University of Glasgow | School of Computing Science

Terrier Team

How Sensitive is Recommendation Systems' Offline Evaluation to Popularity?

Amir H. Jadidinejad, Craig Macdonald, Iadh Ounis
amir.jadidinejad@glasgow.ac.uk

Overview

Datasets used for the offline evaluation of recommender systems are collected through user interactions with an already deployed recommender system. Such datasets can be subject to different types of biases including a system's popularity bias. In this work, we focus on assessing the influence of popularity on the offline evaluation of recommendation systems. Our insights derived from a deep analysis using popularity-stratified sampling reveal that the current offline evaluation of recommendation systems are sensitive to popular items, raising questions about conclusions drawn from the offline comparison of recommendation models.

1. Problem Statement

- The **offline evaluation** of recommendation systems includes:
  - gathering a collection of ***users' interactions*** from a **deployed system**.
  - the use of these interactions to evaluate and compare different recommendation models.
- However, users' interaction can be subject to different types of **biases**.
- Our aim is to investigate the extent to which the offline evaluation of recommendation systems is **sensitive to popularity bias** in the observed users' interactions, which was collected from a deployed system.
- We propose a method based on **stratified sampling** to evaluate the effectiveness of a given recommendation model across groups of items with various **popularity levels**.

2. Recommendation Systems' Offline Evaluation

- The offline evaluation of Recommendation Systems takes a set of random samples ($\mathcal{T}$) of all interactions between users and items ($\mathcal{U}\times\mathcal{I}$) as a held-out test set.
- We report below the results of evaluating 5 well-known recommendation models on the held-out MovieLens-20M and Amazon test sets.

DataSet	Model	MRR@20	nDCG@20
MovieLens-20M	[1] Random	$0.004^{2,3,4,5}$	$0.0007^{2,3,4,5}$
	[2] Popularity	$0.14^{1,3,5}$	$0.067^{1,3,4,5}$
	[3] MF	$0.15^{1,2,4,5}$	$0.06^{1,2,4,5}$
	[4] BPR	$0.14^{1,3,5}$	$0.068^{1,2,3,5}$
	[5] WARP	$\mathbf{0.176}^{1,2,3,4}$	$\mathbf{0.076}^{1,2,3,4}$
Amazon	[1] Random	$0.0002^{2,3,4,5}$	$0.0001^{2,3,4,5}$
	[2] Popularity	$0.0093^{1,3,4,5}$	$0.0089^{1,4,5}$
	[3] MF	$0.0083^{1,2,4,5}$	$0.0091^{1,4,5}$
	[4] BPR	$0.0125^{1,2,3,5}$	$0.0111^{1,2,3,5}$
	[5] WARP	$\mathbf{0.0217}^{1,2,3,4}$	$\mathbf{0.0181}^{1,2,3,4}$

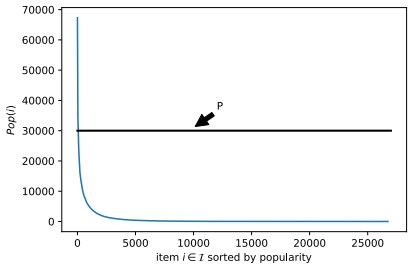
Table 1: The offline evaluation of different models based on 20% held-out samples.

3. Popularity-Stratified Sampling

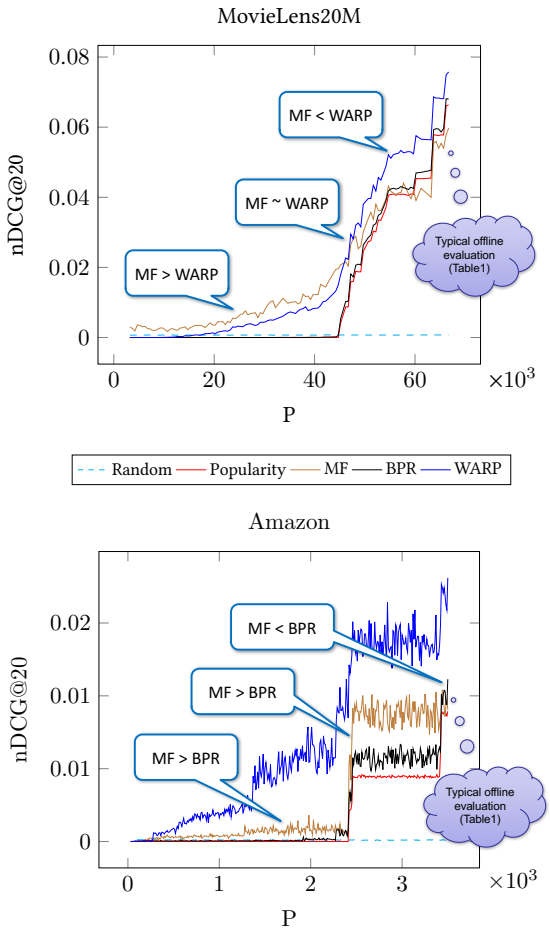
- Our aim is to measure the influence of different groups of items on the offline evaluation of a given model. Therefore, we sample different sub-populations of user-item interactions based on various levels of popularity.
- For a specific popularity **threshold** P, we can randomly sample a fixed number of user-item interactions $\mathcal{T}_P$ containing users' interactions with less popular items depending on the threshold P:

$$\mathcal{T}_P \leftarrow \{\mathcal{U}\times\mathcal{I} : Pop(i) < P\}$$

- Our **popularity-stratified sampling** method allows to neglect the interactions of a few most popular items concerning each P threshold thereby assessing the **sensitivity** of evaluation across groups of items with various popularity levels.



4. Experimental Results



5. Conclusions

- We investigated the **influence of popularity** on the offline evaluation of recommendation systems.
- The proposed **stratified sampling method** provides a detailed analysis to measure the *impact of the recommendation models on different groups of items*.
- Our findings show that indeed the examined **models are sensitive to popularity**, i.e. random sampling from different popularity strata can considerably impact offline evaluation and the subsequent conclusions.
- The current offline evaluation is an *average over all types of users and items* which might not necessarily represent the actual effectiveness of the examined models.
- Popularity is a special type of **closed-loop feedback** that is enforced by the deployed algorithms. We plan to measure the effect of closed-loop feedback in the training and evaluation of recommender systems.

Acknowledgement

The authors acknowledge support from EPSRC grant EP/R018634/1 entitled Closed-Loop Data Science for Complex, Computationally- and Data-Intensive Analytics.

Data & Source