Struggling with popularity bias in RS

Long-tail problem

In RS a small fraction of popular items account for the majority of user interactions. When trained on such data, the model usually gives higher scores to popular items than their ideal values while simply predicting unpopular items as negative.

Agent-Based Modeling

provides a playground for simulation of user interactions on synthetic data and measure the bias without running AB on real users.

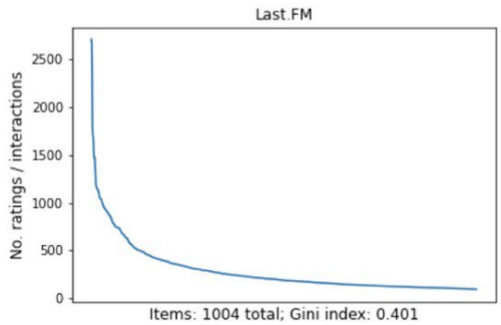
The solution

Using the agent-based model we provide a guide how to setting classical methods of struggling with popularity bias without running AB on real users.

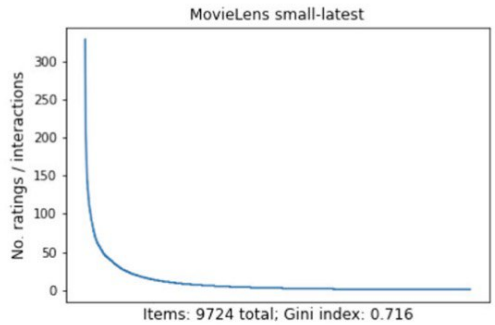
Long-tails in real data

The plots show the interaction counts for each item within the dataset on the x-axis, sorted in descending order.

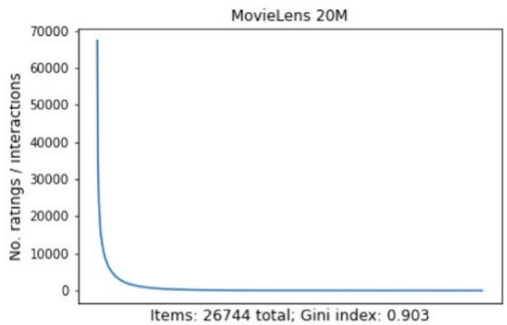
The Gini index expresses the inequality of the distribution, with values closer to 1 indicating a high inequality (range: 0–1)



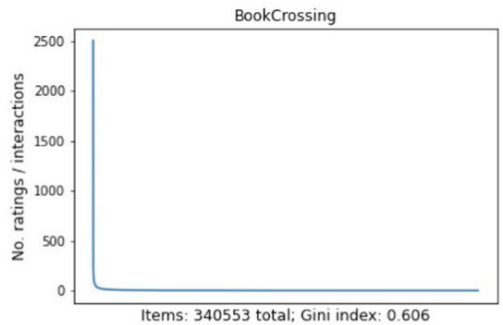
(a) Music dataset: 289 955 interactions, 15 000 users, 1 004 items



(b) Movie dataset: 100 836 interactions, 610 users, 9 724 items



(c) Movie dataset: 20 000 263 interactions, 138 493 users, 26 744 items



(d) Books dataset: 1 149 780 interactions, 105 283 users, 340 553 items

Popularity bias exists in almost all of the popular datasets.