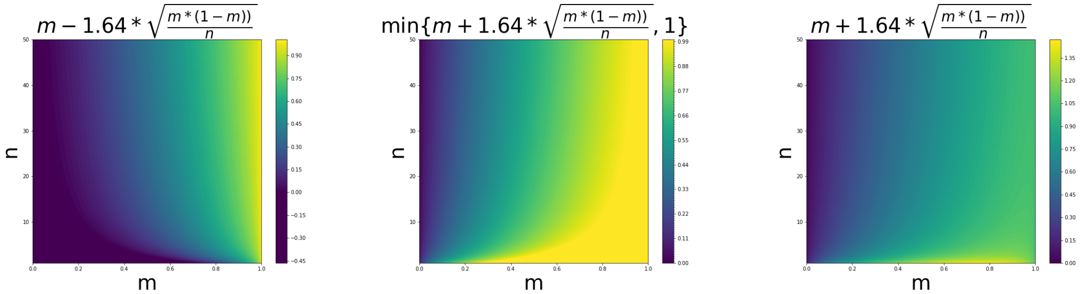
Appendix A. Discussion: Impact of n and m on Each Model

To give the reader a feel for the mathematical impact of the choice between these two models, we share some hopefully informative plots in Figure 2.



(a) Values of $m - 1.64\sqrt{\frac{m(1-m)}{n}}$

(b) Values of $\min\{m+1.64\sqrt{\frac{m(1-m)}{n}}, 1\}$

(c) Values of $m + 1.64\sqrt{\frac{m(1-m)}{n}}$

Figure 2: Showing the relationship between m (metric), n (number in subgroup), and c (edge of confidence interval). Hues in 2(a) shows the values of c in the pessimist's model. Hues in 2(b) shows the values of c in the optimist's model, but with an upper limit of 1 (since no proportion can be larger than 1). Hues in 2(c) shows the values of c in the optimist's model, without limiting the value at 1 (so that we can see more easily where it is very difficult to reject the optimist's hypothesis that the model is fair).

The horizontal axis of each of these plots is m, the metric value, which is assumed to be a proportion for which higher values are preferred (such as accuracy). The vertical axis is n, the size of the subgroup. The hue at (m, n) in Figure 2(a) is the corresponding value of $m - 1.64\sqrt{\frac{m(1-m)}{n}}$. Here we can visually see that, for a fixed metric value m, the subgroup size must be reasonably large in order to reject the hypothesis that the model is "unfair above c" for c near m.

Similarly, the hue at (m, n) in Figure 2(b) is the corresponding value of $m+1.64\sqrt{\frac{m(1-m)}{n}}$, capped at a value of 1 (since no proportion can be larger than 1). Here we can visually see that, for a fixed metric value m, the subgroup size must be reasonably large in order to reject the hypothesis that the model is "fair up to c" for c near m. To further understand the impact of small groups in this optimist's model, we include Figure 2(c). In Figure 2(c), the hue simply gives the value of $m + 1.64\sqrt{\frac{m(1-m)}{n}}$, even if it is larger than 1. This plot further highlights the fact that, in the optimist's model, it is very difficult to reject the hypothesis that the model is perfectly fair for very small subgroups.

Appendix B. Limitations

We note that the issue of multiple hypothesis testing is one which we do not address in depth. If membership in the different groups in question is independent, one can use the

Bonferroni correction to address the multiple hypothesis tests. Under this strict type of multiple hypothesis testing, the p-values that are calculated are using significance level $\frac{\alpha}{n}$, where n is the number of hypotheses that we are testing. This correction guarantees that the probability that we reject *one or more* null hypotheses is no more than α. Considering overlapping subgroups (such as considering fairness both for Black Women and for Latina Women) requires more care, and we do not delve into the issue of overlapping subgroups here. We thus, effectively, assume—counterfactually—that each person is a member of exactly one group. For the purposes of our empirical study (below), we fix the number of protected attributes to be as large as possible, as described in Section C.1.

Appendix C. Methods

We here provide further details on our empirical methods.

For starters, we choose the lale library and its accompanying datasets for two reasons:

1. The number of "fairness datasets" in the lale library is larger than any other conglomeration of fairness datasets that we are aware of.

2. Because the lale library has built-in models, we can apply a consistent type of model to each dataset, so that our experiments are not muddied by differing model constructions.

C.1. Subgroup Identification

The models we created use a forest of boosted trees from the XGBoost library; the functions to easily create these models are also part of the lale library. We created three models using the lale pipeline, using 3-fold cross-validation. The three models can be accessed to evaluate their accuracy on various subgroups. However, since lale requires sklearn version 1.2, we do not have access to the train/test indices of each of the models. Thus, to evaluate the accuracy on group G, we do so on all of the members of G in the dataset.[18]

The set of subgroups G on which we calculated the model accuracy come in part from the fairness data that lale provides, and also from attributes that are well-understood to be sensitive. Specifically, all of the attributes that the lale library lists as "protected" are included in our master list of protected attributes. If the rows in the dataset correspond to individuals, and any of $\{\text{age}, \text{sex}, \text{race}\}$ were not in lale's list of protected attributes, we added them to the master list. From this master list, we created *all* subgroups using *all* categories in the master list. For example, if a dataset had race, sex, and age category, we included in G each triple (r, s, a), where r was a race in that dataset's race column, s was a sex in that dataset's sex column, and a was an age category for that dataset.

C.2. Data Pre-processing

For each of the 20 fairness datasets, we used the built-in lale data pre-processing with small adjustments.

[18] Ideally, we would like to evaluate only on the members of G in the test set for that fold. However, our goal here is to assess our two proposed ideas to address small-sized subgroups, not to assess true model accuracy. Averaging the subgroup accuracy across the three folds provides appropriate information to do that. Thus, for this analysis, we set $m(G)$ to be the average accuracy of the three models for subgroup G.

We used the simple methods for imputing missing data which are provided with the sample notebook at IBM/lale (2023).

In order to use XGBoost, we needed to change some of the predicted categories to integer type.

In order to make the results more understandable, we re-named some of the categories (for example, changing the 'sex' categories from 0/1 to male/female).

The "race" categories in the `nlsy` dataset were atypical, including both categories such as 'GERMAN' and 'BLACK.' We did not attempt to clean that data but left the categories as given.

We created groupings by age for those datasets that don't already come with age groupings (see Appendix C.3).

C.3. Age Grouping

For the age attribute, some of the datasets already come with age groupings. In those cases, we directly used those groupings as the age categories. For the datasets where age was a strictly numerical attribute, we used the following heuristic to create categories:

- If age was already listed by lale as a protected attribute, we used the ranges provided by lale (for priviledged/unpriviledged groups) to create the categories.
- If age was not already listed as a protected attribute:
  - We grouped by decade in all datasets where this produced at least 5 people of each decade.
  - The `law_school` dataset had fewer than 5 members of the [0,9] decade, and fewer than 5 members of the [10, 19] decade, so those were grouped into a 0-19 group

After this initial analysis, we tossed out two of the datasets: `law_school` and `speeddating`. The standard lale models created by XGBoost were 100% accurate on those models, and thus did not provide interesting analysis for us.[19]

C.4. Analysis

For each such subgroup $G \in \mathcal{G}$, we calculated $m(G)$: the average accuracy of the three models on that subgroup. We then calculate the c values associated with each of those subgroups; indexed by c_1 for the optimist's and c_2 for the pessimist's metric. Specifically, for group G we calculate

$$c_1^G = m(G) + 1.64\sqrt{\frac{m(G)(1-m(G))}{n_G}}$$

from the optimist's model and

$$c_2^G = m(G) - 1.64\sqrt{\frac{m(G)(1-m(G))}{n_G}}$$

[19] We suspect that these datasets might be included in lale's list because they have low scores on other fairness metrics, such as the "symmetric class imbalance" metric in the sample notebook at IBM/lale (2023), or because the model must use protected attributes in order to be accurate.

from the pessimist's model.

Once these are calculated for all subgroups, we calculate

$$\begin{aligned} acc_{min} &= \min\{m(G) : G \in \mathcal{G}\} \\ c_1 &= \min\{c_1^G : G \in \mathcal{G}\} \\ c_2 &= \min\{c_2^G : G \in \mathcal{G}\} \end{aligned}$$

We also find their corresponding subgroups:

$$\begin{aligned} G_{min\_acc} &= \text{argmin}\{m(G) : G \in \mathcal{G}\} \\ G_1 &= \text{argmin}\{c_1^G : G \in \mathcal{G}\} \\ G_2 &= \text{argmin}\{c_2^G : G \in \mathcal{G}\} \end{aligned}$$

The group G_{min_acc} is the group with minimum estimated accuracy, while group G_1 (G_2) is on the cusp of rejecting the hypothesis that the model is fair (not being able to reject the hypothesis that the model is unfair) up to accuracy c_1 (c_2). Thus, we call groups G_{min_acc}, G_1, and G_2 the *critical subgroups* for a dataset. For some datasets, there are three distinct critical subgroups, while for other datasets, some of the critical subgroups are the same; see Tables 1, 2, 3, and 4 in Appendix D for details.

Once we had the (up to) three critical subgroups of each dataset, we did two additional analyses.

C.4.1. Subsample Just the Critical Group

Suppose G is a critical subgroup of a dataset. We then created 10 models (each a set of three 3-fold cross-validated models), where we include 10%, 20%, . . . , 100% of the subgroup in the dataset used to create the model. We then evaluated that group's critical value (whether it be $m(G)$, c_1^G, or c_2^G) on each of those 10 models, to see how those values change. The intention here is to mimic increasing samples from just the critical group, and how that additional data collection impacts the fairness evaluation of the model. These results of this analysis were in Figure 1.

C.4.2. Subsample the Entire Dataset

Suppose G is a critical subgroup of a dataset. We also created 10 models (each a set of three 3-fold cross-validated models), where we included 10%, 20%, . . . , 100% of the entire dataset to create the model. We then evaluated that group's critical value (whether it be $m(G)$, c_1^G, or c_2^G) on each of those 10 models, to see how those values change. The intention here is to mimic increasing sampling overall, and how that additional data collection impacts the fairness evaluation of the model. We note that, for the `nursery` dataset, one of the predicted categories (recommend) had only two data points with that category. In order for XGBoost to successfully create a model, we needed to add back both of those two data points into each subsample (if they had been removed in that random subsample). The results of this analysis are in Figure 3.

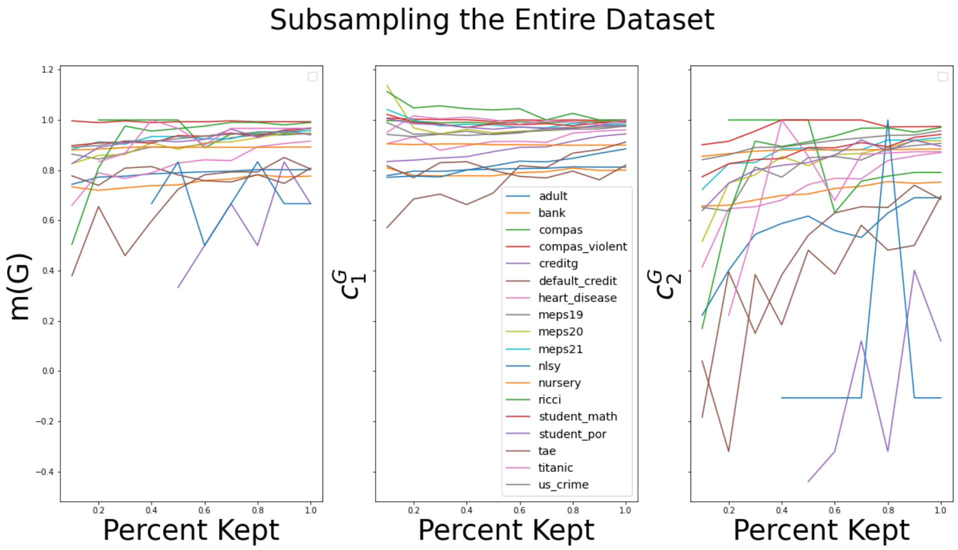


Figure 3: Plots of $m(G), c_1^G$, and c_2^G of critical subgroups G for each dataset. Here we subsampled the entire dataset, and the x-axis corresponds to the percentage of the entire dataset that is kept. Legend lists the dataset name.

Appendix D. Tables

Dataset	Min $m(G)$ Subgroup	Min c_1 Subgroup	Min c_2 Subgroup
adult	(White, Male, 50's)	(White, Male, 50's)	(Other, Male, 60's)
bank	<=24	<=24	<=24
compas	(Male, Native American, 25 - 45)	(Male, African-American, 25 - 45)	(Male, Native American, 25 - 45)
compas_violent	(Female, African-American, <25)	(Male, African-American, 25-45)	(Male, Other, >45)
creditg	(male div/sep, male, <=25)	(male single, male, >25)	(male div/sep, male, <=25)
default_credit	(male, 40's)	(male, 40's)	(female, 70's)
heart_disease	(female, >54)	(female, >54)	(female, >54)
meps19	(White, 80's, female)	(White, 80's, female)	(Non-White, 80's, male)
meps20	(Non-White, 80's, female)	(Non-White, 80's, female)	(Non-White, 80's, female)
meps21	(Non-White, 80's, female)	(Non-White, 80's, female)	(Non-White, 80's, female)
nlsy	(Female, <18, GREEK)	(Female, >=18, GERMAN)	(Female, <18, HAWAIIAN)
nursery	great_pret	great_pret	great_pret
ricci	W	B	W
student_math	(M, <18)	(M, <18)	(M, <18)
student_por	(M, >=18)	(M, <18)	(M, >=18)
tae	1.0	2.0	1.0
titanic	(female, 60's)	(female, 30's)	(female, 60's)
us_crime	TRUE	TRUE	TRUE

Table 1: Subgroups with minimum $m(G)$, c_1, or c_2 for each dataset.

Dataset	Subgroup	Subgroup Category	n	$m(G)$
adult	(White, Male, 50’s)	[race, sex, age_cat]	4256	0.8020050125313280
bank	<=24	age_cat	809	0.7766790276060980
compas	(Male, Native American, 25 - 45)	[sex, race, age_cat]	6	0.9444444444444450
compas_violent	(Female, African-American, <25)	[sex, race, age_cat]	95	0.9929824561403510
creditg	(male div/sep, male, <=25)	[personal_status, sex, age_cat]	2	0.6666666666666670
default_credit	(male, 40’s)	[sex, age_cat]	2771	0.8078912546613740
heart_disease	(female, >54)	[sex, age_cat]	103	0.9158576051779940
meps19	(White, 80’s, female)	[RACE, age_cat, SEX]	184	0.947463768115942
meps20	(Non-White, 80’s, female)	[RACE, age_cat, SEX]	146	0.9474885844748860
meps21	(Non-White, 80’s, female)	[RACE, age_cat, SEX]	142	0.9577464788732400
nlsy	(Female, <18, GREEK)	[gender, age_cat, race]	2	0.666666666666667
nursery	great_pret	parents	4320	0.8922839506172840
ricci	W	race	68	0.9901960784313730
student_math	(M, <18)	[sex, age_cat]	134	0.9676616915422890
student_por	(M, >=18)	[sex, age_cat]	73	0.9406392694063930
tae	1.0	whether_of_not _the_ta_is_a_native _english_speaker	29	0.8045977011494250
titanic	(female, 60’s)	[sex, age_cat]	10	0.9666666666666670
us_crime	TRUE	blackgt6pct	970	0.9663230240549830

Table 2: Subgroups with minimum accuracy value $m(G)$

Dataset	Subgroup	Subgroup Category	n	c_1
adult	(White, Male, 50's)	[race, sex, age_cat]	4256	0.8120224969943970
bank	<=24	age_cat	809	0.8006925092362380
compas	(Male, African-American, 25 - 45)	[sex, race, age_cat]	1563	0.994278290695671
compas_violent	(Male, African-American, 25 - 45)	[sex, race, age_cat]	932	0.999219907516058
creditg	(male single, male, >25)	[personal_status, sex, age_cat]	492	0.94429531762294
default_credit	(male, 40's)	(sex, age_cat)	2771	0.820164957561691
heart_disease	(female, >54)	(sex, age_cat)	103	0.96071630150011
meps19	(White, 80's, female)	[RACE, age_cat, SEX]	184	0.974437789794083
meps20	(Non-White, 80's, female)	[RACE, age_cat, SEX]	146	0.977763384001414
meps21	(Non-White, 80's, female)	[RACE, age_cat, SEX]	142	0.985432236390149
nlsy	(Female, >=18, GERMAN)	[gender, age_cat, race]	179	0.88480628727837
nursery	great_pret	parents	4320	0.9000195456124770
ricci	B	race	27	1.0
student_math	(M, <18)	(sex, age_cat)	134	0.992723469193729
student_por	(M, <18)	(sex, age_cat)	193	0.978258629541195
tae	2.0	whether_of_not_the_ta_is_a_native_english_speaker	122	0.912074688695555
titanic	(female, 30's)	(sex, age_cat)	86	0.99964644295967
us_crime	TRUE	blackgt6pct	970	0.975822194666121

Table 3: Subgroups with minimum c_1 value

Dataset	Subgroup	Subgroup Category	n	c_2
adult	(Other, Male, 60’s)	[race, sex, age_cat]	10	0.6903719639038720
bank	<=24	age_cat	809	0.7526655459759590
compas	(Male, Native American, 25 - 45)	[sex, race, age_cat]	6	0.7910815916767240
compas_violent	(Male, Other, Greater than 45)	[sex, race, age_cat]	49	0.9739395574376140
creditg	(male div/sep, <=25)	[personal_status, age_cat]	2	0.12000000000000000
default_credit	(female, 70’s)	[sex, age_cat]	12	0.6973855176357850
heart_disease	(female, >54)	[sex, age_cat]	103	0.8709989088558780
meps19	(Non-White, 80’s, male)	[RACE, age_cat, SEX]	67	0.9066848240754210
meps20	(Non-White, 80’s, female)	[RACE, age_cat, SEX]	146	0.9172137849483570
meps21	(Non-White, 80’s, female)	[RACE, age_cat, SEX]	142	0.9300607213563300
nlsy	(Female, <18, HAWAIIAN)	[sex, age_cat, race]	1	-0.10643674743062500
nursery	great_pret	parents	4320	0.8845483556220900
ricci	W	race	68	0.9706008691220500
student_math	(M, <18)	[sex, age_cat]	134	0.9425999138908480
student_por	(M, >=18)	[sex, age_cat]	73	0.8952823458833590
tae	1.0	whether_of_not_the_ta_is_a_native_english_speaker	29	0.6838443819651760
titanic	(female, 60’s)	[sex, age_cat]	10	0.8735726878662690
us_crime	TRUE	blackgt6pct	970	0.9568238534438450

Table 4: Subgroups with minimum c_2 value

Appendix E. Analysis of Metric from Kearns et al.

As noted in Section 4.3, we hypothesize that the fairness metric outlined by Kearns et al. (2018) violates *Incentive Compatibility.* The fairness metric likely "looks worse" as additional data is gathered about a small subgroup (i.e., a group whose size in proportion to the entire dataset is small). The fairness metric includes a factor which is the proportion of the subgroup within the dataset. Thus, as additional data is collected from that subgroup alone, this proportion increases, making the model more likely to violate the fairness criteria, hence potentially disincentivizing additional data collection on that subgroup. Here we empirically examine this hypothesis.

E.1. Study Description

Using the same datasets, subgroups, pre-processing, cleaning, and models outlined in Appendix C, we calculate the value of the following expression from Equation (2):

$$\alpha(G)|m(G) - m(\cdot)| \tag{6}$$

Recall that $\alpha(G)$ is the proportion of group G within the total population, and that m is some model performance metric (as in Appendix C, we use accuracy as our sample metric m for this study). The value $m(G)$ is the model performance metric evaluated only on subgroup G, while $m(\cdot)$ is the value of the model performance metric on the entire dataset.

Kearns et al. (2018) use an auditing process wherein the value calculated from expression (6) must be below some threshold ϵ in order for a model to be considered fair. Thus, we can think of expression (6) as describing *unfairness* for group G.[20]

E.2. Methods

We calculate the value in expression (6) on increasing subsamples of each dataset. We concentrate on small sugbroups G that comprise no more than 10% of the total population. Just as in the experiments described in Appendix C, we examine two subsampling scenarios: We subsample the subgroup in question only (to simulate gathering more subgroup data), and we subsample the entire dataset (to simulate gathering more population data). Note that, of course, the values of m depend on the model created, which depends on the subsample of the data used to create the model.

Many of the datasets (`heart_disease, nursery, ricci, student_math, student_por, tae` and `us_crime`) don't have any subgroups comprising less than 10% of the total population, and thus we exclude those datasets from this analysis. From the other datasets, we concentrate on four: the `adult, bank, meps20`, and `titanic` datasets. The results for all other datasets are similar.

In Section 4.3, we hypothesize that the protocol of Kearns et al. violates *Incentive Compatibility.* Specifically, we hypothesize that when subsampling only small groups, the *unfairness* value of expression (6) would *increase*. Since the value $\alpha(G)$ does not change significantly when subsampling the entire population, we do not expect expression (6) to change much when subsampling the entire dataset, aside from the fact that potentially a better model might make expression (6) decrease.

[20] All typical ways of defining "fairness" can be interpreted this way. A higher ϵ in (1) is interpreted as a *de*crease in fairness and thus an increase in unfairness.

E.3. Results and Discussion

The results of subsampling just the small subgroup can be found in Figure 4, and the results of subsampling the entire dataset can be found in Figure 5.

When only the small subgroup is subsampled (as in Figure 4), we see the value of (6) increasing for all subgroups in the `adult` and `bank` datasets. The picture is slightly more muddled in the `meps20` and `titanic` datasets, but these still show either a consistent increase or an initial increase for nearly all of the small subgroups. In other words, the value (6) of unfairness *increases*, indicating that the Kearns et al. auditing process discourages additional data collection of small subgroups, and thus violates *Incentive Compatibility*.

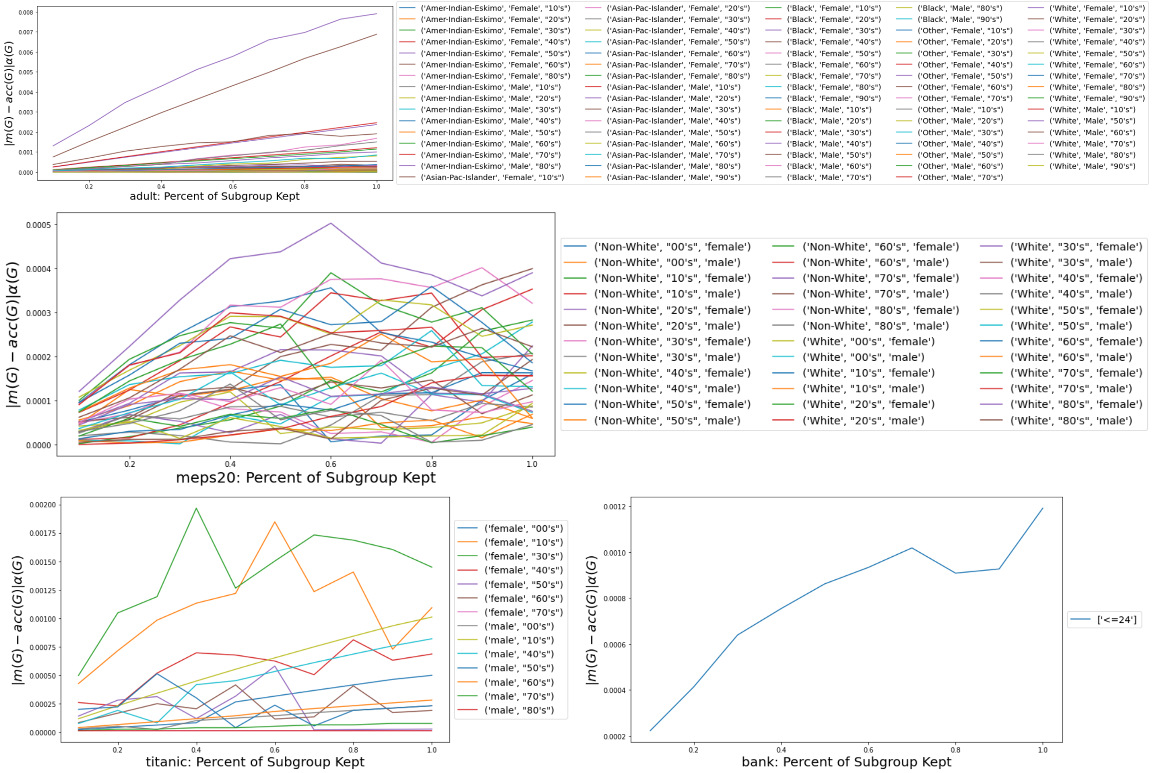


Figure 4: Values of expression (6) on the `adult, meps20, titanic`, and `bank` datasets. Horizontal axis is the percent of the subgroup, vertical axis is unfairness (i.e., the value of expression (6)).

When the entire dataset is subsampled (as in Figure 5), values of (6) remain remarkably consistent in the `adult` dataset, and tend to decrease in the `bank, meps20`, and `titanic` datasets. We can thus conclude that the Kearns et al. approach, while it discourages collecting additional data from only the smallest subgroups in a dataset (thereby not satisfying *Incentive Compatibility*), does not appear to discourage additional data collection when each subgroup's proportion within the population stays consistent.

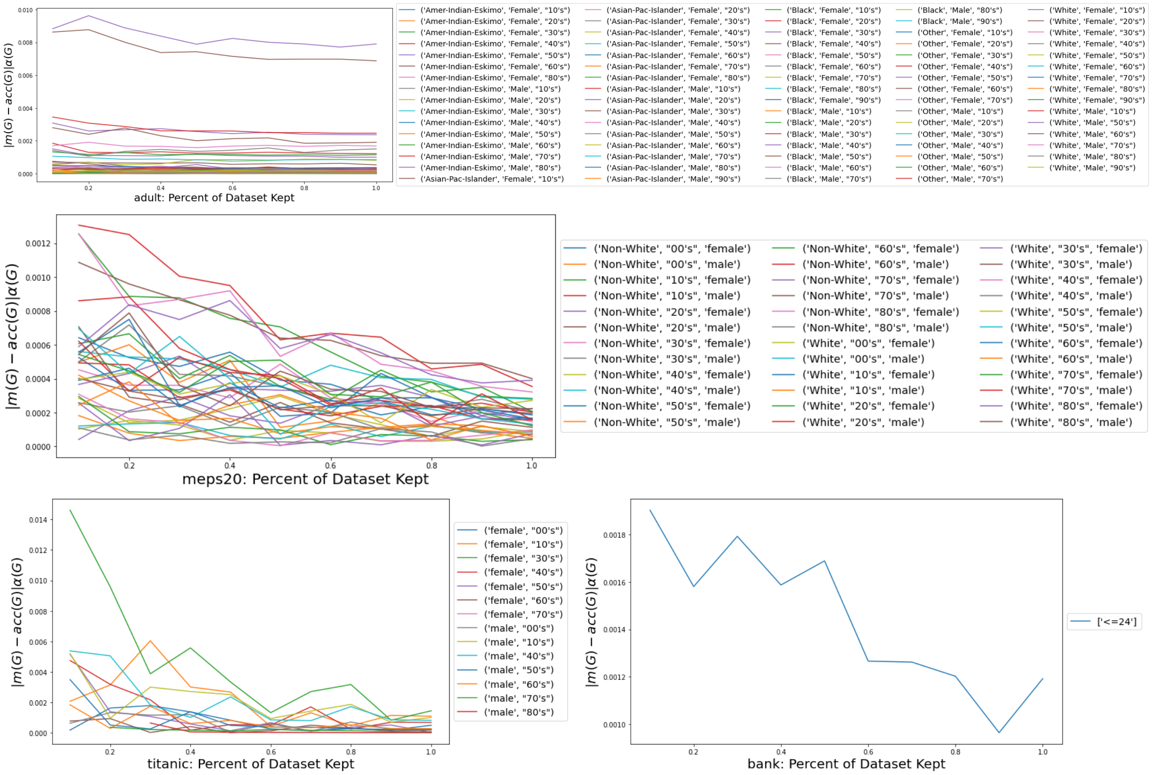


Figure 5: Values of expression (6) on the `adult`, `meps20`, `titanic`, and `bank` datasets. Horizontal axis is the percent of the entire dataset kept, vertical axis is unfairness (i.e., the value of expression (6)).