

Improving Fairness Assessments with Synthetic Data: a Practical Use Case with a Recommender System for Human Resources

SUBMISSION NO. 6

In human resources, recommender systems have been widely adopted for selecting sets of candidates for a job. In such a high impact application, algorithmic bias may create large scale discrimination with life-changing impact. We present a set of methods for assessing such algorithmic bias by using synthetic data to improve the size and representativity of the test set. We focus on bias in the estimation of a candidate's relevance for a job, and arising from the systematic under-estimation of certain groups of candidates. This study finds that obtaining high-quality and privacy-preserving synthetic data remains challenging. However, the ability to synthesise as many data points as desired provides opportunities for improving the statistical significance of fairness metrics. We report our initial results when applying a synthetic data model in a practical use case with real-life data. Finally, we highlight key research challenges for assessing the statistical validity of fairness assessments based on synthetic data.

CCS Concepts: • **Information systems**;

Additional Key Words and Phrases: Fair AI, Recommender System, Human Resources, Synthetic Data

ACM Reference Format:

Submission No. 6., . 2018. Improving Fairness Assessments with Synthetic Data: a Practical Use Case with a Recommender System for Human Resources. In *Woodstock '18: ACM Symposium on Neural Gaze Detection, June 03–05, 2018, Woodstock, NY*. ACM, New York, NY, USA, 5 pages. <https://doi.org/10.1145/1122445.1122456>

1 INTRODUCTION

Recommender Systems (RS) shape much more than our information consumption. For instance, they pre-select and rank the people we could befriend or date, grant resources to, or hire for a job. In such applications, algorithmic bias can have critical impacts on society and perpetuate harmful discrimination at large scale. For instance, discrimination can occur when specific social groups are systematically considered less relevant than other. In Human Resources (HR) and recruitment, implicit bias is known to be at play, i.e., humans unconsciously associate prejudices to social groups [1]. Ground-truth datasets may reflect such prejudices, and RS trained on them can make the discrimination larger and more systematic.

To curb such algorithmic harm, an important first step is to measure the bias that impacts vulnerable populations, e.g., those identifiable by their *protected* attributes such as gender or ethnicity [2]. Discrimination may be further amplified for populations combining several *protected* attributes. This issue of *intersectionality* [3] arises from social systems that harm marginalized groups, and have interlocking effects for people at the intersection of several groups.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

© 2018 Association for Computing Machinery.

Manuscript submitted to ACM

Two key issues arise when collecting datasets to assess algorithmic bias and intersectionality: sensitive information on *protected* features may be inaccessible due to privacy or legal constraints; when accessible, limited data samples may be available e.g., for minorities. Synthetic data generation [4] may help overcoming such limited data availability. Once a synthetic data model is built from real data, new samples can be generated without limits on sample sizes, and without disclosing the personal data of actual people.

In this paper, we explore the potential of synthetic data for assessing RS fairness. We focus on bias in the relevance scores estimating a candidate's suitability for a job. We apply a Copula-based synthetic data model [5] to the original data of an international HR company, and evaluate a test version of their RS that aims to rank candidates for a vacancy.

Our results show that synthetic data may reliably capture the relationships between numerical features, but categorical features seem more challenging. The synthetic data may provide additional samples to address data scarcity (e.g., for minorities) and data imbalance. But these may not suffice to improve the statistical reliability of fairness assessments. With synthetic data, we may assess differences in relevance scores using larger samples. But the synthetic data samples may not provide more reliable assessments. We conclude by highlighting the challenges for assessing the statistical reliability of fairness assessments performed on synthetic data.

2 USE CASE

We used a pre-trained RS from an international recruitment company. The RS predicts the relevance (scores \hat{y}_{ij}) of candidate i for job j given the candidate features X_i with $\hat{y}_{ij} = f(X_i)$. The predicted scores can be under- or over-estimated compared to the ground-truth scores y_{ij} as measured by the residual error $e_{ij} = \hat{y}_{ij} - y_{ij}$. The ground-truth scores are $y_{ij} = 1$ when a candidate i had a successful interview for a job j (i.e., when the first interview leads to a follow-up, term negotiation, or second interview). The ground-truth scores are $y_{ij} = 0$ when a candidate i had the basic requirements for a job j (e.g., with rules checking the required diploma or years of experience) but had no successful interview. Candidates without the basic requirements for a job j are not included in the ground-truth, and would not be proposed to recruiters by the RS. The ground-truth scores are indicators of the promising and relevance of candidates as described by recruiters, and are sufficient for the purpose of our study. However, the ground-truth scores may be biased in many ways (e.g., sampling bias, implicit bias, recruitment errors).

The job candidates have 3 *protected* features for their age, gender, and nationality. They are not included in the 111 features X_i used by the RS to predict relevance scores. We considered two genders, six age groups [6], and two nationalities (Dutch or non-Dutch). We thus had 24 possible combinations of protected features, representing social groups G_k with $k \in \{1, 2, \dots, 24\}$. We conducted a basic fairness assessment that estimate how relevance scores differ between candidates who are equivalently qualified (i.e., same true score $y_{ij} = s$ with $s \in \{0, 1\}$), but are from different social groups G_k . We consider that potential discrimination occurs if the mean error $\bar{e}_{ij,k,s}$ for candidates from group G_k with true score s differ from that of candidates with the same true scores s but from another group G_l . To identify significant differences between two groups, basic *t-tests* can assess the null hypothesis H_0 that their mean errors are equal ($\bar{e}_{ij,k,s} = \bar{e}_{ij,l,s}$) against the alternative hypothesis H_1 that mean errors differ ($\bar{e}_{ij,k,s} \neq \bar{e}_{ij,l,s}$).

3 SYNTHETIC DATA

We studied the recommendations generated for two job types in the Netherlands: "technicians and associate professionals", and "clerical support workers". We trained a synthetic data model using a random sample of 5 830 job offers j , and 4 3595 candidates i who had successful interviews (true score $y_{ij} = 1$), or not but met the basic job requirements (true score $y_{ij} = 0$).

We used the Synthetic Data Vault (SDV) method based on multivariate Gaussian Copulas [5]. It requires that categorical features be made numerical within $[0,1]$, but one-hot encoding is not required for ordinal or multiclass features [5, Fig.6]. Missing values are considered revealing information and modelled as null values.

We generated synthetic data that represented job candidates from all combinations of protected features and true score (0 or 1). We aimed at generating enough samples of candidates to observe significant differences in mean residual errors $\bar{e}_{ij,k,s}$ between social groups G_k , when performing independent two-sample t-tests (for unequal sample sizes and similar variances). We used a power analysis to determine the sample size required for a significance level $\alpha = 0.05$ and power $\beta = 0.2$. We used Cohen's d for the effect size, and the residual errors e_{ij} were measured in the original data.

We generated a total of 521,074 synthetic data points representing job candidates. Due to privacy and security concerns, we are unable to report the sample sizes for each group G_k and true score y_{ij} , in the original or synthetic data.

4 EVALUATION METRICS

Utility metrics measure whether the synthetic data is similar to the original data, and provides similar data analysis results [4] [7]. We applied several utility metrics, which can be grouped into *global* and *analysis-specific* utility metrics.

Global Utility Metrics estimate the similarity of the feature distributions in the original and synthetic data [7]:

- *Statistical metrics* quantify the probability that a feature in the synthetic data follows the same distribution as in the original data. We applied statistical tests that are specific to numerical and categorical variables, i.e., respectively, the Kolmogorov–Smirnov (KSTest) and Chi-Squared (CSTest) tests.

- *Likelihood metrics* fit to a probabilistic model and calculate the probability that the synthetic data belong to the fitted distribution. We applied Gaussian Mixture models fitted to the original data (GMLogLikelihood).

- *Distinguishability metrics* are based on classifiers trained to distinguish the synthetic data from the original data. We trained logistic regression (LogisticDetection) and SVC (SVCDetection) classifiers, and measured the proportion of synthetic data points that are classified as real data.

Analysis-Specific Utility Metrics [4] focus on specific analysis to perform with synthetic data. If the results obtained from original and synthetic data are equivalent, then synthetic data obtains a high *utility*. Analysis-specific metrics include privacy metrics and the Pearson correlation coefficient. We computed the latter for the 20 most relevant features in X_i , as selected by a domain expert from an international HR company that uses such RS.

Privacy metrics estimate the possibility to deduce real feature values of the original data from the synthetic data. We applied the Categorical Correct Attribution Probability (CategoricalCAP) for binary protected features (gender, nationality). For age, we used the Multi-Layer Perceptron regression privacy metric (NumericalMLP) [8].

5 EVALUATION RESULTS

We evaluated the synthetic data with utility and privacy metrics measured in a sample of 10,315 job candidates from the original data. For **analysis-specific utility**, the heatmap of Pearson correlation coefficients (Fig. 1 right) shows that the relations between data features are well-preserved in the synthetic data. For **global utility**, the results vary notably. Statistical tests show that the distributions of categorical features differ between the original and synthetic data (CSTest), while the distributions of numerical features are much more similar (KSTest). But the Gaussian Mixture model (GMLogLikelihood) fitted on the original data would fit the synthetic data with a probability of 0.5 only. The *distinguishability metrics* reflect these mixed results: one classifier (LogisticDetection) finds the synthetic data entirely indistinguishable from the original, while only about 59% of the synthetic data is indistinguishable for the other classifier (SVCDetection). These inconsistencies may arise from the lower quality of categorical features in the synthetic data

(CSTest), e.g., due to missing or erroneous values in the ground-truth, or to the encoding of multiclass features [5, Fig.6]. They must be investigated in future work.

For **privacy**, the results are very poor. Binary features (age, gender) are not protected (CategoricalCAP) while multiclass features offer little protection (NumericalMLP). However, this may be addressed by adding more noise (e.g., with differential privacy) and assessing the risks of attacker to access the information necessary for deducing the protected features.

Global Utility Metrics			
Metric	Value	Min-Max	Goal
CSTest	0.2879	0 to 1	Maximize
KSTest	0.7557	0 to 1	Maximize
GMLogLikelihood	0.5	0 to 1	Maximize
LogisticDetection	1	0 to 1	Maximize
SVCDetection	0.5874	0 to 1	Maximize
Privacy Metrics			
CategoricalCAP	0	0 to 1	Maximize
NumericalMLP	0.2391	0 to 1	Maximize

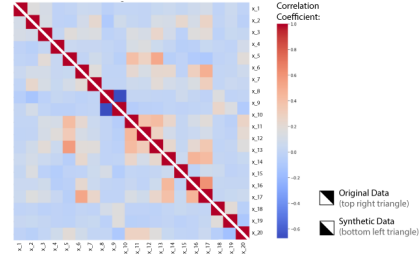


Fig. 1. Evaluation metrics (left) and correlation matrix (right) for the original and synthetic data (bottom and top triangle, respectively).

6 RELIABILITY OF FAIRNESS ASSESSMENTS

The synthetic data we obtained is not of optimal quality, and may not improve the reliability of fairness, nor protect the privacy of job candidates. However, even if the synthetic data had excellent results on these utility metrics, the reliability of fairness assessments would not be guaranteed. First of all, the original ground-truth data may be biased, and its true scores (0 or 1) are only simple indicators of the past interactions between recruiters and candidates. They do not convey the real-life and complex qualities of candidates, nor their real-life suitability for a specific job.

Beside these fundamental and critical issues, we examine bias assessment from a statistical perspectives. It remains challenging to identify significant differences in mean errors using synthetic data (e.g., to test whether $\bar{e}_{ij,k,s} \neq \bar{e}_{ij,l,s}$). Fundamental questions remains, even under the assumption that the ground-truth is representative of the actual populations of job candidates (and jobs): Are statistical tests valid when performed on synthetic data? If a statistical tests give different results with synthetic and original data, which test is most reliable?

The answers to these questions depend on the statistical tests and the synthetic data models that are applied, and their underlying assumptions. The t-test we used requires normally distributed samples, which may be assumed from the central limit theorem as we deal with means $\bar{e}_{ij,k,s}$. However, this test does not control for score variance across jobs j , and the synthetic data may better represent the variance of mean errors $\bar{e}_{ij,k,s}$ across jobs. Kolmogorov-Smirnoff test is an alternative that does not assume the normal distribution of the sample means. An analysis of variance (ANOVA) may be more appropriate in this case, to control score variance across for jobs j , but the number of candidates per job may be very limited (e.g., only a few candidates with ground-truth score $y_{ij} = 1$).

It is critical to choose the right statistical tests, and to verify that their assumptions (thus their interpretation) are compatible with the factors of variability in the specific use case. This applies to tests performed on either original or synthetic data, but for the latter, specific factors of variability should be considered. The type of synthetic data model (e.g., based on Copulas [5], Bayesian networks [9], regressions [10], non-parametric trees [11]), and the quality of the model fitting, influence the variability of the synthetic data features.

Specific statistical metrics may be designed to assess the variability of synthetic data features, and its impact on the predicted scores \hat{y}_{ij} . The variability of candidates with ground-truth scores y_{ij} of either 0 or 1 should be assessed separately. These errors have different signs (candidates with scores $y_{ij} = 1$ can only be under-estimated, and $y_{ij} = 0$ over-estimated) and social impacts (candidates with scores $y_{ij} = 1$ suffer a different loss of opportunity from bias, and systematically lower scores and rankings, than candidates with $y_{ij} = 0$).

7 DISCUSSION

Synthetic data offers interesting perspectives for enabling fairness assessments while preserving the privacy of the data subjects in the original data. In HR applications, algorithmic bias in scoring and ranking job candidates must be assessed, and this may be legally required. The use of RS for HR must also comply with GDPR restrictions regarding the data collection, for any AI model trained using personal data, including synthetic data models.

The international HR company who manages the ground-truth data used in this study has high standards regarding the privacy and security of the job candidates. It was thus not permitted to reveal information that may facilitate attacks, or undermine the application of differential privacy. Hence, to research and publish detailed reports on the statistical validity of bias assessments from synthetic data, it may be preferable to use datasets that have few or no privacy risks (e.g., open-source data, no personal data, or other domains than HR).

Further research is needed to provide insights on the statistical validity of fairness assessments performed on synthetic data. The problem is complex because it combines two AI models: one generating synthetic data and one predicting relevance scores from synthetic and real data. Furthermore, different *utility* metrics and statistical tests may be considered to frame and describe the problem. One can even train the synthetic data model to generate predicted relevance scores as well, for additional comparisons.

Until further statistical research is conducted, synthetic data can be used to obtain samples that address data imbalance, with rare class or intersectional social groups. Synthetic data may improve the representativity of social groups, to limit discrimination. Synthetic data can also be used simply for anonymisation, without re-balancing the distributions and proportions of social groups. The requirements for anonymising and balancing data may be deemed less strict than the requirements for applying statistical tests for fairness assessment. However, caution must be applied there too, as synthetic data may introduce additional bias (e.g., due to quality issues, as shown with *utility* metrics).

REFERENCES

- [1] H. A. Vuletich and B. K. Payne, "Stability and change in implicit bias," *Psychological science*, vol. 30, no. 6, pp. 854–862, 2019.
- [2] S. Hooker, "Moving beyond "algorithmic bias is a data problem"," *Patterns*, vol. 2, no. 4, p. 100241, 2021.
- [3] J. R. Foulds, R. Islam, K. N. Keya, and S. Pan, "Bayesian modeling of intersectional fairness: The variance of bias★," in *Proceedings of the 2020 SIAM International Conference on Data Mining*. SIAM, 2020, pp. 424–432.
- [4] M. Hittmeir, A. Ekelhart, and R. Mayer, "On the utility of synthetic data: An empirical evaluation on machine learning tasks," in *Proceedings of the 14th International Conference on Availability, Reliability and Security*, 2019, pp. 1–6.
- [5] N. Patki, R. Wedge, and K. Veeramachaneni, "The synthetic data vault," in *2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*. IEEE, 2016, pp. 399–410.
- [6] CBS, "Arbeidsdeelname; kerncijfers," 2021, <https://opendata.cbs.nl/>, Last accessed on 2021-08-15.
- [7] M.-J. Woo, J. P. Reiter, A. Oganian, and A. F. Karr, "Global measures of data utility for microdata masked for disclosure limitation," *Journal of Privacy and Confidentiality*, vol. 1, no. 1, 2009.
- [8] Documentation, "SDV metrics library," 2021, https://sdv.dev/SDV/user_guides/evaluation/single_table_metrics.html, Last accessed on 22-12-2021.
- [9] H. Ping, J. Stoyanovich, and B. Howe, "Datasyntesizer: Privacy-preserving synthetic datasets," in *Proceedings of the 29th International Conference on Scientific and Statistical Database Management*, 2017, pp. 1–5.
- [10] B. Nowok, G. M. Raab, and C. Dibben, "Synthpop: Bespoke creation of synthetic data in r," *Journal of statistical software*, vol. 74, no. 1, pp. 1–26, 2016.
- [11] B. Nowok, "Utility of synthetic microdata generated using tree-based methods," *Administrative Data Research Centre, University of Edinburgh*, 2015.