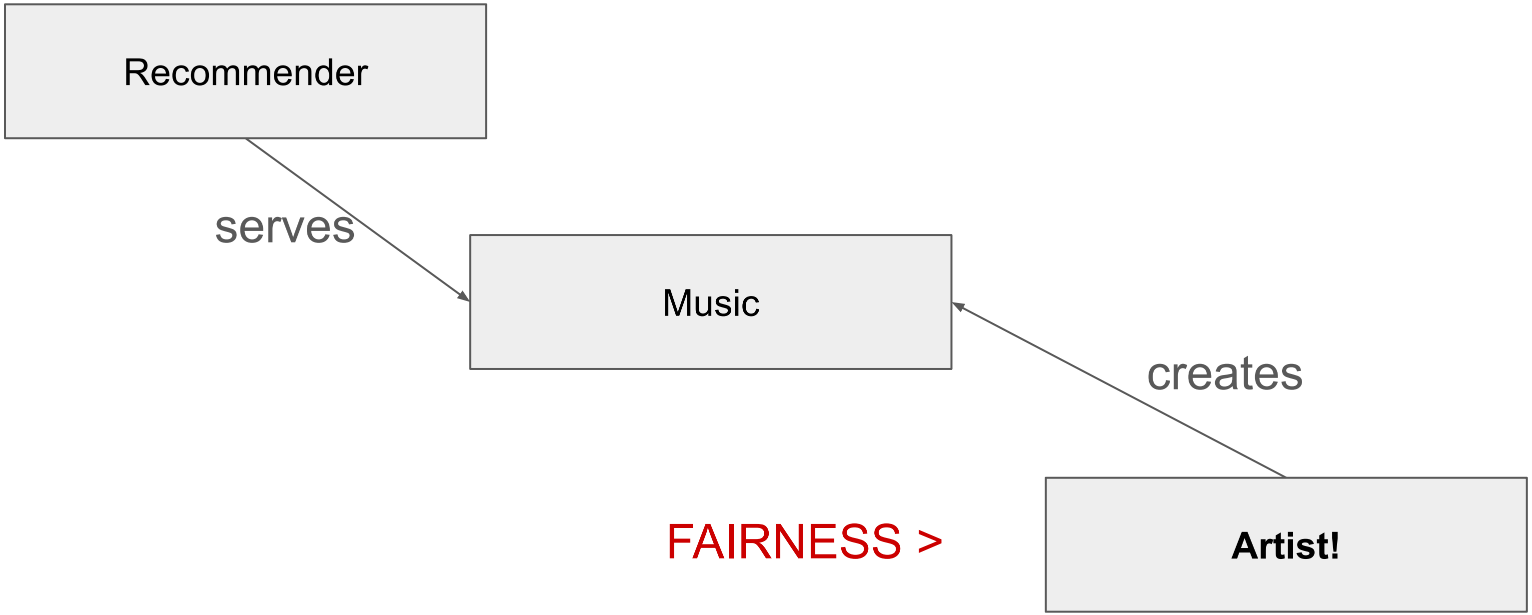
Auditing Gender Bias in Music Recommendation

Analysis of KKBox Recommendation System

Jeffrey Gordon and Cedric Lam

Is a Music Recommender an ADS?



Chosen ADS

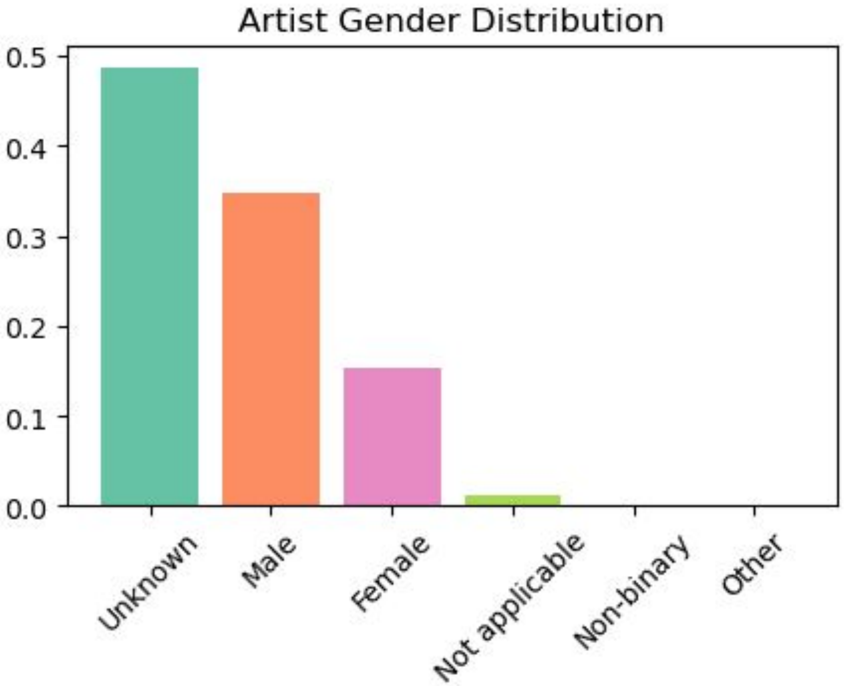
- **Audited system**: Top solution to KKBox's 2017 music recommendation challenge using real listening history.

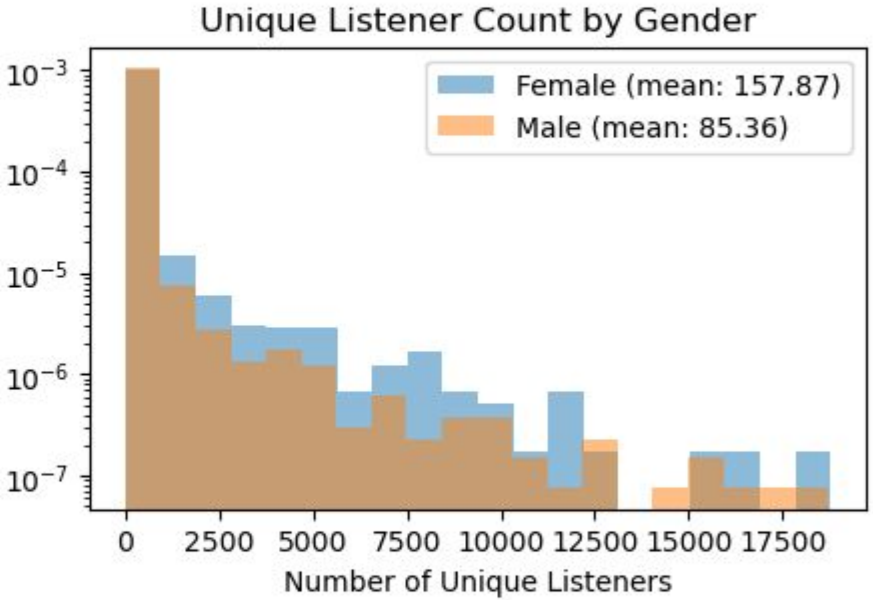
KKBOX

- Issue: the dataset has no artist demographic information!
    - Scraped gender labels from MusicBrainz
    - Focus on solo artists only

MusicBrainz

Gender Skew





listening history data: **skew washes out!**

System Implementation

- Prediction nearly 400 features in total.
- Hybrid ensemble model of LightGBM (60% weight) and Deep Neural Network (40% weight).
- Average AUC of 0.85!!
- Took forever to run…

How to measure performance for an artist?

$$Avg\ Accuracy = \frac{1}{N_{artists}} \sum_{artists} accuracy_{artist}$$

Accuracy	0.77033
Precision	0.74116
Recall	0.56596
FNR	0.40297
FPR	0.14089

Do these metrics differ by artist gender?

Gender	Accuracy	Precision	Recall	FNR	FPR
Male	0.7743	0.7423	0.5659	0.4001	0.1382
Female	0.7645	0.7386	0.5718	0.4033	0.1420

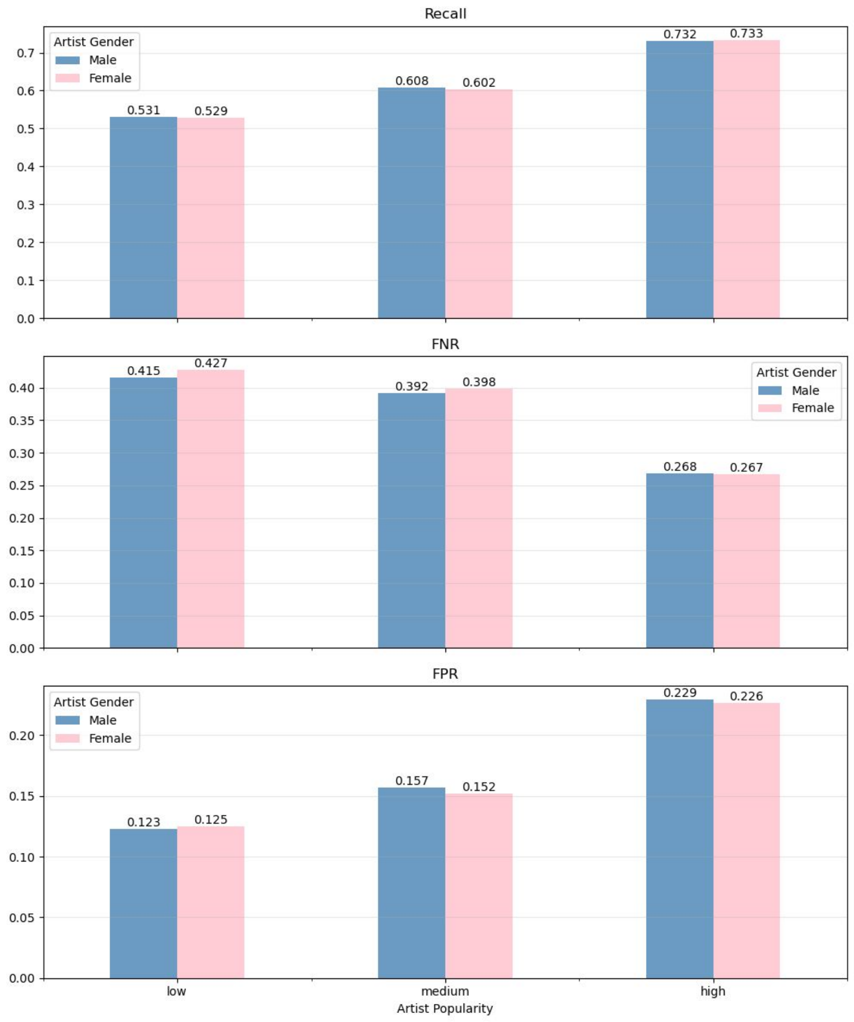
FNR Difference	0.00321
FPR Difference	0.00382
Demographic Parity Ratio	0.97406
Equalized Odds Ratio	0.96849
Selection Rate Difference	0.00818

“Fairness through Blindness”

Popularity Confounds

- Controlling for popularity, fairness metrics are equal for both genders.
- **Minimum gender bias in the predictions.**

- More “generous” as popularity increases.
- Higher recall and FPR. Lower FNR.



🌈 *The ADS exhibits little gender bias!* ✨