# End-to-End Bias Mitigation in Candidate Recommender Systems with Fairness Gates

Arafan, A.M.; Graus, D.; Santos, F.P.; Beauxis-Aussalet, E.

**Publication date**
2022

**Document Version**
Final published version

**Published in**
Proceedings of the 2nd Workshop on Recommender Systems for Human Resources (RecSys-in-HR 2022)

# End-to-End Bias Mitigation in Candidate Recommender Systems with Fairness Gates

Adam Mehdi Arafan[1,†], David Graus[2], Fernando P. Santos[3] and Emma Beauxis-Aussalet[4]

[1]University of Amsterdam, Amsterdam, The Netherlands
[2]Randstad Groep Nederland, Diemen, The Netherlands
[3]University of Amsterdam, Amsterdam, The Netherlands
[4]Vrije Universiteit Amsterdam, Amsterdam, The Netherlands

Abstract

Recommender Systems (RS) have proven successful in a wide variety of domains, and the human resources (HR) domain is no exception. RS proved valuable for recommending candidates for a position, although the ethical implications have recently been identified as high-risk by the European Commission. In this study, we apply RS to match candidates with job requests. The RS pipeline includes **two fairness gates** at two different steps: pre-processing (using GAN-based synthetic candidate generation) and post-processing (with greedily searched candidate re-ranking). While prior research studied fairness at pre- and post-processing steps separately, our approach combines them both in the same pipeline applicable to the HR domain. We show that the combination of gender-balanced synthetic training data with pair re-ranking increased fairness with satisfactory levels of ranking utility. Our findings show that using only the gender-balanced synthetic data for bias mitigation is fairer by a negligible margin when compared to using real data. However, when implemented together with the pair re-ranker, candidate recommendation fairness improved considerably, while maintaining a satisfactory utility score. In contrast, using only the pair re-ranker achieved a similar fairness level, but had a consistently lower utility.

**Keywords**

Fair Artificial Intelligence, Generative Modelling, Information Retrieval, Recommender Systems



## 1. Introduction

Machine learning (ML) applications have proven to be useful in many domains over recent years. However, despite the many benefits of ML-enabled tools, biases can occur and be amplified through the highly scalable nature of ML-enabled systems. Algorithms used in applications such as recidivism prediction, predictive policing, or facial recognition, have revealed bias towards either race, gender or both [1, 2]. These biases can also be expressed through proxy (unobservable) correlations expressed via sensitive attributes such as gender and poorly defined decision boundaries [3, 4].

We are focusing on fairness issues with candidate recommender systems (CRS). The goal of such a system is to recommend the best candidates for a specific job, often computing ranked lists of candidates in descending order of relevance. A variety of fairness issues may arise from the large and diverse pools of candidates and job offers.

In the case of the HR industry, bias in recommendations comes with a high risk of harm as candidates can perpetually face discrimination in finding employment. The risk of harm is especially great considering the scalable nature of recommender systems. Here we focus on a CRS to support a recruiter in finding the best matching candidates for a client job request (e.g., a factory requesting 20 technicians).

As most ML algorithms perform predictions in a discriminative fashion using historical data, it is not trivial to guarantee that discrimination is not (unfairly) influenced by proxies that might be correlated with protected characteristics. The fairness in ML problem has been approached by many researchers such as Rajabi and Garibay who tackled the problem by synthesizing data, or Li et al. constraining recommendations, and Geyik et al. by re-ranking recommendations. These researchers produced state-of-the-art (SOTA) algorithms tackling specific fairness techniques, from which we distinguish two: **pre-processing** (enforcing fairness at the data level) and **post-processing** (enforcing fairness after predictions were made).

These two approaches have traditionally been researched separately in RS and fairness literature, ignoring potential synergistic effects of applying fairness mechanisms at different stages of the ML pipeline. To the best of our knowledge, we found no prior work experimenting with more than one processing technique in a single pipeline. We aim to close this gap by testing SOTA bias mitigation methods in both **pre-** and **post-**processing, and observing the impact on the fairness of candidate

*RecSys in HR'22: The 2nd Workshop on Recommender Systems for Human Resources, in conjunction with the 16th ACM Conference on Recommender Systems, September 18–23, 2022, Seattle, USA.*

[†]Work done while on internship at Randstad Groep Nederland.

adammehdiarafan@gmail.com (A. M. Arafan); david.graus@randstadgroep.nl (D. Graus); f.p.santos@uva.nl (F. P. Santos); e.m.a.l.beauxisaussalet@vu.nl (E. Beauxis-Aussalet)





CEUR Workshop Proceedings (CEUR-WS.org)

ranking. We propose a pipeline for a CRS that integrates two bias mitigation mechanisms (called ***Fairness Gates***, FG) at the pre- and post-processing steps. By FG, we refer to the enforcement of bias mitigation techniques within the pipeline. The FGs are a **synthetic data generator** and a **greedy re-ranker**.

The **synthetic data generator** enforces gender balance in the sampling size while the **greedy re-ranker** optimizes for both utility (the quality or usefulness of candidate recommendations) and gender balance in candidate ranking. In this paper, we explore the fairness-utility trade-offs among re-ranked CRS outputs trained using synthetic data or only real data. Therefore, we focus on exploring **what are the impacts and trade-offs between utility and fairness that arise** from combining synthetic data generation at pre-processing and greedy pair re-ranking at a post-processing level.

Our experimental results show that the best compromise between fairness and utility is achieved when combining the two FGs rather than using just one.

## 2. Background and Related Work

Before presenting the experiments conducted within our novel candidate recommendation pipeline, essential terminology needs to be defined alongside the state of the art in the (sub)task(s) at hand. More specifically, we will first introduce synthetic candidate synthesis which serves as our first FG, before introducing fairness and specifying the relevant techniques used in the CRS pipeline. Finally, we will conclude with the research gap and a summary of how the discussed techniques fit in our CRS.

### 2.1. Data Synthesis

Originally proposed by Rubin in 1993, the synthetic data solution was initially tasked to overcome confidentiality concerns during surveys [8]. Although confidentiality issues have become more important with new stricter European regulations such as the General Data Protection Regulation (GDPR), the current applications of synthetic data have also shown their strength in generating fair and private synthetic data. In fact, synthetic data applications extend far beyond survey data synthesis, use cases range from missing data imputation as well as data augmentation solutions in semi-supervised learning, media applications with image-to-image translation and finally image super-resolution [9].

Data synthesis has evolved from Bayesian bootstrapping methods and predictive posterior distributions to deeper techniques such as Autoencoders (AE), Variational Autoencoders (VAEs), autoregressive models, Boltzmann machines, deep belief networks, and generative adversarial networks (GANs) after the advent of deep learning [10]. These deeper models, more specifically GANs, afforded the synthesis of more complex unstructured data such as images and videos. In the context of this thesis project, GANs will be used to generate tabular (structured) synthetic candidate data.

Despite their popularity, GANs are mainly used for unstructured data synthesis tasks such as image and video synthesis, the generation of synthetic tabular data such as job candidates is not only uncommon from a domain perspective but also from a technical perspective. This is caused by the difficulty of learning discrete features with potentially imbalanced classes. A challenge for which Xu et al. found a solution by integrating a Gumbel Softmax (GS) activation function in their *CTGAN*. The GS is based on the Gumbel-Max trick, a common method for discrete approximation [12].

With the ability to generate categorical features, other issues can hinder the tabular candidate synthesis process. Issues such as input datasets with mixed distributions (as is the case for our input data) can severely affect generative performance. For these problems, Xu et al. propose two solutions: mode-specific normalization for continuous column normalization and conditional sampling to enforce class balancing, both are known problems in discriminatory generative modelling. Therefore, *CTGAN* is an ideal generator for the task at hand as it can balance imbalanced datasets and handle mixtures of data types. Before outlining the fairness-related work, we relate *CTGAN* to our CRS pipeline and discuss its contribution to both the academic and domain gap.

Candidate synthesis is uncommon, although fairness research showed successful use of tabular GANs to generate fair data and more domain-relevant research showed the use of Gaussian copulas for synthetic candidate generation, considerations using *CTGAN*s to support downstream tasks are rare if not unavailable [5, 13]. In the synthetic candidate generation domain, van Els et al. is the unique example in our high risk of harm task. Therefore, the use of GANs, more specifically *CTGAN*s to generate candidates will greatly improve the fairness of our CRS pipeline.

In fact, as outlined by Xu et al., conditional sampling will allow us to synthesize balanced training data with ease which can be used downstream as a fair balanced basis to train candidate-scoring algorithms and mitigate bias; the use of conditional sampling alongside reject sampling (to be introduced in the methodology section) **is how we link candidate synthesis with fairness and ultimately bias mitigation in our end-to-end CRS pipeline**. Therefore, the use of *CTGAN*s is novel in the candidate recommendation domain. With the synthetic **pre-processing** techniques outlined, we will provide an outline of the fairness literature, by focusing more specifically on **post-processing** methods.

## 2.2. Fairness

With the relevant background and related work on candidate synthesis introduced, we now proceed further down our CRS pipeline towards the second FG which will mitigate bias at the **post-processing** level, therefore, after the models are trained on synthetic data to score real candidates. The scored candidates are then evaluated according to a relevant fairness metric and re-ranked using a relevant **post-processing** technique.

Currently, multiple fairness metrics exist, each with their respective strengths and weaknesses. In our case, we only consider demographic parity, which was defined by Kusner et al. as:

- **Demographic Parity**: "A predictor $\hat{Y}$ satisfies demographic parity if $P(\hat{Y}|A = 0) = P(\hat{Y}|A = 1)$." For $A$ representing a sensitive attribute with $a$ levels.

Many other fairness techniques exist, namely the removal of any sensitive attributes. We stress that simply removing sensitive attributes is not guaranteed to remove bias. This process of simply removing protected attributes is known as fairness through unawareness and was shown to perpetuate unfairness [14]. In fact, in our CRS pipeline, we are using the opposite logic to achieve fairness through awareness by explicitely using gender to re-rank candidates in the **post-processing** step.

### 2.2.1. Fairness in Rankings

While demographic parity is useful for quantifying fairness, the enforcement of such rules has yet to be defined. Fairness can be enforced either through a data cleaning process verifying for class imbalances and the existence of sensitive (proxy) variables (**pre-processing**) or modifying model output post-training with approaches such as re-ranking (**post-processing**)[7]. Although we consider the two approaches in this project, the evaluation of our model will follow the SOTA **post-processing** techniques which are presented below.

For our CRS pipeline we will use Geyik et al.'s approach considering it is already used in the HR domain (the task at hand was the recommendation of candidates in *LinkedIn*). Additionally, Geyik et al. achieved SOTA performance with more than a 4-fold reduction in unfairness and a reduction in utility of only 6%. From a research gap perspective, candidate re-ranking is widely used in the industry and researched in Information Retrieval literature. However, despite not being novel in this sub-task, our CRS pipeline fills the research gap by performing the re-ranking of candidates on synthetically trained scoring models.

This is where our end-to-end CRS pipeline contributes to both the domain and the relevant literature, by testing how the combination of candidate synthesis for scoring model training combines with re-ranking methods for a better bias mitigation end-to-end process. **This combination is novel in both the HR domain and in the literature for fairness and generative modelling**.

## 2.3. Summary and Research Gap

The above mini-literature review outlined the different key areas of (candidate) synthesis and fairness processing techniques. As shown, the combination of multiple processing techniques within one CRS pipeline has never been attempted. Therefore, our pipeline is presented as a combination of the presented related work and it will be evaluated based on the output of the candidate rankings. For the evaluation, we will not be comparing our CRS pipeline's *CTGAN* to Xu et al. nor will we be comparing our re-ranker to Geyik et al. as we are using drastically different datasets. Instead we will be developing our own evaluation framework for the candidate data at hand which we will outline in section 3.

The goal of this section was to provide a high-level overview of the literature and techniques used all while exposing the academic gap where our pipeline resides. In the following section, we use the provided background to introduce our experiments with in-depth technical detail and apply the SOTA related work to the candidate recommendation problem with our novel CRS pipeline.

# 3. Methodology

Our CRS follows a point-wise learning to rank approach, where for a given job $j$, we fetch and rank candidates $i$, much like given a query, the goal is to rank documents in the traditional document retrieval scenario. In other words, our recommender system predicts relevance scores $\hat{y}_{i,j}$ given the candidate and job features $X_{i,j}$.

We use real data from an international HR company. For training purposes, the candidate features $X_i$ are associated with a ground truth label $y_{i,j}$ where $y_{i,j} = 1$ if the candidate $i$ has been recruited or shortlisted for a job $j$, and 0 otherwise.

The data used for training is of a structured nature, spanning real-valued, categorical, and binary features. Features correspond to *candidate features* (e.g., job seekers' preferences such as minimum salary, preferred working hours, or maximum travel distance, in addition to data related to their work experience or level of education). *Job features* (e.g., industry of the company, company size, geographical location), and finally *candidate-job features* that represent their overlap (e.g., geographical distance between candidate and job, or a binary feature indicating whether candidate has worked in job's industry before). Much in the same vein that query, document, and query-

document features are designed in a traditional learning to rank for information retrieval-scenario.

### 3.1. Gender balance and synthetic data

Imbalanced data is very common in CRSs, and we focus on gender imbalance for our case, which is common in the job market. To effectively study the issue of imbalance, we construct various explicitly (im)balanced scenarios through a rejection sampling algorithm based on John V. Neumannn's technique [15]. We first sampled rebalanced subsets of the original training data,considering gender as the sensitive attribute *a*. We only considered 2 genders (female, male) as unfortunately our dataset does not contain enough samples of non-binary genders.

To construct our (im)balanced subsets, we randomly sampled job candidates from each job request $j$ with a constrained proportion of candidates from each gender. We generated two datasets with *heavy imbalance* (one with 20% of female candidates, one with 20% of males); two datasets with *minor imbalance* (one with 45% of female candidates, one with 45% of males); and a *balanced* dataset (with 50% of male and female candidates). For each training dataset, 10% of the data points were kept as a held-out test set. To avoid data leakage, all job requests $j$ were unique to the test set. The test dataset sizes in number of unique $< j, i >$-pairs after rejection sampling are shown in Table 1.

| Test Data | Sample Size |
|---|---|
| *heavy imbalance* (20% males) | 38 701 |
| *heavy imbalance* (20% females) | 40 975 |
| *minor imbalance* (45% males) | 48 195 |
| *minor imbalance* (45% females) | 41 972 |
| *balanced* | 48 178 |

**Table 1**
Test set sizes after rejection sampling.

We trained 5 synthetic data models, using each rebalanced dataset as training data for the CTGAN algorithm [11]. We were able to generate balanced synthetic data using the models' conditional sampling parameters. We generated balanced synthetic data where each gender represents 50% of the dataset, for both positive ($y_{i,j} = 1$) and negative ($y_{i,j} = 0$) examples.

The synthetic data generation is our first **fairness gate** (FG) in the CRS pipeline. This FG aims to improve the fairness of candidate scoring $\hat{y}_{i,j}$ by training the CRS on balanced data. The full overview of the experimental pipeline is shown in Figure 1.

### 3.2. Candidate scoring and re-ranking

We trained CRS models to score candidates $i$ by estimating their relevance score $\hat{y}_{ij}$ for the jobs $j$. We trained a total of 10 CRS models, using real or synthetic job candidates as training data (5 datasets each respectively). The jobs for which candidates are scored remain those of the real data, more specifically, the real holdout test data.

We tested the CRS models with their respective holdout test sets, comprising real data with the same gender balance. For each test set, we scored candidates using either the CRS trained with synthetic data or with real data (of the same gender balance), i.e., we use 2 CRS models per each of the 5 test sets, and thus obtain a total of 10 sets of scores. After scoring candidates we rank candidates by descending order of relevance scores, and obtain 10 sets of rankings.

After the candidates are scored and ranked, we introduce our **second Fairness Gate** (FG) at the **post-processing** level of the CRS pipeline. This FG aims to improve the fairness of candidate ranking by using a re-ranking algorithm that interleaves males and females equally at the top ranks (e.g., Figure 2). For our experimental CRS pipeline, we reused the re-ranking algorithm from Geyik et al. [7], and obtained 10 sets of re-rankings (Figure 1).

### 3.3. Metrics and Evaluation

The impact of the re-ranking is evaluated in terms of utility using Normalised Discounted Cumulative Gain ($NDCG$), a common ranking metric to maximise [16]. To measure the impact of the re-ranking, we compared the $NDCG$ scores before re-ranking (by considering the initial ranking as the ideal ranking) and after re-ranking. A lower $NDCG$ score means re-ranking had a negative impact on the original rankings. A higher $NDCG$ score means re-ranking had less impact. As we are considering the impact of the ranking, the $NDCG$ score was calculated after ranking, hence the appearance of only one score. Therefore, we used the $NDCG$ as a single impact metric. The original predicted ranks were used as ground truth (ideal ranking) which was measured against the re-ranked candidates. To ensure the ideal ranks are valid, we have used common classification metrics such as F1 and AUC.

In terms of fairness, we used $NDKL$ (normalized discounted cumulative Kullback-Leibler divergence), a distance metric comparing distribution dissimilarity, such as rank distributions [7].

Here, $NDKL$ calculates the dissimilarity between the distributions of males and females, especially at the top ranks. We consider that demographic parity is achieved when the rank distributions of males and females are similar (i.e., $NDKL = 0$).

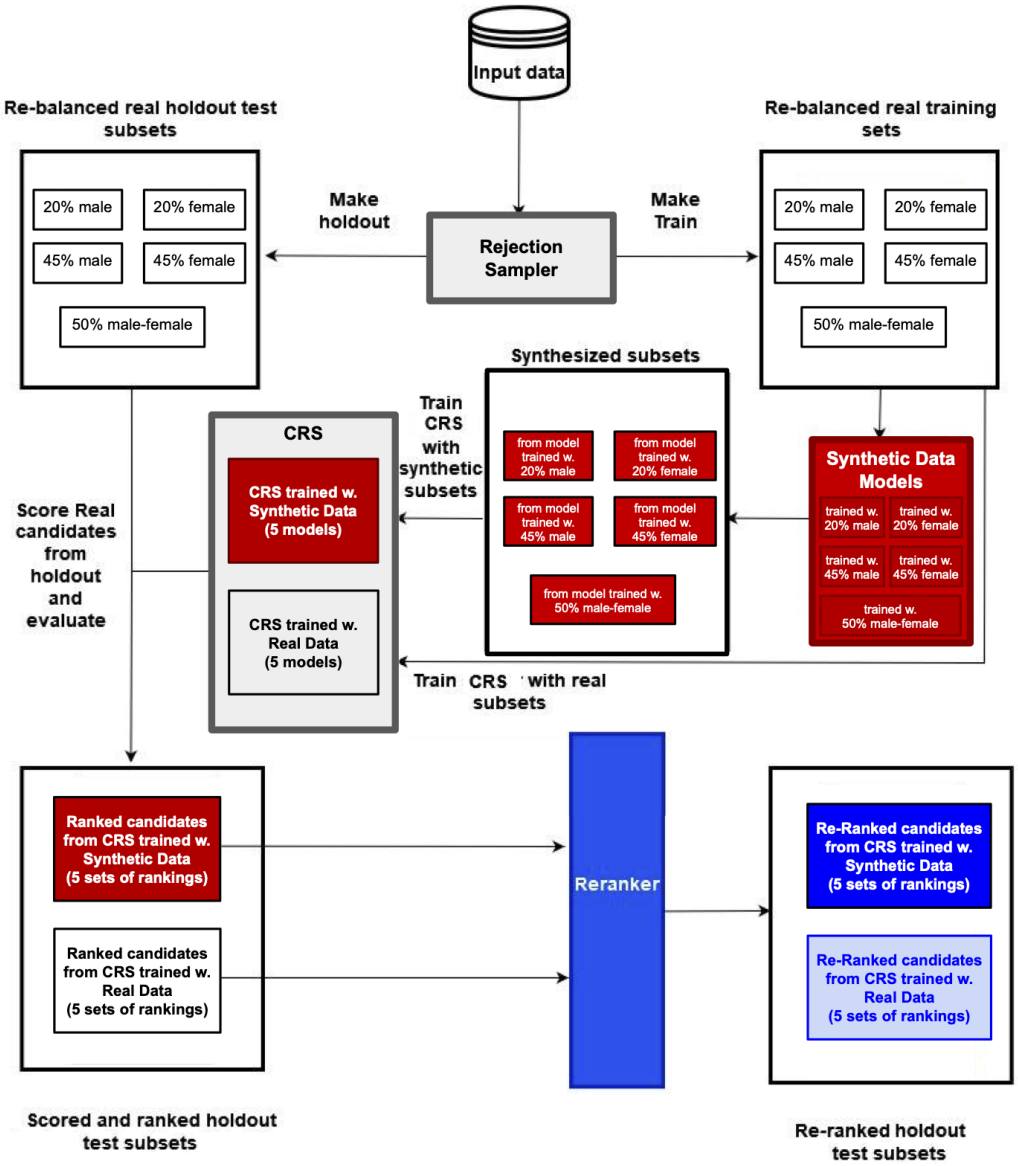


**Figure 1:** Experimental CRS pipeline including bias mitigation techniques at pre-processing and post-processing steps.

## 4. Results and Analysis

We present the results of the CRS that include one, two, or none of our Fairness Gates (FG): re-balancing the training set with synthetic data (1st FG), and re-ranking the job candidates (2nd FG). We consider 3 levels of data imbalance, and summarise the NDCG and NDKL for each level in Table 2.

The *NDCG* difference is noticeable between CRS models trained with real or synthetic datasets (i.e., between pairs of rows in Table 2). For the *heavy imbalance* case, the increase in utility is almost two-fold (+45%).

The *NDKL* difference is very small between CRS models trained with real or synthetic datasets, and shows a negligible improvement of fairness. These results show that **using balanced synthetic data to train CRS mnodels (1st FG) considerably improved utility (*NDCG*) while maintaining the same level of fairness (*NDKL*).**

The *NDKL* decreases before and after ranking (i.e., last two columns in Table 2), showing that the

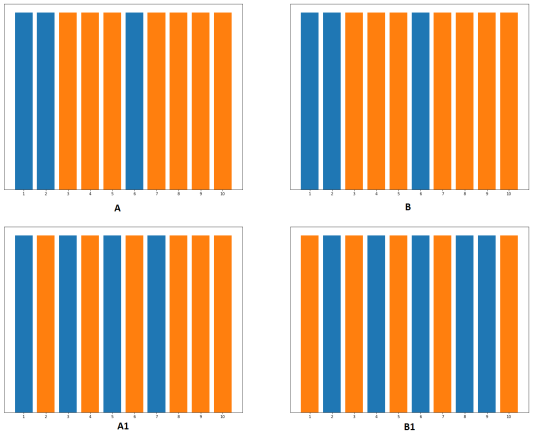


**Figure 2:** Plot displaying the rankings of the top 10 candidates before re-ranking and after re-ranking. The ranks of the candidates are on the x-axis. Female candidates are blue bars, and male candidates are orange bars. The ranking **A** are from a CRS trained on *heavily imbalanced* data, and **A1** represents the re-ranked candidates from **A**. Similarly, **B** and **B1** are the initial and re-ranked rankings for a CRS trained on the *balanced* dataset.

| Ranked Lists | NDCG | NDKL Before re-ranking | NDKL After re-ranking |
|---|---|---|---|
| *Heavy imbalance*: CRS trained w. real data | 0.384 | 0.366 | 0.200 |
| *Heavy imbalance*: CRS trained w. synthetic data | 0.693 (+45%) | 0.358 | 0.197 |
| *Minor imbalance*: CRS trained w. real data | 0.403 | 0.217 | 0.126 |
| *Minor imbalance*: CRS trained w. synthetic data | 0.647 (+38%) | 0.213 | 0.126 |
| *No Imbalance* : CRS trained w. real data | 0.403 | 0.213 | 0.124 |
| *No Imbalance*: CRS trained w. synthetic data | 0.633 (+36%) | 0.206 | 0.124 |

**Table 2**
Average *NDCG* and *NDKL* for ranked list obtained at each level of data imbalance, using CRS trained with real or synthetic data (1st FG), with or without re-ranking (2nd FG).

rank distributions of male and female candidates are more similar after re-ranking. The decrease is of similar magnitude for each level of data imbalance, i.e., whether the CRS model is trained with real or synthetic data. These results show that **using re-ranking at post-processing (2nd FG) equally improved fairness (*NDKL*) whether or not synthetic data was used to train CRS models (1st FG)**.

We also explored the score distributions for male and female candidates. Those attributed by CRS models trained with real data are unevenly skewed toward the left, even in cases where the real data is balanced (*balanced dataset*). However, for CRS models trained with synthetic data, **the score distributions of both genders shift more to the right, creating a more normally-shaped score distribution across both studied genders**.

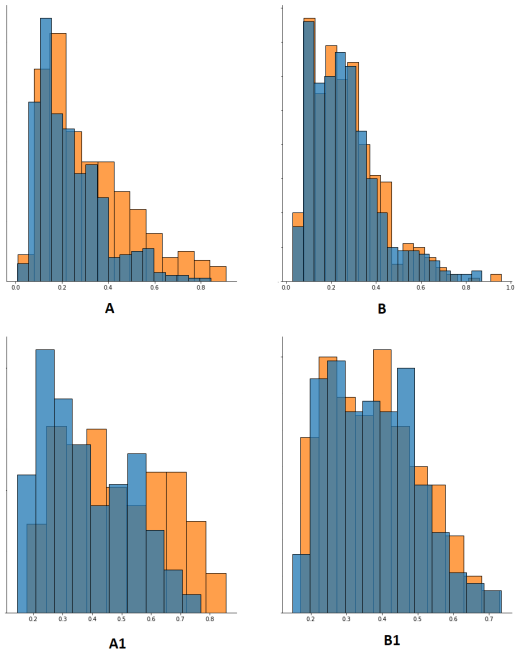


**Figure 3:** Score distribution for male and female candidates. The score assigned to the candidates is on the x-axis, female candidates are in blue while male candidates are in orange. **A** represents a CRS model trained with *heavily imbalanced* real data, and **A1** a CRS trained with synthetic data learned (from a generator trained on *heavily imbalanced* data). **B** and **B1** are the the *balanced* dataset.

## 5. Discussion

Despite the promising results shown in section 4, our CRS pipeline has shown some pitfalls. More specifically, the computation of *NDCG* using ranked candidates as ground truth and only evaluating the re-ranked perfor-

mance can come with additional validity issues. However, it should be noted that these validity issues can be easily averted by adding another $NDCG$ calculation evaluating also non-re-ranked candidates against a ground truth constructed from another holdout set for example.

Additionally, supplementary validation methods could have been considered. For instance, it could have been beneficial to use future $j$, not included in the data, in further evaluations. Statistical tests could have also been conducted, while other user-based approaches, such as an evaluation with recruiters, could have contributed to reinforce the validity of this project. These extra validation steps should be implemented before deploying the fairness mechanisms proposed

Furthermore, some findings were unexplainable with the current analysis. For instance, the $NDKL$ scores for CRSs trained on real *minor imbalanced* datasets are lower than those trained on real *balanced* datasets, which also applies after re-ranking. Although the scores vary by a small margin, such behaviour is difficult to explain considering the complexity of our pipeline, rendering de-bugging tasks equally complex.

Additional unexplainable results are also visible on the synthetic to real comparison with CRSs trained on synthetic datasets such as *heavy imbalance* showing more unfairness by a small margin when compared to real-trained counterparts. These unexplainable findings between real and synthetic subsets are even more puzzling considering, figure 3 shows more balanced scoring for all synthetically-trained CRSs which should result in a lower $NDKL$ score before re-ranking.

Finally, the implementation of demographic parity to enforce equal proportions between genders oversimplifies the complexity of the candidate hiring landscape. This oversimplification can be resolved in future research with a lesser degree of generalizability. Future research can be more specific by adjusting fairness rules to the domain of the job request $j$. For instance, certain jobs such as security personnel can show real-world skewness towards a certain gender. A future CRS pipeline needs to adjust its fairness rules at $j$ level.

Despite these limitations and suggestions for future work, overall, our research successfully showed that the combination of synthetic data and re-ranking was a combination contributing to both fairness and utility even when compared to CRSs trained on real balanced data such as the *balanced* dataset. Therefore, as expected, a combination of pre-processing and post-processing FGs proved to be useful.

## 6. Conclusion

The goal of our CRS pipeline was never to produce SOTA synthetic candidates and recommendations, despite our satisfactory results. The goal was to build a recommendation pipeline using both real and synthetic data to be able to experiment with fair processing techniques and as a result, mitigate bias in candidate recommendations. From this perspective, the double fair-gated CRS pipeline was successfully built and the generation of synthetic candidates was successful, valid and accurate throughout the pipeline.

The generated data has shown to be accurate on all (im)balance levels, validating the expectations on mode-specific normalization and conditional sampling in CTGANs, while also demonstrating the benefits of rejection sampling methods in re-balancing imbalanced data and using the synthetic candidates generated from it to score real (im)balanced test subsets fairly. From a fairness perspective, it was also shown how scorers trained on synthetic candidates outperform scorers trained on balanced real data from a utilitarian perspective.

Although the issues outlined in section 5 concerning the lack of measurement of pre-re-ranked utility raise some minor validity concerns, the evidence shows **how synthetically-trained CRSs provide fair, useful candidate recommendations when integrated in such a pipeline**.

## 7. Future Work

In future work, the recommendations shared in the discussion can be considered. More specifically, the use of additional evaluation methods with human-in-the-loop evaluation using recruiters or the use of future requests to test the CRS pipeline.

Additionally, future researchers should also consider the use of less data-greedy rejection sampling techniques as we have lost more than 80% the amount of the holdout information we had at the start of the pipeline. This can either be resolved with more elegant rejection sampling constraints, the use of larger datasets or data-augmentation techniques through synthetic data for instance. The latter could have been considered in this project if it was within the scope of our research.

Finally, with a solved data scarcity problem future researchers can consider the discussed domain-adjustable fairness rules for more specific fairness constraints to overcome real-world skewness.

## 8. Acknowledgements

We acknowledge the University of Amsterdam - Master programme Information Studies for creating the conditions to perform this research and for financially supporting this publication.

## References


[1] A. Chouldechova, Fair prediction with disparate impact: A study of bias in recidivism prediction instruments, Big Data 5 (2017) 153–163. URL: https://doi.org/10.1089/big.2016.0047. doi:10.1089/big.2016.0047. arXiv:https://doi.org/10.1089/big.2016.0047, pMID: 28632438.

[2] J. Buolamwini, T. Gebru, Gender shades: Intersectional accuracy disparities in commercial gender classification, in: S. A. Friedler, C. Wilson (Eds.), Proceedings of the 1st Conference on Fairness, Accountability and Transparency, volume 81 of *Proceedings of Machine Learning Research*, PMLR, 2018, pp. 77–91. URL: https://proceedings.mlr.press/v81/buolamwini18a.html.

[3] S. Hajian, J. Domingo-Ferrer, A methodology for direct and indirect discrimination prevention in data mining, IEEE Transactions on Knowledge and Data Engineering 25 (2013) 1445–1459. doi:10.1109/TKDE.2012.72.

[4] A. Prince, D. Schwarcz, Proxy discrimination in the age of artificial intelligence and big data, Iowa Law Review 105 (2020) 1257–1318. Publisher Copyright: © 2020 University of Iowa. All rights reserved.

[5] A. Rajabi, O. O. Garibay, Tabfairgan: Fair tabular data generation with generative adversarial networks, arXiv preprint arXiv:2109.00666 (2021).

[6] Y. Li, H. Chen, S. Xu, Y. Ge, Y. Zhang, Towards personalized fairness based on causal notion, CoRR abs/2105.09829 (2021). URL: https://arxiv.org/abs/2105.09829. arXiv:2105.09829.

[7] S. C. Geyik, S. Ambler, K. Kenthapadi, Fairness-aware ranking in search & recommendation systems with application to linkedin talent search, 2019. URL: https://doi.org/10.1145/3292500.3330691. doi:10.1145/3292500.3330691.

[8] D. B. Rubin, Discussion statistical disclosure limitation, Journal of Official Statistics 9 (1993) 461–468.

[9] I. Goodfellow, Nips 2016 tutorial: Generative adversarial networks, 2017. URL: https://arxiv.org/abs/1701.00160. doi:10.48550/ARXIV.1701.00160.

[10] A. C. Ian GoodFellow, Yoshua Bengio, Deep Learning, 1st ed., MIT Press, Cambridge, Massachusetts, United States, 2016.

[11] L. Xu, M. Skoularidou, A. Cuesta-Infante, K. Veeramachaneni, Modeling tabular data using conditional gan, 2019. URL: https://arxiv.org/abs/1907.00503. doi:10.48550/ARXIV.1907.00503.

[12] E. Jang, S. Gu, B. Poole, Categorical reparameterization with gumbel-softmax, 2016. URL: https://arxiv.org/abs/1611.01144. doi:10.48550/ARXIV.1611.01144.

[13] S.-J. van Els, D. Graus, E. BeauxisAussalet, Improving fairness assessments with synthetic data: a practical use case with a recommender system for human resources, 2022.

[14] M. J. Kusner, J. R. Loftus, C. Russell, R. Silva, Counterfactual fairness, 2018. arXiv:1703.06856.

[15] J. Neumann, Various techniques used in connection with random digits, National Bureau of Standards, Applied Math Series 12 (1951) 768–770.

[16] K. Järvelin, J. Kekäläinen, Cumulated gain-based evaluation of ir techniques, ACM Transactions on Information Systems (TOIS) 20 (2002) 422–446.