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

Andy Gray
445348
445348@swansea.ac.uk

INTRODUCTION

[Project Discription]
We aim to remove the potential gender bias in a suggested pay to an employee from data with a clear gender bias within the dataset.

[Motivation]
In 2018, women, no matter their background, on average earned just 82 cents for every $1 earned by men [?]. ML requires vast amounts of past data to inform future events, with AI and machine learning 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.

Using fairness techniques at preprocessing stages [?] of supervised learning, we will aim to remove the bias of someone's gender from a suggested pay salary for an individual.

[Summary of existing lit]

[Problems with lit]

[Project Spec]

[Result findings]

[Overview]
We will next look into the background of this topic and review existing literature. We will then explain the study design, the libraries used to create to solution and the dataset used, and the pre and post data processing techniques used. We will then present the results and analyse them. Finally, we will be discussing the over empirical study and concluding the overall findings.

BACKGROUND & LITERATURE REVIEW

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 [3]

Libraries

We used Python 3 [4] to create the empirical experiment. Additional libraries used were Pandas [2] to allow us to load in the data and wrangle the data frames. Seaborn [5] was also used to visualise the data. XGBoost [1] 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: 1), 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.

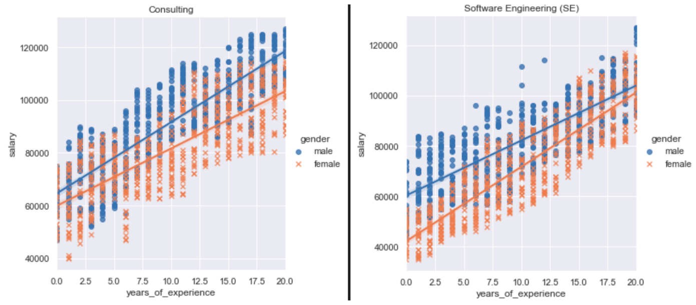

Figure 1. The dataset's datapoints and initial regression fit with the data being unmodified, split by gender and career option.

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: 2). 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.

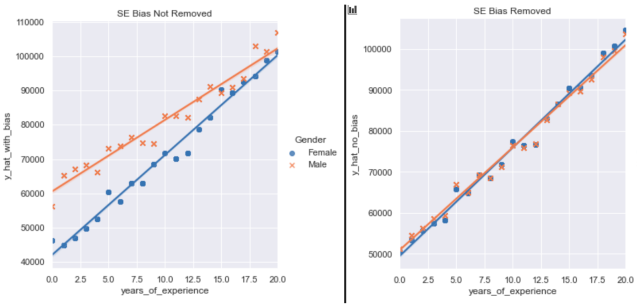

Figure 2. The model's 'Software Engineering' initial $\hat{y}$ regression prediction with bias and the $\hat{y}$ predition after the bias has been removed.

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: 3). 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.

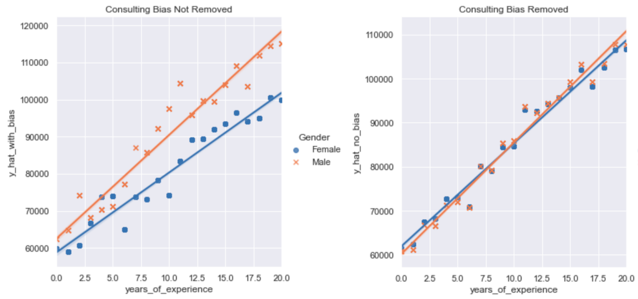

Figure 3. The model's 'Consulting' initial $\hat{y}$ regression prediction with bias and the $\hat{y}$ predition after the bias has been removed.

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

[discussion]

[conclusion]

I would like to point out from the outset that there is no question that this approach will lead to a decrease in model performance on your validation data. In our contrived example, the RMSPE is 12% for predictions that encode bias and 14% for predictions where we removed the gender contribution. Nonetheless, this decrease in performance is acceptable and encouraged in many settings. After all, the purpose of your models is not only to make good predictions but to also allow you to identify ways to pull levers, such as modify user behavior on your website or prevent harmful things from happening when diagnosing a disease. Hence, if you want to build a model that is

not prejudiced by your data, you can't go wrong with letting the model first measure the amount of prejudice and then resetting all the bias contributing factors to zero.

REFERENCES

[1] 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

[2] 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.

[3] 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.

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

[5] 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