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]

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.

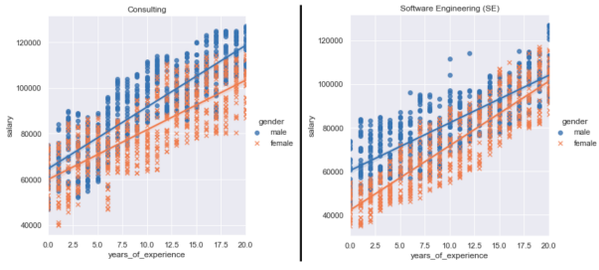
Dataset and Data Preprocessing

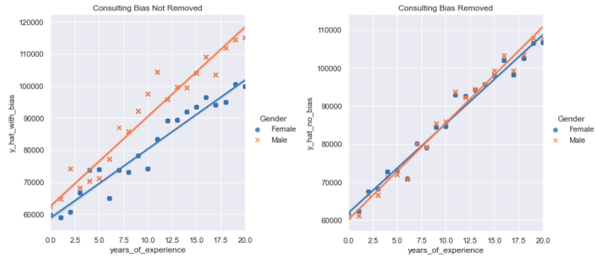
Synthetic dataset based around two job types -> why because was hard to find data set that does not provide "just an overview".

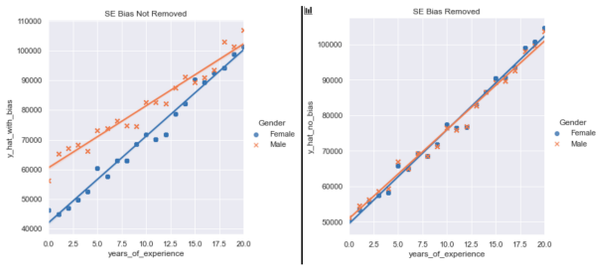
pandas.get_dummies

RESULTS & ANALYSIS

Unmodified dataset:







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