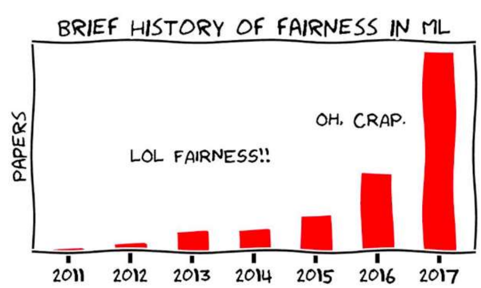
Reducing Age Bias in Machine Learning:
An Algorithmic Approach

Adriana Solange Garcia de Alford,
Steven Hayden, and Nicole Wittlin
Amy Atwood, Ph.D,
Senior Data Scientist – T-Mobile, Capstone adviser

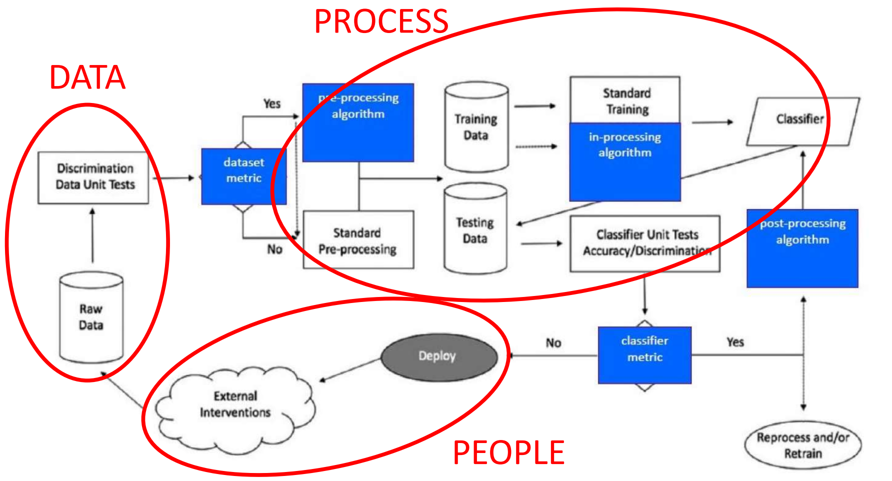
Bias in Machine Learning



Taken from Moritz Hardt lecture notes.

Is it BIASED because it is UNFAIR?
Is it FAIR because it is UNBIASED?

Bias in Machine Learning



https://aif360.mybluemix.net/

Bias in Machine Learning

PROCESS

Intervention Points

Pre-Processing

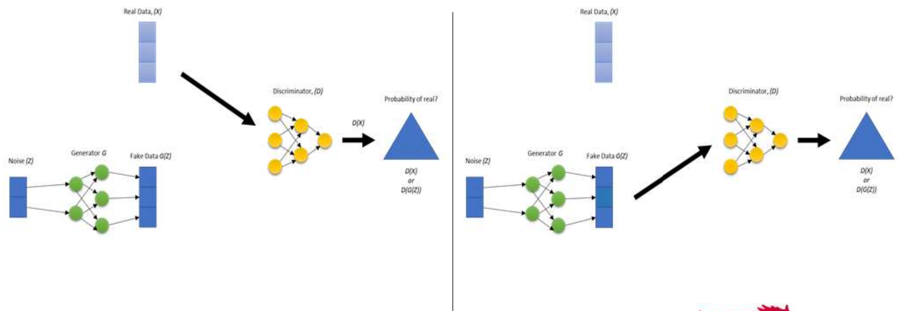
In-Processing

Post-Processing

- Classification
- Regression
- AI Fairness 360
  Prejudice Remover
- **Adversarial Debiasing**

GAN Overview

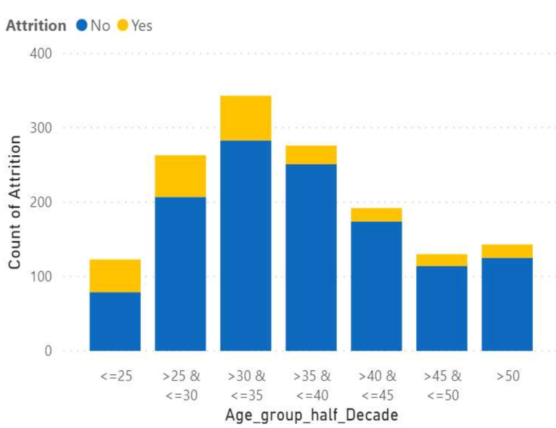
https://ya-webdesign.com/imgdownload.html

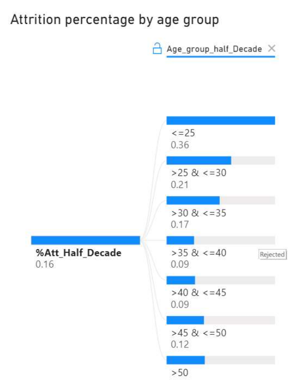


DataScience@SMU 5

Experiment

Understanding Employee Attrition





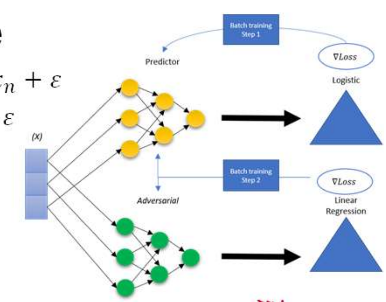
DataScience@SMU 6

Experiment

- Overview Experiment
  - Age(Z) is protected
  - Age(Z) Correlated with explanatory (X) of predictor model
- Goals
  - Good Accuracy
  - Demographic Parity
    - Both protected and unprotected classes receive a positive outcome at equal rates.
    - Demographic Parity = True Positives + False Positives

Experiment

- Baseline
  - Logistic model
  - $\hat{y}_{Attrition} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_n x_n + \varepsilon$
- Adversarial Architecture
  - $\hat{y}_{Attrition} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_n x_n + \varepsilon$
  - $\hat{A}_{Age} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_n x_n + \varepsilon$
  - $Loss = -\alpha \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$



Experiment Results: Fairness

Goal: Improve group fairness based on demographic parity

- Evaluated differences in **accuracy** and **demographic parity** between the baseline model and a GAN model
- Calculated standard metrics to evaluate performance
- Calculated several other metrics to evaluate group fairness
- Metrics were calculated for 7 age groups in 5-years increments

Experiment Results: Accuracy

- Initial step to evaluate **Accuracy** of both models when predicting Attrition
- Accuracy from GAN model was compared to accuracy baseline model

ACCURACY	<= 25	(25, 30]	(30, 35]	(35, 40]	(40, 45]	(45, 50]	>50	OVERALL
Baseline/Pre-Gan	0.7576	0.8400	0.8315	0.9138	0.9464	0.8696	0.7941	0.8587
Post-Gan	0.6970	0.800	0.8314	0.8966	0.8929	0.9130	0.7941	0.8315

Table 1: Accuracy Comparison of Baseline and Adversarial Models by Age Group

- Accuracy in both models was expected to be similar
- Accuracy from GAN was lower across all groups
- Groups less than 35 and older population over 50, resulted in a lower accuracy on Attrition
- Attributed this to larger number of observations in the middle age groups

Experiment Results: Demographic Parity

- **Demographic Parity (DP) is achieved when:**
  - Each group has equal likelihood to be assigned a positive outcome
  - Proportion of positive predictions in the subgroups is close to each other

DEMOGRAPHIC PARITY	<= 25	(25, 30]	(30, 35]	(35, 40]	(40, 45]	(45, 50]	>50	OVERALL
Baseline/Pre-Gan	0.9394	0.9600	0.9438	0.9655	0.9464	0.9565	1.000	0.9375
Post-Gan	1.000	1.000	0.9888	0.9828	1.0000	1.000	1.000	0.9973

Table 2: Demographic Parity Comparison of Baseline and Adversarial Models by Age Group

- **Improved <u>DP</u> range across all groups:**

  Baseline between 94-100%; **GAN** range 98-100%

- **Small trade-off between Accuracy and Fairness GAN:**

  Accuracy decreased 2% but <u>DP</u> increased 6%.

Conclusions

- **Most adversarial debiasing work focused** on protected groups such as race, sex and gender bias; we considered binned data
- Achieved **Demographic Parity** based on results from a comparative analysis between the baseline model and the GAN model
- Our focus was on <u>Age debiasing</u>, and how age bias can be <u>prevented in deep learning models</u>

Conclusions

- **Bias must be addressed in advance and throughout the ML lifecycle – NOT as an afterthought**
- Mitigating bias using adversarial network architecture shows promise, yet we cannot be confident that systems are unbiased and fair

We <u>cannot</u> and <u>must not</u> replace the inquisitiveness, skepticism, moral imagination, compassion, and the sensitivity to foresee consequences that <u>humans</u> bring to bear on machine learning.

DataScience@SMU 13

Bias in Machine Learning: An Adversarial Approach

Solange Garcia de Alford, Steven Hayden, Nicole Wittlin
Master of Science in Data Science
Southern Methodist University, Dallas, TX 75275, USA

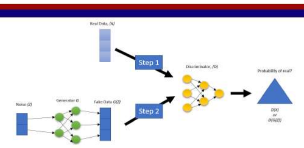
DataScience@SMU

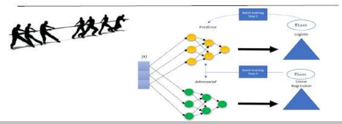
Introduction

- Bias is very prevalent, occurring in ML models at pre-process, in-process, post-process stages.
- Examples of ML bias are widely known – COMPAS, Amazon hiring algorithm/resume scan, Word2Vec. Most are binary: protected class vs unprotected class.
- Our study focuses on eliminating bias stemming from AGE when predicting employee attrition.

Main Topics

- Adversarial learning can be leveraged to mitigate bias and unfairness.
- Competing models of GAN, where Predictor (P) tries hinder Discriminator (D) with fake data, while feedback from D tries to hinder P prediction ability.
- Our study: P -- predict employee prediction; D -- predict age.
- Goal: improve group fairness via demographic parity (DP) (all equally likely of positive outcome (TP + FP)).





Model Architecture

Baseline Model

- Logistic Model
- $\hat{y}_{Attrition} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_n x_n + \varepsilon$

Adversarial Models

- $\hat{y}_{Attrition} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_n x_n + \varepsilon$
- $\hat{A}_{Age} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_n x_n + \varepsilon$
- $Loss = -\alpha \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$

Results

- **Improved DP**: range Pre-GAN for all groups between 94-100%; Post-GAN range 98-100%.
- **Small trade-off between Accuracy and Fairness Post-GAN:** accuracy decreased 2% but DP increased 6%.
- **Demonstrate work beyond binary classes**: can work toward having more than one unprotected group.
- **See Results Chart**

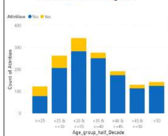
Data Overview

IBM Employee Attrition Dataset

Attrition: 84% NO / 16% YES

Age binned in 5-year ranges

More attrition <= age 35



ACCUR	< = 25	(25, 30]	(30, 35]	(35, 40]	(40, 45]	(45, 50]	> 50	Overall
Pre GAN	0.7575	0.84	0.8315	0.9138	0.9464	0.8696	0.7941	0.8587
Post GAN	0.6970	0.8	0.8315	0.8966	0.8929	0.9130	0.7941	0.8315

DEMO PARITY	< = 25	(25, 30]	(30, 35]	(35, 40]	(40, 45]	(45, 50]	> 50	Overall
Pre GAN	0.9394	0.96	0.9438	0.9655	0.9464	0.9565	1.0	0.9375
Post GAN	1.0	1.0	0.9888	0.9828	1.0	1.0	1.0	0.9972

Conclusions / Future Work

- Bias must be addressed in advance and throughout – NOT as an afterthought.
- Re-run study with larger, real dataset and/or pre-processed data that balances attrition % or sampled differently.
- Refine code with Early Stop, when the adversary has sufficiently mitigated bias and correlation is no longer detected in the adversarial model for Z(x), Age.
- We CANNOT and MUST NOT replace the inquisitiveness, skepticism, mortal imagination and compassion that humans bring to bear on Machine Learning.