Preferential Sampling For Data Debiasing

Atharva Tidke
Department of Computer Science
University of Durham
Durham, England

I. Project Proposal

In recent years job recruitment for all types of jobs e.g., part-time, full-time, seasonal, etc. has almost completely moved online whether it is filling out application forms on a company's website or a general job posting/recruitment website like Indeed or Glassdoor. Professional networking tools like LinkedIn have become the de facto way of getting in touch with recruiters. All of which generates an influx of data which can be used by recruiters to implement AI solutions to automate or simplify numerous tasks.

As per a 2018 report by LinkedIn [1] which investigates the top global trends in recruiting, 'Diversity', 'Data' and 'Artificial Intelligence' are amongst the top four new trends in recruitment. Diversity is most embraced trend with over half of companies are already tackling it head-on. Of the 8800+ recruiters surveyed 71% said they focus their diversity efforts on ensuing gender diversity and 49% on ensuing racial and ethnic diversity. In addition, 56% say they use data to increase retention, 50% say they use data to evaluate skills gaps, and 50% say they use it to build better offers.

The report also states that recruiters and hiring managers, globally, shared the sentiment that AI is a bold disrupter in the field with 76% saying that AI's impact of recruiting will be at least somewhat significant. Only 8% of recruiters surveyed have adopted the use of AI but these numbers expected to grow in the future and AI will be used to automate or aid with several recruitment tasks as seen in Figure 1.

Figure 1 - Areas where AI will impact recruiting, source [1]

So, whilst we are still in the early stages of widespread adoption of AI in recruitment, it is the perfect time to develop and implement fair AI solution for recruitment related tasks as seen in figure 1, because there is huge potential for fair AI solutions in this field.

Discrimination based on race and sex with respect to wage whilst narrowing is still present. According to the Office for National Statistics (ONS) the gender pay gap has decreased from 27.5% in 1997 to 15.5% in 2020 but clearly still persistent [2]. Furthermore, the gender pay gap is still absurdly high in several industries even as of 2020, e.g., Precision instrument makers and repairers (46%), Production managers and directors in mining and energy (38%), IT engineers (36%), Financial institution managers and directors (34%) just to name a few. Also, according to ONS most minority ethnic groups earn less on average (up to 16% less) than White British people as of 2019, though some groups earned more than their White British counterparts [3] Furthermore, ONS states that men earned a higher hourly median wage than women in all, but three ethnic groups (as of 2019) as seen in Figure 2.

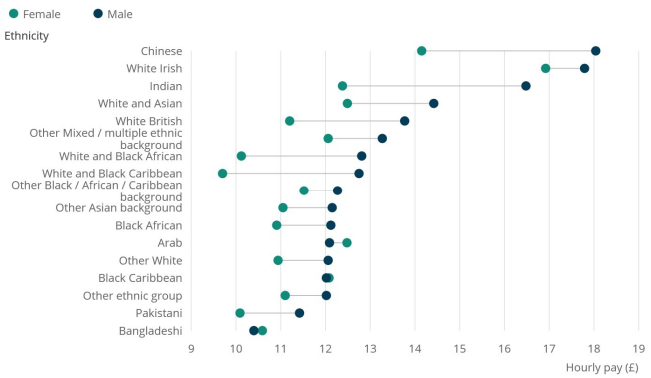

Figure 2 - Median hourly pay, 17 ethnic groups by sex, England and Wales, 2019, source ONS [3]

In summary, recruiters are embracing diversity and there still work that need to be done to reduce gender and racial pay gap. With the increasing use of data and AI in recruitment it is likely that a naively trained system will inherit this bias and discriminate which would not be acceptable for most recruiters. Therefore, I would like to implement a bias mediation

technique, that can be used to implement machine learning models that do not discriminate individuals based on their sex or race and can be used by recruiters in the recruitment process.

The Python Pandas Scikit-learn stack is an industry standard and widely used by many major recruiters for various purposes, therefore, I will be using these tools to implement my solution as well.

II. Project Progress

A. Data Analysis

For a human-centric dataset I chose Adult Census Income data set from UCI machine learning repository [4]. This dataset includes demographic information such as age, sex, race and country of origin of individuals. It is split into training data and test data on UCI's website, but I combined the two. The transformation done on the dataset is detailed in Section 3.2 subsection 'Data Preparation' of the submitted notebook.

The distribution of continuous features is as follows

	age	education-num	capital-loss	capital-gain	hours-per-week
Mean	38.6	10.1	1079	87.5	40.4
Std	13.7	2.57	7452	403.0	12.4
Var	188.0	6.61	5.55e7	1.62e5	153.5

The most frequent values of categorical features

Feature	**3 most frequent values**
sex	Male, Female
race	White, Black, Asian-Pac-Islander
native-country	United-States, Mexico, Philippines
relationship	Husband, Not-in-family, Own-child
occupation	Prof-specialty, Craft-repair, Exec-managerial
marital-status	Married-civ-spouse, Never-married, Divorced
education	HS-grad, Some-college, Bachelors
workclass	Private, Self-emp-not-inc, Local-gov

After cleaning the data, the following bias was observed.

The proportion of males earning >50K a year (30.4%) is roughly three times as much as females (10.9%), this is indicative of the historic bias towards women not being paid equal to their male counterparts. The privileged group (Male) makes up 73% percent of the data set. Females earning >50K a year are on average 2.4 years younger, have spent 0.27 more years in formal education and work roughly 6 hours less per week than their male counterparts.

The proportion of the White population earning >50K a year (25.4%) is on average twice as large as the other ethnicities (except 'Asian-Pac-Islander'). This is again indicative of the historic racial bias towards people of colour in the US. The privileged group (White) makes up 85% percent of the data set. People of colour earning >50K a year are on average 2 to 4 years younger (except 'Black'), have spent roughly the same or a greater number of years in formal education (except 'Amer-Indian-Eskimo') and work roughly 1 to 2 hours less per week compared to their White American counterparts.

The proportion of Americans earning >50K a year (24.4%) is lower than several countries e.g., France (42.2%), India (41.1%), Taiwan (40%), etc. and higher compared to a number Central and South American, Caribbean and Southeast Asian countries. This representation is likely because high skilled workers from European countries, India, Taiwan etc. who migrate to the US are likely to have high paying jobs and individuals from Central and South American, Caribbean and Southeast Asian countries are likely to experience systemic injustice and as a result lower pay. The privileged group (United States) makes up 90% of the dataset.

B. Conventional Implementation

This dataset was created to train a classification algorithm to predict whether an individual earns more than 50K a year. Such a classifier has many applications for recruiters, for example,

- To build better offers by checking which salary class the candidate in question is assigned by the classifier
- Employee retention - Evaluate bias within their own organization by checking whether the employees they currently employ are being paid fairly. This can be achieved by comparing their predicted and actual income class.

Therefore, I will implement an unbiased classifier that will predict the income class ('salary') of an individual based on the attributes provided in the dataset.

After reviewing papers related to this dataset, most of them used a decision tree classifier and after some experimentation I found that Scikit-Learn's random forest classifier gives comparable results, so I decided

to use it for the conventional and fair implementation. Scikit-learn does have a decision tree classifier which uses an optimized version of the CART algorithm (which is very similar to C4.5), but I found its accuracy is not comparable to the C4.5 decision tree classifier in the paper I want to implement the bias mediation technique from and therefore I did not use it.

My chosen paper uses rate of discrimination to quantify discrimination of the classifier. The rate of discrimination for an attribute is defined as the difference between the of the rate of positive outcomes of the privileged and the deprived group. The ideal value is 0% and any value above 10% indicates bias towards the privileged class. The attributes 'sex', 'race', and 'native-country' of the dataset can be a basis for discrimination which are within the scope of this project therefore, I will be measuring and mediating the discrimination that occurs with regards to these attributes.

After training the classifier on 70-30 naively split data the rate of discrimination was 16.1% (sex), 8.8% (race) and 3.9% (native-country). The disparate impact was 0.31 (sex), 0.54 (race) and 0.78 (native-country). This indicates the model heavily favors the privileged class with regards to their sex and race. On the other hand, there was no significant bias with respect to an individual's native country.

The same classifier was then trained on 70-30 stratified split of data with regards to race and sex so that the training data was representative of the demographic groups in the dataset ('native-country' was not considered in the representation criteria because the naïve model showed no significant discrimination with regards to an individual's native-country).

The rate of discrimination was 16.4% (sex), 6.7% (race), 1.4% (native-country). The disparate impact was 0.26 (sex), 0.62 (race) and 0.92 (native-country). Because of the stratified train-test split, the rate of discrimination w.r.t. sex increased and discrimination w.r.t race and native-country decreased. This may be because the deprived classes had more representation in the training data which led to less bias w.r.t. race and native-country but exposed the underlying bias w.r.t. sex even more.

The reason for this bias w.r.t. sex is likely to be the imbalanced rate of desired outcome (salary >50K) for favored and deprived groups in the train data as detailed in Section II A. This classifier is also used as the first baseline mode referred to as 'No' here after.

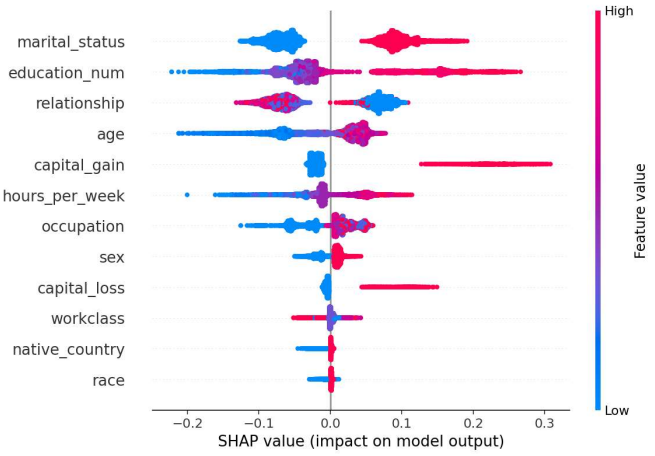

Figure 3 SHAP value impact summery plot for the baseline classifier

This bias is further apparent after examining the SHAP value of summery plot in Figure 3, high feature value for sex i.e., if an individual's sex is Male, it has a positive impact of the classifiers output value i.e., the classifier predicts salary is >50K but, when sex is Female it has a negative impact on the classifiers output.

C. Fair Machine Learning Implementation

For the fair machine learning implementation, I decided to implement a pre-processing technique of 'Preferential Sampling' (PS) as described in Kamiran, Toon (2011) [5] because it was the most effective data debiasing technique in the experiment results. The paper states that the data objects close to the decision boundary are more prone to have been discriminated or favored due to discrimination in the dataset and to give preference to them for sampling. I used a Naive Bayes (NB) classifier to assign the data objects in the test set their probability of being predicted as the positive class. Deprived Positive (DP) and Favored Positive (FP) objects are sorted in ascending order, and the objects of Deprived Negative (DN) and Favored Negative (FN) in descending order to identify the objects close to the decision boundary. The top objects in DN and FP are deleted from the train set and top objects in DP and FN are duplicated until a given threshold of disparate impact is met.

As suggested in the paper I first trained further two baseline classifiers, one with the 'sex' attribute hidden (No_SA) and one with 'race' attribute hidden (No_RA). The original experiment only applies PS to reduce bias based on the 'sex' attribute, but I also resampled it mediate bias based on the 'race' attribute. The classifier (No_PS) was trained on this resampled data.

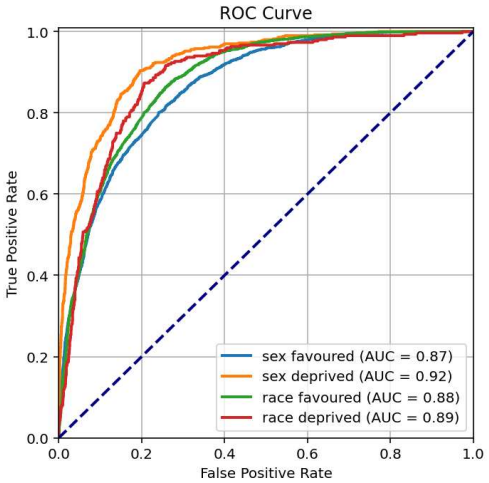

Figure 4 ROC curve of the No_PS classifier

As seen in Figure 4, the are under the curve of ROC is comparable for the privileged and deprived class for both the sex and race attribute.

The results of my implementation and the J48 classifier trained of preferentially sampled data from the paper are as follows

		No	No_SA	No_RA	No_PS
J48	Acc (%)	86	86	-	84.6
	Disc, sex (%)	16.5	17	-	4
My Imp	Acc (%)	84.10	83.97	84.08	84.10
	Disc, sex (%)	16.4	16.2	16.3	3.1
	Disc, race (%)	6.7	6.5	6.3	1.6

Overall, the rate of discrimination with regards to sex of my implementation is comparable to that of experimental results in the research paper. The exact number of substitutions and deletions when PS was not detailed in the paper, so I made these modifications until the desired disparate impact was reached. Also, the discrimination accuracy tradeoff is not as significant in my implementation, this may be because my implementation uses a different classifier.

In conclusion, PS has been successful in mediating bias based on race and gender and can be used by recruiters for classification tasks detailed at the beginning of the section.

REFERENCES

[1] News.linkedin.com. 2021. LinkedIn 2018 Report Highlights Top Global Trends in Recruiting. [online] Available at: https://news.linkedin.com/2018/1/global-recruiting-trends-2018

[2] Ons.gov.uk. 2021. Gender pay gap in the UK - Office for National Statistics. [online] Available at: https://www.ons.gov.uk/employmentandlabourmarket/peopleinwork/earningsandworkinghours/bulletins/genderpaygapintheuk/2020

[3] Ons.gov.uk. 2021. Ethnicity pay gaps - Office for National Statistics. [online] Available at: https://www.ons.gov.uk/employmentandlabourmarket/peopleinwork/earningsandworkinghours/articles/ethnicitypaygapsingreatbritain/2019

[4] C. Blake and C. Merz. Adult Income Dataset, UCI repository of machine learning databases, 1998. [online] Available at : https://archive.ics.uci.edu/ml/datasets/adult

[5] Kamiran, Faisal & Calders, Toon. (2011). Data Pre-Processing Techniques for Classification without Discrimination. Knowledge and Information Systems. 33. 10.1007/s10115-011-0463-8.