On Inherited Popularity Bias in Cold-Start Item Recommendation (Supplementary Material)

Anonymous Author(s)

Table 1: Statistics for datasets used. The size of the feature vector for each data mode is indicated in parentheses.

Dataset	Users	Items	Interactions	Density	Modes (dim)
Clothing	39,387	23,033	278,677	0.031%	Text (384)
Electronics	192,403	63,001	1,689,188	0.014%	Image (4096) Text (384)
Microlens	98,129	17,228	705,174	0.042%	Image (1024) Text (1024) Video (768)

1 Dataset Statistics

Table 1 contains the user and item counts and data mode information for the datasets used in our experiments.

2 Inherited Popularity Bias: Other Datasets

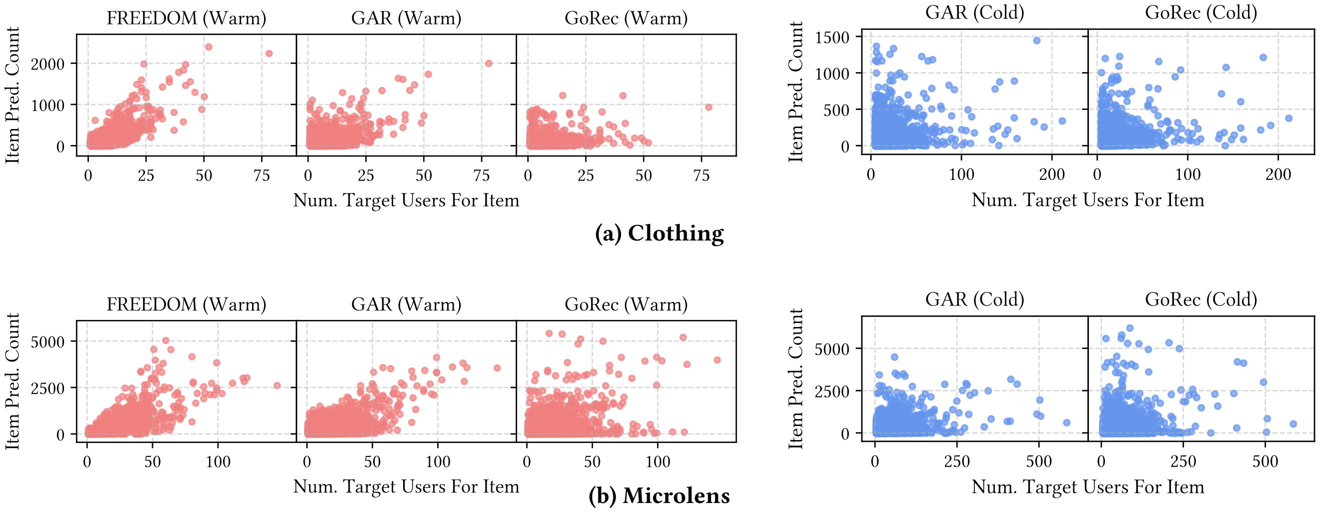
In Figure 1 we visualize top 20 prediction counts against holdout set counts for the items in the Clothing and Microlens datasets (i.e. the equivalent of Figure 1 in the main paper). Since these datasets have fewer items than Electronics, their outlier behavior is less extreme, but there is still evidence of the same problematic behaviors. In the warm predictions, a small number of items have much larger prediction counts than the rest of the population, and the cold models mirror this pattern. However, the overexposed cold items are often not popular, meaning that these biased behaviors are lowering recommendation quality as well as item fairness. We see this concretely for Clothing and Microlens in the main results in the paper, where our mitigation method actually improves accuracy alongside item fairness in some cases.

3 Impact of α

In this section we provide further insight into the effect of the magnitude scaling hyperparameter α on model predictive behavior. In Figure 2 we visualize the impact of α on prediction counts for the cold item test set. We can see that the scaling increasingly balances the distribution as α grows, and that the curves converge towards the limit where all vectors have a magnitude of μ_w. While the unscaled models leave up to 20% of items out of the top 20 predictions for all users, all of the scaled versions provide at least 10 predictions to the vast majority of items. This illustrates why, by applying this scaling, we are able to achieve significant increases in Gini Diversity seen in Table 1 in the main paper.

In Figure 3, we plot the effect of α on user accuracy and low-end item accuracy. In Clothing and Microlens, both user and item accuracy increase for smaller values of α, although this levels off as α gets larger, and an inverse relationship between the two metrics starts to appear. Nonetheless, by careful selection of α we are able to achieve a material increase in the performance of under-served items for these datasets while preserving recommendation quality from the user perspective. In the Electronics dataset, the inverse relationship between NDCG and MDG-Min80% is stronger, and it seems that the much larger item population makes it significantly more difficult to promote correct items into the top 20. There is therefore a larger compromise between user and item-level performance, and the appropriate α value may depend on the application and other fairness considerations.

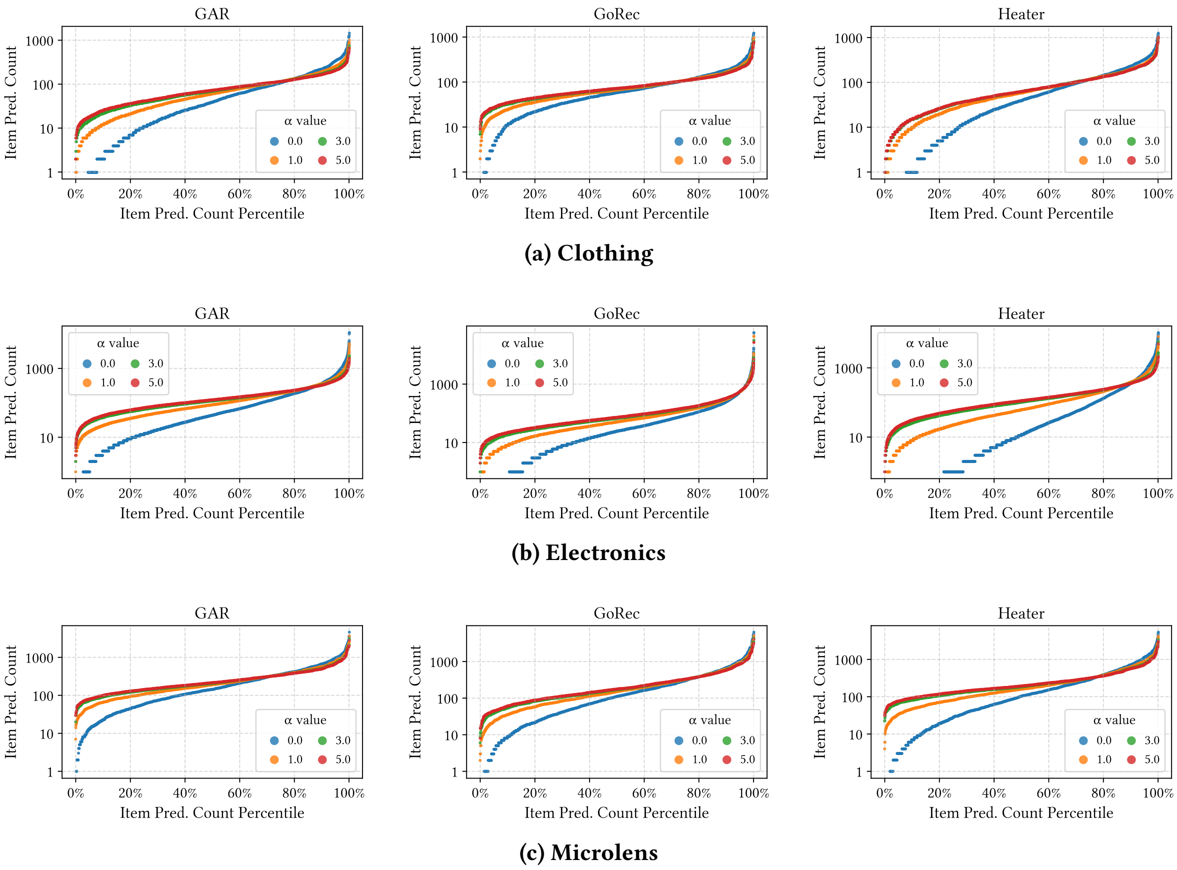
Conference acronym 'XX, Woodstock, NY
2018. ACM ISBN 978-1-4503-XXXX-X/2018/06
https://doi.org/XXXXXXX.XXXXXXX



(a) Clothing

(b) Microlens

Figure 1: Warm and cold item prediction counts with $k = 20$ against the number of target users (i.e. the number of times an item appears in the validation or test set interactions) for Clothing (top) and Microlens (bottom). Each dot represents an item.

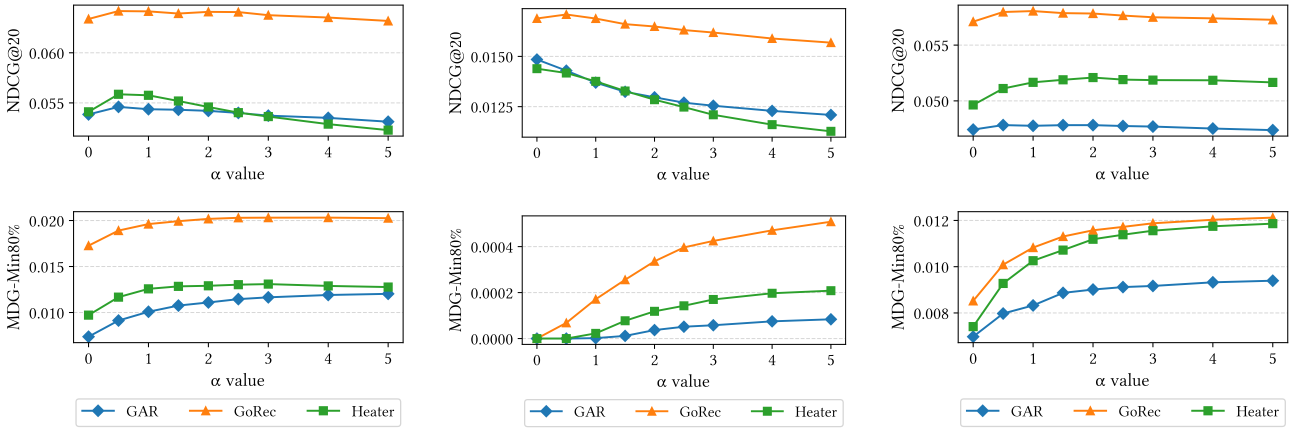


(a) Clothing

(b) Electronics

(c) Microlens

Figure 2: Cold test set item prediction counts at $k = 20$ against item prediction count percentiles (i.e. each item's position in the sorted list of prediction counts). Only items predicted at least once are plotted.



(a) Clothing **(b) Electronics** **(c) Microlens**

Figure 3: Impact of α on cold test set NDCG and MDG-Min80% with $k = 20$.