Preventing Bias in Machine Learning by using Bias Aware Algorithms: An Empirical Experimental Study - Full Report

Andy Gray
445348
445348@swansea.ac.uk

ABSTRACT

In the UK since 1970, many laws have got put into legislation to make employees, whether they are male or female, get the same amount of pay for the same job. The ultimate aim was to reduce the pay difference between men and women, with the equal pay act in 1970 and then the equality act in 2010. However, with these laws getting brought in, women are still getting paid less than men. With the introduction of machine learning (ML) into our everyday lives, these biases will only get more and more, as the models will be learning from the past decisions. Therefore, we have created a solution that will look at finding the bias to mitigate it. Ensuring that the bias for a suggested employees pay would be removed, no matter if they are a man or a woman. Our empirical study achieved very positive results, which we hope that this concept can then get transferred to other situations in the same context.

INTRODUCTION

We aim to remove the potential for gender bias in a suggested pay to an employee from data with a clear gender bias within the dataset. Through using pre and post-processing data methods to identify the bias and then aim to remove it. Due to the years of women and men getting paid different amounts. Even though laws have come into power to prevent such a thing from happening, but yet still does.

In 2018, women, no matter their background, on average earned just 82 cents for every $1 earned by men [3]. Machine Learning (ML) requires vast amounts of past data to inform future events, with AI and ML being the key driver behind many decisions. However, with there being a well-known gap between a person's gender and their pay, the ML models will only learn this and use this as a factor in their decisions making. Therefore, to stop this from happening, a system needs to be put into place to remove this process's bias. Therefire, through using fairness techniques at pre and post-processing stages [10] of supervised learning, we will aim to remove the bias of someone's gender from a suggested pay salary for an individual to make the whole process seem fairer for all.

With mothers in part-time jobs getting hit by a "pay penalty" and often are not given pay rises linked to experience [2], a sense of fairness needs to be take for the skills they have. Additionally, "more than three out of four UK companies pay their male staff more than their female staff, and in nine out of 17 sectors in the economy, men earn 10% or more on average than women" [17].

Therefore, even though by law in the UK, it is illegal to pay men and women different amounts for doing the same job, a gender pay gap still exists, and there is still a difference in equal pay. This difference in equal pay is driven by the use of salary pay scale ranges and, in some cases, a bias to think that women are not as valuable as men within a similar role.

So we are proposing a solution that will help reduce the bias within salary negotiations that removes the elements of gender bias to produce a tool that will look at the person's job role and their years of experience. Additionally, not taking into account previous pay and if they were working full or part-time.

Therefore, we will create a model that will predict a recommended employees wage based on their years of experience, not taking into account their gender or if they had previously worked part-time or not. The differences will be found in the two genders wages to neutralise these differences so that the model will be almost identical for both men and women. Doing this should also take away any potential unconscious bias that the employer might have regarding what amount of pay to give to the employee.

This model will get achieved by using mitigating bias methods in pre and post-processing. By finding out where the bias is, we can then change decision boundaries to present a fairer outcome based on only years of experience.

The model was able to predict both men and women on an almost identical linear prediction. Only very slight differences were between the different genders and their linear plotted regression lines. These differences could get removed entirely with additional minor tweaks to the model parameters. Overall, the result was very positive.

We will next look into the background of this topic and review existing literature. Then an explain of the study design, the libraries used to create the solution, the dataset used, and the pre and post data processing techniques used. We will then present the results and analyse them. Finally, a discussion on the empirical study and concluding the overall findings.

BACKGROUND & LITERATURE REVIEW

In 2018 companies with 250 or more employees were expected to file their first data on their gender pay gap data. With the actions of companies publishing their data, this created a significant discussion around what people earn and ultimately what difference in what men and women earn [8]. In 2017 a UK male earned 18.4% more than a woman [13].

It is important to note that the pay gap is not the same thing as equal pay. Equal pay is a law that got legislated in 1970 that ensures that men and women doing the same job should get paid the same amount [8]. Due to the equal pay act 1970 [1]

and the equality act 2010 [7], it is illegal to pay works different amounts for the same man if they are a man or woman.

While companies like DDB UK, the legal entity, which includes Adam & EveDDB, Tribal Worldwide, Gutenberg Global and Cain & Abel, releases detailed gender pay gap reports each year. The company has, since 2017, enforced a rule that 50% of the candidates for senior positions getting interviewed must be female. Additionally, they are initiating mandatory unconscious bias training [4]. However, while making these changes, a recent slowdown in reducing the gap has occurred. Additionally, the gender pay gap between graduates has not improved since 1993. There have been no improvements despite the gap getting a reduction for non-graduates [2].

Mothers in part-time jobs are getting hit by a "pay penalty" and are often not given pay rises linked to experience [2]. By the time a couple's first child is aged 20, many mothers earn nearly a third less than the fathers. A key factor was women working part-time in motherhood [6]. This situation that mothers find themselves in gets referred to as the "motherhood penalty" [14]. For mothers born in 1970, there is a 34% overall gender pay gap. This gap is primarily due to the impact of parenthood on earnings, with women earning less and men earning more after having children. However, there is a significant but much smaller gender pay gap between childless women and men born at this time, a 12% gap [14].

In 2020, "More than three out of four UK companies pay their male staff more than their female staff, and in nine out of 17 sectors in the economy, men earn 10% or more on average than women" [17]. This is an issue generated when workers get paid, on a scale, based on their "experience". It opens up an opportunity for employees to be paid different amounts based on what is perceived to be their worth, which can have an unconscious bias at the heart of it.

As we have already mentioned, bias exists whether it be conscious or unconsciously within everyday life. However, these biases that we humans have will only make those biases even more prominent when used to train ML models. This bias could be visibly evident or could go without a trace, but the models will only learn from past data to predict the future, and if that passed data is biased, the future will be significantly biased.

There are several ways that a model can become biased. These include: sampling bias, measurement bias, exclusion bias, experimenter or observer bias, prejudicial bias, conformational bias, and bandwagoning or bandwagon effect [12].

We need to make sure that bias does not become ingrained within our ML models. When it does, it can harm our daily lives. The bias can get manifested in exclusion, such as certain groups getting denied loans or technology not working the same for everyone. As AI becomes more a part of our lives, the dangers from bias only grow larger [12].

There are multiple ways to reduce bias (see fig: 1). The first thing that is required is to understand what the bias is [10]. In order to understand the bias, this stage has three sub-sections within it, Socio-technical causes of bias, bias manifestation in data, fairness definition [10]. Another way is by aiming to mitigate the bias [10]. This stage involves preprocessing, in-processing and post-processing the data and model [10]. The final method is accounting for bias [10]. This method involves there being a bias-aware data collection, describing and modelling bias and finally explaining ai decisions [10].

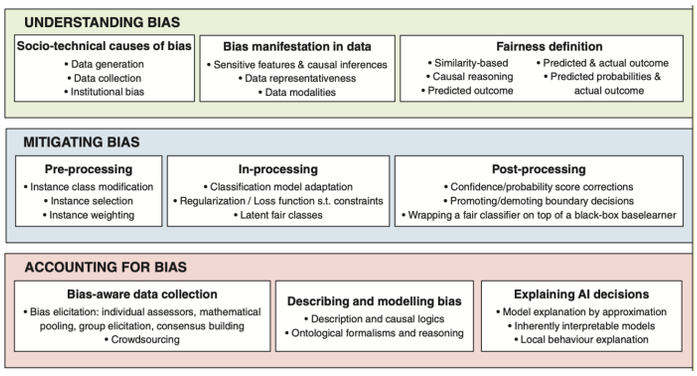

Figure 1. Overview of the areas of bias to consider [10].

Study Design

Our study is around the topic of bias in algorithms and looking at ways to remove these biases. The study will be looking at ways to detect the bias, measure it and then reduce it. Our study got carried out in a manner that follows an empirical experimental study method.

Our study has looked into ways to identify bias within a dataset and then look at ways to remove this bias, ensuring that protected characteristics, in our case gender, do not impact or impede a person's proposed suggested salary.

We aimed to try and find out if there was first any bias within the dataset. We did this by first plotting out the dataset based on the characteristic of male and female. We initially created a model that would truthfully represent the gender bias within the model's predictions.

To achieve removing the bias, we extracted the prediction-specific interactions. By getting the interactions, we could cancel out their effect and the influence driven by the gender variable. We then re-calculate our predictions, which immediately shows the removal of any sign of prejudice within the dataset. Additionally, the model conscientiously captures the variation driven by the employee's years of employment and career path.

This bias-aware approach to modelling can be applied to other forms of input types, with a similar approach being used by Google [11]

Libraries

We used Python 3 [15] to create the empirical experiment. Additional libraries used were Pandas [9] to allow us to load in the data and wrangle the data frames. Seaborn [16] was also used to visualise the data. XGBoost [5] to create the model and extract the critical interactions from within the model.

Dataset and Data Pre & Post-processing

We create our dataset to simulate data containing gender, years of experience, and career type. Our overall aim is to predict the salary of someone while taking these features into account. The type of career will be focusing on software engineering (SWE) and consulting. We decided to create this dataset synthetically due to time restrictions and lack of data collection containing real-world figures of these two career options. Most datasets found provided more of an overview than the exact figures, and we believe this is due to the sensitive and personal nature of the required data.

We based the synthesised data on general assumptions about pay. There will be a clear positive relationship between years of experience and a person's salary. An SWE will earn less than a consultant, and being male will earn them more money than females generally. Additional factors within the data are that in SWE roles, women will start at the lower end of the scale while men will be varied and, therefore, women will be over time increasing their pay. However, this increase will be at a faster rate than men but from a lower starting point. While for consulting, both males and females will start at the same rate, but men will get more considerable increases in pay over time compared to their women counterparts.

As some of the columns contained text categorical data types, we first processed the data to convert this to a numerical value. The data got split into a train test set of 70/30%.

We ensured that when the model was making its predictions, the predictions' interactions got provided with the results. We then created a bias variable based on the model's gender values and a bias index based on the values being in the datasets feature names and the bias-variance.

We were then able to create a de-biased $\hat{y}$ value by summing the interaction values and then minus 1. The outputted value then has created the new values with the gender bias removed.

Parameter Settings

We set the learning rate to the model as 0.1, between the range of 0 and 1 that the model expects. The learning task for the model got set to regression squared error, which looks at the regression squared loss. The number of boost rounds got set to 100 as well.

RESULTS & ANALYSIS

When we look at the dataset in its original representation (see fig: 2), we can see a clear divide between the male and female pay salaries. We can see that the regression line for SWE is ~42,000 for women and ~60,000 for men starting, and then both these gender values increase to ~105,000. Therefore both values roughly reaching the same value at 20 years of experience. While for consulting, the starting off values are very close to each other, ~60-65,000. However, these values completely deviate when at the top end of the scale. Men have a value of ~115,000, while women have a value of ~105,000.

Before removing the gender bias, we can see that the model's predictions mirror a similar result to the original data's general trend for SWE (see fig: 3). However, when we look at the values when the gender bias gets removed, the model's predictions almost align perfectly with each other. It is evident that there is some difference between the two gender predictions, but the difference is extremely marginal when concerning SWE employees. With both men and women being ~65,000 starting and ~112,000 at 20 years of experience.

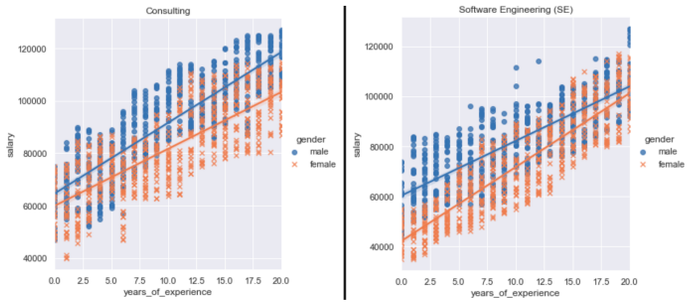

Figure 2. The dataset's datapoints and initial regression fit with the data being unmodified, split by gender and career option.

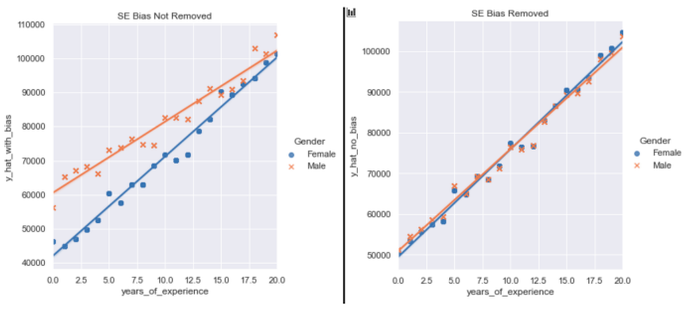

Figure 3. The model's 'Software Engineering' initial $\hat{y}$ regression prediction with bias and the $\hat{y}$ predition after the bias has been removed.

In relations to consulting, we can see that the predictions with the bias still intact, the model's predictions follow the same patterns as the data overall (see fig: 4). The regression representation is close to being equally matched. With the starting off wage for both men and women around ~60-65,000 and the top end of the range being ~109-111,000. It is evident that there is some difference between the two gender predictions, but the difference is very marginal concerning consulting employees.

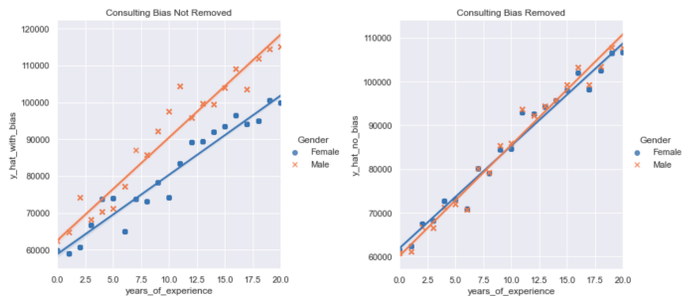

Figure 4. The model's 'Consulting' initial $\hat{y}$ regression prediction with bias and the $\hat{y}$ predition after the bias has been removed.

Therefore, overall we can see that the approach has created almost perfect un-bias results between genders over the different career options. While it has not matched the values up entirely between the genders, it has created a much more even representation than prior.

DISCUSSION & CONCLUSION

Overall, we can see that the gender aspect now has little impact on the model's overall predictions. However, while the bias has not entirely removed, it has dramatically reduced and almost disappeared. Therefore we believe that further investigation is required to eradicate this, but we can see we are taking steps in the right direction.

This model shows is that when using salaries based on the years of experience, when can remove the gender bias within the data. however, further exploration is required to see if this is still the same case when looking at other professions, to see if there might be a slight difference in salaries but maybe not a significant difference, to whether this technique will still have the same positive outcome. We believe it will still be able to be used across the board on similar data sets.

While we can say, more work needs to get done to determine the approaches robustness in other contexts. It is clear that this approach will harm the model's accuracy metric scores. Especially when we compare the results, our de-biased model predicts compared to our testing data's actual results. We believe this decrease in the model's performance gets justified due to the model's ability to create a more levelled playing field between men and women. Using their years of experience as the main driving force and nothing else when predicting an employee's potential pay amount.

As women get penalised for working part-time and their years of experience not taken into account, this has lead to women getting paid a lot less than men. Therefore, we believe this method would stop any individual, whether it be men or women, who decide to go part-time at any point in their lives get treated with the same opportunity in potential pay. Especially compared to full-time workers, as they might be working full time, the experience is still the same. However, ultimately the model will aim to remove years of deep bias in our methods and history in pay amounts awarded to men and women. Therefore, by building a model that does not prejudice through its training data, we must first let the model first measure the amount of prejudice. We can then reset the bias contributing factors to zero.

REFERENCES

[1] UK Public General Acts. 1970. Equal Pay Act 1970. (29 May 1970). legislation.gov.uk, Retrieved May 5, 2021 from: https://www.legislation.gov.uk/ukpga/1970/41/enacted.

[2] BBC. 2018. Mothers suffering 'pay penalty' at work, report suggests. (5 February 2018). BBC News, Retrieved May 5, 2021 from: https://www.bbc.co.uk/news/business-42939584.

[3] Robin Bleiweis. 2020. Quick Facts About the Gender Wage Gap. (2020). https://www.americanprogress.org/issues/women/reports/2020/03/24/482141/quick-facts-gender-wage-gap.

[4] Sam Bradley. 2021. Here's how UK agencies are improving their gender pay gaps. (5 May 2021). The Drum, Retrieved May 5, 2021 from: https://www.thedrum.com/news/2021/05/05/here-s-how-uk-agencies-are-improving-their-gender-pay-gaps.

[5] Tianqi Chen and Carlos Guestrin. 2016. XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16)*. ACM, New York, NY, USA, 785–794. DOI: http://dx.doi.org/10.1145/2939672.2939785

[6] Monica Costa Dias, Robert Joyce, and Francesca Parodi. 2018. Mothers suffer big long-term pay penalty from part-time working. (5 February 2018). BBC News, Retrieved May 6, 2021 from: https://www.ifs.org.uk/publications/10364.

[7] UK Statutory Instruments. 2010. The Equality Act 2010 (Equal Pay Audits). (29 May 2010). legislation.gov.uk, Retrieved May 5, 2021 from: https://www.legislation.gov.uk/uksi/2014/2559/contents/made.

[8] Lora Jones. 2018. What is the gender pay gap? (5 February 2018). BBC News, Retrieved May 5, 2021 from: https://www.bbc.co.uk/news/business-42918951.

[9] Wes McKinney and others. 2010. Data structures for statistical computing in python. In *Proceedings of the 9th Python in Science Conference*, Vol. 445. Austin, TX, 51–56.

[10] Eirini Ntoutsi, Pavlos Fafalios, Ujwal Gadiraju, Vasileios Iosifidis, Wolfgang Nejdl, Maria-Esther Vidal, Salvatore Ruggieri, Franco Turini, Symeon Papadopoulos, Emmanouil Krasanakis, and others. 2020. Bias in data-driven artificial intelligence systems—An introductory survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 10, 3 (2020), e1356.

[11] Ben Packer, Yoni Halpern, Mario Guajardo-Céspedes, and Margaret Mitchell. 2018. Text Embedding Models Contain Bias. Here's Why That Matters. (1 May 2018). Google AI, Retrieved April 27, 2021 from https://developers.googleblog.com/2018/04/text-embedding-models-contain-bias.html.

[12] Ronald Schmelzer. 2020. 6 ways to reduce different types of bias in machine learning. (10 June 2020). TechTarget, Retrieved May 8, 2021 from https://searchenterpriseai.techtarget.com/feature/6-ways-to-reduce-different-types-of-bias-in-machine-learning.

[13] Roger Smith. 2017. Annual Survey of Hours and Earnings: 2017 provisional and 2016 revised results. (26 October 2017). BBC News, Retrieved May 5, 2021 from: https://www.ons.gov.uk/employmentandlabourmarket/peopleinwork/earningsandworkinghours/bulletins/annualsurveyofhoursandearnings/2017provisionaland2016revisedresults/gender-pay-differences.

[14] Trade Union Congress. 2016. The Motherhood Pay Penalty - Key Findings from TUC/IPPR Research March 2016. (29 March 2016).

[15] Guido Van Rossum and Fred L. Drake. 2009. *Python 3 Reference Manual*. CreateSpace, Scotts Valley, CA.

[16] Michael L. Waskom. 2021. seaborn: statistical data visualization. *Journal of Open Source Software* 6, 60 (2021), 3021. DOI: http://dx.doi.org/10.21105/joss.03021

[17] Aleksandra Wisniewska, Billy Ehrenberg-Shannon, and Sarah Gordon. 2020. Gender pay gap: how women are short-changed in the UK. (25 September 2020). Financial Times, Retrieved May 5, 2021 from: https://www.legislation.gov.uk/uksi/2014/2559/contents/made.