Comparing Equity and Effectiveness of Different Algorithms in an Application for the Room Rental Market

David Solans
david.solans@upf.edu
University Pompeu Fabra,
Barcelona, Spain

Francesco Fabbri
francesco.fabbri@eurecat.org
Eurecat, Barcelona, Spain
University Pompeu Fabra,
Barcelona, Spain

Caterina Calsamiglia
caterina.calsamiglia@barcelona-ipeg.eu
ICREA at IPEG,
Barcelona, Spain

Carlos Castillo
chato@acm.org
University Pompeu Fabra,
Barcelona, Spain

Francesco Bonchi
francesco.bonchi@isi.it
ISI Foundation, Turin, Italy
Eurecat, Barcelona, Spain

ABSTRACT

Machine Learning (ML) techniques have been increasingly adopted by the real estate market in the last few years. Applications include, among many others, predicting the market value of a property or an area, advanced systems for managing marketing and ads campaigns, and recommendation systems based on user preferences. While these techniques can provide important benefits to the business owners and the users of the platforms, algorithmic biases can result in inequalities and loss of opportunities for groups of people who are already disadvantaged in their access to housing.

In this work, we present a comprehensive and independent algorithmic evaluation of a recommender system for the real estate market, designed specifically for finding shared apartments in metropolitan areas. We were granted full access to the internals of the platform, including details on algorithms and usage data during a period of 2 years.

We analyze the performance of the various algorithms which are deployed for the recommender system and asses their effect across different population groups.

Our analysis reveals that introducing a recommender system algorithm facilitates finding an appropriate tenant or a desirable room to rent, but at the same time, it strengthen performance inequalities between groups, further reducing opportunities of finding a rental for certain minorities.

CCS CONCEPTS

• **Information systems → Recommender systems**.

KEYWORDS

performance, demographics, fairness, recommender system

AIES '21, May 19–21, 2021, Virtual Event, USA.
© 2021 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 978-1-4503-8473-5/21/05…$15.00
https://doi.org/10.1145/3461702.3462600

ACM Reference Format:
David Solans, Francesco Fabbri, Caterina Calsamiglia, Carlos Castillo, and Francesco Bonchi. 2021. Comparing Equity and Effectiveness of Different Algorithms in an Application for the Room Rental Market. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society (AIES '21), May 19–21, 2021, Virtual Event, USA.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3461702.3462600

1 INTRODUCTION

Two-sided sharing economy platforms have changed how business is conducted in a multitude of domains. They have been particularly disruptive in the real-estate sector where platforms such as Airbnb have changed the status quo. These platforms typically involve three types of stakeholders: (i) providers of items/services, (ii) customers seeking to acquire from the providers, and (iii) the platform itself, which intermediates and matches providers and customers based on their preferences. The explosive growth of these platforms in the real estate sector has been at the core of various political battles at some of the largest cities in the world. Advocates of the sharing economy argue about the benefits they can bring to societies, such as extra income, better distribution and allocation of resources, and the creation of new opportunities for cities and municipalities.[1] On the other hand, critics argue that the costs generated by the platforms surpass their benefits by far: they are very appealing business options so that the main side effect of their wide adoption is that they worsen what is an already troublesome housing shortage in particularly attractive areas, driving up rental prices and, ultimately, boosting gentrification. Concerns also exist about the potentially discriminatory impact of their algorithms.

In this work, we focus on the latter problem. Specifically, we present a comprehensive and independent algorithmic evaluation of a recommender system of a platform used in the real state market,[2] designed specifically for finding shared apartments in metropolitan areas. Our examination enjoys full access to the internals of the platform, including details on algorithms and usage data during a period of 2 years. More in detail, the platform aims to help *listers*, i.e., landlords/landladies or room owners, find appropriate *seekers*,

[1] Airbnb study: Airbnb related activities contributed with up to 175M$ to the city of Barcelona. https://www.airbnb.es/press/news/new-study-airbnb-community-contributes-175-million-to-barcelona-s-economy
[2] Company name omitted.

i.e., users looking for a room to rent. The recommender system facilitates matching and interaction between seekers and listers, with profile-based matching functionalities resembling those of dating platforms [7]. Listers can "like" the profiles of seekers and send a request to them. Seekers can accept such requests in case they like the offered room. If a lister sends a request to a given seeker and the latter gives a positive response to it, then a *match* occurs, which lets them talk through an in-app chat service to arrange a meeting and potentially sign a rental contract.

The platform mediates the connections between providers (listers) and customers (seekers), and as a mediator it has the potential to either facilitate or hamper the emergence of societal biases. Indeed, the bias against certain minorities, if left unmitigated, can be amplified through its recommendations [13]. These biases are particularly dangerous in this sector, where the fundamental right to adequate housing [24] might be compromised.

1.1 Research objectives and findings

The main goal of our algorithmic evaluation is to identify and quantify existing biases in different versions of the platform, showing the trade-offs and potential harms of introducing a machine learning based functionalities, also accounting for the different recommender systems used during the application life-cycle. In contrast with most previous work, our research focuses on the biases exhibited by the system through its recommendations, instead of analyzing how the users behave on the platform [1].

The particular design of the platform, with a baseline method running permanently, executed together with ML-based methods evolving over time, allowed us to extract conclusions in comparison to the baseline.

Our findings show that the introduction of a ML-based algorithm increases the probability of *matching* for the majority of users. This means that the recommendation system effectively facilitate the finding of room-mates or flat-mates.

At the same time, the ranking algorithms utilized in the application exhibit various types of inequalities in terms of performance, significantly affecting the experience and opportunities of some groups of users. Among other differences, the system performance varies across demographic groups based on self-declared gender, sexual orientation, age, and main spoken language. Moreover, we observe that minority groups – groups already disadvantaged or with smaller prevalence in the population – experience lower performance of the system or more differences on its functioning, depending on the particular model they are exposed to.

Section 3 provides the details about the platform and the setting for our analysis. It also present the specific research questions that are addressed in the remainder of this paper. Section 4 presents the methodology and the specific utility metrics adopted. Section 5 describes our dataset and provides some general statistics. Finally, Section 6 present in details our experimental results and findings.

The next section describes previous work related to the analysis presented here and provides some background.

2 BACKGROUND AND RELATED WORK

2.1 Access to housing

Experiments conducted throughout the last decades reveal discriminatory behaviors and practices that negatively affect minorities when trying to buy or rent property. Chambers and Conway (1992) debate the idea of *sustainable livelihoods*, that as they explain, require social equity among other things to achieve sustainability. They expose their ideas with a special focus on the rural poor and other minorities. Turner et al. (2002) describes a series of experiments in 23 metropolitan areas in the United States, revealing serious differences between white and minority citizens on different aspects related to access to housing for renting or buying. Wachter and Megbolugbe (1992) show that there are persistent differences in home ownership rates across racial and ethnic groups in the US.

More recently, an experiment conducted by the Barcelona city hall showed how prejudices decrease the opportunities of finding housing to buy or rent for some groups. In particular, it was observed that LGBTQ seekers or those with Arabic sounding names had a lower chance of being scheduled for visiting a flat [5].

In contrast with these previous works, our experiments are based on an online platform in which the contact between users is mediated and influenced by a recommendation algorithm. Although the observed behaviour in the system could be a mirror of societal biases contained in the training data of the machine learning system, those biases, if not mitigated, can be amplified by an algorithm.

2.2 Algorithmic fairness in double sided markets

Analyzing the case of Airbnb, Quattrone et al. (2016) outlined the difficulties of creating regulatory policies in a changing environment. They collected a set of recommendations for regulating Airbnb, contributing to the general idea of "algorithmic regulation", which advocates for the analysis and use of large sets of data to produce evidence-based regulations that are responsive to real-time demands. Sühr et al. (2019) analyze a double-sided market in the context of ride hailing platforms, giving an special emphasis to the role of the riders (producers). Hutson et al. (2018) analyze a similar setting, in their case online dating apps, revealing different inequities based on race and/or sexual orientation.

Our work contributes to this research as the first work that studies how different versions of a system facilitate the goals and preferences of users in different sides of the market. Also, we quantify the effects of using a ML-based algorithm in comparison with a rules-based random baseline.

2.3 Algorithmic Auditing

In many cases, perhaps in most cases, designers of computational systems fail to include accountability and transparency mechanisms "by design" [12]. In this context, Algorithmic Auditing allows us to uncover and understand potential sources of discrimination driven by such algorithmic-based decision-making, as a post-hoc solution to audit the system behaviour in its past executions.

Early research on this topic includes a set of methods to detect discrimination in online platforms [19]. Eslami et al. (2017) show

how to detect and quantify a rating algorithm's bias using cross-platform audit techniques in the context of hotel rating platforms. Their work identified systematic differences of ratings between three existing platforms and revealed how bias awareness can shift users' attention from their own experience to the system as a whole, even trying to open the black-box by gaming the rating system. This work also introduces a taxonomy of methods for algorithmic auditing that frames our work as a *within-platform* study. Galdon Clavell et al. (2020) describe an auditing of an application used to promote well-being between its users. This work discusses the issue of not collecting sensitive data, as required by the GDPR and the data minimization principle. This might prevent researchers from detecting biases against protected groups.

In the specific context of search engines auditing, Mehrotra et al. (2017) proposes a methodology for measuring differential satisfaction across demographics. Based on the proposed methodology, they conduct an external auditing of a search engine based on a dataset collected by a third party composed by search queries issued to the platform in a short period of time. Their analysis shows significant differences in usage patterns and evaluation metrics for different demographic groups, mainly based in age and gender. In the domain of access to housing, Asplund et al. (2020) perform an algorithmic auditing of received user ads and of the ordering of recommendations in different housing portals in the U.S. They use a strategy based on "sock puppets," creating automatic systems that interact with the platform under fake user profiles, concluding that there are not statistical significant differences between results shown for profiles simulating different age or race. In the topic of Policy Learning in Raking, Singh and Joachims (2019) proposes a theoretical methodology to optimize not only for the utility of the rankings for the users, but also considering fairness constrains of exposure with respect to the ranked items. Their work studies the relation between an allocation metric (normalized cumulative gain) and group disparity, measured in terms of item exposure, proposing the inclusion of exposure-allocation constrains in the learning.

In our work, we use an "open box" (or "white box") algorithmic review approach to evaluate a system's performance across various groups and different algorithms. As main differential feature w.r.t previous work, having access to the internals of the application, allow us to conduct an empirical analysis based on the real profiles of the user, without the necessity of creating "sock puppets" or fake profiles for that. Also, we do not treat the system as a monotonic black-box, but consider different versions and their implications to the users exposed to each of them.

3 SETTING

The platform analyzed in this work corresponds to a system that aims to help matching users having available rooms in their flats, with potential new tenants or flat-mates/room-mates. This setting can then be described as a two sided market, where *listers* supply rooms that are in demand by *seekers*. Most of the interactions are done through a mobile app that offers a recommendation list for the listers.

As depicted in Figure 1, listers receive recommendations in the form of an ordered list of ~20 recommended seeker profiles. These profiles come from a pool of qualifying profiles (e.g., seekers searching for a room in the area where the lister's room is located), from where recommendations are selected. This selection might include profiles selected by the baseline method interleaved with profiles prioritized by a ML-based recommender system. This allows the platform to monitor in real time the performance of each of the RecSys versions compared to the random group, using a within-subjects [18] A-B testing [9]. The same user can be exposed to A or B treatments in different visits or receive recommendations given by A and B in a given recommendation list.

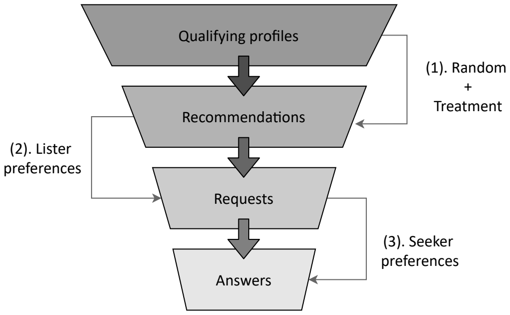


Figure 1: Platform's recommendation pipeline.

In the following, we refer to the baseline system as *random* and to each of the ML-based systems as a *ranking*, given that their main difference is how they rank qualifying seekers. We analyze the performance of *random* and *ranking* separately to understand their differences, and the implications of introducing a ML-based system.

Once the list of recommendations is shown to the listers, they select and can send a request to a subset of seekers according to their own preferences. After the listers send a request, seekers receive a notification for each of them. These requests can, in turn, be accepted or rejected by seekers.

In the following, we consider all the ranking systems together as opposed to the random system. However, we note that different recommendation models that were developed at different points within the life cycle of the platform are used, and their training sets are slightly different. Although it is something outside the scope of the present paper, and a limitation of our work, each new version of the ranking system may have been influenced by the behavior of older versions, and this could lead to feedback loops amplifying biases for each new version of the system [10]. For the purpose of this study, we compare the performance of the different models, including the random system, as isolated instances, whereas their recommendations can appear together in the recommendation lists. However, each ranking model is optimized for the same objective: to maximize the expected probability of a match.

In this setting, biases can be observed directly (i) in the ranking produced by each system, (ii) in the lister preferences when selecting among the ranked items, or (iii) in the seeker preferences within the received requests.

3.1 Research questions

Our research questions are related to the stages of the pipeline depicted in Figure 1 and are the following:

RQ1. How effective are the different recommendation methods? If we consider the baseline *random* recommender as a *control*, and each of the ML-based systems as a *treatment*, we would like to answer this question considering both average effects (treatment versus control) as well as heterogeneous effects (different treatments). In the next section we describe suitable metrics for measuring effectiveness.

RQ2. Are there any disparities arising from the usage of ML-based rankings? This is also a question that we address both at the level of average effects as well as heterogeneous effects through appropriate metrics.

4 METHODOLOGY

The methodology that we use to analyze whether the system leads to biased or discriminatory outcomes follows previous studies [6, 17] and consists of four main steps:

(1) Identification of potentially disadvantaged groups.
(2) Selection of effectiveness and disparity metrics.
(3) Computation of relevant metrics for each stage.
(4) Comparative analysis of treatment and control settings across groups.

4.1 Identification of protected groups

The main purpose of this initial step is to identify potentially discriminated [16], disadvantaged groups whose lack of privileges might be replicated or amplified within the platform. We consider four groupings that can lead to discrimination in this scenario and that we can evaluate with the available data: i) gender, ii) age, iii) languages spoken, and iv) "gay friendly" profiles. We remark that the data made available to us did not include any identifier that allows us to link these attributes to individuals, nor we made any attempt to do so. We also maintained data security by keeping the dataset within our research infrastructure, which can only be accessed by researchers in our team directly involved in this research.

4.1.1 Self-declared gender. Users specify their gender in a binary form (male/female) when registering for the app. The cases where the user did not inform their gender are discarded from the analysis.

4.1.2 Self-declared age. Users also can specify their age. Following Mehrotra et al. [11] we consider the individuals in the range [18–75] and then, looking at the distribution of data, create 4 different groups: (i) 18-34 (millennials), (ii) 35-54 (generation X), (iii) 55-75 (boomers) and (iv) < 18 or > 75 (outlier)

4.1.3 Languages spoken. The city from which we use data (Barcelona) is a cosmopolitan city hosting people from a variety of places. The main languages declared by users of the platform in this city are Catalan, Spanish, English, and Italian. Basically all Catalan-speakers users of the platform in Barcelona also speak Spanish. Hence, we compare these majority languages against cases in which the listers indicated other languages (such as Arabic).

4.1.4 "Gay friendly" profiles. Many descriptions of listed rooms, as well as profiles of individuals, included phrases such as "gay-friendly" or even "only gay-friendly people are welcome." Users are not asked to declare sexual orientation in this platform, but as sexual orientation had been found to be one determinant in access to housing [5], we consider that analyzing this "gay friendly" signal was important. We use a set of phrases that are variants of "gay friendly" to detect descriptions fitting this category.

Understanding that users in each side of the market might have different goals and/or preferences, we additionally *separate people according to their role within the platform* (lister or seeker).

4.2 Utility metrics

To define the metrics, we first need to introduce some notation. Let $\mathcal{U}$ represent the set of all users, with $\mathcal{U}_{\mathcal{L}}$ corresponding to listers, and $\mathcal{U}_{\mathcal{S}}$ corresponding to seekers, in such a way that $\mathcal{U} = \mathcal{U}_{\mathcal{L}} \cup \mathcal{U}_{\mathcal{S}}$. We remark that a small fraction of users ($\approx$ 4%), are listed as both room-owners as well as room-seekers.

Let $\mathcal{H}$ represent the set of rooms, and $H : \mathcal{H} \rightarrow \mathcal{U}_{\mathcal{L}}$ associate each room with its lister. Let $\mathcal{R} \subseteq \mathcal{H} \times \mathcal{U}_{\mathcal{S}}$ describe the recommendations presented to the listers, i.e., the different seekers selected for each room. Let $\mathcal{X} \subseteq \mathcal{R}$ be the requests created from such recommendations, i.e., the instances in which the recommendation was followed by a lister who contacted a seeker, and finally let $\mathcal{A} \subseteq \mathcal{X}$ the instances in which the contacted seeker answered positively to the request.

From the identification of the protected groups we can generate several partitions (e.g by gender, age, language spoken and sexual orientation). Users can be partitioned: (i) by gender ($\mathcal{G}$), (ii) by age ($\mathcal{Y}$), , (iii) by language spoken ($\mathcal{N}$) and (iv) by "gay friendly" ($\mathcal{F}$). We use the symbol $\mathcal{P}$ to reference the complete set of partitions.

Some of our utility metrics are independent of the role that a user has in the system. For instance, we assume that users in both sides want to minimize the effort required to find a roommate. Other metrics recognize that in some cases users may have opposite goals. For instance, listers want to minimize the income they obtain by renting their rooms at the highest possible price, while seekers seek to rent a room at the lowest possible price, all other things being equal.

4.2.1 DCG - Discounted Cumulative Gain (for listers). This is a measure of ranking quality, which in our case measures the value of a list of recommendations given to a lister. The metrics consider the positions in the ranking list of the items that a user finds relevant [8]. In its more general form, given a list of recommendations $R = \langle (r, u_1), (r, u_2), ..., (r, u_{|R|}) \rangle$ for a room $r \in \mathcal{H}$:

$$DCG_R = \sum_{i=1}^{|R|} w_i \cdot v_i$$

where w_i is a discounting factor that decreases with i, and v_i is the relevance of the i-th recommendation in R.

A common choice for the discounting factor is logarithmic discount: $w_i = 1/\log_2(1+i)$. The relevance of the i-th recommendation can be defined as the extent to which $H(r) \in \mathcal{U}_{\mathcal{L}}$, the lister of room r, will consider $u_i \in \mathcal{U}_{\mathcal{S}}$ an appropriate candidate for renting the room. The discounting factor stresses the requirement that the most useful recommendations should appear near the top of the list.

We use the normalized version of *DCG* that is divided by its maximum possible values, so the resulting *nDCG* is in the range $[0, 1]$.

4.2.2 CR - Conversion Rate (for listers). A “conversion” in online marketing indicates a successful traversal through a funnel, e.g., becoming a purchasing customer. In our case, success for a lister means finding of a suitable seeker, hence CR measures the probability that a request sent by a lister is accepted. If $\mathcal{X}^\ell$ are all the requests performed by lister $\ell \in \mathcal{U}_\mathcal{L}$, and $\mathcal{A}^\ell$ are all the requests that are accepted by the recipient seekers, then:

$$CR_\ell = \frac{\mathcal{A}^\ell}{\mathcal{X}^\ell}$$

4.2.3 CTR - Click Through Rate (for seekers). This indicates the probability that a seeker is contacted after being shown to a lister. Similar metrics have been used before to approximate item relevance for users [11], and CTR is a common metric used to evaluate, for instance, the relevance of web pages in personalized advertisement Richardson et al. [15].In our case, for a generic seeker $s \in \mathcal{U}_\mathcal{S}$, we consider the fraction of listers who click on him/her over the total number of listers that saw him/her. Let $\mathcal{R}^s$ be the set of recommendations containing the seeker s and $\mathcal{X}^s$ the set of requests created from such recommendation by the listers:

$$CTR_s = \frac{\mathcal{X}^s}{\mathcal{R}^s}$$

4.2.4 e_s – Exposure (for seekers). Differences in exposure have been recently studied to evaluate whether ranking models used in search and recommendation treat people from different groups similarly [20]. In our setting, we consider $\mathcal{R}^s$, which are all the recommendations including a particular seeker s, and the position $p(s,r)$ of the seeker s within a particular recommendation $r \in \mathcal{R}^s$.

$$e_s = \sum_{r \in \mathcal{R}^s} w_{p(s,r)}$$

where w_i is a discounting factor that decreases with i, as in the computation of *DCG*.

Assuming to consider a subset $S_a \subseteq \mathcal{U}_\mathcal{S}$, where all the seekers considered in the subset are characterized by the property a (e.g. a sensitive attribute), we can quantify the **disparate exposure** received by the group as:

$$DT(S_a) = \frac{\sum_{s \in S_a} e_s}{\sum_{s \in \mathcal{U}_\mathcal{S}} e_s} \times \frac{|\mathcal{U}_\mathcal{S}|}{|S_a|}$$

Where $|S_a|$ and $|\mathcal{U}_\mathcal{S}|$ corresponds to the size of the two sets. This index is inspired by the metrics already introduced by Singh and Joachims (2018). This non-negative metric $DT(S_a)$ is equal to 1 when the exposure generated for the group S_a is proportional to its relative size, if $DT(S_a) < 1$ then the group is *under-exposed* while for $DT(S_a) > 1$ the group is *over-exposed.*

5 DATASET DESCRIPTION

The platform that we study operates in several large cities across the world. We select the city in which the platform has its largest use base, Barcelona. The dataset gathered for this research contains 4,296,000 rows describing recommendations issued during a contiguous 30-months period from January 2017 through June 2019. It contains information about 61,997 unique users. Each recommendation includes a *lister* and *room* for which the recommendation is created, and the *seeker* that is recommended for that room and lister. Including the position in which each seeker was listed and the utility score assigned by the ranking system to it. Additionally, when the lister initiates a *request* from the recommendation, we have information that a request was initiated and about the response from the seeker addressed by the request. Responses by the seekers include accepting or rejecting the request, or leaving it pending, which means the request expires when the room is rented or becomes unavailable. The dataset also contains demographic information about the *age*, *(binary) gender*, *level of studies*, *work occupation*, and *spoken languages* for both seeker and listers.

Model	Listers	Seekers	Recommendations	Requests	Conv. Rate
BSL	35.72K	6.76K	1.78M	343.82K	19.37%
CF	15.35K	794.00	200.59K	45.72K	22.79%
MF	9.78K	7.83K	2.54M	568.37K	22.37%
XGB-1	4.02K	1.21K	396.54K	80.45K	20.29%
XGB-2	10.47K	5.07K	237.66K	84.54K	35.57%
XGB-3	3.80K	3.18K	384.31K	101.22K	26.34%

Table 1: Summary of the number of recommendations created with the different models through the operation of the platform. BSL is the random baseline; the other models are based on Machine Learning.

5.1 General statistics

The dataset contains baseline and ML-based recommendations. The baseline recommendations (BSL) are based on a random selection of available seekers for a room. They have always been provided by the platform, throughout its entire operation, and are used as a control. The ML-based recommendations have gone through several re-design iterations, including the following models:

- **Collaborative filtering (CF).** A collaborative filtering model trained to maximize the probability that listers send requests to the recommended seekers.
- **Matrix factorization (MF).** It corresponds to an instance of a Factorization Machine inspired by the model proposed by Rendle (2010). It included features from the rooms.
- **XG-Boost** During the operation of the platform, different versions of XG-Boost (gradient boosted decision trees) have been used: (i) **XGB-1**, first version of the model, which optimizes the probability of sending a request; (ii) **XGB-2**, second version, which optimizes the probability of a match, following the approach introduced by Volkovs et al. (2017), (iii) **XGB-3**, third version, which optimizes the probability of matches leading to actual rentals.

The number of recommendations generated by each method, as well as the time periods in which they were generated, are presented in Table 1. A summary of demographic information is reported in Table 2. In the following, we will use the acronym RS to refer to all the ML-based ranking systems together, in contrast with the baseline BSL.

6 RESULTS

In this section we report our analysis and our findings w.r.t the research questions introduced in Section 3.

Table 2: Percentage of seekers (S) and listers (L) belonging to different groups.

	Male		Female		Baby-boomer		Generation-X		Millenial		Outlier		Eng-Ita-Spa		Other		No-gay-friendly		Gay-friendly	
Model	L	S	L	S	L	S	L	S	L	S	L	S	L	S	L	S	L	S	L	S
CF	44.04	50.41	55.96	49.59	3.78	1.53	42.58	25.00	53.26	73.3	0.39	0.17	88.55	82.84	11.45	17.19	99.46	99.67	0.54	0.36
MF	43.80	56.22	56.20	43.78	4.13	1.44	37.59	22.7	57.96	75.71	0.32	0.15	98.14	99.05	1.86	0.95	99.30	99.19	0.70	0.81
XGB-1	45.00	59.86	55.00	40.14	3.72	1.15	36.75	22.61	59.15	76.00	0.38	0.25	94.45	96.94	5.55	3.06	99.26	99.15	0.74	0.85
XGB-2	46.69	43.23	53.31	56.77	4.03	1.33	39.49	22.67	55.98	75.90	0.50	0.10	98.88	99.50	1.12	0.50	99.47	99.27	0.53	0.73
XGB-3	43.76	55.22	56.24	44.78	4.03	1.54	37.53	23.4	57.98	75.04	0.46	0.02	99.8	99.76	0.20	0.24	99.52	99.18	0.48	0.82
BSL	43.49	51.22	56.51	48.78	4.10	1.67	38.23	25.02	57.29	73.15	0.38	0.15	94.75	89.52	5.25	10.49	99.42	99.43	0.58	0.58

6.1 Observed performance and disparities in the recommendations

We begin the evaluation by analyzing the first step in the recommendation pipeline (Figure 1). This part of the funnel selects a set of qualifying profiles, i.e. the list of suitable seekers according to the preferences selected by the lister for a room, then it ranks them and shows the top 20.

6.1.1 Lister side. We first compare, from the perspective of the listers, the relevance of recommendations selected by the random baseline (BSL) against the performance of recommendations created by any of the ML-based ranking system (RS). We assess the quality of the recommendations, computing the *nDCG* by considering that the relevant items are the seekers to whom the listers send a request. This utility metric is computed at individual level and then aggregated for each demographic group.

The random BSL exhibits an average *nDCG* score of 0.42. The performance by demographic groups is shown in Figure **??**.

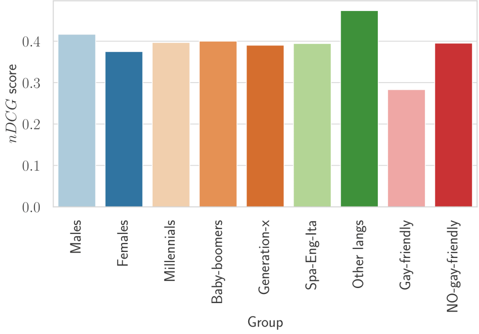


Figure 2: ***nDCG*** **score of recommendations for BSL.**

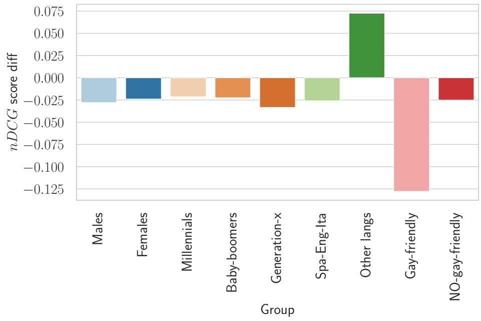


Figure 3: ***nDCG*** **score difference between the RS and BSL across demographic groups.**

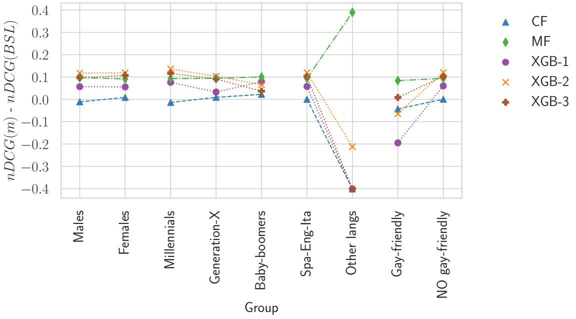


Figure 4: ***nDCG*** **difference between ML-models and BSL.**

The introduction of the ML-based ranking system leads to an increase in the overall performance, with an average *nDCG* score of 0.49. However, the increase in performance of the ranking system is not equal across the different demographic groups, as shown in Figure 3.

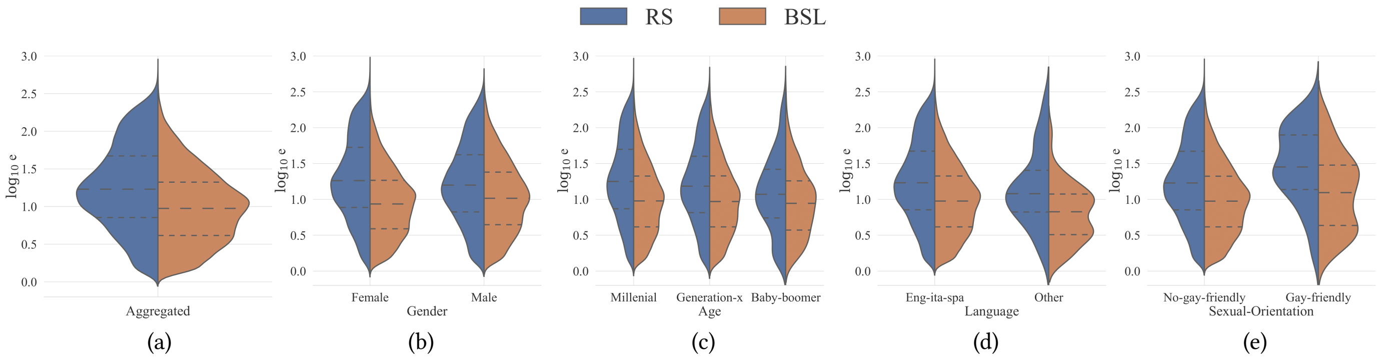
The observed differences in the *nDCG* score indicate that most groups obtain better recommendations, except for the "Gay-friendly" group, one of the minorities considered in our analysis, who got a decrease of 2.3% of the *nDCG* score.

We next compare the performance of individual models. Figure 4 reports the difference in performance between ML-models and BSL across demographic groups.

We observe that *XGB-2* is one of the best in terms of *nDCG* for most of the groups. On the other hand, the *MF* model is the more robust, since the differences in performance among groups are minimal. It is also the only model reporting a gain of performance w.r.t. the random baseline for the "Other languages" group. *CF* is the one showing the larger differences in performance by groups. In particular "Males", "Millenials", "Spa-Eng-Ita" and "No Gay-friendly" obtain better recommendations with the random baseline than with *CF*.

Observation 1. *ML-based ranking models have in general a positive average effect in recommendation performance, but different models lead to heterogeneous effects in terms of quality of recommendations for different groups.*

6.1.2 Seeker side. After analyzing the performance obtained by the listers to whom recommendations are presented, we consider the experience of the recommended users, i.e., the seekers. To evaluate the recommender systems from the seekers' side, we focus on the *exposure* they receive. As in the previous section, we first look at



(a) (b) (c) (d) (e)

Figure 5: Exposure distribution comparison (log-scale) between BSL and RS: total (Aggregated) and by demographics (Gender, Age, Language and Sexual-Orientation). The dashed lines in each violin plot represent the first, second and third quartile.

the average effect of the ML-based ranking systems (BSL vs RS), then perform an analysis per model.

In Figure 5 we report the exposure distribution for RS and BSL. Consistently in all the plots we can observe a heavy tail for RS on the larger values of exposure. This indicates that introducing the ML-based model leads to larger disparities in exposure among seekers. This effect results to be stronger for the groups of "Females", "Millennials", "Other" (language) and "No Gay-friendly".

> Observation 2. *The introduction of the ML-based recommendations increases the disparity in the exposure distribution: some people get much more exposure than the rest.*

We next analyze the exposure for each model across demographic groups.

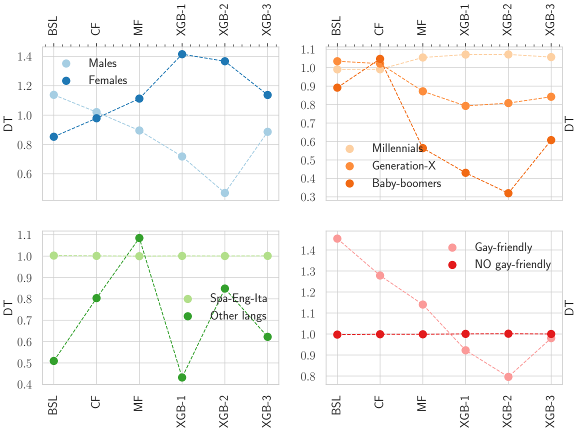


Figure 6: Exposure for the different models across demographic groups.

Figure 6 reports the exposure that different models give to different demographic groups. It shows that on average the exposure is larger for "Females" than for "Males" for most of the models. Regarding the "Age" partition, the group with more members, "Millennials", obtains a fairly constant average exposure with a little increment for last versions of the models where as the other groups obtain lower exposure in general. For the remaining two partitions ("Spoken languages" and "Gay-friendly"), we observe how the majority groups obtain an exposure close to 1, meaning that they are shown a numbers of times closely correlated to the size of their group, where as the two minorities experience more variance on their exposure, depending on the individual model that is recommending them.

6.2 Observed performance and disparities in the requests

6.2.1 Lister side. We next use the *CR* (Conversion Rate) metric to quantify the performance of the system for the listers. In general, the random baseline had a *CR* score of 10.36, which implies that on average, a generic lister needs to send ≈ 10 messages to recommended seekers about a room to get at least one seeker to accept it. By analyzing the *CR* score aggregated by groups, we obtain the results reported in Figure 7. In such plot, we first observe that the system does not present relevant differences of performance along the different subgroups. We can also observe that male listers have lower *CR* score than females, inverting the trend observed for the *nDCG* metric used to evaluate the quality of the recommendations. This means the recommendations show to men appear to be more relevant than those shown to women as they click on the top ones more, but once men issue a request to a seeker they have smaller chances than women of getting their request accepted.

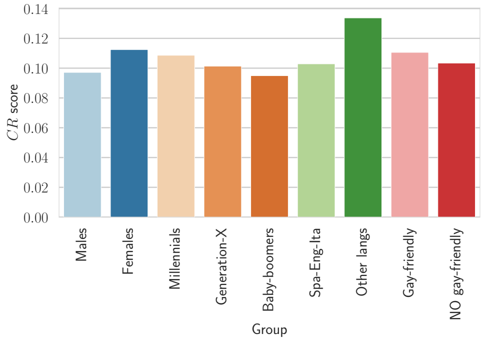


Figure 7: *CR* score for BSL across demographic groups

Looking how the *CR* score changes (Figures 8 and 9) with the addition of the different ML-based models, we observe heterogeneous variation of performances along the groups. In Figures 8

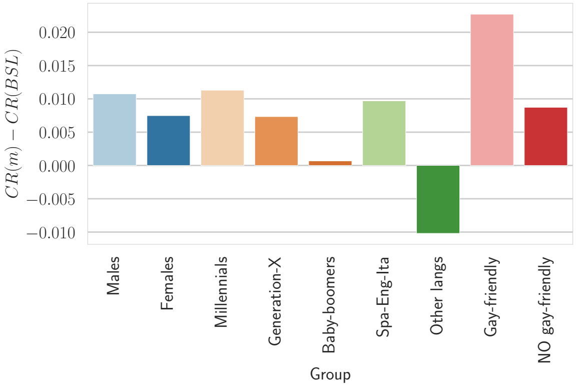


Figure 8: Conversion Rate (*CR*) differences between each ML model and BSL.

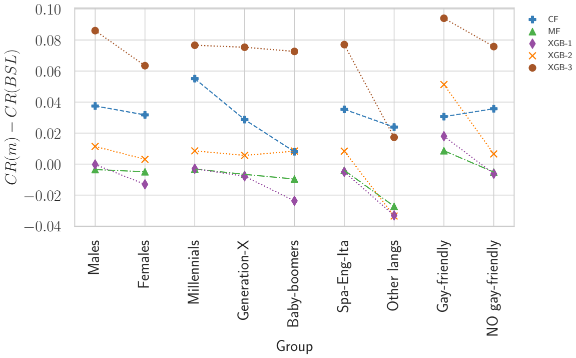


Figure 9: Conversion Rate (*CR*) differences w.r.t. BSL for each model.

we notice how two subgroups do not benefit from the use of the RS ("Baby-boomers" and "Other languages"). Figure 9 shows the relevant differences in CR across models. None of the models is consistent in terms of CR along the groups, where XGB-3 and CF result to be the ones improving the most the baseline. CF is also the one which leads to biggest difference in performance in "Age" group. XGB-2 is capable to generates the smaller differences in CR for the subgroups within "Gender" and "Age". MF and XGB-1 are consistently worse than the baseline, since we observe a CR larger than the baseline, only for the "Gay-Friendly" group.

> Observation 3. *The addition of ML-based rankings leads on average to improvements in terms of conversion rate, but the levels of improvement are substantially different from one model to another.*

6.2.2 Seeker side. To analyze the effectiveness for the seekers we adopt the Click Through Rate (*CTR*): as usual, such metric is evaluated for each model and across demographic groups. Figure 10 reports the results of such analysis. As the baseline model (BSL) is a random selection of seekers, it may better reflect the "raw" preferences of the listers. When comparing against the ML-based models, we notice how the CTR does not improve equally across groups. In all the models, except for *XGB-2*, we observe a significant difference between women and men: different strategies and models do not reduce the gap in the two values of CTR. Only *XGB-2* is able to obtain the same benefits for both.

Additionally, focusing on the "Age" attribute, we see how the "Millenials" obtain higher *CTR* along all the models while, "Generation-X" and "Baby-boomers" subgroups are always less clicked. The gap between the three categories is in some cases partially mitigated (*XGB-2* and *XGB-3*) but never reduced completely to zero. In the partition by "Spoken languages", while the baseline model shows a slightly higher CTR for the "Other languages" group, this distance is strongly reduced along the other models. We also notice how all the *XGB* models flip the order of the two subgroups, in particular this phenomenon appears stronger in *XGB-2*. Eventually, we observe a systematic gap of preferences between the two subgroups in the "Gay-Friendly" partition. The "No Gay Friendly" subgroup experiences an average positive difference in *CTR* of 5%, except for the case of *XGB-3*, which leads to same CTR. Finally, we also see how *XGB-2*, which is optimized for matches, is reflected here with a gain of *CTR* for all the groups, probably explained by the fact that such model is doing a better work on recommending seekers to the listers that will be interested on them.

> Observation 4. *Increases in Click Through Rate (CTR) by the ML-based recommenders are not consistent across groups. The changes in CTR are not aligned with the changes in exposure across groups and models.*

6.2.3 Performance and equity trade-off. In the context of Learning-to-Rank (LTR) Singh and Joachims (2019) defended the necessity of considering not only ranking utility to the users but also enforce the need of utility-aware metrics. Adapting this framework, we evaluate the quality of the recommendations that each model provides for the listers, in comparison to a measure of algorithmic fairness, which we define next for both sides of the market. As a measure of utility or quality, we compute the average of the *nDCG* scores measured for the different groups. We call this new metric **Balanced** *nDCG*, which corresponds for a generic model *m*:

$$Balanced_{nDCG}(m) = \frac{\sum\limits_{\mathcal{P}_i \in \mathcal{P}} \sum\limits_{a \in \mathcal{P}_i} nDCG(m, a)/|\mathcal{P}_i|}{N}$$

where N corresponds to the cardinality of all the possible partitions defined by demographics. In our specific case, considering all the subgroup, we have N = 4. On the other side, we want also to highlight potential discrepancies in performance between demographic groups, in order to compare with the quality measure defined above. To do so, we quantify the algorithmic fairness of each model using a metric inspired by the notion of demographic parity, which states that *each demographic group should receive the positive outcome at equal rates* [3]. We translate this context into two different metrics, one for the listers, one for the seekers. For the listers, we define a measure of disparity based on the average standard deviation of *nDCG* scores across the demographic groups, called $\sigma_n DCG$. To define the disparity metric for the seekers, inspired by Singh and Joachims (2019), we look at the ratio of the exposure and CTR by groups:

$$D_{ind}(a, m) = \frac{e_s(a, m)}{CTR(a, m)},$$

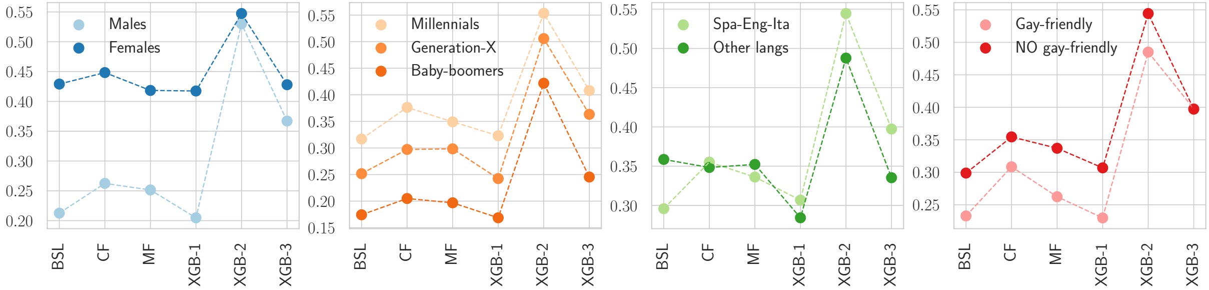


Figure 10: CTR differences across different models

Where m is the model selected and a the specific demographic attribute. This new measure D_{ind} express the alignment between the two measures of quality for the seekers. It is a non-negative index that gets lower values when the exposure given to the group is lower than the merit observed (CTR). Then, for each model we compute D_{model}, which corresponds to the average standard deviation of $D_{ind}(a, m)$, computed as follows:

$$D_{model}(m) = \frac{\sum\limits_{\mathcal{P}_i \in \mathcal{P}} \sigma_{\mathcal{P}_i}}{|\mathcal{P}|}$$

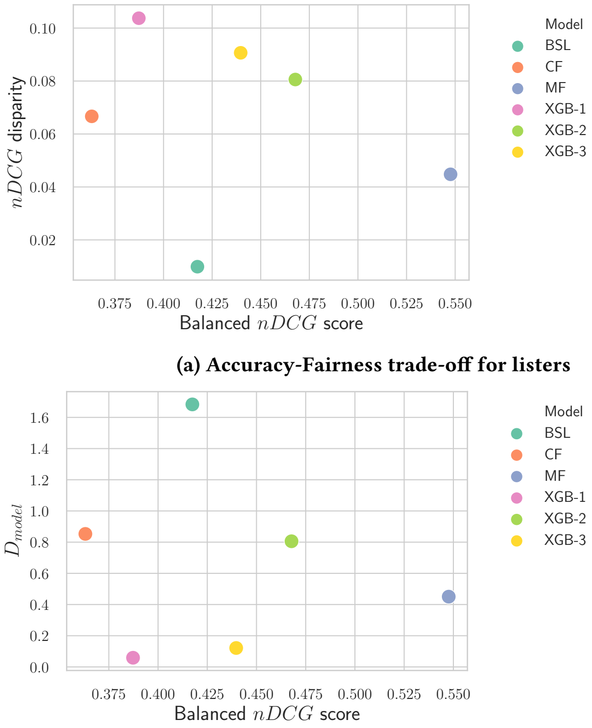
Where $\sigma_{\mathcal{P}_i}$ is the standard deviation of D_{ind} observed for all the sensitive attributes belonging to the $\mathcal{P}_i$. Through the use of this three new metrics we can focus on two different trade-offs concerning accuracy and fairness: (i) one for the listers, which is measured comparing $Balanced_{nDCG}$ and $\sigma_n DCG$ and (ii) the other for the seekers comparing $Balanced_{nDCG}$ and $D_{model}(m)$. In both cases the measure of accuracy is the same, since it is the metric which quantifies most the quality of performance for the platform.

Figure 11 reports the resulting trade-off plots. We observe in Figure 11a that for most of the models improving the average accuracy decreases also the level of unfairness for the listers. Only the BSL shows lower disparity with lower performances. While, observing (Figure 11b), we notice that improving the average accuracy of the system tends to slightly increase the level of unfairness for the seekers. CF model is the only one which results to be lower in accuracy but also unfair towards the seekers. The BSL model is interestingly unfair towards the seekers too.

Observation 5. *Improving the accuracy of the system benefits the listers, increasing fairness (as defined above) among groups. On the other side, the disparity in exposure seems to be weakly affected by the improvement in accuracy.*

6.3 Disparities in the answers: inequality of incomes

To evaluate the performance and equity for the last step of the recommendations pipeline, we use a different perspective. Understanding that this last step might result in an economic transaction in the form of a rental contract, we evaluate potential inequalities in the incomes users get across different models.



(a) Accuracy-Fairness trade-off for listers

(b) Accuracy-Fairness trade-off for seekers

Figure 11: *nDCG*-Fairness tradeoff. Listers between *nDCG* and disparities per across groups for *nDCG* (top). Seekers trade-off between *nDCG* and Exposure (bottom).

6.3.1 *Listers side.*

This analysis is three-fold: first, we consider the room price assigned to the rooms uploaded by listers, then, we evaluate for potential differences in such distribution with respect to the room price distribution in the requests that were accepted; finally, each model is considered individually to evaluate for potential differences between them. For this analysis, we discard the rooms with prices that are outliers (> 1000 EUR or < 200 EUR).

From this assessment, we first can observe that the ranking system does not imply a significant difference of room price in the accepted requests, with an average difference of ≈ 2% between the ML-based ranking and the random baseline: the average for the random system is 414.77 EUR per month, whereas lister make 425.04 EUR per month when exposed to the ranking system in average.

Nevertheless, as result of the analysis, we detect few cases with differences bigger than 3% in average between the room price in the uploaded rooms and the resulting room price in the accepted requests (figures were omitted for brevity). Among these few cases, none of them reach more than 10% of relative difference in average. Executing a statistically significance test, we find that none of those detected case is statistically significant for $p - value = 0.05$.

6.3.2 Seekers side. We then assess whether there are potential differences in the average price of the rooms accepted by seekers across different models. This analysis does not reveal any significant difference across any of the demographic groups or models. The performed evaluation reports an average of 412.77 EUR and 423.78 EUR per month for the random baseline and ML-based ranking systems respectively.

> Observation 6. *The observed disparities in the quality of recommendations shown to listers, probability of listers sending a requests to seekers, and probability of seekers accepting those requests, do not seem to lead to substantial differences in the prices at which rooms are rented.*

7 DISCUSSION AND CONCLUSIONS

Understanding the performance of a sharing-economy platform across all its users involved is an arduous task that requires to consider multiple aspects in the assessment. In this paper, we approach this task by first considering the role of different users inside of the platform. To perform our analysis, we need to consider the different goals that users might have depending on the side of the market where they are located. In this context, we conduct a layer-by-layer analysis, evaluating not only the system performance but also potential inequities created by such system for each of the steps in the recommendation pipeline. We also evaluate different versions of ranking system used during the platform life-cycle, and compare them to a baseline model based on random recommendations.

Our results show that compared to the random baseline, ML-based ranking systems on average increase the relevance of the provided rankings for the direct consumers of them, i.e., listers, according to the *nDCG* score. Splitting this analysis across demographic groups, we observe how certain groups do not benefit equally from average increases in the system performance, and even may be served worse than the baseline in some cases. Focusing on the other side of the market, i.e., seekers, we observe how incorporating the ML-based system increases disparities in exposure among them, leading to a small fraction of users receiving larger exposure, resulting in yet another example of disparate exposure caused by a ML-based system.

Then, we analyze the requests issues by listers when they find a relevant seeker among the recommendations. From there, we first show disparities in the Conversion Rate (*CR*) metric, a measurement of how easy it is to get accepted by the contacted seekers. During the assessment of the request driven by the random system, we observed small inequities between demographic groups that perhaps merit further analysis. In general, those sub-groups of the population which benefit more according to the *nDCG* (i.e., they find more suitable seekers to contact among the top recommendations), are also the ones with lower gains in the *CR* metric (i.e., they are not accepted as much by the seekers they contact).

After that, we observe how the addition of the ranking system created unevenly distributed gains of performance for the *CR* score across demographic groups. That, after being analyzed per model, showing significant gains for more sophisticated models. However, once again, minorities or already disadvantaged groups, obtained lower performance for that metric.

On the seekers side, we observed that inequities in Click Through Rate (CTR), a metric of interest of ranked users for the listers, were generally consistent across demographic groups. This fact can be interpreted as a systematic failure of recommending certain groups to the listers that would really be interested on them or, as an example of biased user preferences altering a performance metric. Most probably, it could be due to a combination of both aspects.

After analyzing the first two layers in the system, we wanted to empirically validate some ideas introduced by Singh and Joachims (2019), where authors claimed that ranking systems optimized for the utility of the rankings to users, tend to be oblivious on their impact to the ranked items. We assessed this in a fairness-utility analysis for both sides of the market. First, we observed how increasing the utility lead to lower disparities in the same metric for the listers. Nevertheless, we also observed how higher accuracy led at the same time to slightly larger inequities for the seekers exposure, validating the hypothesis. From this analysis, we also observed how the random system was not following the general trend of the rest of the models, most probably because it was not really optimized for the utility of the rankings.

Finally, we assessed whether different models related to inequalities in the amount of the economic transactions facilitated by them (rentals). From this final analysis, we can claim that generally there were no significant differences either for the listers or seekers for each of the models. In other words, while the different inequalities we have observed impact the probability that a user finds a rental, they do not seem to change substantially the price at which rooms are rented, for the cases in which a rental is found.

As a result of this analysis, we conclude that when analyzing such a system, measuring average effects may be quite insufficient, and it is necessary to consider each stage in the process, each algorithm, and each sub-group of people.

8 ACKNOWLEDGMENTS

This work has been partially supported by the HUMAINT programme (Human Behaviour and Machine Intelligence), Centre for Advanced Studies, Joint Research Centre, European Commission. The project leading to these results have received funding from "la Caixa" Foundation (ID 100010434), under the agreement LCF/PR/PR16/51110009. Additionally, this project has been partially funded by project SoBigData++, grant agreement 871042 of the European Commission.

REFERENCES

[1] Joshua Asplund, Motahhare Eslami, Hari Sundaram, Christian Sandvig, and Karrie Karahalios. 2020. Auditing Race and Gender Discrimination in Online Housing

Markets. *Proceedings of the International AAAI Conference on Web and Social Media* 14, 1 (May 2020), 24–35. https://ojs.aaai.org/index.php/ICWSM/article/view/7276

[2] Robert Chambers and Gordon Conway. 1992. Sustainable rural livelihoods: practical concepts for the 21st century. *IDS Discussion Paper* 296 (01 1992).

[3] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. Fairness through Awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference* (Cambridge, Massachusetts) *(ITCS '12)*. Association for Computing Machinery, New York, NY, USA, 214–226. https://doi.org/10.1145/2090236.2090255

[4] M. Eslami, K. Vaccaro, K. Karahalios, and K. Hamilton. 2017. "Be Careful; Things Can Be Worse than They Appear": Understanding Biased Algorithms and Users' Behavior Around Them in Rating Platforms. In *ICWSM*.

[5] Ariadna Fitó, Xavier Espinach, Ramon Gras, and Julenne Ramos. 2020. LA CLAU POT SER UN NOM: Detecció d'evidències de discriminació en l'accés al mercat de lloguer d'habitatge a Barcelona. https://ajuntament.barcelona.cat/dretsidiversitat/sites/default/files/La%20Clau%20pot%20ser%20un%20Nom.pdf

[6] Gemma Galdon Clavell, Mariano Martín Zamorano, Carlos Castillo, Oliver Smith, and Aleksandar Matic. 2020. Auditing Algorithms: On Lessons Learned and the Risks of Data Minimization *(AIES '20)*. Association for Computing Machinery, New York, NY, USA, 265–271. https://doi.org/10.1145/3375627.3375852

[7] Jevan A Hutson, Jessie G Taft, Solon Barocas, and Karen Levy. 2018. Debiasing desire: Addressing bias & discrimination on intimate platforms. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (2018), 1–18.

[8] K. Järvelin and Jaana Kekäläinen. 2002. Cumulated gain-based evaluation of IR techniques. *ACM Trans. Inf. Syst.* 20 (2002), 422–446.

[9] Ron Kohavi and Roger Longbotham. 2017. *Online Controlled Experiments and A/B Testing.* 922–929. https://doi.org/10.1007/978-1-4899-7687-1_891

[10] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2019. A Survey on Bias and Fairness in Machine Learning. arXiv:1908.09635 [cs.LG]

[11] Rishabh Mehrotra, Ashton Anderson, Fernando Diaz, Amit Sharma, Hanna Wallach, and Emine Yilmaz. 2017. Auditing Search Engines for Differential Satisfaction Across Demographics. In *Proceedings of the 26th International Conference on World Wide Web Companion* (Perth, Australia) *(WWW '17 Companion)*. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE, 626–633. https://doi.org/10.1145/3041021.3054197

[12] Frank Pasquale. 2015. *The Black Box society.* Cambridge: Harvard University Press.

[13] Giovanni Quattrone, Davide Proserpio, Daniele Quercia, Licia Capra, and Mirco Musolesi. 2016. Who Benefits from the "Sharing" Economy of Airbnb?. In *Proceedings of the 25th International Conference on World Wide Web* (Montréal, Québec, Canada) *(WWW '16)*. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE, 1385–1394. https://doi.org/10.1145/2872427.2874815

[14] S. Rendle. 2010. Factorization Machines. In *2010 IEEE International Conference on Data Mining*. 995–1000.

[15] Matthew Richardson, Ewa Dominowska, and Robert Ragno. 2007. Predicting Clicks: Estimating the Click-through Rate for New Ads. In *Proceedings of the 16th International Conference on World Wide Web* (Banff, Alberta, Canada) *(WWW '07)*. Association for Computing Machinery, New York, NY, USA, 521–530. https://doi.org/10.1145/1242572.1242643

[16] Salvatore Ruggieri, Dino Pedreschi, and Franco Turini. 2010. Data mining for discrimination discovery. *ACM Transactions on Knowledge Discovery from Data (TKDD)* 4, 2 (2010), 1–40.

[17] Pedro Saleiro, Benedict Kuester, Loren Hinkson, Jesse London, Abby Stevens, Ari Anisfeld, Kit T Rodolfa, and Rayid Ghani. 2018. Aequitas: A bias and fairness audit toolkit. *arXiv preprint arXiv:1811.05577* (2018).

[18] N. J. Salkind. 2017. *Between-subjects design. In Encyclopedia of research design (Vol. 1).* Thousand Oaks, CA: SAGE Publications, Inc., 82–84.

[19] Christian Sandvig, K. Hamilton, K. Karahalios, and C. Langbort. 2014. Auditing Algorithms : Research Methods for Detecting Discrimination on Internet Platforms.

[20] Ashudeep Singh and Thorsten Joachims. 2018. Fairness of Exposure in Rankings. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (London, United Kingdom). ACM, 2219–2228.

[21] Ashudeep Singh and Thorsten Joachims. 2019. Policy Learning for Fairness in Ranking. arXiv:1902.04056 [cs.LG]

[22] Tom Sühr, Asia J. Biega, Meike Zehlike, Krishna P. Gummadi, and Abhijnan Chakraborty. 2019. Two-Sided Fairness for Repeated Matchings in Two-Sided Markets: A Case Study of a Ride-Hailing Platform. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (Anchorage, AK, USA) *(KDD '19)*. Association for Computing Machinery, New York, NY, USA, 3082–3092. https://doi.org/10.1145/3292500.3330793

[23] M. Turner, S. L. Ross, G. Galster, J. Yinger, E. Godfrey, Beata A. Bednarz, C. Herbig, S. Lee, and B. Zhao. 2002. Discrimination in Metropolitan Housing Markets: National Results from Phase I HDS 2000.

[24] General Assembly United Nations. 1948. Universal Declaration of Human Rights.

[25] Maksims Volkovs, Guang Wei Yu, and Tomi Poutanen. 2017. Content-Based Neighbor Models for Cold Start in Recommender Systems. In *Proceedings of the Recommender Systems Challenge 2017* (Como, Italy) *(RecSys Challenge '17)*. Association for Computing Machinery, New York, NY, USA, Article 7, 6 pages. https://doi.org/10.1145/3124791.3124792

[26] Susan M. Wachter and Isaac F. Megbolugbe. 1992. Racial and Ethnic Disparities in Homeownership.