Broadening the Scope: Evaluating the Potential of Recommender Systems beyond prioritizing Accuracy

Vincenzo Paparella*
Politecnico di Bari
Bari, Italy
vincenzo.paparella@poliba.it

Dario Di Palma*
Politecnico di Bari
Bari, Italy
d.dipalma2@phd.poliba.it

Vito Walter Anelli
Politecnico di Bari
Bari, Italy
vitowalter.anelli@poliba.it

Tommaso Di Noia
Politecnico di Bari
Bari, Italy
tommaso.dinoia@poliba.it

ABSTRACT

Although beyond-accuracy metrics have gained attention in the last decade, the accuracy of recommendations is still considered the gold standard to evaluate Recommender Systems (RSs). This approach prioritizes the accuracy of recommendations, neglecting the quality of suggestions to enhance user needs, such as diversity and novelty, as well as trustworthiness regulations in RSs for user and provider fairness. As a result, single metrics determine the success of RSs, but this approach fails to consider other criteria simultaneously. A downside of this method is that the most accurate model configuration may not excel in addressing the remaining criteria. This study seeks to broaden RS evaluation by introducing a multi-objective evaluation that considers all model configurations simultaneously under several perspectives. To achieve this, several hyper-parameter configurations of an RS model are trained, and the Pareto-optimal ones are retrieved. The ***Quality Indicators*** **(QI)** of Pareto frontiers, which are gaining interest in Multi-Objective Optimization research, are adapted to RSs. QI enables evaluating the model's performance by considering various configurations and giving the same importance to each metric. The experiments show that this multi-objective evaluation overturns the ranking of performance among RSs, paving the way to revisit the evaluation approaches of the RecSys research community. We release codes and datasets in the following GitHub repository: https://github.com/sisinflab/RecMOE.

CCS CONCEPTS

• **Information Systems** → **Recommender systems**; • **Computing methodologies** → Pareto Optimality.

KEYWORDS

Recommender Systems, Multi-Objective Evaluation, Pareto optimality

ACM Reference Format:
Vincenzo Paparella, Dario Di Palma, Vito Walter Anelli, and Tommaso Di Noia. 2023. Broadening the Scope: Evaluating the Potential of Recommender Systems beyond prioritizing Accuracy. In *Seventeenth ACM Conference on Recommender Systems (RecSys '23), September 18–22, 2023, Singapore, Singapore.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3604915.3610649

1 INTRODUCTION AND MOTIVATION

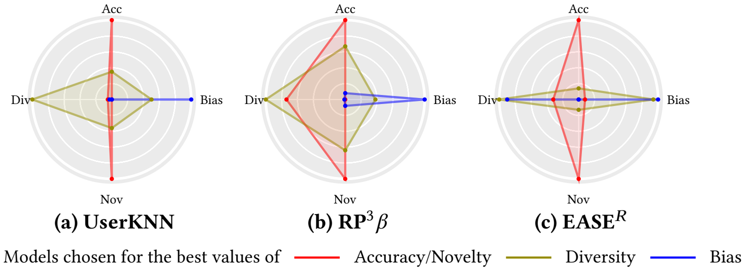
The success of Recommender Systems (RSs) is often measured by its ability to accurately predict a user's preferences and suggest relevant items. However, other beyond-accuracy metrics have been proposed to capture different aspects of recommendation quality, such as diversity and novelty of suggestions [21, 23, 26], and fairness issues [7, 16, 31].

While beyond-accuracy metrics have gained momentum in the RecSys research community, accuracy of suggestions is still consistently prioritized over the other facets of recommendation [4, 6]. The common practice is to select the best model solely based on the accuracy metrics (e.g., nDCG, Recall, or Precision), which limits the consideration of performance on beyond-accuracy metrics. Consequently, the best model in terms of accuracy may not guarantee the best performance in terms of diversity, novelty, or fairness, and vice versa. This limitation in choosing the best models may result in a lack of information on the actual behavior of RS models across multiple perspectives of recommendation. In this regard, we provide a motivating example by training 32 hyper-parameter settings of three baselines (i.e., EASE^R [24], RP$^3\beta$ [19], and UserKNN [20]) on the Goodreads dataset[1]. Figure 1 shows the min-max normalized values of recommendation algorithm performance by selecting the best hyper-parameter settings for each baseline. We do this based on the best values of various metrics representing accuracy (nDCG), novelty (EPC) [26], diversity (1 - Gini coefficient) [14], and algorithmic bias (APLT) [1] evaluation perspectives. When selecting the model based on the highest value of a given metric, a larger shape area on the resulting graph indicates reasonably high values of the other metrics. As expected, we find that the selection strategy for the best model tremendously impact the other metrics. Namely, selecting the best hyper-parameter setting according to accuracy guarantees the best value of novelty, but leads to sub-optimal value of diversity and worse value of algorithmic bias, and vice versa.

*Corresponding authors.

[1] More details on the experimental settings will be provided in Section 3.

RecSys '23, September 18–22, 2023, Singapore, Singapore
© 2023 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-0241-9/23/09.
https://doi.org/10.1145/3604915.3610649



(a) UserKNN **(b) RP$^3\beta$** **(c) EASE^R**

Figure 1: Kiviat diagrams indicating the performance of the models on the Goodreads dataset. The models are selected according to different metrics for each objective (i.e., Accuracy/Novelty, Diversity, and Bias). Higher means better.

Then, assessing a model's performance for each metric, for example after selecting it based solely on accuracy, results in a lack of knowledge about the potential of the model on beyond-accuracy metrics. Hence, the need of a *multi-objective evaluation* emerges to *simultaneously* assess the models' performance on several criteria, even though the training of such models could still aim to maximize the accuracy of recommendation (e.g., to choose the best iteration, or trigger a stopping condition in the training phase).

To address this problem of multi-objective evaluation, we exploit the definition of Pareto optimality from the Multi-Objective Optimization (MOO) theory [18]. Given a set of objectives to maximize, we define a specific hyper-parameter setting of a model as a Pareto-optimal solution if there is no other setting that improves at least one objective function without hurting another one. The set of such Pareto-optimal configurations composes the so-called Pareto frontier [28]. An approach to consider simultaneously more metrics in the evaluation would be to select a solution from the Pareto frontier through well-known methods (e.g., hypervolume [30]). However, evaluating a specific configuration of a model only provides information on that particular setting and fails to provide insights into the overall potential of the model. Therefore, to enhance the multi-objective evaluation of RSs, we need to assess the entire set of Pareto-optimal configurations of a model. Simply visualizing the Pareto frontier only enables qualitative analysis, being challenging when multiple objectives are involved. We propose to introduce in RSs research the ***Quality Indicators*** , previously adopted in the literature of MOO [15], which are designed to evaluate Pareto frontiers by providing a real number to quantify and rank the performance of a model corresponding to a Pareto frontier under different perspectives. To the best of our knowledge, QIs have already been exploited to evaluate Pareto frontiers — mostly their relative dominance — generated by evolutionary algorithm [8, 10, 11] applied in the context of Multi-Objective RSs [29, 30]. In contrast, we aim to use them to offer insights into unexplored aspects of traditional RSs. In detail, the contributions of our work are:

- We experimentally show the negative impact of prioritizing recommendation accuracy over other important metrics and motivate the need of a multi-objective evaluation of RSs models. The results emphasize the importance of a more comprehensive evaluation approach to ensure a thorough understanding of RS behavior across multiple dimensions.
- We train 32 hyper-parameter settings of 5 state-of-the-art recommendation models using 3 public datasets. We compute the Pareto frontier in two multi-objective scenarios to provide a exhaustive evaluation of the recommendation models.
- To enhance the multi-objective evaluation of RSs, we evaluate various models under different scenarios simultaneously by utilizing the ***Quality Indicators*** of Pareto frontiers to enable an even more comprehensive analysis of RSs.

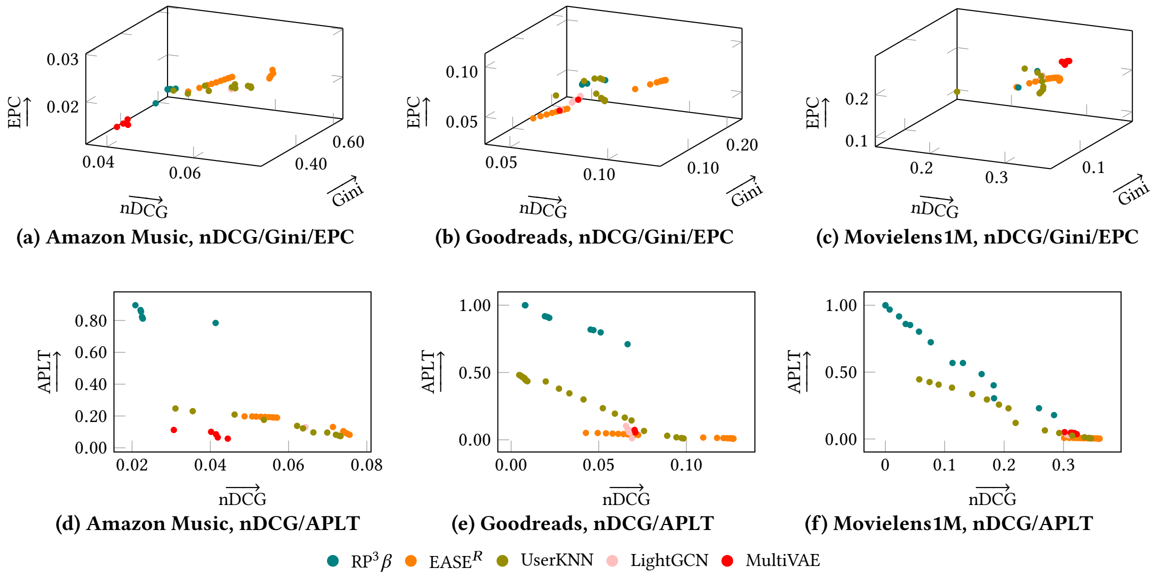
2 QUALITY INDICATORS

In this Section, we present the Quality Indicators (QIs) to assess the Pareto frontiers corresponding to an RS model. They can be classified according to the quality they assess.

Spread QI. The QIs for Spread indicate the range of the Pareto-optimal solutions on the Pareto frontier. For our study, we use the Maximum Spread ($\mathcal{MS}$) [32]. Specifically, this spread indicator measures the range of a Pareto frontier by considering the maximum extent of each objective. Given the Pareto-optimal solutions set A and the number of objectives m, $\mathcal{MS}$ is defined as $\mathcal{MS}(A) = \sqrt{\sum_{j=1}^{m} \max_{a,a' \in A}(a_j - a'_j)^2}$, where a and a' are solutions belonging to A. The higher the value, the better the extensiveness of the curve.

Uniformity QI. The uniformity of a Pareto frontier provides information about the distribution of the solutions. A higher uniformity of the curve denotes that the solutions are less dispersed, while a low uniformity indicates more diversity within the set. In the case of RSs, having low uniformity leads to a wide range of options for decision-makers. Specifically, we employ the Spacing metric ($\mathcal{SP}$) [22] that measures the variation in the Manhattan distances between the Pareto-optimal solutions. Given the N Pareto-optimal solutions $a_i \in A$ and the number of objectives m, $\mathcal{SP}$ is defined as $\mathcal{SP}(A) = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(\bar{d} - d_1(a_i, A/a_i))^2}$ with $d_1(a_i, A/a_i) = \min_{a \in A/a_i} \sum_{j=1}^{m} \left|a_{ij} - a_j\right|$, where $\bar{d}$ is the mean of all the Manhattan distances $d_1(a_1, A/a_1)), \ldots, d_1(a_N, A/a_N))$ and a_{ij} represents the j-th objective of the solution a_i. The lower the value, the more concentrated the solutions are on the Pareto frontier. However, an $\mathcal{SP} = 0$ indicates that all the solutions could be equidistant. The interpretation of $\mathcal{SP}$ is strictly related to $\mathcal{MS}$.

Cardinality QI. Given K generic solutions belonging to the set B, the QIs for cardinality determine the proportion of Pareto-optimal solutions in this set. Specifically, the Error Ratio ($\mathcal{ER}$) [25] is defined as $\mathcal{ER}(B) = \frac{\sum_{b \in B} e(b)}{K}$ with $e(b) = 1$ if b is a Pareto-optimal solution, 0 otherwise. A higher $\mathcal{ER}$ value indicates greater Pareto-optimal solutions in the set B.



(a) Amazon Music, nDCG/Gini/EPC **(b) Goodreads, nDCG/Gini/EPC** **(c) Movielens1M, nDCG/Gini/EPC**

(d) Amazon Music, nDCG/APLT **(e) Goodreads, nDCG/APLT** **(f) Movielens1M, nDCG/APLT**

Figure 2: Pareto optimal solutions plots for Amazon Music, Goodreads, and MovieLens1M. The first row refers to the nDCG/Gini/EPC scenario, and the second row refers to the nDCG/APLT scenario. The arrows indicate the optimal directions.

All quality aspects QI. The QIs included in this category provide insights into the spread, uniformity, and cardinality of the Pareto frontiers simultaneously. Among them, the Hypervolume ($\mathcal{HV}$) [33] is a volume-based QI that measures the volume of the objective function space dominated by the Pareto frontier. Given the Pareto-optimal solutions $a \in A$ and a reference point r, $\mathcal{HV}$ is defined as $\mathcal{HV}(A) = \lambda(\bigcup_{a \in A}\{x \mid a \prec x \prec r\})$, where λ denotes the Lebesgue measure. The larger the hypervolume, the better the solution set is.

3 EXPERIMENTS

Given a set of multiple metrics to assess simultaneously, we aim to answer the following research questions: **RQ1**: *To what extent can the models provide Pareto-optimal configurations? Are these configurations uniformly distributed, or are they dispersed enhancing diverse solutions to the trade-off?* **RQ2**: *Which model has the Pareto frontier that simultaneously offers better solutions on multiple metrics?*

3.1 Experimental Setup

Datasets. We select three different datasets to cover several domains. Specifically, we use *Amazon Music* (music domain), *Goodreads* [27] (book domain), and *Movielens1M* [12] (movie domain). Regarding Goodreads (18892 users, 25475 items, 1378033 interactions, 0.99 sparsity) and Movielens1M (6040 users, 3706 items, 1000209 interactions, 0.95 sparsity), we do not apply any pre-processing step, while we obtain a pre-processed version of the Amazon Music dataset from work by Anelli et al. [3] (14354 users, 10027 items, 145523 interactions, 0.99 sparsity).

Baselines and Hyper-parameters Settings Exploration. We train five recommendation algorithms, i.e., EASE^R [24], MultiVAE [17], LightGCN [13], RP$^3\beta$ [19], and UserKNN [20]. Specifically, we train 32 hyper-parameter values combinations of each model by exploiting the Elliot framework [2]. We define the set of hyper-parameters values for these baselines from previous works [4, 5]. We provide complete information on the explored values in the GitHub repository. We set nDCG@10 as the optimization target. MultiVAE and LightGCN are trained with a batch size of 256 and 300 epochs by applying the early stopping strategy with patience of 10.

Metrics. We assess the baselines' performance under several perspectives. We compute nDCG, Precision, and Recall for the accuracy of recommendations. From the final user point of view, we evaluate the diversity (with Gini index [14] and Item Coverage) and novelty (with EPC and EFD [26]). Finally, we measure the popularity bias of the recommendations with APLT [1] – the greater, the better – and ARP [14] – the less, the better. All these metrics refer to cutoff 10.

Multi-Objective Evaluation Methodology. We clarify how we obtain the Pareto frontiers corresponding to each baseline to evaluate them through the quality indicators described in Section 2. Given the experimental setup described above, we can identify a subset of the computed metrics to compose a multi-dimensional objective function space. Each single hyper-parameters configuration of a model represents a solution in this space since we have computed their performance values regarding such metrics. As a result, we obtain 32 points in the objective function space for each baseline. Among these points, we can identify the Pareto-optimal configurations, which lay on the Pareto frontier. Consequently, given an objective function space designated by a set of metrics, we gather five Pareto frontiers, each corresponding to one trained baseline. Once the Pareto-optimal solutions composing the Pareto frontiers are identified, we can exploit the QIs to evaluate the Pareto frontiers of the models.

We carry out the multi-objective evaluation by identifying two different evaluation scenarios. On the one hand, we focus on user-centered objectives (accuracy, diversity, and novelty of recommendations). This scenario leads to a three-dimensional space in which the axes are nDCG, Gini index, and EPC. On the other hand, we

Table 1: Classical analysis of the baselines' results in terms of Accuracy, Diversity, Novelty, and Bias of recommendations. The arrows indicates the descending or ascending order for the best solution. Best values are in bold. Second best values are underlined.

	Model	nDCG↑	Recall↑	Precision↑	Gini↑	ItemCV↑	EPC↑	EFD↑	APLT↑	ARP↓
Amazon Music	EASE^R	**0.07560**	**0.09481**	**0.02049**	0.25846	8891	**0.02863**	**0.34370**	0.08196	37.6760
	UserKNN	0.07329	0.09424	0.02004	0.21426	8361	0.02741	0.32669	0.07363	42.7840
	MultiVAE	0.04446	0.06264	0.01269	0.22379	6556	0.01606	0.19478	0.05773	28.4834
	LightGCN	0.06433	0.08632	0.01797	0.33387	**9121**	0.02355	0.28666	0.12980	28.1607
	RP$^3\beta$	0.04136	0.05070	0.01071	**0.44327**	8973	0.01521	0.20087	**0.78420**	**4.46494**
Goodreads	EASE^R	**0.12685**	**0.08278**	**0.09680**	0.04144	6842	**0.10599**	**1.23522**	0.00882	475.874
	UserKNN	0.09842	0.06533	0.07416	0.02873	6434	0.08117	0.92929	0.01021	587.527
	MultiVAE	0.07090	0.04812	0.05718	0.05126	7387	0.05974	0.69948	0.05533	443.142
	LightGCN	0.06896	0.04835	0.05352	0.06434	7729	0.05722	0.68752	0.01176	356.040
	RP$^3\beta$	0.06645	0.04177	0.05066	**0.19076**	**14941**	0.05759	0.78194	**0.71016**	**64.3545**
Movielens1M	EASE^R	**0.36075**	**0.15574**	**0.32462**	0.06152	980	**0.27472**	**3.22977**	0.00260	1198.44
	UserKNN	0.34603	0.14980	0.31189	0.04556	920	0.25320	3.01901	0.00462	1305.30
	MultiVAE	0.32223	0.14189	0.29147	**0.12550**	**1836**	0.25631	3.00231	0.03657	1002.73
	LightGCN	0.31087	0.13204	0.28113	0.09899	1481	0.24170	2.84602	0.02806	1046.17
	RP$^3\beta$	0.28403	0.12287	0.27017	0.09266	1588	0.21789	2.58115	**0.17851**	**961.877**

compare the accuracy of recommendations against the algorithmic bias, by obtaining a two-dimensional objective function space (nDCG vs. APLT). Figure 2 depicts the Pareto frontiers of the models trained on each datasets for the two evaluation scenarios.

3.2 Results and Discussion

To commence the experimental assessment, we establish a benchmark for the upcoming investigation. In detail, a preliminary analysis of the baselines' performance is conducted by reporting the results of the best configurations according to the values of nDCG@10 in Table 1. This analysis serves as context and motivates the subsequent exploration where QIs of the Pareto frontiers are utilized to answer the research questions (Table 2).

A "traditional" analysis of recommendation performance. The results in Table 1 corroborate the recent literature findings [3, 9]. For the three datasets, EASE^R and UserKNN are the models providing the most accurate recommendations. Observing the novelty metrics, the accuracy and novelty of recommendations exhibit a positive correlation. However, we arrive at very different conclusions by examining the other beyond-accuracy metrics. On the one hand, concerning the diversity of recommendations, the remaining models (LightGCN, MultiVAE, RP$^3\beta$) generally perform better than EASE^R and UserKNN across all datasets. On the other hand, RP$^3\beta$ consistently outperforms its competitors in addressing the popularity bias. This peculiar performance puzzle does not offer insight into the general behaviour of the model or whether other instances of it follow a similar performance trend. To unravel this puzzle, we shift to a multi-objective evaluation-based analysis aimed at assessing the recommendation performance under several criteria simultaneously.

Distribution of Pareto-optimal configurations. To answer RQ1, we examine the values of **Error Ratio** ($\mathcal{ER}$), **Maximum Spread** ($\mathcal{MS}$), and **Spacing metric** ($\mathcal{SP}$). Different scenarios may arise when examining the behaviour of a model. Firstly, when the model yields higher $\mathcal{ER}$, $\mathcal{MS}$, and $\mathcal{SP}$ values, it suggests that the model's configurations are widely spread and varied, implying that it can provide multiple solutions on the Pareto frontier. Secondly, suppose the model exhibits higher $\mathcal{ER}$ and $\mathcal{MS}$ values but lower $\mathcal{SP}$ values. In that case, it indicates that the model's settings are dispersed but concentrated in certain areas of the objective function space. This behaviour could result in fewer solutions on the Pareto frontier. Thirdly, if the model has higher values of $\mathcal{ER}$ and lower values of $\mathcal{MS}$ and $\mathcal{SP}$, it implies that the model can offer various Pareto-optimal settings, which are all concentrated in the same area of the objective function space. Finally, a low number of Pareto-optimal configurations can indicate some issues with the solutions' characteristics, regardless of the $\mathcal{MS}$ and $\mathcal{SP}$ values.

Our investigation begins with the nDCG/APLT metrics for the Movielens1M dataset (as shown in Table 2), with Figure 2f illustrating the results for a better understanding. Within this context, RP$^3\beta$ provides a broad range of acceptable solutions ($\mathcal{ER}$=0.47) with a wide dispersion (highest value of $\mathcal{MS}$), and the solutions are dispersed along the entire Pareto frontier (highest value of $\mathcal{SP}$). Therefore, RP$^3\beta$ offers various solutions for an optimal trade-off between recommendation accuracy and algorithmic bias. UserKNN exhibits similar behaviour, with the second highest values for $\mathcal{ER}$, $\mathcal{MS}$, and $\mathcal{SP}$ (0.5, 0.53, and 0.02, respectively). In contrast, EASE^R offers a limited choice, featuring a not extensive and highly concentrated frontier (low values of $\mathcal{MS}$ and $\mathcal{SP}$), despite having numerous solutions on the frontier (highest value of $\mathcal{ER}$). Finally, MultiVAE and LightGCN present a limited number of Pareto-optimal configurations (lowest $\mathcal{ER}$ values), which influence the quality of their Pareto frontiers regarding range and spacing. As illustrated in Figure 2f, QIs provide an adequate and quantitative depiction of the models' behaviour. We can then extend our scrutiny to the remaining datasets. UserKNN, RP$^3\beta$, LightGCN, and MultiVAE maintain their respective performance across the Amazon Music (Figure 2d) and Goodreads (Figure 2e) datasets. Upon examination of Table 2, for these datasets, EASE^R demonstrates higher $\mathcal{MS}$ values than the one for Movielens1M. The corresponding Pareto frontiers are broader (higher $\mathcal{MS}$), but the solutions are concentrated into two well-separated clusters (lower $\mathcal{SP}$). This outcome emphasizes that EASE^R leaves the intermediate area between these clusters uncovered, being incapable of offering a balanced optimal trade-off between the two objectives. Let us focus on the user-centric scenario, where our objectives include nDCG/Gini/EPC, as shown in Figures 2a, 2b, and 2c. It is worth noting that UserKNN has

Table 2: Quality Indicators of the Pareto frontiers results for the identified scenarios. The arrow indicates the descending or ascending order for the best solution. $\mathcal{SP}$ has no specific order of solutions, since its interpretation is strictly connected with the MS indicator. C counts how many solutions lay on the Pareto frontier.

	Model	Objectives									
		Accuracy / Novelty / Diversity					Accuracy / Bias				
		$\mathcal{HV}\uparrow$	$\mathcal{ER}\uparrow$	$\mathcal{MS}\uparrow$	$\mathcal{SP}$	$C\uparrow$	$\mathcal{HV}\uparrow$	$\mathcal{ER}\uparrow$	$\mathcal{MS}\uparrow$	$\mathcal{SP}$	$C\uparrow$
Amazon Music	EASE^R	**0.00095**	**0.46875**	0.24986	0.01476	**15**	0.01355	**0.43750**	0.11886	0.00669	**14**
	UserKNN	0.00082	0.34375	**0.29452**	0.00496	11	0.01448	0.34375	**0.17871**	0.00980	11
	LightGCN	0.00051	0.06250	0.01335	0.00000	2	0.00835	0.03125	0.00000	0.00000	1
	MultiVAE	0.00022	0.12500	0.09656	0.01738	4	0.00468	0.15625	0.05629	0.00351	5
	RP$^3\beta$	0.00039	0.18750	0.20753	0.05888	6	**0.03489**	0.21875	0.11336	0.01173	7
Goodreads	EASE^R	0.00074	**0.59375**	0.09910	0.00227	**19**	0.00439	0.65625	0.09433	0.00214	21
	UserKNN	**0.00110**	0.31250	**0.19889**	0.01287	10	0.02267	**0.71875**	**0.48042**	0.01471	**23**
	LightGCN	0.00051	0.18750	0.06743	0.00783	6	0.00696	0.18750	0.09180	0.01536	6
	MultiVAE	0.00043	0.06250	0.05022	0.00000	2	0.00521	0.06250	0.01827	0.00000	2
	RP$^3\beta$	0.00083	0.12500	0.05584	0.01213	4	**0.05544**	0.28125	0.29529	0.02657	9
Movielens1M	EASE^R	0.00865	**0.68750**	0.09833	0.00446	**22**	0.00281	**0.65625**	0.06001	0.00196	**21**
	UserKNN	**0.01296**	0.28125	**0.30929**	0.03641	9	0.08191	0.50000	0.52723	0.01810	16
	LightGCN	0.00807	0.18750	0.01012	0.00287	6	0.00974	0.15625	0.00617	0.00181	5
	MultiVAE	0.01216	0.21875	0.03419	0.00427	7	0.01639	0.18750	0.02528	0.00293	6
	RP$^3\beta$	0.00839	0.06250	0.03796	0.00000	2	**0.14014**	0.46875	**0.86913**	0.03228	15

proven its proficiency in generating several well-diversified hyperparameter configurations across all datasets. This model boasts the best or second-best values of $\mathcal{ER}$ and $\mathcal{MS}$, along with high $\mathcal{SP}$ values, particularly for the Goodreads and Movielens1M datasets. However, LightGCN and MultiVAE exhibit subpar performance considering the number of Pareto-optimal configurations and their distribution, while EASE^R boasts a wide Pareto frontier but is confined to specific regions, failing to cover the central (and more balanced) area. In contrast, RP$^3\beta$ behaves differently from the previous scenario, providing fewer solutions on the Pareto frontier for the accuracy/diversity/novelty trade-off.

In summary, in response to RQ1, we can assert that UserKNN provides several diversified optimal solutions that effectively balance the two scenarios. Conversely, EASE^R, while offering numerous optimal solutions, tends to provide solutions that are concentrated and clustered. RP$^3\beta$ is effective in balancing accuracy and bias but struggles in disentangling user-centred metrics. Finally, it is worth noting that LightGCN and MultiVAE yield inferior performance in this regard.

Performance on all quality metrics. In response to RQ2, we can utilize the Hypervolume ($\mathcal{HV}$) measure. $\mathcal{HV}$ evaluates the performance of models from multiple objectives simultaneously, as shown in Table 2. By considering the cardinality and dispersion of the Pareto-optimal solutions and the dominance among the Pareto frontiers, $\mathcal{HV}$ provides us with valuable insights. The higher the volume or area under the frontier, the greater the $\mathcal{HV}$. The results show that UserKNN outperforms the other models by achieving the best or second-best values of $\mathcal{HV}$ for all datasets and scenarios. This result indicates that UserKNN generates an extensive and diversified Pareto frontier while performing well across all metrics. While EASE^R has the highest value of $\mathcal{HV}$ for the Amazon Music dataset in the user-centred scenario, it does not dominate or get dominated in the remaining cases. This result highlights the model's limited reliance on accounting for multiple metrics. LightGCN shows no distinctive trends, while MultiVAE's $\mathcal{HV}$ decreases when dealing with sparser datasets. RP$^3\beta$ confirms its capability in managing the nDCG/APLT trade-off by achieving the highest values of $\mathcal{HV}$ and visual dominance of its Pareto frontiers against the others in Figures 2d, 2e, and 2f.

In summary, to answer RQ2, our findings indicate that in terms of multi-objective evaluation, UserKNN is the superior model overall. However, when considering the accuracy/bias trade-off, RP$^3\beta$ emerges as a noteworthy contender.

Final observations. In evaluating recommendation systems, accuracy is typically given top priority. Thus, in our initial analysis, EASE^R emerged as the frontrunner due to its impressive accuracy. However, when subjected to our multi-objective evaluation, EASE^R was often outperformed by other models. UserKNN, on the other hand, demonstrated superior performance across diverse metrics. Surprisingly, RP$^3\beta$ ranked the lowest in terms of accuracy but proved to be particularly effective in finding a balance between nDCG and APLT (bias) performance. These findings challenge the traditional ranking of recommendation systems, paving the way for new research in model evaluation.

4 CONCLUSION AND FUTURE WORK

In our study, we utilize Quality Indicators of Pareto frontiers to conduct a multi-objective evaluation of Recommender Systems (RSs). Our experiments aim to assess RSs with three (Accuracy / Novelty / Diversity) and two (Accuracy / Bias) conflicting objectives. While EASE^R exhibits superior accuracy, our evaluation has unveiled a new ranking of the baselines. UserKNN stands out as it provides several diverse solutions which perform well in both multi-objective scenarios. Additionally, RP$^3\beta$ proved to be highly effective in the accuracy/algorithmic bias scenario. Moving forward, we plan to extend this evaluation to other baselines. Furthermore, we intend to leverage the Pareto frontiers' quality indicators to evaluate the impact of the models' hyper-parameters in a multi-objective scenario.

Acknowledgements. This work was partially supported by the following projects: Secure Safe Apulia, Casa delle Tecnologie Emergenti Comune di Matera, LUTECH DIGITALE 4.0, OVS Fashion Retail Reloaded, CT_FINCONS_III, KOINÈ, MOST - Centro Nazionale per la Mobilità Sostenibile.

REFERENCES

[1] Himan Abdollahpouri, Robin Burke, and Bamshad Mobasher. 2019. Managing Popularity Bias in Recommender Systems with Personalized Re-Ranking. In *Proceedings of the Thirty-Second International Florida Artificial Intelligence Research Society Conference, Sarasota, Florida, USA, May 19-22 2019*, Roman Barták and Keith W. Brawner (Eds.). AAAI Press, 413–418. https://aaai.org/ocs/index.php/FLAIRS/FLAIRS19/paper/view/18199

[2] Vito Walter Anelli, Alejandro Bellogín, Antonio Ferrara, Daniele Malitesta, Felice Antonio Merra, Claudio Pomo, Francesco Maria Donini, and Tommaso Di Noia. 2021. Elliot: A Comprehensive and Rigorous Framework for Reproducible Recommender Systems Evaluation. In *SIGIR '21: The 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, Virtual Event, Canada, July 11-15, 2021*, Fernando Diaz, Chirag Shah, Torsten Suel, Pablo Castells, Rosie Jones, and Tetsuya Sakai (Eds.). ACM, 2405–2414. https://doi.org/10.1145/3404835.3463245

[3] Vito Walter Anelli, Alejandro Bellogín, Tommaso Di Noia, Dietmar Jannach, and Claudio Pomo. 2022. Top-N Recommendation Algorithms: A Quest for the State-of-the-Art. In *UMAP '22: 30th ACM Conference on User Modeling, Adaptation and Personalization, Barcelona, Spain, July 4 - 7, 2022*, Alejandro Bellogín, Ludovico Boratto, Olga C. Santos, Liliana Ardissono, and Bart P. Knijnenburg (Eds.). ACM, 121–131. https://doi.org/10.1145/3503252.3531292

[4] Vito Walter Anelli, Alejandro Bellogín, Tommaso Di Noia, and Claudio Pomo. 2021. Reenvisioning the comparison between Neural Collaborative Filtering and Matrix Factorization. In *RecSys '21: Fifteenth ACM Conference on Recommender Systems, Amsterdam, The Netherlands, 27 September 2021 - 1 October 2021*, Humberto Jesús Corona Pampín, Martha A. Larson, Martijn C. Willemsen, Joseph A. Konstan, Julian J. McAuley, Jean Garcia-Gathright, Bouke Huurnink, and Even Oldridge (Eds.). ACM, 521–529. https://doi.org/10.1145/3460231.3475944

[5] Vito Walter Anelli, Yashar Deldjoo, Tommaso Di Noia, Daniele Malitesta, Vincenzo Paparella, and Claudio Pomo. 2023. Auditing Consumer- and Producer-Fairness in Graph Collaborative Filtering. In *Advances in Information Retrieval - 45th European Conference on Information Retrieval, ECIR 2023, Dublin, Ireland, April 2-6, 2023, Proceedings, Part I (Lecture Notes in Computer Science, Vol. 13980)*, Jaap Kamps, Lorraine Goeuriot, Fabio Crestani, Maria Maistro, Hideo Joho, Brian Davis, Cathal Gurrin, Udo Kruschwitz, and Annalina Caputo (Eds.). Springer, 33–48. https://doi.org/10.1007/978-3-031-28244-7_3

[6] Vito Walter Anelli, Tommaso Di Noia, Eugenio Di Sciascio, Claudio Pomo, and Azzurra Ragone. 2019. On the discriminative power of hyper-parameters in cross-validation and how to choose them. In *Proceedings of the 13th ACM Conference on Recommender Systems, RecSys 2019, Copenhagen, Denmark, September 16-20, 2019*, Toine Bogers, Alan Said, Peter Brusilovsky, and Domonkos Tikk (Eds.). ACM, 447–451. https://doi.org/10.1145/3298689.3347010

[7] Ludovico Boratto, Gianni Fenu, and Mirko Marras. 2021. Interplay between upsampling and regularization for provider fairness in recommender systems. *User Model. User Adapt. Interact.* 31, 3 (2021), 421–455. https://doi.org/10.1007/s11257-021-09294-8

[8] Zheng-Yi Chai, Ya-Lun Li, and Sifeng Zhu. 2021. P-MOIA-RS: a multi-objective optimization and decision-making algorithm for recommendation systems. *J. Ambient Intell. Humaniz. Comput.* 12, 1 (2021), 443–454. https://doi.org/10.1007/s12652-020-01997-x

[9] Maurizio Ferrari Dacrema, Simone Boglio, Paolo Cremonesi, and Dietmar Jannach. 2021. A Troubling Analysis of Reproducibility and Progress in Recommender Systems Research. *ACM Trans. Inf. Syst.* 39, 2 (2021), 20:1–20:49. https://doi.org/10.1145/3434185

[10] Bingrui Geng, Lingling Li, Licheng Jiao, Maoguo Gong, Qing Cai, and Yue Wu. 2015. NNIA-RS: A multi-objective optimization based recommender system. *Physica A: Statistical Mechanics and its Applications* 424 (2015), 383–397. https://doi.org/10.1016/j.physa.2015.01.007

[11] Mounir Hafsa, Pamela Wattebled, Julie Jacques, and Laetitia Jourdan. 2022. A Multi-Objective E-learning Recommender System at Mandarine Academy. In *Proceedings of the 2nd Workshop on Multi-Objective Recommender Systems co-located with 16th ACM Conference on Recommender Systems (RecSys 2022), Seattle, WA, USA, 18th-23rd September 2022 (CEUR Workshop Proceedings, Vol. 3268)*, Himan Abdollahpouri, Shaghayegh Sahebi, Mehdi Elahi, Masoud Mansoury, Babak Loni, Zahra Nazari, and Maria Dimakopoulou (Eds.). CEUR-WS.org. https://ceur-ws.org/Vol-3268/paper9.pdf

[12] F. Maxwell Harper and Joseph A. Konstan. 2016. The MovieLens Datasets: History and Context. *ACM Trans. Interact. Intell. Syst.* 5, 4 (2016), 19:1–19:19. https://doi.org/10.1145/2827872

[13] Xiangnan He, Kuan Deng, Xiang Wang, Yan Li, Yong-Dong Zhang, and Meng Wang. 2020. LightGCN: Simplifying and Powering Graph Convolution Network for Recommendation. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval, SIGIR 2020, Virtual Event, China, July 25-30, 2020*, Jimmy X. Huang, Yi Chang, Xueqi Cheng, Jaap Kamps, Vanessa Murdock, Ji-Rong Wen, and Yiqun Liu (Eds.). ACM, 639–648. https://doi.org/10.1145/3397271.3401063

[14] Dietmar Jannach, Lukas Lerche, Iman Kamehkhosh, and Michael Jugovac. 2015. What recommenders recommend: an analysis of recommendation biases and possible countermeasures. *User Model. User Adapt. Interact.* 25, 5 (2015), 427–491. https://doi.org/10.1007/s11257-015-9165-3

[15] Miqing Li and Xin Yao. 2019. Quality evaluation of solution sets in multiobjective optimisation: A survey. *ACM Computing Surveys (CSUR)* 52, 2 (2019), 1–38.

[16] Yunqi Li, Hanxiong Chen, Zuohui Fu, Yingqiang Ge, and Yongfeng Zhang. 2021. User-oriented Fairness in Recommendation. In *WWW '21: The Web Conference 2021, Virtual Event / Ljubljana, Slovenia, April 19-23, 2021*, Jure Leskovec, Marko Grobelnik, Marc Najork, Jie Tang, and Leila Zia (Eds.). ACM / IW3C2, 624–632. https://doi.org/10.1145/3442381.3449866

[17] Dawen Liang, Rahul G. Krishnan, Matthew D. Hoffman, and Tony Jebara. 2018. Variational Autoencoders for Collaborative Filtering. In *Proceedings of the 2018 World Wide Web Conference on World Wide Web, WWW 2018, Lyon, France, April 23-27, 2018*, Pierre-Antoine Champin, Fabien Gandon, Mounia Lalmas, and Panagiotis G. Ipeirotis (Eds.). ACM, 689–698. https://doi.org/10.1145/3178876.3186150

[18] R. Marler and Jasbir Arora. 2004. Survey of Multi-Objective Optimization Methods for Engineering. *Structural and Multidisciplinary Optimization* 26 (04 2004), 369–395. https://doi.org/10.1007/s00158-003-0368-6

[19] Bibek Paudel, Fabian Christoffel, Chris Newell, and Abraham Bernstein. 2017. Updatable, Accurate, Diverse, and Scalable Recommendations for Interactive Applications. *ACM Trans. Interact. Intell. Syst.* 7, 1 (2017), 1:1–1:34. https://doi.org/10.1145/2955101

[20] Paul Resnick, Neophytos Iacovou, Mitesh Suchak, Peter Bergstrom, and John Riedl. 1994. GroupLens: An Open Architecture for Collaborative Filtering of Netnews. In *CSCW '94, Proceedings of the Conference on Computer Supported Cooperative Work, Chapel Hill, NC, USA, October 22-26, 1994*, John B. Smith, F. Donelson Smith, and Thomas W. Malone (Eds.). ACM, 175–186. https://doi.org/10.1145/192844.192905

[21] Marco Túlio Ribeiro, Anísio Lacerda, Adriano Veloso, and Nivio Ziviani. 2012. Pareto-efficient hybridization for multi-objective recommender systems. In *Sixth ACM Conference on Recommender Systems, RecSys '12, Dublin, Ireland, September 9-13, 2012*, Padraig Cunningham, Neil J. Hurley, Ido Guy, and Sarabjot Singh Anand (Eds.). ACM, 19–26. https://doi.org/10.1145/2365952.2365962

[22] Jason R Schott. 1995. *Fault tolerant design using single and multicriteria genetic algorithm optimization.* Technical Report. Air force inst of tech Wright-Patterson afb OH.

[23] Dusan Stamenkovic, Alexandros Karatzoglou, Ioannis Arapakis, Xin Xin, and Kleomenis Katevas. 2022. Choosing the Best of Both Worlds: Diverse and Novel Recommendations through Multi-Objective Reinforcement Learning. In *WSDM '22: The Fifteenth ACM International Conference on Web Search and Data Mining, Virtual Event / Tempe, AZ, USA, February 21 - 25, 2022*, K. Selcuk Candan, Huan Liu, Leman Akoglu, Xin Luna Dong, and Jiliang Tang (Eds.). ACM, 957–965. https://doi.org/10.1145/3488560.3498471

[24] Harald Steck. 2019. Embarrassingly Shallow Autoencoders for Sparse Data. In *The World Wide Web Conference, WWW 2019, San Francisco, CA, USA, May 13-17, 2019*, Ling Liu, Ryen W. White, Amin Mantrach, Fabrizio Silvestri, Julian J. McAuley, Ricardo Baeza-Yates, and Leila Zia (Eds.). ACM, 3251–3257. https://doi.org/10.1145/3308558.3313710

[25] David Allen Van Veldhuizen. 1999. *Multiobjective evolutionary algorithms: classifications, analyses, and new innovations.* Air Force Institute of Technology.

[26] Saul Vargas and Pablo Castells. 2011. Rank and relevance in novelty and diversity metrics for recommender systems. In *Proceedings of the 2011 ACM Conference on Recommender Systems, RecSys 2011, Chicago, IL, USA, October 23-27, 2011*, Bamshad Mobasher, Robin D. Burke, Dietmar Jannach, and Gediminas Adomavicius (Eds.). ACM, 109–116. https://dl.acm.org/citation.cfm?id=2043955

[27] Mengting Wan, Rishabh Misra, Ndapa Nakashole, and Julian J. McAuley. 2019. Fine-Grained Spoiler Detection from Large-Scale Review Corpora. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, Anna Korhonen, David R. Traum, and Lluís Màrquez (Eds.). Association for Computational Linguistics, 2605–2610. https://doi.org/10.18653/v1/p19-1248

[28] Haolun Wu, Chen Ma, Bhaskar Mitra, Fernando Diaz, and Xue Liu. 2022. Multi-FR: A Multi-objective Optimization Framework for Multi-stakeholder Fairness-aware Recommendation. In *Transactions on Information Systems (TOIS)*. ACM.

[29] Fatima Ezzahra Zaizi, Sara Qassimi, and Said Rakrak. 2023. Multi-objective optimization with recommender systems: A systematic review. *Information Systems* 117 (2023), 102233. https://doi.org/10.1016/j.is.2023.102233

[30] Yong Zheng and David (Xuejun) Wang. 2022. A survey of recommender systems with multi-objective optimization. *Neurocomputing* 474 (2022), 141–153. https://doi.org/10.1016/j.neucom.2021.11.041

[31] Ziwei Zhu, Jianling Wang, and James Caverlee. 2020. Measuring and Mitigating Item Under-Recommendation Bias in Personalized Ranking Systems. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval, SIGIR 2020, Virtual Event, China, July 25-30, 2020*, Jimmy X. Huang, Yi Chang, Xueqi Cheng, Jaap Kamps, Vanessa Murdock, Ji-Rong Wen, and Yiqun Liu (Eds.). ACM, 449–458. https://doi.org/10.1145/3397271.3401177

[32] Eckart Zitzler, Kalyanmoy Deb, and Lothar Thiele. 2000. Comparison of multiobjective evolutionary algorithms: Empirical results. *Evolutionary computation* 8, 2 (2000), 173–195.

[33] Eckart Zitzler and Lothar Thiele. 1998. Multiobjective optimization using evolutionary algorithms—a comparative case study. In *Parallel Problem Solving from Nature—PPSN V: 5th International Conference Amsterdam, The Netherlands September 27–30, 1998 Proceedings 5*. Springer, 292–301.