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

Andy Gray
445348
445348@swansea.ac.uk

HYPOTHESIS OR CONJECTURE

To remove the potential gender bias in suggested pay to an employee from data with a clear gender bias within the dataset. Using bias-aware algorithms to figure out how much data exists in the data to measure the bias then correctly and then remove the outcome's bais' effects.

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.

Additional libraries like Shapley Additive Exploration or possibly LIME will get used to gain some insights into the explainability of the models.

RESEARCH LANDSCAPE AND SOCIAL SIGNIFICANCE EVIDENCE

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 [] of supervised learning, we will aim to remove the bias of someone's gender from a suggested pay salary for an individual.

OUTLINE THE EXPERIMENTAL METHOD(S)

The experiment will aim to plot the initial dataset to see where the decisions are for the different genders in questions. We will then aim to remove this gender bias by first identifying the bias and then removing it.

We will also aim to use additional libraries to gain insights and explainability from the outputs to see how much of an 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. An 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 look at removing gender bias from a predictive model and using tools to look at the explainability of the model and what gets used to create the predictions.

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`.