On the Need for Fairness in Financial Recommendation Engines

Sirui Yao
Department of Computer Science
Virginia Tech
Blacksburg, VA, 24060

Bert Huang
Department of Computer Science
Virginia Tech
Blacksburg, VA, 24060

Abstract

Recommender systems are a key technology in all areas of applied machine learning. In finance, they can be used to support human decisions, such as approving loans, managing assets, and assessing risks. Because these financial applications can have a direct monetary impact on users, issues of unfairness in machine-learning-based recommendation are critical. In this paper, we argue that the study of fairness in machine learning for finance applications must include a special focus on recommender systems. We discuss the potential impact that various form of unfairness can have in financial applications. Finally, we highlight previous and recent research on fairness in recommendation and open research directions.

1 Introduction

Machine learning is playing an important role in many domains of financial services, ranging from financial products including loans, insurance, real estate, and stocks to management of portfolios composed of various types of financial assets [22]. These applications are usually in the form of filtering large amounts of information and suggesting the most reasonable and profitable options. One of the most important forms of machine learning in the industry today is the recommendation engine, which supports human decision making by learning user preferences from historical decision data. Recommendation engines are used in assisting insurance agents to offer suitable insurance products to their clients based on various criteria such as price, co-pay, and benefits [1]; products whose profitability are considered hard to predict are also recommended through recommender systems that analyze economic events, market news, and investors' risk-aversion preference and trading behavior [22]. Therefore, as a key technology of applied machine learning, recommender systems have tremendous potential to promote more efficient and effective financial decisions, and study of machine learning's impact on finance should pay close attention to them.

Practitioners must be aware of the potential impact of applying such technologies. Since they are trained on data from the real world, which already has a long history of human bias, data can be severely contaminated, and historical biases can be reflected or reinforced by algorithms. Careless application of recommender systems may put certain subgroups of people into disadvantaged positions by systematically reducing their exposure to some resources. For example, if highly profitable stocks or real-estate properties are consistently less recommended to one subgroup of people than to others, the recommender is biased against the disadvantaged subgroup, posing both ethical and legal concerns.

According to the Federal Fair Lending Regulations and Statutes [16], "the Equal Credit Opportunity Act (ECOA), which is implemented by the Board's Regulation B (12 CFR202), prohibits discrimination in any aspect of a credit transaction. It applies to any extension of credit, including residential real estate lending and extensions of credit to small businesses, corporations, partnerships,

NIPS 2018 Workshop on Challenges and Opportunities for AI in Financial Services: the Impact of Fairness, Explainability, Accuracy, and Privacy, Montréal, Canada. This is a non-archival publication - the authors may submit revisions and extensions of this paper to other publication venues.

and trusts." Therefore, enforcing fairness in financial recommender systems is of both ethical and legal significance.

In this paper, we first discuss different notions of fairness and their implication in recommender systems in the financial domain. Then we list key research problems needing further study and finally the technical challenges in solving these problems.

2 Fairness Notions

In this section, we discuss three notions of fairness: demographic parity, equality of opportunity, and counterfactual fairness. We show their implications in the task of loan recommendations, which seeks to pair the appropriate lenders and individuals who apply for loans. As an example, we consider gender as the sensitive feature and discuss the effects of recommendation on subgroups of male and female (or non-binary) users. Demographic parity and equality of opportunity are visualized in Figure 1.

Demographic Parity Demographic parity is perhaps the most direct measure of fairness. Its goal is to ensure that an algorithm treats members of a protected group equivalently to members outside the group. Kamishima et al. [10, 12, 11] have studied methods to reduce unfairness in recommender systems by adding a regularization term that enforces demographic parity, penalizing differences among the average predicted ratings of user groups. In loan recommendations, if the ratio of female and male lenders who successfully acquired loans is disproportionate to the ratio of applicants in these two subgroups, the demographic parity rule is violated. Demographic parity is only appropriate when preferences are unrelated to the sensitive features.

Equal Opportunity As an alternative to demographic parity, Hardt et al. [9] propose to measure unfairness with the true-positive rate and true-negative rate. This idea encourages what they refer to as equal opportunity, a concept that upholds fairness while respecting group differences. These approaches to fair machine learning all depend on the learning algorithm having access to a sensitive feature or division of users. Yao and Huang [19] applied this notion of fairness to collaborative filtering systems by evaluating bias based on the deviation of predicted ratings from ground truth. In a loan recommendation application, unequal opportunity would occur if one subpopulation receives lower-quality recommendations than another, causing this disadvantaged group to have lower true-positive (or true-negative) rates. The disadvantaged group would then receive less benefit from the data-driven recommendations, while an advantaged group receives high-quality recommendations. If these groups are defined by protected identity attributes, such as gender, deploying the recommendation engine could amount to providing a service to certain customers based on gender.

Counterfactual Fairness Counterfactual fairness [7, 17, 15] is inspired by the question: How would the prediction change if the sensitive attribute were a different value? For example, consider a

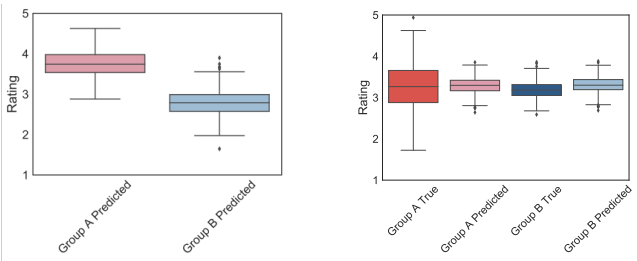

Figure 1: Demographic parity and equal opportunity. Each box plot compares the statistics of true and predicted feedback (e.g. ratings) of users in two population groups—A and B. Left: Demographic parity. The distribution of predicted ratings for Group A has a much higher probability for higher rating values than the distribution of predicted ratings for Group B. Right: Equal opportunity. Though averages across both groups for true and predicted ratings are approximately equal, the variance of Group A's true ratings indicates that the predicted ratings make more significant absolute errors for Group A than Group B.

suitable loan lender being recommended to a male borrower but not if he registered in the system with exactly the same information with his gender changed to female. This scenario would indicate direct gender-based disparate treatment. Counterfactual fairness is different from demographic parity and equal opportunity because it enforces individual fairness [3] while parity and equal opportunity evaluate fairness across groups. A variation of counterfactual fairness has been proposed [20], which defines fairness with the idea that "given the choice between various sets of decision treatments or outcomes, any group of users would collectively prefer its treatment or outcomes, regardless of the (dis)parity as compared to the other groups."

3 Research Directions

Almost all research on fairness in recommender systems seeks answers for three questions: (1) how to quantify bias in recommender systems, (2) how to identify the causes of bias, and (3) how to mitigate bias. Specifically,

1. One big challenge of studying fairness is that fairness may not be well-defined. Even speaking from a legal perspective, there is no universally applicable definition of fairness. Instead, it is usually studied case by case. Finding general metrics for recommender fairness will help us better understand and monitor bias. The notions of fairness, as introduced in Section 2, only provide guidelines for evaluating bias and require consideration of contextual details during implementation.
2. The causes of unfair recommendation should be studied to facilitate the development of approaches that directly counteract these causes; Garcia-Gathright et al. [6] proposed that bias could be "introduced at different levels of data gathering and usage, including: user biases, societal biases, data processing biases, analysis biases, and biased interpretation of results", further, biases can even be amplified through "the interplay between data bias and algorithmic bias: biased training data results in biased algorithms, which in turn produce more biased data in a feedback loop." Efforts in designing models that are more interpretable help trace the source of bias and reduce the risk of bias introduced through black-box processes [8, 18].
3. Mitigation of bias is often the most important goal of research on fair recommendation. Many approaches have been proposed and intervene in different stages of modeling [5]: pre-, mid- and post-processing. As mentioned in Section 3, training data may be the cause of discrimination. Pre-processing algorithms may therefore be crucial to reverse the biases from the source. For example, Calmon et al. [2] and Zemel et al. [21] learn data transformations through optimization with multiple utility and fairness goals; mid-processing approaches take actions during the process of modeling by modifying objective functions or add additional steps to enforce fairness. Yao and Huang [19] explicitly optimize over the fairness metric as part of the objective; Lee et al. [14] developed a fairness-aware model for microfinance services based on a Bayesian personalized ranking optimization criterion coupled with matrix factorization (BPRMF). This method not only maximizes the utility of matching but also increases diversity in loan providers. Post-processing takes the output of models and modifies that output to be fair [9]. Kim et al. [13] develop multiaccuracy boosting, a post-processing framework for black-box models which iteratively identifies subpopulations that the model systematically makes more mistakes on and enforces equality constraints on each identifiable subgroup.

4 Technical challenges

In this section, we discuss some challenges in enforcing fairness in recommender systems. The problem of bias in recommender systems needs to be studied separately from those in classical prediction/classification tasks because it cannot be solved by simply borrowing from the research on fairness for other forms of machine learning. First, Recommender models have different structures than classical prediction in that, instead of learning from independent and identically distributed (iid) examples, we have to simultaneously consider two entities—users and items—and the interaction between them. This dependency is especially important in evaluating the causes of unfairness in recommender systems, where we not only consider the distribution of feedback within the subgroups but how these groups affect other.

Another key difference is that recommender systems usually suffer from the problem of sparsity, where the available data is scarce compared to the number of entities to learn. The majority of users only have very limited data on their historical behavior. Recommendation further suffers from the cold-start problem, which is an extreme case of sparsity where no data is available for a new user or a new item. This poses challenges in evaluating unfairness and the effectiveness of fairness mitigation approaches because the models may overfit and be mislead by noise.

Third, recommenders serve as tools to promote decision making, but they do not have control over the outcome since the recommendations still can be filtered by user preference. Therefore, building a fair recommender does not guarantee fairness in users' decisions. This is important when we consider the feedback loop in recommender systems [4].

5 Conclusion

In this paper, we discussed the significance of studying fairness in recommender systems to promote responsible decision making in finance domains. We describe the implications of various notions of fairness using the example of loan recommendation. We listed several key research directions that can work together to promote fairness in recommender systems. Lastly, we discussed the technical challenges that should be addressed by future research.

Acknowledgments

We thank Deloitte for partially supporting this work.

References

[1] A. Abbas, K. Bilal, L. Zhang, and S. U. Khan. A cloud based health insurance plan recommendation system: A user centered approach. *Future Generation Computer Systems*, 43:99–109, 2015.

[2] F. Calmon, D. Wei, B. Vinzamuri, K. N. Ramamurthy, and K. R. Varshney. Optimized preprocessing for discrimination prevention. In *Advances in Neural Information Processing Systems*, pages 3992–4001, 2017.

[3] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pages 214–226. ACM, 2012.

[4] D. Ensign, S. A. Friedler, S. Neville, C. Scheidegger, and S. Venkatasubramanian. Runaway feedback loops in predictive policing. *arXiv preprint arXiv:1706.09847*, 2017.

[5] S. A. Friedler, C. Scheidegger, S. Venkatasubramanian, S. Choudhary, E. P. Hamilton, and D. Roth. A comparative study of fairness-enhancing interventions in machine learning. *arXiv preprint arXiv:1802.04422*, 2018.

[6] J. Garcia-Gathright, A. Springer, and H. Cramer. Assessing and addressing algorithmic bias-but before we get there. *arXiv preprint arXiv:1809.03332*, 2018.

[7] S. Garg, V. Perot, N. Limtiaco, A. Taly, E. H. Chi, and A. Beutel. Counterfactual fairness in text classification through robustness. *arXiv preprint arXiv:1809.10610*, 2018.

[8] L. H. Gilpin, D. Bau, B. Z. Yuan, A. Bajwa, M. Specter, and L. Kagal. Explaining explanations: An approach to evaluating interpretability of machine learning. *arXiv preprint arXiv:1806.00069*, 2018.

[9] M. Hardt, E. Price, and N. Srebro. Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems*, pages 3315–3323, 2016.

[10] T. Kamishima, S. Akaho, H. Asoh, and J. Sakuma. Correcting popularity bias by enhancing recommendation neutrality. In *Poster Proceedings of the 8th ACM Conference on Recommender Systems (RecSys)*, 2014.

[11] T. Kamishima, S. Akaho, H. Asoh, and J. Sakuma. Recommendation independence. In *Conference on Fairness, Accountability and Transparency*, pages 187–201, 2018.

[12] T. Kamishima, S. Akaho, H. Asoh, and I. Sato. Model-based approaches for independence-enhanced recommendation. In *IEEE 16th International Conference on Data Mining Workshops (ICDMW)*, pages 860–867. IEEE, 2016.

[13] M. P. Kim, A. Ghorbani, and J. Zou. Multiaccuracy: Black-box post-processing for fairness in classification. *arXiv preprint arXiv:1805.12317*, 2018.

[14] E. L. Lee, J.-K. Lou, W.-M. Chen, Y.-C. Chen, S.-D. Lin, Y.-S. Chiang, and K.-T. Chen. Fairness-aware loan recommendation for microfinance services. In *Proceedings of the 2014 International Conference on Social Computing*, page 3. ACM, 2014.

[15] R. Mehrotra, J. McInerney, H. Bouchard, M. Lalmas, and F. Diaz. Towards a fair marketplace: Counterfactual evaluation of the trade-off between relevance, fairness & satisfaction in recommendation systems. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, pages 2243–2251. ACM, 2018.

[16] F. Reserve. Federal fair lending regulations and statutes: Fair housing act consumer compliance handbook, 2008.

[17] C. Russell, M. J. Kusner, J. Loftus, and R. Silva. When worlds collide: integrating different counterfactual assumptions in fairness. In *Advances in Neural Information Processing Systems*, pages 6414–6423, 2017.

[18] K. R. Varshney, P. Khanduri, P. Sharma, S. Zhang, and P. K. Varshney. Why interpretability in machine learning? an answer using distributed detection and data fusion theory. *arXiv preprint arXiv:1806.09710*, 2018.

[19] S. Yao and B. Huang. Beyond parity: Fairness objectives for collaborative filtering. In *Advances in Neural Information Processing Systems*, 2017.

[20] M. B. Zafar, I. Valera, M. Rodriguez, K. Gummadi, and A. Weller. From parity to preference-based notions of fairness in classification. In *Advances in Neural Information Processing Systems*, pages 229–239, 2017.

[21] R. Zemel, Y. Wu, K. Swersky, T. Pitassi, and C. Dwork. Learning fair representations. In *International Conference on Machine Learning*, pages 325–333, 2013.

[22] D. Zibriczky12. Recommender systems meet finance: a literature review. 2016.