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

Andy Gray
445348
445348@swansea.ac.uk

HYPOTHESIS OR CONJECTURE

To remove the potential gender bias in a suggested pay to an employee from data with a clear gender bias within the dataset.

FOCUS OF THE STUDY

The study aims to remove bias within algorithms. This aim is within a context that there is an awareness of bias within the data. It is well documented and known that women are paid less than men for doing the same role. This situation is known as the gender pay gap. When companies are looking at how much to offer new workers or performance reviews to current employees, when current employees' data get used to forming how much a person should get paid, a man and women will receive different amounts. An ML model will likely figure this out. Therefore we aim to remove the bias and prejudice in the data to build a model representing the employees more fairly.

Using bias-aware algorithms to figure out how much bias exists in the data and then measure it to remove its effects. This de-bias will get achieved by finding how much bias exists in the dataset. Once this gets identified, the model is used to measure it correctly and then subtract that bias's effect on the outcome.

Additional libraries like Shapley Additive Exploration and LIME will get used to gain insights into the models' explainability. These explainability tools will allow us to see what impacts the model's predictions and check that gender bias is removed from the model's predictions.

RESEARCH LANDSCAPE AND SOCIAL SIGNIFICANCE

In 2018, women, no matter their background, on average earned just 82 cents for every $1 earned by men [1]. ML requires many 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.

Through using fairness techniques at preprocessing stages [3] of supervised learning, we will aim to remove the bias of someones gender from a suggested pay salary for an individual.

OUTLINE THE EXPERIMENTAL METHOD(S)

The empirical experimental study will aim to plot the initial dataset to see where the decisions are for the different genders in questions. Through visualising techniques and predicting the potential outputs for each gender and their years of service, we will have a benchmark as to what the data, when un-biased, would have produced for the employee.

We will then aim to remove this gender bias by first identifying the bias and then removing it. While again visualising the de-biased data and then using the exact predictions as previously used to see what the new results would be and demonstrate a de-bias in the model's output. We will use additional libraries to gain insights and explainability from the outputs to see how much impact the methods have had on removing gender bias.

DATA TO BE USED

We will be using self-created simulated data containing gender, years of experience, and career type. The overall aim is to predict the salary of someone while taking these features into account. We will also predict the salary of an employee while also removing any gender bias within the results. The type of career will be focusing on software engineering (SWE) and consulting.

As the initial dataset will be synthetic, based on general assumptions about pay, which are well known, there will be a clear positive relationship between years of experience and a person's salary. A SWE will earn less than a consultant, and being male will earn them more money than females. Additional considerations 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.

CONCEPTS TO BE DISCUSSED

The study's concepts will be to remove gender bias from a predictive model and use tools to look at the model's explainability and what gets used to create the predictions. This removal of bias will get done by using preprocessing de-biasing methods. As human society had a long history of suffering from cognitive biases leading to social prejudices and mass injustice [4], we will aim to remove the bias that the models will gain from our cognitive biases.

This study will aim to make pay reviews fairer for each gender and standardise. As the outputted model will aim to allow frequently review salaries for parity between genders and races. When recruiting, set the pay range offered on years' experience with some leeway for notable achievements, not how well the candidate negotiated their last pay package [2].

REFERENCES

[1] 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.

[2] Jessica Fuhl. 2020. 10 ways to eliminate gender bias in the workplace. (2020). https://www.sage.com/en-gb/blog/eliminate-gender-diversity-workforce/.

[3] 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.

[4] Procheta Sen and Debasis Ganguly. 2020. Towards Socially Responsible AI: Cognitive Bias-Aware Multi-Objective Learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 2685–2692.