Pricilla Nakyazze Research Discussion 3

2025-06-18

In What Ways Do Recommender Systems Reinforce Human Bias?

As machine learning becomes increasingly embedded in decision-making processes across sensitive domains—such as advertising, credit, employment, education, and criminal justice—concerns about fairness, ethics, and discrimination are gaining prominence. While these systems can enhance accuracy and efficiency, they can also reinforce or even amplify existing human biases. In some cases, they may introduce new forms of bias, effectively encoding discrimination into automated decisions.

Do Recommender Systems Reinforce or Prevent Unethical Targeting?

Reflecting on the techniques we've studied, it's clear that recommender systems can both reinforce bias and contribute to unethical customer segmentation if not carefully designed and monitored.

Evan Estola, a lead engineer at Meetup, offers an example of Amazon's recommendation system: "Customers who bought this item also bought..." This is a marketing-driven recommender aimed at boosting sales. While useful, such systems can subtly nudge consumers toward certain products or brands, reinforcing existing consumption patterns that may reflect biased data.

Estola also highlights how Google's recommendations shape the visibility of opportunities. For instance, a search for "Data Science" may return only certain types of courses or providers, reinforcing popularity and limiting diversity in educational options.

Recommender systems don't just affect what we see—they influence what we pay, the news we consume, the job openings we find, and even the advertisements we're exposed to. For example, Orbitz was found to direct Mac users toward more expensive hotel listings, based on the assumption that they are wealthier. This is a clear case of unethical segmentation based on inferred socioeconomic status.

Worse still, a study showed that Google ads served to users with Black-sounding names were more likely to suggest a criminal record—an egregious example of racial bias in algorithmic targeting.

Preventing Discrimination in Recommender Systems

To prevent unethical targeting, recommender systems must be explicitly designed to avoid prejudice. A core principle should be that protected attributes (like race, gender, or age) must not be inferable from the features used for recommendations.

Identifying vulnerable or misrepresented groups is crucial. For example, in tech-focused meetup groups, an algorithm may infer from historical data that women are less likely to attend tech events. This is a harmful stereotype that should not be propagated. Transparent and interpretable models—like logistic regression—can help us audit what the system is learning. In this case, the model should not infer that a user is uninterested in tech based on gender.

Techniques to mitigate bias include:

Data segregation: Separate sensitive features (like gender or race) from behavioral and interest data. Different models can be used for different purposes.

Ensemble modeling: Use multiple algorithms to balance prediction performance while minimizing bias.

Controlled diversity modeling: Introduce mechanisms to ensure diversity in recommendations. Diverse test profiles can be used to probe algorithmic behavior across different demographic groups.

Equalized Odds and Fair Thresholding: Use fairness-aware metrics such as Equalized Odds, which seek to equalize false positive and false negative rates across different groups. When ROC curves are used, predictors can be adjusted by mixing thresholds (from convex hulls of ROC points) to balance fairness and accuracy.

Ultimately, if data on user attributes and outcomes are available, algorithms can be modified or re-trained to ensure equitable treatment and remove discriminatory patterns in predictions to remove discrimination or targeting.