Evaluating Sequential Recommendations in the Wild: A Case Study on Offline Accuracy, Click Rates, and Consumption

Anastasiia Klimashevskaia$^{1[0000-0002-8946-667X]}$, Snorre Alvsvåg$^{2[0009-0004-8037-4965]}$, Christoph Trattner$^{1[0000-0002-1193-0508]}$, Alain D. Starke$^{3[0000-0002-9873-8016]}$, Astrid Tessem$^{2[0009-0007-0610-5427]}$, and Dietmar Jannach$^{4[0000-0002-4698-8507]}$

[1] MediaFutures, University of Bergen, Norway
{Anastasiia.Klimashevskaia@uib.no,christoph.trattner}@uib.no
[2] Amsterdam School of Communication Research, University of Amsterdam, Netherlands a.d.starke@uva.nl
[3] TV2, Norway
{Snorre.Alvsvag,Astrid.Tessem}@tv2.no
[4] University of Klagenfurt, Austria dietmar.jannach@aau.at

Abstract. Sequential recommendation problems have received increased research interest in recent years. Our knowledge about the effectiveness of sequential algorithms in practice is however limited. In this paper, we report on the outcomes of an A/B test on a video and movie streaming platform, where we benchmarked a sequential model against a non-sequential, personalized recommendation model, as well as a popularity-based baseline. Contrary to what we had expected from a preceding offline experiment, we observed that the popularity-based and the non-sequential models led to the highest click-through rates. However, in terms of the adoption of the recommendations, the sequential model was the most successful one in terms of viewing times. While our work points out the effectiveness of sequential models in practice, it also reminds us about important open challenges regarding (a) the sometimes limited predictive power of classic offline evaluations and (b) the dangers of optimizing recommendation models for click-through-rates.

Keywords: Sequential Recommendation · A/B Test · Field Test · Offline-Online Comparison

1 Introduction

Next-item recommendations are a common feature on today's music and video streaming sites. Such recommendations are shown to users—usually when the streaming of the current item ends—with the goal of pointing them to further relevant content on the site. Thereby, users are being kept active and continue to be engaged with the service.

The automated selection of recommendations to display to the user can be achieved through various methods. Assuming that the past preferences of the current user are known, any traditional recommendation approach, e.g., based on matrix factorization [42], can in principle be used to identify further relevant content for the user. In most recent years, however, the potential value of considering sequential patterns in the available interaction data has been highlighted in the literature [37]. Correspondingly, a rich variety of technical *sequential recommendation* approaches has been proposed in the recent literature, including prominent models such as SASRec [25] or BERT4Rec [40]. Furthermore, a multitude of *session-based* techniques [44,20] have been put forward, which particularly focus on the most recent interactions in anonymous sessions.

Although there are various works that describe technical models for sequential recommendation, little is documented about the effectiveness of these models in practice. While there are some industry reports on the value of *session-based* techniques [9,26,28], the literature on *sequential models*, i.e., ones that both consider long-term user preferences and sequential information, is even more scarce. This work aims to contribute to a better understanding of the value of sequential recommendation models in real-world environments. Our goal is to understand whether it is possible to successfully recommend a next item with both high accuracy and diversity on a real-life streaming platform, compared to the currently employed non-sequential mechanism. We report on the outcomes of an A/B test on a video and movie streaming platform. We compared a traditional sequence-agnostic model with a hybrid model that puts strong emphasis on the last watched item to generate next-item recommendations. The main finding of our study is that the sequential model is indeed effective in terms of increasing *consumption* (i.e., watch time) on the platform. The traditional sequence-agnostic collaborative filtering model, however, led to a higher *CTR*. This latter result corroborates findings from an earlier real-world study [15] that algorithms that lead to high CTR values may not necessarily be the best ones from the business perspective. Hence, optimizing for the CTR may be misleading, depending, of course, on the business model of the platform. Considering our findings, we argue that we narrow the gap between academic and industrial research in recommender systems, providing insights useful for both sides.

2 Background

In this section, we first discuss examples of works that emphasize the importance of considering short-term and longer-term user preferences in sequential recommendation models. Then, we review a selected set of previous publications that compare offline and online experimental results.

2.1 The Role of Long-term and Short-term User Preferences

Recent research in sequential and session-based recommendation has focused on leveraging time-ordered user interaction logs to predict users' next actions [37]. Instead of considering the traditional user-item interaction matrix, such methods try to learn user behavior patterns in sequences of interactions, suggesting a tighter connection between consumed items that closely follow each other in time. These patterns can subsequently be leveraged to predict a user's next action during an ongoing session and help to suggest a next item more effectively. In addition, such patterns can be useful at detecting and considering preference changes and interest drifts in user profiles.

In the domain of *sequential* recommendation methods, SASRec [25] is one example of a well-known approach that implements Transformer-based self-attention based sequential model. Specifically, the authors of SASRec suggest to combine long-term preference modeling capabilities and prediction accuracy in sparse data settings with their architecture. As a result, the proposed model significantly outperformed previous RNN-based techniques. Another prominent example of sequential recommendation methods is BERT4Rec [40], which allowed to successfully discover more insightful historical sequence representations through bidirectional encoding.

At the same time, *session-based* recommendation systems [44,20] have gained attention for their ability to provide personalized suggestions without relying on long-term user histories. These systems utilize various neural network architectures to model user behavior within short sessions. RNNs have shown promise in this domain, with GRU4Rec [13] representing a pioneering work in this domain. Multiple alternatives were proposed in recent years, including models that extend GRU4Rec in different ways, e.g., GRU4RecBE [36], which incorporates BERT-extracted item features. In such models, the key is often to focus on the very last interactions observed by a user. The authors of [30], however, point out that while immediate short-term trends are crucial, considering longer-term user preferences—even within an ongoing session—can actually also be important. They proposed STAMP as a way to combine both longer-term user interests and short-term interests in one recommendation model.

Considering both long-term and short-term preference information has also shown to be beneficial for *session-aware* recommendation scenarios, i.e., in settings where previous session information is retained for non-anonymous users. In [17], for example, the authors combine a traditional recommendation model based on the user-item interaction matrix with different heuristics that consider the most recently observed user interactions in an ongoing session.

2.2 Evaluation Perspectives

The landscape of research on recommender systems evaluation is largely dominated by offline experiments [12]. Reports on A/B test outcomes can sometimes be found in the literature, and in very rare cases these A/B tests also involve sequential recommendation models [27,3]. Unfortunately, however, the discussion of the specifics of the A/B test setup and outcomes is often limited to a few sentences or paragraphs in the papers.

In A/B tests involving traditional, non-sequential recommendation models, the worrying observation can be made that the outcomes of offline experiments are not aligned with online results. Researchers at Netflix, for example, mention that they do not find *"[offline experiments] to be as highly predictive of A/B* *test outcomes as we would like.”* [10], see also [39]. Generally, while there are some works that report that offline results and the outcomes observed in testing with actual users are aligned [2,5,24], there are multiple reports in the literature where this was not found to be the case [1,4,8,9,16,34,35], including works that compare user-study results with online outcomes.

In the context of session-based recommendation, notable insights regarding the transition of models “from the lab to production” can be found in [28]. Here, the authors compared both offline results, feedback from a user-centric analysis and online results. One of the reported findings is that the best offline model not necessarily generates the most useful recommendations in production. Similarly to [28], we compare in this study offline results with metrics that are relevant from a business perspective. A particular aspect of our work is that we consider two commonly used metrics (click-through rates and engagement), which are, however, not necessarily aligned. Furthermore, our work is related to [19], as we will compare the effectiveness of a traditional non-sequential model with a sequential model.

To deal with the problem of the offline-online gap, in recent years, researchers have explored alternative evaluation approaches based on *counterfactual estimation and off-policy evaluation* [23]. The promise of such approaches is that they are better able to deal with potentially biased historical data, thus leading to more accurate predictions of the online effectiveness of different algorithms using only offline data. In our research, we were not relying on such approaches because we had the opportunity to explore the true effects of different algorithms in the production system. Furthermore, some recent work [22] finds that existing off-policy evaluation schemes and certain metrics have their limitations in predicting online effectiveness. According to the authors, only specific offline estimators—in the current work Discounted Cumulative Gain (DCG)—in certain settings can be considered unbiased, and that involves non-trivial calculation of Inverse Propensity Scores (IPS) for further weighting.

3 Methodology – Study Design

Our study was performed in collaboration with the video and media streaming platform TV 2, which belongs to a national broadcaster with a country-wide user base. TV 2 service is actively logging millions of user interactions every month. The platform provides various types of content, including movies, series, news and linear TV. In this study, we focus solely on the recommendation of movies and series. In this section, we first describe the context of the platform where the A/B testing was set up, then discuss the choice of algorithms and implementation details, and finally report specifics of both offline and online testing conditions.

3.1 Application Context

The overall user interface of the platform is similar to the interface of other streaming services like Netflix. It contains multiple rows of movies and series on the landing page, some of them providing personalized content. In our study, we, however, do not focus on the landing page, but on the "next-item" recommendations that are displayed when a movie or the last available episode of a series ends (typically with rolling credits). Figure 1 shows a stylized version of this screen, where on the bottom three recommendations are usually displayed. In the case of streaming series, we note that no recommendation is displayed when the next episode of the series is available to watch. In such scenario, the next episode is automatically played.

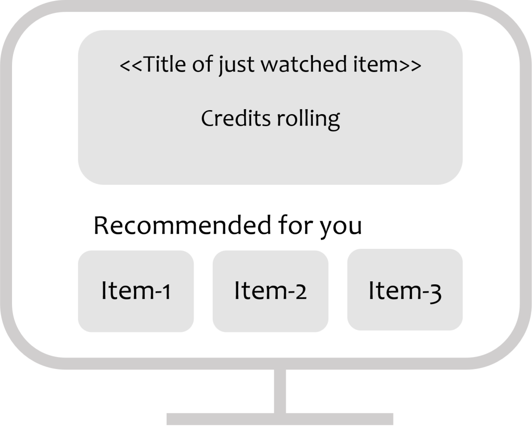

Fig. 1. Stylized screen with recommendations. After the user is finished with watching a movie or a season of a series, they should be provided with a relevant recommendation to possibly extend the watching session and keep them engaged.

3.2 Algorithms

For our study, together with our collaboration partner, we considered three recommendation models. a sequence-agnostic top-n recommendation model, a sequential model, and a hybrid that combines the scores of the two models. The non-sequential model is a traditional matrix factorization approach that is used to generate recommendations at other places of the platform. For the sequential model, the goal of the platform provider was to gauge the effectiveness of a lightweight, yet effective model that can be effectively implemented in an industrial setting at a limited cost. The hybrid model was included because the literature has shown, as discussed above, that it may be important to consider both short- and long-term user interests in sequential recommendation. All three models were evaluated in the offline tests; for the online test, the matrix factorization model and the hybrid sequential model were deployed to production.

Clearly, various other models could have been tested, and will be considered in future A/B tests. A particular focus in the future will be to consider alternative sequential models in the A/B test. One main goal of the current study however was to gauge the value of using a sequential model compared to a traditional top-n recommendation model as already deployed at the platform. Overall, there were thus certain practicalities which constrained us in the choice of the tested models.

*Matrix Factorization Model (*ALS*)* The main recommendation algorithm currently used on the platform is the Alternating Least Squares [14,41] matrix factorization model as implemented in the *Implicit* recommendation library.[5] The model was trained with data starting from 2022. No additional tuning of the hyperparameters was necessary in the offline study, because the model had been used in production for other purposes for quite some time already, and tuned extensively on the data. We simply reused the current best-performing ALS parameters for the offline evaluation in this work.

*Hybrid Sequential (*HSEQ*)* Our sequential model is a hybrid one. It considers the users' long-term preferences through the mentioned *ALS* model and combines it with the efficient and simple, yet effective Markov Chain (MC) model proposed in [31].

Given an item j, the Markov Chain model computes the relevance score of each possible next item i as follows [31]:

$$\text{score}_{MC}(i, j) = \frac{\Sigma_{p \in S_p} \Sigma_{x=1}^{|p|-1} \mathbf{1}_{EQ}(j, p_x) \times \mathbf{1}_{EQ}(j, p_{x+1})}{\Sigma_{p \in S_p} \Sigma_{x=1}^{|p|-1} \mathbf{1}_{EQ}(j, p_x)} \tag{1}$$

In the equation, S_p is the set of all sessions, $\mathbf{1}_{EQ}(x, y)$ is a function that returns 1 if $x = y$ and 0 otherwise. To counteract the given popularity bias in the data and to ensure that the MC component does not dominate the hybrid too much, log normalization was applied to the formula.

The final hybrid score $H_{u,j}$ for user u and item j is then a weighted combination of the scores returned by the *ALS* and the MC model. Since the scores may not be compatible in terms of their absolute values, a *softmax* operation is applied before combining them:

$$H_{u,j} = argmax_{i \in I}^{k}(\text{softmax}(ALS_{u,i}^{n}) \times w + \text{softmax}(MC_{u,i}^{n}) \times (1 - w)), \tag{2}$$

where I is the set of items. The parameter n denotes how many items to take from each recommender and w is a weight factor for the individual components. In the experiments, we used $n = 20$ and $k = 3$, as our goal was to present exactly three recommendations. The best value for w was 0.1, as determined through grid search in an offline evaluation.

*Popularity-Based (*POP*) Fallback* When there is a new user on the platform, neither the *ALS* nor the *HSEQ* can be effectively applied. This situation is commonly referred to as user "cold-start", see, e.g. [29], and require a special

[5] `https://benfred.github.io/implicit/`

treatment. In our case, both models revert to non-personalized suggestions of items that are *currently* popular on the platform. Specifically, the items that have been watched the most during the last 15 minutes are recommended as the most popular. Such a strategy has proven effective previously also in other recommendation domains [33].

3.3 Offline and Online Experimental Setups

Offline Experiments Before deploying the models online, we conducted a series of offline experiments on a subset of the interaction data that is collected from the production platform. The viewing data from five consecutive months was used for training, and the data from the subsequent month served as evaluation data. In total, the dataset involves more than 2 million active user profiles and about 5000 unique items. The viewing sequence data was filtered by removing sequences of episodes of the same series, as these are auto-play events. Furthermore, only sequences involving movies and series were taken into account as other types of content, such as news, are not relevant in the application setting.

To compute the evaluation metrics, 1 million sequences were sampled from the data. We evaluated both recommendation accuracy with the Mean Reciprocal Rank (MRR) metric, as well as beyond-accuracy aspects, using such metrcis as the Average Popularity Score (APS) and Coverage (Cov) [11] at different cutoff lengths. Furthermore, we report the Gini index [7], which helps us understand to what extent an algorithm is concentrating on a small set of items in its recommendations. A higher Gini index value means higher concentration and, thus, lower personalization [18]. The dataset that the experiments were conducted on is proprietary, but we are able to share the codeof the algorithms online[6]

Online Experiments The online experiment was conducted for a period of 19 days in May 2024. Ten days were used for collecting the experimental data; the preceding nine days were used for testing and for ramping up the changed recommendation service, to avoid surprise and novelty effects.

Using the A/B testing environment of the platform, a subset of the user base was involved in the experiment and randomly assigned into two experiment groups: one group received "watch next" recommendations through the *ALS* model, and the other through the *HSEQ* model. The models were trained beforehand with the previously collected data.

On average, we registered around 100,000 impressions per day from the users included in the test[7]. Approximately 45% of these impressions originated from users who were exposed to *ALS*-generated recommendations, around 40% from the user group with *HSEQ*-based recommendations, and roughly 15% of the

[6] Our code can be found at `https://github.com/sfimediafutures/Evaluating-Sequential-Recommendations`. Implementing these algorithms for other real-life platforms will commonly require certain adjustments.

[7] We unfortunately cannot disclose the exact number of active users who participated in the study.

impressions originated from the fallback *POP* model. We emphasize that in both groups non-personalized popularity-based recommendations were made in the (less frequent) case of new users.

In terms of online performance indicators, we measured the CTR as well as viewing-time-based metrics. The CTR is defined as usual as the number of clicks compared to the number of impressions (i.e., how often recommendations were displayed). The viewing-time-based metrics represent a measure for the success of the recommendations, similar to the "Long CTR" used by YouTube [6]. Specifically, we measure $V@[t]$, where $[t]$ represents the viewing time after a click on a recommendation. For example, $v@3min$ means that the user has watched at least 3 minutes of the video; $v@50\%$ indicates that at least half of the video was watched. In addition, we also measure popularity (APS), Coverage (Cov), and the Gini Index like in the offline experiments.

4 Results

We present the results of the offline and online experiments next in Section 4.1 and Section 4.2, respectively. We first provide the metrics results from both online and offline evaluation phases.

4.1 Offline Experiments

The main results of our offline evaluations are shown in Table 1 and Table 2. We recall that the recommendations shown to the users consisted of exactly three items. Thus, we report the metric values for the cut-off length of one and three. In addition, we report the MRR value also for the cut-off length of 20 to analyze if the algorithm ranking would be different for longer recommendation lists. Besides accuracy in terms of MRR, we also report the popularity score APS and Coverage. In terms of the compared models, we include the two models that were tested in production (*ALS* and *HSEQ*) as well as the "pure" MC model for reference.

Table 1. Offline evaluation – Mean Reciprocal Rank (MRR) at varying cut-off values.

Model	**MRR@1↑**	**MRR@3↑**	**MRR@20↑**
ALS	0.0028	0.0049	0.0082
MC	**0.1597**	**0.2152**	**0.2505**
HSEQ	0.1473	0.2019	0.2396

The results show that the pure MC model is *by far* better in terms of accuracy than the *ALS* model. This result is not too surprising, given that the *ALS* model returns items that are assumedly of *general* interest to the user, whereas the MC model considers the user's *current* behavior (in terms of the

Table 2. Offline evaluation – Popularity (APS), Coverage and Gini Index at varying cut-off values.

Model	**APS@1↓**	**APS@3↓**	**Cov@1↑**	**Cov@3↑**	**Gini@1↓**	**Gini@3↓**
ALS	**0.0802**	**0.0821**	**0.2240**	**0.3061**	**0.8654**	**0.8610**
MC	0.4694	0.4169	0.1763	0.2896	0.9835	0.9834
HSEQ	0.4625	0.4033	0.1400	0.2240	0.9786	0.9785

just watched item). Another factor that contributes to this huge performance gap is the popularity distribution of the data, where the top-25 items in the catalog account for more than half of the overall viewing time. The MC model is known to have a strong popularity bias [31], as can be seen also in Table 2.[8] Thus, it successfully recommends popular content in our offline experiment, whereas matrix factorization models like *ALS* are known to often focus (too) strongly on niche content, see [8]. With regard to beyond-accuracy metrics, it however turns out that *ALS* is consistently favorable over the MC model. Overall, the *HSEQ* model represents a sort of middle ground in terms of a trade-off between accuracy and beyond-accuracy aspects of the recommendations.

Since one of our goals was not to further increase the popularity bias on the platform, we decided to evaluate the hybrid model online in an A/B test. We recall that with the available A/B testing environment on the platform, only two models can be deployed in parallel. We configured the hybrid model *HSEQ* in a way ($w = 0.9$) that high offline accuracy is retained, while the popularity bias is reduced at least to some extent.

4.2 Online A/B Testing Results

The results of the online study are shown in Table 3 and Table 4. In Table 3, we report the CTR results and the viewing-time-based success metric. Besides the two main groups, *ALS* and *HSEQ*, we report the numbers for those cases where the models defaulted to popularity-based recommendations because of a lack of user data. Since this is not an additional treatment in the experiment and apply for a certain group of users, we separate the numbers for the popularity-based recommendations for the others. Still, we believe these numbers may represent an interesting reference point.

Looking at the CTR values, we find that the sequence-agnostic *ALS* model is leading to a better result than the sequential *HSEQ* model. This was quite surprising to us, given that the *ALS* model was performing quite poorly in the offline evaluation. If CTR would have been the main target measure from a business perspective, the MRR results in the offline experiments would have misled us in choosing the *HSEQ* model for production.

For the given platform, CTR is, however, not of major interest, and the focus is much more on actual viewing times, which reflect engagement and which

[8] Higher APS and Gini Index values indicate stronger popularity bias.

Table 3. Online evaluation – Click-Through Rate (CTR) and Viewing Rate (V).

Model	CTR↑	V@1min↑	V@3min↑	V@50%↑
ALS	**0.0173**	0.3554	0.3063	0.2421
HSEQ	0.0165	**0.3837**	**0.3364**	**0.2669**
POP	0.0178	0.2910	0.2462	0.1918

are assumed to be an indicator for satisfaction and continued use of the service. In terms of viewing times, the *HSEQ* model is in fact more effective than the *ALS* model. The increases are substantial: for the most relevant v@50% metric, we observe an increase in viewing times of about 10%. For comparison, recommending popular items in an unpersonalized way leads to a substantial drop in viewing times, even when compared to the *ALS* model.[9]

Table 4. Online evaluation – Popularity (APS), Coverage, Gini Index

Model	APS↓	Cov↑	Gini↓
ALS	**0.1795**	**0.1070**	**0.8275**
HSEQ	0.2040	0.0908	0.9038
POP	0.2912	0.0188	0.9875

The popularity, coverage and Gini results in Table 4 are rather unsurprising. The *HSEQ* model has a stronger tendency than the *ALS* model to recommend more popular items in general, as expected from the offline experiment. In terms of coverage, the difference between the models is small, with the *ALS* model covering a slightly larger fraction of the item space. In addition, the *ALS* model also leads to lower Gini values, which is in line with the slightly higher coverage. recommendation. We observe that the Gini Index values are lower for both models than in the offline experiments. [10]

5 Discussion

Implications Overall, our results show that sequential recommendation models—even when implemented in a relatively simple form—can effectively help users find relevant content. This is evidenced through the increased viewing times that

[9] We would like to point out that this comparison needs to be taken with a grain of salt, because new users might generally watch less than returning ones.

[10] The absolute differences are generally small, but we point out that Gini index values for recommendation algorithms usually only cover a small part of the theoretical 0-1 range [18].

we observed in our study. The proposed hybrid approach furthermore allows us—through the weight factor—to strike a balance between recommendations that are appealing to a larger fraction of the audience and recommendations of less popular items to support the discovery of novel items for users.

The good results of the *ALS* model in terms of CTR are arguably surprising, given the poor performance of the model in the offline experiments. On the one hand, this can be yet another example of a discrepancy between offline and online testing of a recommender system, proving classic offline evaluation unreliable. On the other hand, the results can be interpreted as an additional evidence that the use of the CTR as an optimization target must be well justified [21], and that higher CTR values may not correlate with the usefulness of the service. Namely, CTR can point towards *engaging* recommendations, but not necessarily accurate and relevant ones. Our interpretation for the observed findings is that the recommendations by the *ALS* model (and by the popularity-based technique) raised the users' interest, and they frequently explored these recommendations. However, it then apparently turned out that these recommendations were not as relevant for the users as those provided by the *HSEQ* model. Generally, like in search and information retrieval, it is important to understand the possible meanings of high CTR values for a given application. Did users explore many items because they were all considered interesting; or did they explore several items because they were not immediately finding what they are looking for? Is the goal of the recommendation to provide items for exploration or allow the user to reach the relevant content with fewest clicks possible?

We also emphasize the need for thorough and detailed evaluations with multiple metrics in different dimensions and recommendation qualities [45]. Given our results, there is no clear winner model across the examined metrics. The MC model led to the best accuracy results in the offline testing and the *ALS* and *HSEQ* models are competing in online accuracy depending on whether CTR or viewing rates prioritized. At the same time, the *ALS* model was the least popularity biased. It is therefore of utmost importance to carefully evaluate and understand what metric should be prioritized for deployment, as there is often a trade-off between different recommendation qualities. Moreover, the benefits of emphasizing certain qualities like the novelty of the recommendations may only become visible when considering a longer observation horizon.

In terms of limitations, we recall that our study only lasted for 19 days. This did not allow us to analyze such longitudinal effects of the sequential and sequence-agnostic recommendations. We, however, believe that it is highly important to analyze the effects over longer periods of time. In the current configuration of the *HSEQ* model, the recommendations are often including relatively popular items. In the long run, it may, however, be desirable that also lesser known items receive more exposure through the recommendations. Similar goals are also mentioned for the recommendation service by Netflix to increase what they call the "effective catalog size" [10]. In such a longitudinal study, our hybrid sequential model could be configured with different weight factors to find the optimal balance relevance and novelty for this application use case. Additionally, we acknowledge that other baselines could be used in the A/B test, however, involving real-life streaming platform involves certain restrictions and limitations, thus we operated with ALS as a baseline since it has been successfully used by the company for several years. Real-life studies on industry platforms tend to have additional costs and limitations when it comes to implementing further baselines for comparison, and it is up to the business to decide whether it is a worthy trade-off. Last but not least, real-life studies often suffer from additional limitations such as setup restrictions, which can potentially hinder the generalizability of the observed results. Certain distinct features of the platform, such as user interface, user base, demographics, and many others can have their own influence on the observations and affect the transferability of the results. However, we believe that our observations did not suffer strongly from such effects and the results are possible to reproduce to a certain extent at least within the same application domain.

Related Studies and Findings Some of the observations made in our present study are in line with insights from earlier field tests. The problem of the sometimes limited alignment of the results obtained in offline experiments with relevant business metrics in online tests is mentioned in a number of papers, as discussed in more depth earlier. Furthermore, the fact that some algorithms are well suited to raise the users' interest and attention, as measured through the CTR, but which do not lead to increased business value was reported previously, for example, in [15]. In that study, some of the algorithms that were compared in an A/B test have been successful in enticing users to inspect various mobile games and download free demo versions. However, these were not the algorithms that ultimately led to the best business value in terms of increased sales. Similarly, in our present study we found that models with a high CTR value were not optimal in terms of the more important viewing times metric.

Another finding of our offline-online comparison was that the *ALS* model was performing very poorly in the offline test. In the online study, the gap between the sequential model and the *ALS* model was, however, much smaller, and the *ALS* model even led to higher click-through-rates. This again points to the problem that it remains difficult to predict the utility or value of a set of recommendations for end users from accuracy metrics in offline evaluations. A related finding was reported [32], where the music recommendations returned by Spotify's API were performing very poorly in an offline comparison. In a subsequent user study, however, the recommendations by Spotify were considered largely relevant and particularly useful for discovery. These outcomes should encourage researchers to combine not only various metrics in an evaluation, but to also explore alternative evaluation methods for more robust and interpretable results.

6 Conclusion

Academic research in the area of recommender systems is challenged by the fact that scholars usually do not have access to a real-world system. This, in turn, leads to a certain over-reliance on offline experiments and abstract, application-independent computational measures. With this work, we aim to contribute to the body of evidence about the effectiveness of certain approaches—in our case sequential models—in practice and in terms of relevant business metrics.

In our future work, we plan to study alternative models from the academic literature without the restrictions from the industry platform requirements. This includes both modern deep learning techniques that have proven to be effective in such tasks, and also simpler, lightweight models that still remain competitive in the field [38] and may have a lower negative environmental impact [43]. Despite possible limitations that were discussed earlier, another potential area for future work will lie in the exploration of other evaluation techniques such as counterfactual estimation. In addition, running an extended A/B test for a longer period of time in the future shall help us obtain a deeper understanding of potential longitudinal effects of the deployed methods and will allow us to observe whether the observed metrics change over time. Last but not least, repeating the described field study in other application domains will help us understand to what extent our observations are generalizable.

References

1. Beel, J., Langer, S.: A comparison of offline evaluations, online evaluations, and user studies in the context of research-paper recommender systems. In: Research and Advanced Technology for Digital Libraries: 19th International Conference on Theory and Practice of Digital Libraries, TPDL 2015. pp. 153–168 (2015)
2. Brovman, Y.M., Jacob, M., Srinivasan, N., Neola, S., Galron, D., Snyder, R., Wang, P.: Optimizing similar item recommendations in a semi-structured marketplace to maximize conversion. In: Proceedings of the 10th ACM Conference on Recommender Systems. pp. 199–202 (2016)
3. Chen, Q., Zhao, H., Li, W., Huang, P., Ou, W.: Behavior sequence transformer for e-commerce recommendation in Alibaba. In: Proceedings of the 1st International Workshop on Deep Learning Practice for High-Dimensional Sparse Data. DLP-KDD '19 (2019)
4. Cremonesi, P., Garzotto, F., Turrin, R.: Investigating the persuasion potential of recommender systems from a quality perspective: An empirical study. Transactions on Interactive Intelligent Systems **2**(2), 11:1–11:41 (2012)
5. Cremonesi, P., Garzotto, F., Turrin, R.: User-centric vs. system-centric evaluation of recommender systems. In: Proceedings INTERACT 2013. vol. 8119, pp. 334–351 (2013)
6. Davidson, J., Liebald, B., Liu, J., Nandy, P., Van Vleet, T., Gargi, U., Gupta, S., He, Y., Lambert, M., Livingston, B., Sampath, D.: The YouTube Video Recommendation System. In: Proceedings of the 4th ACM Conference on Recommender Systems. pp. 293–296 (2010)
7. Dorfman, R.: A formula for the Gini coefficient. The Review of Economics and Statistics pp. 146–149 (1979)
8. Ekstrand, M.D., Harper, F.M., Willemsen, M.C., Konstan, J.A.: User perception of differences in recommender algorithms. In: Proceedings of the 8th ACM Conference on Recommender Systems. pp. 161–168 (2014)

9. Garcin, F., Faltings, B., Donatsch, O., Alazzawi, A., Bruttin, C., Huber, A.: Offline and online evaluation of news recommender systems at swissinfo.ch. In: Proceedings of the 8th ACM Conference on Recommender Systems. pp. 169–176 (2014)
10. Gomez-Uribe, C.A., Hunt, N.: The Netflix recommender system: Algorithms, business value, and innovation. ACM Transactions on Management Information Systems **6**(4), 13:1–13:19 (2015)
11. Herlocker, J.L., Konstan, J.A., Terveen, L.G., Riedl, J.T.: Evaluating collaborative filtering recommender systems. ACM Transactions on Information Systems (TOIS) **22**(1), 5–53 (2004)
12. Hidasi, B., Czapp, Á.T.: Widespread flaws in offline evaluation of recommender systems. In: Proceedings of the 17th ACM Conference on Recommender Systems. pp. 848–855 (2023)
13. Hidasi, B., Karatzoglou, A., Baltrunas, L., Tikk, D.: Session-based recommendations with recurrent neural networks. In: 6th International Conference on Learning Representations (2016)
14. Hu, Y., Koren, Y., Volinsky, C.: Collaborative filtering for implicit feedback datasets. In: Proceedings International Conference on Data Mining (ICDM '08). pp. 263–272 (2008)
15. Jannach, D., Hegelich, K.: A case study on the effectiveness of recommendations in the mobile internet. In: Proceedings of the 3rd ACM Conference on Recommender Systems. pp. 205–208 (2009)
16. Jannach, D., Lerche, L.: Offline performance vs. subjective quality experience: A case study in video game recommendation. In: ACM Symposium on Applied Computing (ACM SAC 2017) (2017)
17. Jannach, D., Lerche, L., Jugovac, M.: Adaptation and evaluation of recommendations for short-term shopping goals. In: Proceedings of the 9th ACM Conference on Recommender Systems. pp. 211–218 (2015)
18. Jannach, D., Lerche, L., Kamehkhosh, I., Jugovac, M.: What recommenders recommend: an analysis of recommendation biases and possible countermeasures. User Modeling and User-Adapted Interaction (UMUAI) **25**(5), 427–491 (2015)
19. Jannach, D., Ludewig, M., Lerche, L.: Session-based item recommendation in e-commerce: On short-term intents, reminders, trends, and discounts. User-Modeling and User-Adapted Interaction (UMUAI) **27**(3–5), 351–392 (2017)
20. Jannach, D., Quadrana, M., Cremonesi, P.: Session-based recommendation. In: Ricci, F., Shapira, B., Rokach, L. (eds.) Recommender Systems Handbook. Springer US (2021)
21. Jannach, D., Zanker, M.: Impact and value of recommender systems. In: Ricci, F., Shapira, B., Rokach, L. (eds.) Recommender Systems Handbook. Springer US (2021)
22. Jeunen, O., Potapov, I., Ustimenko, A.: On (normalised) discounted cumulative gain as an off-policy evaluation metric for top-n recommendation. In: Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. pp. 1222–1233 (2024)
23. Joachims, T., Swaminathan, A.: Counterfactual evaluation and learning for search, recommendation and ad placement. In: Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 1199–1201 (2016)
24. Kamehkhosh, I., Jannach, D.: User Perception of Next-Track Music Recommendations. In: Proceedings of the 25th Conference on User Modeling, Adaptation and Personalization. pp. 113–121. UMAP '17 (2017)

25. Kang, W.C., McAuley, J.: Self-attentive sequential recommendation. In: 2018 IEEE international Conference on Data Mining (ICDM). pp. 197–206 (2018)
26. Kersbergen, B., Sprangers, O., Schelter, S.: Serenade - low-latency session-based recommendation in e-commerce at scale. In: SIGMOD '22: International Conference on Management of Data. pp. 150–159 (2022)
27. Kim, Y., Kim, K., Park, C., Yu, H.: Sequential and diverse recommendation with long tail. In: Proceedings of the 28th International Joint Conference on Artificial Intelligence. p. 2740–2746. IJCAI'19 (2019)
28. Kouki, P., Fountalis, I., Vasiloglou, N., Cui, X., Liberty, E., Al Jadda, K.: From the lab to production: A case study of session-based recommendations in the home-improvement domain. In: Proceedings of the 14th ACM Conference on Recommender Systems. pp. 140–149. RecSys '20 (2020)
29. Lam, X.N., Vu, T., Le, T.D., Duong, A.D.: Addressing cold-start problem in recommendation systems. In: Proceedings of the 2nd International Conference on Ubiquitous Information Management and Communication. pp. 208–211 (2008)
30. Liu, Q., Zeng, Y., Mokhosi, R., Zhang, H.: STAMP: short-term attention/memory priority model for session-based recommendation. In: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. pp. 1831–1839 (2018)
31. Ludewig, M., Jannach, D.: Evaluation of session-based recommendation algorithms. User Modeling and User-Adapted Interaction (UMUAI **28**, 331–390 (2018)
32. Ludewig, M., Jannach, D.: User-centric evaluation of session-based recommendations for an automated radio station. In: Proceedings of the 13th ACM Conference on Recommender Systems. Copenhagen (2019)
33. Ludmann, C.A.: Recommending news articles in the CLEF news recommendation evaluation lab with the data stream management system odysseus. In: Working Notes of CLEF 2017 - Conference and Labs of the Evaluation (2017)
34. Maksai, A., Garcin, F., Faltings, B.: Predicting online performance of news recommender systems through richer evaluation metrics. In: Proceedings of the 9th ACM Conference on Recommender Systems. pp. 179–186 (2015)
35. McNee, S.M., Albert, I., Cosley, D., Gopalkrishnan, P., Lam, S.K., Rashid, A.M., Konstan, J.A., Riedl, J.: On the recommending of citations for research papers. In: Proceedings of the 2002 ACM Conference on Computer Supported Cooperative Work. pp. 116–125. CSCW '02 (2002)
36. Potter, M., Liu, H., Lala, Y., Loanzon, C., Sun, Y.: GRU4RecBE: a hybrid session-based movie recommendation system (student abstract). In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 36, pp. 13029–13030 (2022)
37. Quadrana, M., Cremonesi, P., Jannach, D.: Sequence-aware recommender systems. ACM Computing Surveys (CSUR) **51**(4), 1–36 (2018)
38. Rendle, S., Freudenthaler, C., Schmidt-Thieme, L.: Factorizing personalized markov chains for next-basket recommendation. In: Proceedings of the 19th International Conference on World Wide Web. pp. 811–820 (2010)
39. Steck, H., Baltrunas, L., Elahi, E., Liang, D., Raimond, Y., Basilico, J.: Deep Learning for Recommender Systems: A Netflix Case Study. AI Magazine **42**(3), 7–18 (2021)
40. Sun, F., Liu, J., Wu, J., Pei, C., Lin, X., Ou, W., Jiang, P.: BERT4Rec: Sequential recommendation with bidirectional encoder representations from transformer. In: Proceedings of the 28th ACM International Conference on Information and Knowledge Management. pp. 1441–1450 (2019)

41. Takács, G., Pilászy, I., Tikk, D.: Applications of the conjugate gradient method for implicit feedback collaborative filtering. In: Proceedings of the 5th ACM Conference on Recommender Systems (2011)
42. Takács, G., Tikk, D.: Alternating least squares for personalized ranking. In: Proceedings of the 6th ACM Conference on Recommender Systems. pp. 83–90 (2012)
43. Vente, T., Wegmeth, L., Said, A., Beel, J.: From clicks to carbon: The environmental toll of recommender systems. In: Proceedings of the 18th ACM Conference on Recommender Systems. pp. 580–590 (2024)
44. Wang, S., Cao, L., Wang, Y., Sheng, Q.Z., Orgun, M.A., Lian, D.: A survey on session-based recommender systems. ACM Computing Surveys (CSUR) **54**(7) (2021)
45. Zangerle, E., Bauer, C.: Evaluating recommender systems: Survey and framework. ACM Computing Surveys **55**(8) (2022)