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

Andy Gray
445348
445348@swansea.ac.uk

ABSTRACT

Author Keywords

Authors' choice; of terms; separated; by semicolons; include commas, within terms only; this section is required.

CCS Concepts

•**Human-centered computing** → **Human computer interaction (HCI);** *Haptic devices;* User studies; Please use the 2012 Classifiers and see this link to embed them in the text: https://dl.acm.org/ccs/ccs_flat.cfm

INTRODUCTION

Hypothesis or Conjecture

To remove the potential gender bias in suggested pay to an employee from data that has 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.

MAIN

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.

CHI'20, April 25–30, 2020, Honolulu, HI, USA

ISBN 978-1-4503-6708-0/20/04…$15.00

DOI: https://doi.org/10.1145/3313831.XXXXXXX

Research Landscape and Social Significance Evidence

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.

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 simulated data containing gender, years of experience and type of career initially. 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.

ACKNOWLEDGMENTS

REFERENCES

[1] ACM. 1998. How to Classify Works Using ACM's Computing Classification System. (1998). http://www.acm.org/class/how_to_use.html.

[2] R. E. Anderson. 1992. Social Impacts of Computing: Codes of Professional Ethics. *Social Science Computer Review December* 10, 4 (1992), 453–469. DOI: http://dx.doi.org/10.1177/089443939201000402

[3] Anna Cavender, Shari Trewin, and Vicki Hanson. 2014. Accessible Writing Guide. (2014). http://www.sigaccess.org/welcome-to-sigaccess/resources/accessible-writing-guide/.

[4] @_CHINOSAUR. 2014. "VENUE IS TOO COLD" #BINGO #CHI2014. Tweet. (1 May 2014). Retrieved Febuary 2, 2015 from https://twitter.com/_CHINOSAUR/status/461864317415989248.

[5] Morton L. Heilig. 1962. Sensorama Simulator. U.S. Patent 3,050,870. (28 August 1962). Filed Februrary 22, 1962.

[6] Jofish Kaye and Paul Dourish. 2014. Special issue on science fiction and ubiquitous computing. *Personal and Ubiquitous Computing* 18, 4 (2014), 765–766. DOI: http://dx.doi.org/10.1007/s00779-014-0773-4

[7] Scott R. Klemmer, Michael Thomsen, Ethan Phelps-Goodman, Robert Lee, and James A. Landay. 2002. Where Do Web Sites Come from?: Capturing and Interacting with Design History. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '02)*. ACM, New York, NY, USA, 1–8. DOI:http://dx.doi.org/10.1145/503376.503378

[8] Nintendo R&D1 and Intelligent Systems. 1994. *Super Metroid*. Game [SNES]. (18 April 1994). Nintendo, Kyoto, Japan. Played August 2011.

[9] Psy. 2012. Gangnam Style. Video. (15 July 2012). Retrieved August 22, 2014 from https://www.youtube.com/watch?v=9bZkp7q19f0.

[10] Marilyn Schwartz. 1995. *Guidelines for Bias-Free Writing*. ERIC, Bloomington, IN, USA.

[11] Ivan E. Sutherland. 1963. *Sketchpad, a Man-Machine Graphical Communication System*. Ph.D. Dissertation. Massachusetts Institute of Technology, Cambridge, MA.

[12] Langdon Winner. 1999. *The Social Shaping of Technology* (2nd ed.). Open University Press, UK, Chapter Do artifacts have politics?, 28–40.