors in a shared low-dimensional space. Recommendations are generated by measuring the alignment between these vectors, allowing the model to capture hidden patterns in user preferences. MF remains a widely used and effective baseline for collaborative filtering(Koren et al.,2009).
- *Most Popular Items (Most Pop).* A non-personalized base line that ranks items by overall interaction counts. Despiteits simplicity—and known popularity bias—it can be competitive on some benchmarks (Abdollahpouri, 2019; Ferrari Dacrema et al.,2019;Ji et al.,2020).
- *Multinomial Variational Autoencoder (Multi-VAE).* A generative neural model for implicit feedback that uses a multinomial likelihood with control lable regularization to learn user preference patterns from sparse data, often out performing classical ML approaches (Lianget al.,2018).
- *Neural Collaborative Filtering (NeuMF).* A hybrid neural model that merges two complementary components: Generalized Matrix Factorization (GMF), which extends MF with flexible interaction weights, and a Multi-Layer Perceptron (MLP), which captures complex nonlinear relationships. By combining these path ways, NeuMF effectively models diverse user–item interaction patterns and has become one of the most influential deep learning approaches in recommendation (Ferrari Dacremaet al.,2019;He et al.,2017).
- *Neural Factorization Machines (NFM).* An extension of FM by adding a neural layer on top of FM's interaction representation to capture higher-order, nonlinear effects, improving accuracy over FM and "Wide&Deep"-style models (Cheng et al.,2016; He and Chua,2017).
- *Neural Graph Collaborative Filtering(NGCF).* graph-based recommendation model that propagates user and item embeddings through the user-item interaction graph. By capturing higher-order connectivity and relational patterns, NGCF enriches representation learning beyond direct interactions, leading to stronger recommendation performance compared to MF and NeuMF(He et al.,2017; Koren et al.,2009; Wang et al.,2019).
- *Matrix Completion (Pure SVD).* An adaptation of singular value decomposition designed for top-n recommendation tasks. It treats missing ratings as zeros and reconstructs the user-item matrix using truncated SVD, from which ranked item lists are derived. Pure SVD provides a simple yet competitive alternative to both latent factor and neighborhood-based methods (Cremonesi et al.,2010; Koren et al.,2009).
- *Sparse Linear Methods (SLIM).* A machine learning approach that learns a sparse item-item similarity matrix to reconstruct user interactions. By directly modelling item associations from implicit feedback, SLIM produces accurate and interpretable recommendations, often outperforming traditional neighborhood-based and factorization methods (Ning and Karypis,2011).

3. Cold-Start Challenge in Online Advertising

Generally speaking, the cold-start issue is a critical challenge for recommender systems, particularly with in the domain of online advertising. Unlike e-commerce platforms (e.g.,Amazon), where users are typically logged in and individual profiles can be accumulated over time, online advertising platforms often deal with anonymous, first-time visitors, making it significantly more difficultto model user preferences. This problem is exacerbated by privacy regulations, SEO (Search Engine Optimization) constraints, and other operational factors(Kant,2021). As aresult, cold-start recommendations refer to scenarios where the system must deliver relevant suggestions to users without prior interactions. For$_{a,i}$ instance, on Appvizer [1], a leading European online media platform for SaaS (Software-as-a-Service), over one million unique visitors are recorded monthly—yet more than 95% of them exhibit cold-start behavior.

For online media marketers, choosing an effective recommendation strategy under cold-start conditions remains a complex challenge. The performance of different algorithms can vary significantly depending on the degree of data sparsity. A key motivation of this study is to provide empirical guidance by examining the performance of 13 state-of-the-art recommendation models (cf.Section2). This comparative analysis aims to support informed decision-makingin choosing the most suitable recommendation strategy for a traffic-driven, anonymous-user advertising environment. Additionally, we investigate how model performance evolves as training data becomes denser. To this end, we propose a straight forward strategy (outlined below) that transforms a highly sparse cold-start problem into a more learn able one by reformulating the user-item interaction matrix to reduce its sparsity.

Formally, let M denote an online magazineor media platform comprising a set A of articles[2]. Given a catalogue of items (i.e., advertisements) I and a set of users (visitors) U, the goal of a recommender system $Rec$is to generate a ranked list $LRec\ (u,i)$ of relevant items $i \in I$for each user $u \in U$, based on the context of an article $a \in A$ with which the user is engaged. The conventional approach reliesona user-item interaction matrix $X \in \mathrm{R}|U| \times |I|$, where Xu, i=1ifuser u interacted with item i, and 0 otherwise. In real-world advertising environments with millions of anonymous visitors, this matrix is typically extremely sparse, limiting model learnability and recommendation accuracy. To address this, we propose a shift in the modeling granularity: from visitor-level interactions to article-level agg regations. Specifically, we construct adenser matrix $X' \in \mathrm{R}|A| \times |I|$, where X'=1if item i was interacted with when displayed on article page a. Since the Number of articles is typically orders of magnitude smaller than the number of users ($|A| \ll |U|$), this transformation increases the matrix density substantially. This reformulation offers several advantages: (i) it enables the use of powerful matrix-based recommendation techniques even in sparse environments; (ii) it retains context-awareness through article-level conditioning; and (iii) it aligns with the ultimate goal of ad delivery, which takes place within content pages—thereby validating the article-level perspective when ever ads are embedded at the page level.

In the following sections,we present the experimental design and finding so four study, which aim to address the following research questions:

- *RQ1:* Which recommendation models perform best under cold-start conditions in online advertising platforms with anonymous user traffic?
- *RQ2:* To what extent does the proposed interaction matrix reformulation mitigate sparsity and enhance recommendation performance across different model types?

4. Experimental Protocol

4.1 Data set

The experiments conducted in this study are based on a real-world dataset collected through Appvizer's tracking system over a six-month period (January to June 2024). It captures visitor interactions—specifically, advertisement clicks—across all article pages published on the Appvizer platform during that period.As highlighted in Section 3, we evaluated 13 state-of-the-art recommendation models under two recommendation scenarios:

Highsparsity, based on the original *visitor-level* interaction matrix X

Lowersparsity, using the *article-level* interaction matrix X'

[1]https://www.appvizer.com

[2]While we use the term articles to reflect our case study on online magazines, the proposed strategy generalizes to any online advertising context where ads are displayed within specific web pages. In this broader view, each "article" simply represents a distinct content page, such as product pages, blog posts, or landing pages, onto which advertisements are served.

Table 2. *Dataset statistics under two recommendation scenarios*

Recommendation Scenario	# Users	# Items	# Interactions	Sparsity
visitor-level	57 902	444	79 530	99.69%
article-level	2 985	444	79 530	93.99%

Note: *In the article-level setting, "users" refer to article pages rather than individual visitors*

4.2 Recommender Constructions

Constructing recommendation models from scratch can be both time-intensive and inefficient, particularly when mature libraries off errobust, ready-to-use implementations. Several open-source frameworks have emerged to streamline the development of recommender systems, including Surprise (Hug,2020), Microsoft recommenders (Graham et al.,2019), and Daisy (Sun et al.,2023). These libraries provide abroad suite of pre-implemented, state-of-the-art models, enabling researchers and practitioners to focus on experimentation, tuning, and deployment rather than model implementation.

In this study, we adopted the *Daisy* framework to implement the 13 recommendation models under comparison. Each model was instantiated in tw oversions, corresponding to the two data set scenarios: visitor-level (high sparsity) and article-level (reduced sparsity). Specifically, the data set was partitioned into three subsets: (i) a training set (80% of the data), (ii) a validation set (10%of the training data), and (iii) a test set (20% of the data).This split was carried out in a time- aware manner: user interactions were chronologically ordered by time stamp, and the earliest 80% were used for model training. The validation set was then used to tune model hyper-parameters (described in the next subsection), and final performance was evaluated on the held-out test set.

Hyper-parametertuning. Hyper-parameter optimization (HPO) is a critical phase in the training of ML models, aiming at identifying the combination of hyper-parameters that yields the best predictive performance. This process is particularly important for modern recommender systems, whose effectiveness can vary significantly depending on the data set characteristics and the selected hyper-parameter configuration (Jannachetal.,2015). In our study, we employed Optuna (Akiba etal.,2019), a widely adopted and actively maintained HPO framework. For each recommendation model evaluated in our experiments, we executed 50 HPO iterations with Optuna, targeting improved recommendation accuracy as the primary objective.

4.3 Evaluation Metrics

Evaluating the performance of are commender system is essential for understand ingits effectiveness across different dimensions of recommendation quality. In this study, we employ a comprehensive set of standard metrics to assess both the accuracy and diversity of the models under comparison:

- *Precision:* Measures the proportion of recommended items that are relevant to the user. It reflects the accuracy of the recommended list.
- *Recall:* Measures the proportion of relevant items that are successfully retrieved by the system. It indicates the system's ability to identify all relevant items.
- *F1-score:* The harmonic average of precision and recall, offering a balanced view of both metrics in a single score.
- *Mean Reciprocal Rank (MRR):* A rank-sensitive metric that captures the average inverse rank of the first relevant item in each recommendation list.
- *Normalized Discounted Cumulative Gain (NDCG):* Evaluates not only whether relevant items appear in the list but also how highly they are ranked. Higher-ranked relevant items contribute more to the score.
- *Hit Ratio*: Measures the proportion of times that at least one relevant item appears in the recommendation list provided to users.
- *Mean Average Precision (MAP):* evaluates the quality of ranked recommendations by averaging the precision scores obtained at the ranks where relevant items occur, across all users.
- *Coverage:* Indicates the proportion of items in the catalogue that are recommended at least once across all users. A higher coverage score signifies greater diversity in recommendations. For example, a coverage of 1means that every item in the catalogue has be enrecommended at least once.

Together, these metrics provide amultifaceted evaluation framework, allowing for robust comparisons of different recommendation algorithms in terms of both predictive accuracy and recommendation diversity.

5. Results and Discussions

In this section, we first presentand analyze the performance comparison of the tested recommendation models across both accuracy-oriented metrics (e.g.,F1-score, Hit Ratio) and coverage. We begin by examining model performance under the original high-sparsity scenario—corresponding to the *visitor-level* scenario described in Table 2—to address our first research question (RQ1) defined in Section 3.

Subsequently, we evaluate model performance on the lower-sparsity data set resulting from our proposed interaction matrix reformulation strategy (i.e.,the*article-level* recommendation scenario described in Table 2). This enables us to explore the effect of sparsity reduction and to compare model robustness across both scenarios, there by answering our second research question (RQ2). Finally, we discuss in Section 5.3 the limitations and implications of the present study.

5.1 Performance Comparison and Analyse of the Tested Recommendation Models

Figure 1 presents the comparative performance of the 13 tested state-of-the-art recommendation models across eight evaluation metrics, based on the original visitor-level (i.e., high-sparsity) data set. Each metric is reported in a *Metric@K* format, where the horizontal axis denotes the size of the recommendation list (*K*). For all reported metrics, higher values indicate better model performance.

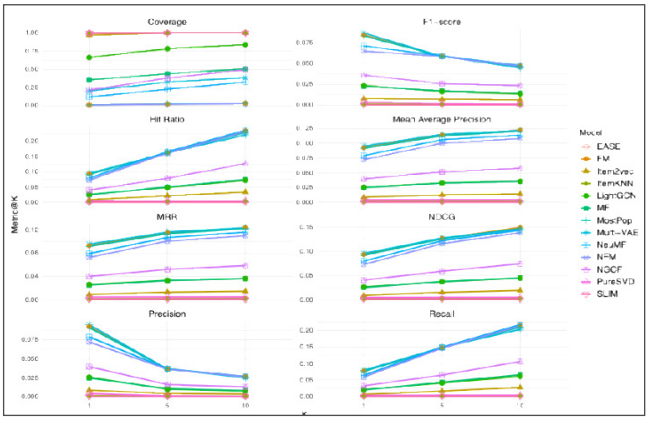

Figure 1. *Performance comparison of the 13 recommendation models on the original sparse dataset across 8 evaluation metrics.*

In terms of overall recommendation accuracy, we observe that in this cold-start setting, i.e., with a sparsity level of 99.69%, the absolute performance of all 13 evaluated models remains relatively low. For example, even the best-performing model achieves a *HitRatio @10* below 0.25, highlighting the inherent difficulty of the task. At the individual model level, there is notable variation acrossmetrics. For instance, interms of NDCG@1,the observed scoresrange from as lowas≈0.001 for EASE and Item KNN to≈0.10 for Multi-VAE, Most Pop and FM, representing a 100-fold difference in ranking quality.

Overall, DL-based recommendation models (e.g.,Multi-VAE, Neu MF, NFM, and NGCF) consistentlyoutperformML-basedmodels(e.g.,SLIM, EASE,and Pure SVD) across the21 evaluated accuracy configurations (i.e., 7 accuracy metrics × 3 list sizes in Figure 1). Surprisingly, the simplerule-based Most Popmodel, based solely on item popularity, achieves accuracy comparable to (andinsomecases surpassing) that of DL-based models. This finding, consistent with prior observations (Ferrari Dacremaetal.,2019), suggests that in cold-startre commendation scenarios, straight forward popularity-based strategies can perform competitively—even against advanced DL models.

Let us now turn to the evaluation of the 13 models with respect to the coverage metric, which measures are commende rsystem's ability to suggest adiverserange of items. One might reasonably expect that models like Most Pop, designed to recommend only them ostpopular items,would yield lower coverage. As shown in Figure 1, this intuition is generally confirmed: the most accurate model stend to exhibitlimited coverage. For instance, interms of *Coverage@10*, highly accurate models such as NeuMF, NGCF, Multi-VAE, Most Pop, and NFM all score below 0.4, indicatinga relatively narrow item recommendation spread. In contrast, models like SLIM, EASE, and Pure SVD achieve them aximum coverage value of 1, reflecting abroad erability to expose users to A wider portion of the item catalogue. This contrast illustrates a commontrade-off in recommender systems between accuracy and diversity, and emphasizes the importance of aligning model choice with the specific strategic goals (Du et al.,2021).

5.2 Impact of the sparsity reduction on recommendation performance

As shown in Table 2, shifting the recommendation focus from individual visitors to article pages leads to a reduction in dataset sparsity of ≈6%. This subsection investigates how this sparsity reduction affects model performance, as visualized in Figure 2. Each subplotin Figure 2 corresponds to a specific evaluation metric,with every recommendation model (x-axis) represented by two bars: one for the original sparsesetting (blue) and one for the densersetting (red). Since similar trends were observed across all three evaluated recommendation list lengths (i.e., Metric@1, @5, and @10),we focus our discussion belowon Metric@5results (cf.Figure 2).

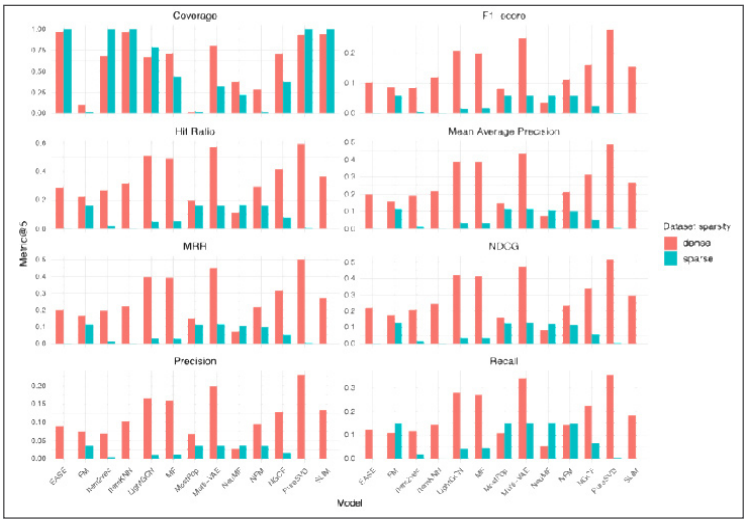

Figure 2. *Comparison of model performances using Metric@5 under sparse (visitor-level) and dense (article-level) recommendation scenarios (cf. Table 2)*

In terms of recommendation accuracy, the performances of nearly all tested models, except NeuMF, improved significantly under the article-level, denser recommendation scenario. Take, forinstance, the *Hit Ratio* metric: Multi-VAE's best value increased from ≈0.15to≈0.60, making a 300% improvement. As visually depicted in Figure 2, the red bars (densescenario) consistently exceed the blue bars (sparsescenario) across even accuracy metrics, supporting the effective nesss of our proposed sparsity-reduction strategy. Interestingly, this improvement appears more pronounced for ML-based models (e.g., EASE, SLIM, PureSVD) and classic rule-based models (e.g., ItemKNN) than for DL-based approaches (e.g., NFM, Multi-VAE, NGCF). For example, where as Multi-VAEsawa 300% boost in *Hit Ratio*, Item KNN experienced an increase of over 16 000% (from 0.002 to 0.323), and Pure SVD improved by 11 900% (from 0.005 to 0.603). These results rein force the widely held view that DL models are well-suited for highly sparse data, making them a strong choice in high-sparsity cold-start scenarios. However, when data sparsity is reduced, simpler ML models can capitalize on the denser signal to match or even exceed the performance of more complex DL counter parts.

Interestingly, the coverage performance of the tested model sreveals a distinctly different trend. As discussed in the previous section, under the high-sparsity scenario, the most accurate models generally exhibited low coverage. However, Figure 2 suggests that in the denser, article-level scenario, improvements in recommendation accuracy do not necessarily come at the expense of coverage. In fact, it is possible to achieve high accuracy and high coverage simultaneously. For instance, the Pure SVD model not only achieved top-tier accuracy scores but also reached near maximal coverage, demon strating its ability to recommend relevant items while spanning abroad portion of the catalogue. From a marketing stand point, this is particularly note worthy: the model seems capable of widely distributing the advertisement catalogue (coverage) while effectively matching relevant items to appropriate article contexts (accuracy)—a highly desirable combination for campaign optimization. In contrast, the Most Pop model, which consistently ranks among the most accurate insparse settings, continues to suffer from low coverage,even in the denser scenario. This underscores the inherent limitation of popularity-based recommendation strategies, which tend to reinforce a small subset of popular items at the expense of diversity.

Finally, as noted at the beginning of this section, we examine the ranking dynamics of the tested recommendation models—specifically, whether the top-performing models in the sparse (visitor-level) scenario (cf.Figure 3) retain their leading positions when sparsity is reduced. Figure 3 presents the comparison result in the denser, article-level setting. In this context, the DL-based Multi-VAE model remains among the top performers, confirming its robustness across different levels of data density. However, a notable shift in relative performance emerges: classic matrix factorization models such as Pure SVD and MF out perform several DL approaches, including NFM, NGCF, and NeuMF. This shift highlights the sensitivity of model effectiveness to data sparsity and suggests that certain models may be better suited to specific operational conditions.

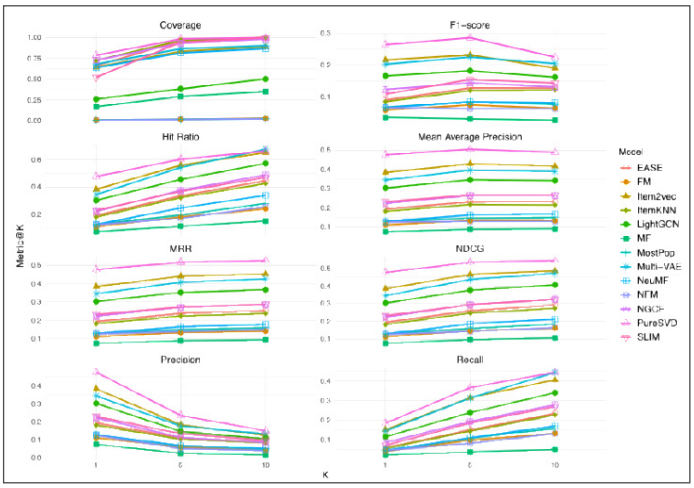

Figure 3. *Performance comparison of the 13 recommendation models on the denser dataset across 8 evaluation metrics.*

These findings lead us to the answers to ourtwore search questions:

- *RQ1: Which recommendation models perform best under cold-start conditions in online advertising platforms with anonymous user traffic?*

Answer: DL-based models, particularly Multi-VAE (Liangetal.,2018), demonstrate strong performance under extreme cold-start conditions. Interestingly, a simple rule-based strategy such as recommending the most popular items can also yield competitive accuracy, confirming findings from prior studies and highlighting its practicalutility insuch sparse scenarios (Ferrari Dacrema et al.,2019).

- *RQ2: To what extent does the proposed interaction matrix reformulation mitigate sparsity and enhance recommendation performance across different model types?*

Answer: The proposed interaction matrix reformulation offers a general strategy for reducing data sparsity. Inourstudy, it reduced sparsity by approximately 6% and significantly improved recommendation accuracy for most models. This effect is especially pronounced for classic ML-based models (e.g., Pure SVD, SLIM), which benefit more than DL models. Notably, this accuracy gain is not achieved at the expense of coverage, as many models were able to simultaneously maintain or even improve their coverage performance.

6. Discussion, Limitation and Implication

This study assessed the performance of 13 state-of-the-art recommender systems on a real-world data set from the online magazinead vertising domain, evaluated across two recommendation scenarios. Our proposed approach reduces data sparsity by shifting the recommendation focus from the visitor level (i.e.,anonymous users from web traffic) to the article page level. While this strategy led to no table gains in recommendation accuracy, as demonstrated in the previous section, it also introduces an important limitation. Specifically, there formulated recommendation task no longer targets individual users. Instead, recommendations are generated at the content (webpage) level—that is, relevant advertisements are selected to be displayed with in specific webpages. This design prioritizes enhancing user engagement with the article page itself, potentially at the cost of reduced personalization for individual visitors. However, as discussed in Section 3, traffic-driven online advertising platforms often lack sufficient information to build reliable preference profiles due to user anonymity. In such contexts, the trade-off appears acceptable: the gain in contextual relevance and accuracy at the page level mayout weigh the loss in individual-level personalization. Another important limitation lies in the staticnature

of the evaluation set up. The recommendation models in our study were trained and tested on historical interaction data, without accounting for temporal dynamics or session-level context. User interests and ad relevance often fluctuate over time or across browsing sessions, and static models may struggle to adapt to these short-term behavioral shifts. Recent advances in session-based and context-aware recommender systems underscore the importance of capturing sequential patterns and contextual signals (deSouzaPereira Moreira et al.,2021). Incorporating these dynamic factors in future work could further enhance ad targeting effectiveness and user engagement in time-sensitive scenarios.

From apractical stand point, our findings yield actionable in sights for developers and decision-makers in online advertising. First, in typical cold-start settings, DL-based approaches, such as Multi-VAE (Liangetal.,2018), consistently offer strong performance. Alternatively, in particularly sparse scenarios, simple rule-based methods like Most Pop (Jietal., 2020) can achieve surprisingly competitive results. Second, for content-driven platforms that monetize through on-page advertising, shifting the recommendation unit from individuals to content pages may be a pragmatic and effective solution. This approach can significantly enhance recommendation accuracy without incurring major trade-offs in diversity or coverage.

7. Conclusion

Improving user engagement, such as increasing click-through rates, is acentral objective form any online advertising platforms. Recommender systems play apivotrole in this process by predicting user interests based on interaction histories. However, for traffic-driven platforms dealing with anonymous visitors, the cold-start problem remains a significant challenge due to the sparsity of user interaction data.

In this study, we present a systematic evaluation of 13 state-of-the-art recommender systems— spanning rule-based, ML-based, and DL-based approaches— using real-world interaction data from an online magazine advertising platform. The goal is to provide practitioners with actionable insights into model performance under real-world constraints, particularly incold-start scenarios. To address extreme sparsity, we propose a simple but effective reformulation of the interaction matrix that improves data density without requiring user identity.

The main findings of the study are summarized as follows. In a typical cold-start setting with 99.7% sparsity, none of the considered models, regard less of type, achieve high recommendation accuracy. Even the mostad vanced DL models, such as Multi-VAE (Liangetal.,2018), Neu MF (HeandChua,2017), and NFM (He and Chua,2017), achieve only modest performance. Surprisingly, a simple popularity-based method (MostPop) performs comparably to DL-based models, suggesting that when interaction data is minimal, complex architectures may not provide substantial benefit.

In addition, when applying our sparsity reduction strategy, all model types demonstrate significant performance improvements in terms of recommendation accuracy. In particular, ML-based models (e.g.,Pure SVD (Cremonesietal.,2010), EASE (Steck, 2019)) benefit the most and often out perform DL models under the denser recommendation scenario. Interestingly, this improvement in accuracy does not come at the cost of diversity. Several models (e.g., Pure SVD) achieve high coverage while also ranking among the most accurate, indicating that relevance and diversity can coexist in less sparse recommendation scenarios.

We hope these insights will help advertising platforms better understand the trade-offs between model complexity, data sparsity, and recommendation quality, ultimately guiding them toward more effective and scalable solutions.

Declarations

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

Originality and Copyright Compliance. The authors confirm that the submitted manuscript is original, has not been published else where in any form, and is not under consideration for publication in any other journal or out let. We further confirm that the manuscript does not infringe upon any copy right or intellectual property rights. All necessary permissions have been obtained for any copy righted material included in the manuscript.

8. References

1. Abdollahpouri,H.(2019).Popularity bias in ranking and recommendation. In *Proceedings oft he 2019 AAAI/ACM Conference on AI, Ethics, and Society*, AIES ’19, page 529–530, New York, USA. ISBN: 9781450363242, DOI: 10.1145/3306618.3314309.

2. Akiba, T., Sano, S., Yanase, T., Ohta, T., and Koyama, M. (2019).Optuna:A next-generation hyper parameter optimization framework. In *Proceedings of the 25th ACM SIGKDD Inter- national Conference on Knowledge Discovery & Data Mining*, KDD '19, page 2623–2631, NewYork, NY,USA. Association for Computing Machinery, ISBN:9781450362016, DOI: 10.1145/3292500.3330701.

3. Barkan,O. and Koenigste in, N. (2016). Item 2vec:Neural item embedding for collaborative filtering. In*2016IEEE26 th International Works hopon Machine Learning for Signal Processing(MLSP)*, pages 1–6. DOI: 10.1109/MLSP.2016.7738886.

4. Binns, R. (2016).Self-authored interest profiles for personalised recommendations. *International Journal of Internet Marketing and Advertising*, 10(3):207–222, DOI: 10.1504/IJIMA.2016.080168.

5. Bobadilla, J.,Ortega,F.,Hernando, A., and Gutiérrez, A. (2013). Recommender systems survey. *Knowledge-Based Systems*, 46:109–132, ISSN: 0950-7051, DOI: 10.1016/j.knosys.2013.03.012.

6. Cheng,H.-T.,Koc,L.,Harmsen, J.,Shaked,T., Ch andra, T.,Aradhye, H., Anderson, G.,Corrado, G., Chai, W., Ispir, M., Anil, R., Haque, Z., Hong, L., Jain, V., Liu, X., and Shah, H. (2016). Wide&deep learning for recommender systems. In *Proceedings of the 1st Work shop on Deep Learning for Recommender Systems*, DLRS2016,page7–10,NewYork, NY, USA. Association for Computing Machinery, ISBN :9781450347952, DOI: 10.1145/2988450.2988454.

7. Cremonesi, P., Koren, Y., and Turrin, R. (2010). Performance of recommender algorithms on top-nrecommendation tasks. In *Proceedings of the Fourth ACM Conferenceon Recommender Systems*, RecSys'10, page39–46, NewYork, NY, USA. Association for Computing Machinery, ISBN: 9781605589060, DOI: 10.1145/1864708.1864721.

8. de Souza Pereira Moreira, G., Rabhi, S., Lee, J. M., Ak, R., and Oldridge, E. (2021).

9. Transformers 4rec:Bridging the gap between nlp and sequential/session-based recommendation. In *Proceedings of the15th ACM Conferenceon Recommender Systems*, RecSys'21,page143–153, NewYork,NY, USA. Association for Computing Machinery, ISBN: 9781450384582, DOI: 10.1145/3460231.3474255.

10. Dong,Z.,Wang,Z.,Xu,J.,Tang,R.,andWen,J. (2022). Abrief history of recommender systems. DOI:10.48550/arXiv.2209.01860.

11. Du, Y., Ranwez, S., Sutton-Charani, N., and Ranwez, V. (2021).Is diversity optimization always suitable? toward a better understanding of diversity within recommendation ap- proaches. *Information Processing & Management*, 58(6):102721, ISSN: 0306-4573, DOI: 10.1016/j.ipm.2021.102721.

12. Ferrari Dacrema, M., Cremonesi, P., and Jannach, D. (2019). Are we really making much progress?a worrying analysis of recent neural recommendation approaches. In *Proceed-ings of the 13th ACM Conferenceon Recommender Systems*, RecSys'19, pages101–109.DOI: 10.1145/3298689.3347058.

13. Graham,S., Min,J.-K., and Wu,T.(2019). Microsoft recommenders:Tools to accelerate developing recommender systems. In *Proceedings of the 13th ACM Conference on Recommender Systems*, RecSys'19,page542–543,NewYork,NY,USA. Association for Computing Machinery, ISBN: 9781450362436, DOI: 10.1145/3298689.3346967.

14. He, X. and Chua, T.-S. (2017).Neural factorization machines for sparse predictive analytics.In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '17, pages 355–364. Association for Computing Machinery, ISBN: 978-1-4503-5022-8, DOI: 10.1145/3077136.3080777.

15. He, X., Deng,K.,Wang, X.,Li,Y., Zhang,Y., and Wang, M. (2020). Lightgcn:Simplifying and powering graph convolution networkfor recommendation.In *Proceedings of the 43rd Interna- tional ACMSIGIR Conferenceon Researchand Developmentin Information Retrieval*, SIGIR '20, pages 639–648. Association for Computing Machinery, ISBN: 978-1-4503-8016-4, DOI: 10.1145/3397271.3401063.

16. He,X., Liao,L., Zhang, H., Nie,L., Hu,X., and Chua,T.-S.(2017). Neural collaborative filtering. In *Proceedings of the 26th International Conference on World Wide Web*, WWW '17, pages 173–182. ISBN: 978-1-4503-4913-0, DOI: 10.1145/3038912.3052569.

17. Hug, N. (2020). Surprise:A python library for recommender systems. *Journal of Open Source Software*, 5(52):2174, DOI: 10.21105/joss.02174.

18. Jannach, D., Lerche,L., Kamehkhosh, I., and Jugovac, M.(2015). What recom-menders recommend: An analysis of recommendation biases and possible countermeasures. *User Modeling and User-Adapted Interaction*, 25(5):427–491, ISSN: 0924-1868, DOI: 10.1007/s11257-015-9165-3.

19. Javed, U., Shaukat, K., Hameed, I. A., Iqbal, F., Alam, T. M., and Luo, S. (2021).A review of content-based and context-based recommendation systems. *International Journal of Emerging Technologies in Learning (iJET)*, 16(3):274–306, DOI: 10.3991/ijet.v16i03.18851

20. Ji,Y.,Sun, A., Zhang, J., and Li, C.(2020). Are -visit vof the popularity base line in recommender systems. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Developmentin Information Retrieval*, SIGIR'20, pages 1749–1752. Association for Computing Machinery, ISBN: 978-1-4503-8016-4, DOI: 10.1145/3397271.3401233.

21. Kant, T.(2021). Identity, Advertising, and Algorithmic Targeting: Or How Not) to TargetYour "Ideal User".*MIT Case Studies in Social and Ethical Responsibilities of Computing*, (Summer 2021), DOI: 10.21428/2c646de5.929a7db6.

22. Liang,D., Krishnan, R. G.,Hoffman, M. D., and Jebara,T. (2018). Variational autoencoders for collaborative filtering. In *Proceedings of the 2018 World Wide Web Conference*, WWW '18, pages 689–698. International World Wide Web Conferences Steering Committee, ISBN: 978-1-4503-5639-8, DOI: 10.1145/3178876.3186150.

23. Malthouse,E.C.,Hessary,Y.K.,Vakeel,K.A.,Burke,R .,andFudurić,M.(2019). An algorithm for allocating sponsored recommendations and content:Unifying program- matic advertising and recommender systems.*Journal of Advertising*, 48(4):366–379, DOI: 10.1080/00913367.2019.1652123.

24. Mikolov,T., Sutskever, I., Chen, K., Corrado, G.S., and Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In Burges, C., Bottou, L.,Welling, M., Ghahra- mani, Z., and Weinberger, K., editors, *Advances in Neural Information Processing Systems*, volume 26 of *NIPS'13*, page 3111–3119. Curran Associates, Inc., ISBN: 9781632660244.

25. Mohan, G. (2020). The role of retargeted advertisements in dealing with deflecting customers and its impact on the online buying process. *International Journal of Internet Marketing and Advertising*, 14(4):417–432, DOI: 10.1504/IJIMA.2020.111050.

26. Najafi-Asadolahi, S. and Fridgeirsdottir, K. (2014). Cost-per-click pricing for display advertising. *Manufacturing & Service Operations Management*, 16(4):482–497, DOI: 10.1287/msom.2014.0491.

27. Ning,X.and Karypis,G.(2011). Slim:Sparse linear methods fortop-n recommender systems. In *2011IEEE11th International Conferenceon Data Mining*, pages 497–506.ISSN:2374-8486, DOI: 10.1109/ICDM.2011.134.

28. Pan, F.,Li,S., Ao,X.,Tang, P., and He,Q.(2019). Warm up cold-start advertisements:Improving ctr predictions via learning to learn id embeddings. In *Proceedings of the 42nd International ACMSIGIR Conference on Research and Development in Information Retrieval*, pages 695–704. DOI: 10.1145/3331184.3331268.

29. Rendle, S. (2010). Factorization machines. In *2010 IEEE International Conference on Data Mining*, pages 995–1000. ISSN: 2374-8486, DOI: 10.1109/ICDM.2010.127.

30. Sarwar, B., Karypis, G., Konstan, J., and Riedl, J. (2001). Item-based collaborative filtering recommendation algorithms. In *Proceedings of the 10th International Conference on World Wide Web*, pages 285–295. DOI: 10.1145/371920.372071.

31. Steck, H. (2019). Embarras singly shallow auto encoders for sparse data.In *The World Wide Web Conference*, WWW '19, pages 3251–3257. Association for Computing Machinery, DOI: 10.1145/3308558.3313710.

32. Sun,Z., Fang, H.,Yang, J.,Qu,X., Liu, H.,Yu,D., Ong,Y.,and Zhang,J.(2023). Daisyrec 2.0: Bench marking recommendation for rigorous evaluation. *IEEET ransactionson Pattern Analysis and Machine Intelligence*, 45(7):8206–8226, DOI: 10.1109/TPAMI.2022.3231891.

33. Wang, X., He,X.,Wang,M., Feng, F., and Chua,T.-S. (2019). Neural graph collaborative filtering. In *Proceedings of the 42nd International ACMSIGIR Conferenceon Research and Development in Information Retrieval*, SIGIR'19, pages 165–174. Association for Computing Machinery, DOI: 10.1145/3331184.3331267.

34. Wu,Z.,Pan,S.,Chen,F.,Long,G.,Zhang,C.,andYu,P. S.(2021). A comprehensive survey on graph neural networks. *IEEE Transactions on NeuralNetworks and Learning Systems*,32(1):4– 24, DOI: 10.1109/TNNLS.2020.2978386.

35. Zhang,G.,Liu,Y., and Jin, X. (2020). A survey of auto encoder-based recommender systems.

36. Frontiers of Computer Science,14:430–450,DOI:10.1007/s11704-018-8052-6.

37. Zhang,S., Yao,L., Sun,A., and Tay,Y.(2019). Deeplearning based recommendersys- tem: A survey and new perspectives. *ACM computing surveys (CSUR)*, 52(1):1–38, DOI: 10.1145/3285029.

38. Zhao, X., Zheng, X., Yang, X., Liu, X., and Tang, J. (2020). Jointly learning to recommend and advertise.In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '20, page 3319–3327, New York, NY, USA. Association for Computing Machinery, ISBN: 9781450379984, DOI: 10.1145/3394486.3403384.