Reducing Age Bias in Machine Learning
An Algorithmic Approach

Adriana Solange Garcia de Alford[1], Steven Hayden[1,2], Nicole Wittlin[1], and Amy Atwood[2]

[1] Master of Science in Data Science, Southern Methodist University, Dallas TX 75275 USA {sgarciadealford, skhayden, nwittlin}@smu.edu
[2] T-Mobile, 3255 160th Avenue SE, Bellevue, WA 98008 USA
email{steven.hayden, amy.atwood}@t-mobile.com

Abstract. In this paper, we study the prevalence of bias in machine learning; we explore the life cycle phases where bias is potentially introduced into a machine learning model; and lastly, we present how Adversarial Learning can be leveraged to remove unwanted bias and unfair behavior from a machine learning algorithm. This study focuses particularly on the topics of age bias in predicting employee attrition and presents a practical approach for how Adversarial Learning can be successful in mitigating age bias.

1 Introduction

In today's age of Big Data, data are driving decision-making in almost every field imaginable. At the foundation of this decision-making is a wide range of algorithms churning through endless data to determine if an individual might be a good employee, a worthy loan candidate, a possible repeat criminal, or the university's next star graduate. But how are these decisions made? What criteria are considered? What if the outcomes are biased?

These algorithmic processes may seem like black box magic, but there are increasing trends to bring the work out of the shadows and assess machine learning for bias and fairness, as well as accuracy and efficiency. In fact, over the last few years, "the majority of the machine learning community is finally publicly acknowledging the prevalence and consequences of bias in machine learning models" [19].

Many widely studied and cited examples of bias in machine learning highlight disparity based on race and sex. One of the most frequently noted examples of biased machine learning is COMPAS, a regression model used by judges to predict if a perpetrator was likely to recidivate. The model was proven to double the predicted false positives for recidivism for African American ethnicities than for Caucasian ethnicities [13]. Another example from 2018 was discovered in Amazon's machine learning model built to evaluate candidate resumés for software engineering roles, where the algorithm explicitly penalized resumés with the word "women's" and downgraded candidates from two all-women's colleges [10]. Research teams and practitioners have been exploring approaches, techniques, and models to account for and mitigate bias in machine learning in recent years. One intriguing option is the use of adversarial learning. Our team will utilize two competing adversarial models – similar in nature to GANs (generative adversarial networks) [7] – to reduce age bias in a sample data set related to a common human resources business problem.

In the domain of human resources, employee attrition garners significant attention. The driving factors of employee attrition have been well studied, and the goal of this work is not to duplicate those efforts. Instead, we seek to explore how age may be influencing employee attrition and whether adversarial learning can reduce bias related to age when predicting if employees will stay or leave a company. The data set we will use is the IBM HR Analytics Employee Attrition and Performance data, found on Kaggle.

Age, similar to race and sex, is considered a protected class and should not legally be used as an influential factor in many situations, particularly those related to employment and the workplace. Yet, the Equal Employment Opportunity Commission (EEOC) acknowledges that "research on ageist stereotypes demonstrates that most people have specific negative beliefs about aging and that most of those beliefs are inaccurate. These stereotypes often may be applied to older workers, leading to negative evaluations and/or firing, rather than coaching and retraining" [11]. The lack of mixed-age teams and reverse-age mentoring, coupled with age stereotypes and bias, can cause varying degrees of tension in the workplace between the generations. Ultimately, our goal is to simulate exploring age bias within the context of employee attrition and demonstrate how an adversarial learning methodology could reduce bias.

2 Related Research

2.1 Overview of Bias in Machine Learning

The concept of bias – and fairness as a related, complementary topic – has been studied extensively for decades in philosophy and psychology, well "before computer science started exploring it" [17]. There is no universal, commonly accepted definition of fairness; however, it is often defined in relationship to individuals or groups, where individual fairness seeks to treat similar individuals the same, and group fairness means different groups are treated equally. Bias can be defined as a personal and sometimes unreasoned judgement [1] that can impede fairness from being achieved. For this paper, we seek to achieve group fairness, where different age groups are treated equally, through the reduction of bias due to age.

In the last few years, understanding bias in the context of machine learning has become a more widely discussed topic. One goal in machine learning work is to strive for the highest accuracy possible. But even with good accuracy, has consideration been given to how it could impact the populations (or subpopulations) the work is intended to serve or reflect? A well-documented example is COMPAS, which upon further examination was found to double predict false positives for recidivism for African American ethnicities. The developers did not set out to create a biased system, but it was an unfair and unanticipated outcome.

When defining a question or problem, IBM's AI Fairness 360 toolkit – an open source software toolkit that can help detect and remove bias in machine learning models – recommends three key questions to consider: what type of fairness (i.e. individual vs group) is trying to be achieved; what are appropriate tools and metrics depending on the type of fairness and the data; and what algorithms to use [3].

Simply removing the "protected variable," say race, gender, or age, from the data will not yield a debiased machine learning model. In fact, because other variables can be highly correlated with the protected variables and serve as proxy data, a model could even amplify bias [12]. It is important to understand where bias may be introduced in the machine learning process and the appropriate metrics to measure bias and fairness.

There are three points in the machine learning pipeline where bias can be addressed: during the data collection and preprocessing; during the selection and creation of models; and when implementing results. There are opportunities to intervene at each of these points to mitigate bias. It is recommended that intervention occurs as early in the machine learning process as possible, but it will depend on the answers to the questions about fairness and the specific data set [3].

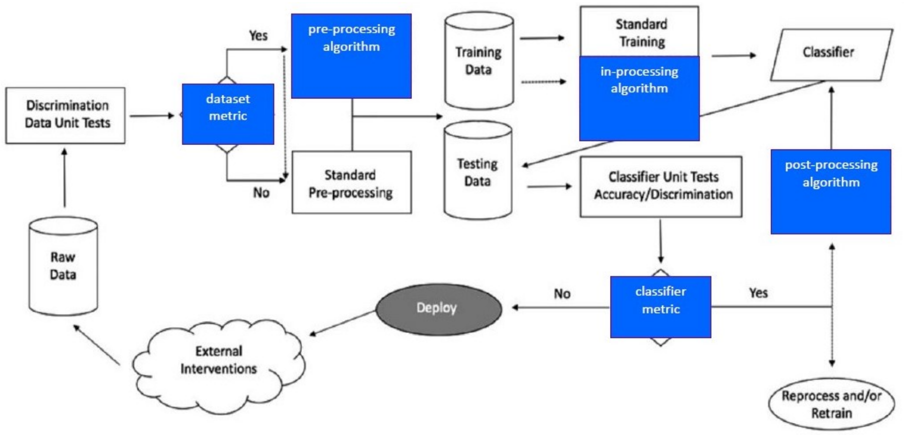


Fig. 1: IBM AI 360 Machine Learning Model graphic with Intervention Points

The IBM AI 360 tool kit offers a good visual to understand the overall process and points for intervention. See Figure 1. The first intervention point is during preprocessing. Often this relates to the origin of the data and how it was collected or sampled. For example, groups may be excluded or under represented based on population characteristics. Older generation may utilize social media less than younger adults, so sampling from social media would under sample the older generation compared to the rest of the population. Understanding population distributions and protected groups (or protected attributes) will dictate how to make adjustments, such as resampling or collecting more data, during the preprocessing stage. Preprocessing intervention modifies the training data in an attempt to mitigate bias.

When preprocessing is not an option, it may be necessary to rely on techniques to implement during the machine learning process (known as "in process"). This stage seeks to modify the learning algorithms to remove bias or discrimination during the training process, often by changing the objective function or imposing a new constraint [17]. Of particular interest to this team is the use of adversarial networks, where an adversarial algorithm attempts to remove bias during the training process.

In situations where the machine learning process takes place in a "black box" with no opportunity to intervene in the data or the algorithms, postprocessing techniques can be an option. At this stage, the model training is complete, so a bias mitigation algorithm could be applied to the resulting predicted labels. This intervention would occur before implementation of the model, and the data scientist needs to consider the impact of the results.

2.2 Overview of Employee Attrition Research

The topics of employee turnover and factors influencing attrition are well-studied by both the research community and human resources (HR) practitioners. To understand the evolution of the field, Hom, Lee, Shaw, and Hausknecht [9] conducted a comprehensive and systematic review of the most seminal publications about employee turnover in the Journal of Applied Psychology over the last 100 years. Additionally, they assessed other key theoretical and methodological contributions to the field as part of the review.

Since 2000 and forward, 21st century theory and research has focused on: the idea of embeddedness, or factors and forces influencing attrition; the evolution of the job search process; the influences of employee-organization relationships and HR management systems on collective turnover; the impact of HR practices on good-performers vs bad-performers; and the relationship of collective turnover on organizational performance. The field also saw "significant scholarly attention to employee turnover at the group, team, work unit, and organizational levels" [9].

With a basic, foundational understanding of the field of employee attrition, the team turned to exploring common predictors of turnover. Rubenstein, Eblerly, Lee, and Mitchell followed up the comprehensive work of Hom et al. to "assess the progress made in research on voluntary employee turnover" [17]. The aspect of their work that is of most interest to this team is understanding the predictors of employee attrition, especially related to the role of age.

Rubenstein, et al., explored 57 predictors across 1,800 side effects (drawn from a study population of 316 articles). They categorized these specific predictors into several groups: individual attribute predictors (tenure, age, conscientiousness); job aspects (characteristics, security, complexity, as well as traditional characteristics of pay, role ambiguity, role conflict); traditional job attitudes (involvement, satisfaction); newer personal conditions (coping, stress); organizational context (climate, size, prestige); person-context interface (fit, influence, peer/group relations); existing job market (alternatives); attitudinal withdrawal; and employee behavior (absenteeism, performance).

Frye, et al., conducted an analysis on two distinct data sets to determine the predictors of employee attrition in those two populations [5]. The first data set was anonymized information from the U.S. Office of Personnel Management (OPM), reflecting a full year of separation data from October 2014 and September 2015. Attributes of particular interest in this data set were: age data group, captured categorically in 5-year increments; a lower limit age (lower bound for age); and years to retirement (based on the Federal Employee Retirement system eligibility base of 57 years old). The features of the final model with best prediction accuracy included: lower age limit; length of service; various age bins (age 20-24, 40-44, 45-49, 50-54, or 55-59). They also utilized IBM HR Analytics Employee Attrition and Performance data, a fictional data set created by IBM data scientists, to compare the strongest features of that set with the features selected in the OPM model. Ultimately, the model results were comparable for both sets of data with the relationship between length of service, pay scale, and age being strong predictors of attrition [5].

From this background, it is evident that age does play a role in employee attrition; our team seeks to understand impact of this and assess whether or not age bias may be causing undue influence in predicting attrition. The goal of the work described in this paper is not to predict employee attrition, per say but rather to show how age bias could be reduced through adversarial learning models when attempting to predict attrition.

3 Understanding Generative Adversarial Networks (GANs)

There have been many techniques, such as Deep Neural Networks, Generative Adversarial Networks (GAN), and Convolutional Neural Network, used to classify images [8]. One technique – GAN – stands out in image classification and is the foundation for this paper's approach to reduce bias. In 2014, Ian Goodfellow and his colleges proposed a new framework to estimate generative models (a model that includes the distribution of the data itself and the likelihood of a given example) via an adversarial process, in which two models are trained simultaneously: a generative model G that captures the data distribution, and a discriminative model D that estimates the probability that a sample came from the training data rather than G. [7] The adversarial framework makes this classification into a supervised machine learning model, since it knows if the data is real or fake.

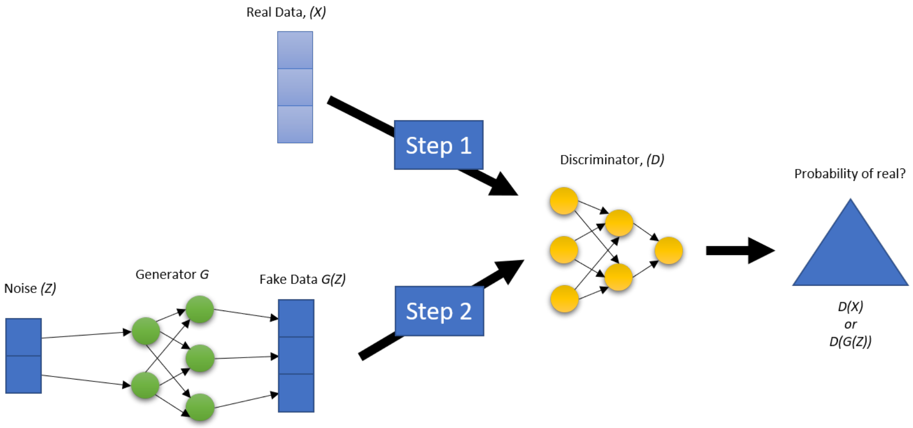


Fig. 2: Data flowing from the Generator G

Figure 2 illustrates the basic structure of GANs in two steps that happen simultaneously. For explanation purposes, each step is described separately. In Step 2, the distribution of the real data set is used to generate a fake data set by combining real data and noise, through generator (G) algorithm thus, the vector size is the same for both G(Z) and Real Data(X). Once it is done generating fake data, it passes Fake Data G(Z) to the discriminator (D).

The discriminator (D) is a classification algorithm used to determine if the data came from Fake Data G(Z) or Real Data(X). In Step 1, the same done, but this time the real data (X) gets passed to the discriminator (D). The discriminator (D) and the generator (G) update their parameters iteratively, based on a loss function (explained below) and through back propagation. Back propagation is the process of modifying the parameters (weights) of the network between neurons, so the next iteration is more accurate. If the discriminator (D) classified the data correctly as real or fake then the generator (G) gets updated. If the discriminator (D) incorrectly classified the data real or fake, then the discriminator (D) gets updated.

$$V(D,G) = E_{x\ P_{data(x)}}[logD(x)] + E_{z\ P_{z(z)}}[log(1-D(G(z)))]$$

(1)

Equation 1 describes the loss function. The first term $E_{x\ P_{data(x)}}[logD(x)]$ represents an expected output of the log of discriminator (G) when the input is from the real data. The second term $E_{z\ P_{z(z)}}[log(1-D(G(z)))]$ is the expected value of the log of the quantity of 1 minus the discriminator (G), which is making predictions about the fake samples. This function makes the generator (G) want to minimize the G(Z) and the discriminator (D) maximize it. The conflicting elements of this function make the generator (G) seek to maximize G(Z) while the discriminator (D) works to minimize G(Z). This creates the adversarial "game."

To train the discriminator (D), the process loops through the data created by the generator (G) and the real data k times before the generator (G) is updated. This allows the discriminator (D) to optimize given k, where k is a parameter chosen prior to training. GANs use the gradient of the loss function to change the parameters of the discriminator (D) to maximize the loss function that sets up the optimization between both models.

Training the generator (G) takes noise data and then transforms it to get the fake data for k times. This time it allows the generator (G) to optimize given k. The gradient of the loss function is then used to change the parameters of the generator (G) to minimize the loss function. This is a complete iteration of training a GAN model.

Goodfellow also discussed when to stop training the model. Once the discriminator (D) consistently cannot distinguish between fake or real data then training is done. In other words, once the distribution of the generator (G) matches the real data. [7] The discriminator algorithm can then be implemented in the desired tool for classifying.

4 Data Set: IBM Employee Attrition Data

The data set used for this study is fictitious employee data simulated by IBM data scientists to represent employee attrition and performance [2]. There are approximately 1,470 entries representing a specific employee, with 35 descriptive features, where 'Attrition' is the dependent, target feature to be predicted. We are treating 'Age' as a protected variable as it should not have undue influence when predicting attrition.

From the related research, we know that age does have an impact on employee attrition. Given the trends noted by the EEOC, we anticipate that attrition may increase with age. Additionally, we believe that the predicted attrition rate of our machine learning model should be approximately equal across all age groups; therefore, we are looking for evidence that attrition related to age bias may be occurring more frequently with older employees.

In the data set, 1,233 employees are classified as "no" with regards to attrition, meaning they remained at the company, and the remaining 237 (with "yes") did leave. This breakdown of 84% remaining vs 16% departing is in line with what we anticipated, and we will be cognizant of the imbalanced distribution as we proceed with the analysis.

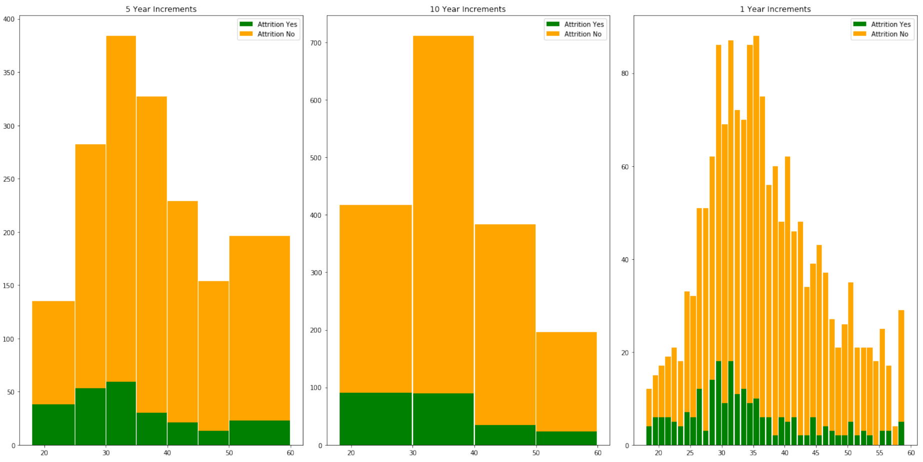
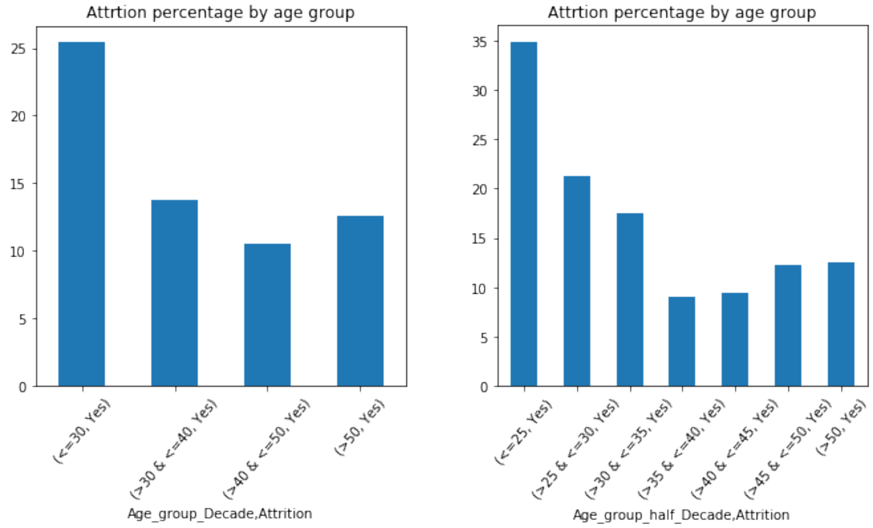


Fig. 3: Attrition vs Non-attrition of Employees by Age Groups



(a)

(b)

Fig. 4: Attrition Percentage by Age Groups

Next, we explored attrition across age range, in one-year, five-year and 10-year increments. Combining observations into buckets provides a reasonable population size for each group; the five- and 10- year increments are most useful for our purposes. At both ends of the age distribution, there were fewer records (i.e. age 18-19 reflected 17 employees; only five employees were age 60.) Therefore, we combined the youngest employees as 25 and under and the oldest employees as above 50 for analysis purposes. The distribution of employees in this data set generally follows a normal distribution, with a slight right skew toward younger employees, as illustrated in Figure 3. The green segments of the chart indicate the employees that did leave (attrition = yes).

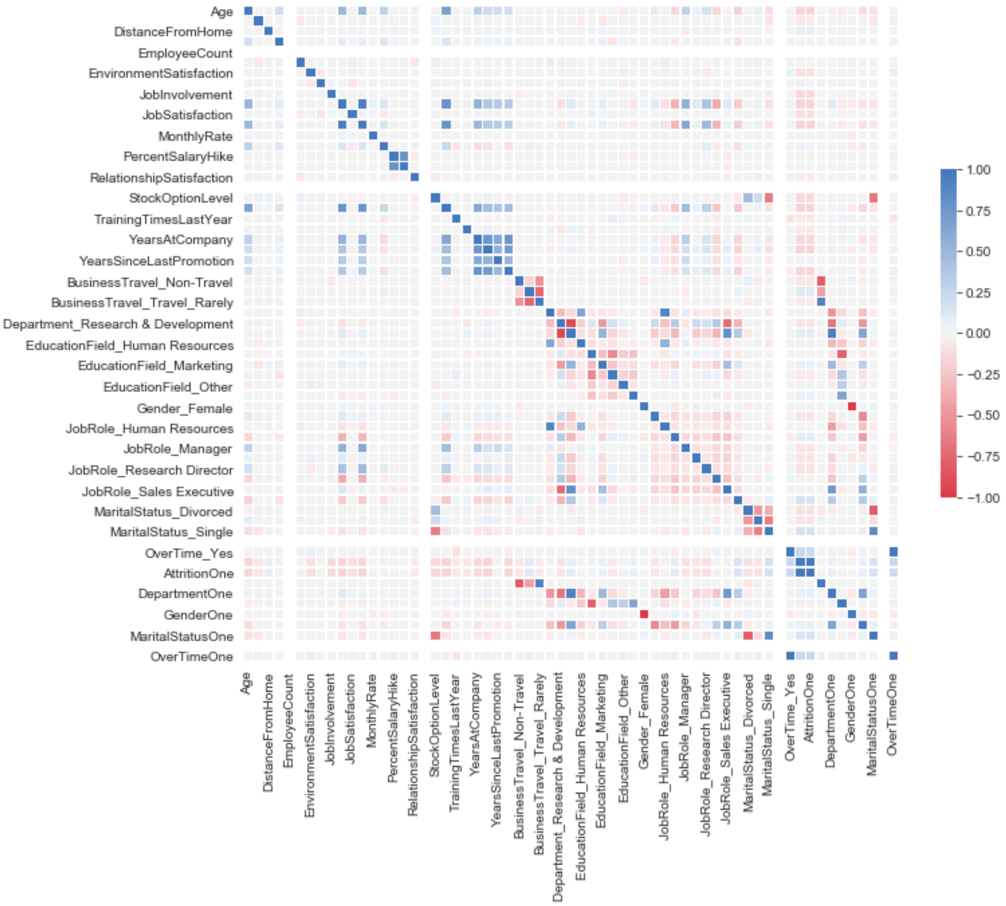


Fig. 5: Correlation Matrix for IBM HR Employee Attrition Data Set

We also looked at the percentage of attrition in each age category, for the five-year and 10-year increments. See Figure 4. The 10-year histogram shows the most valuable information: the percentage of attrition is highest for younger workers (35% for 25 and under) and then decreases over the next ten years. The percentage of older employees – age 45 and over – remains relatively stable at just above 10%. Understanding the percentage distribution of attrition across the age groups is key, since during the experiment stage we expect equal predictions across all ages to represent a fair model.

Finally, the team explored a correlation matrix of the data set to understand the relationships between variables, particularly age. In Figure 5, positive correlation is displayed in blue, and negative correlation is displayed in red. We observed that employee age is strongly related to: job level, monthly income, total number of working years, the number of years worked at the company, and number of companies he or she has worked at. A weaker relationship exists between age and educational level, number of years in the current role, number of years since last promotion and number of years with current manager. The correlation matrix also reinforces the inverse relationship between age to attrition seen in the histograms: older employees show lower levels of attrition while young groups demonstrate higher levels. These insights are critical to understand as we conduct further analysis as bias can be amplified through proxy variables that encode similar information to the protected attribute.

The results above are the most important aspect of our data analysis. Other steps taken as part of the Exploratory Data Analysis are noted below.

1. EDA was executed in Python language and libraries.
2. The Python Jupyter notebook is found in the project's GitHub URL: "https://github.com/skhayden/SMU-Capstone-Age-Bias-in-Predictive-Modeling-.git".
3. Data was extracted in Excel format and converted to CSV format.
4. The data was loaded into a Data Frame to calculate simple statistics for all features.
5. Found the value counts for unique ages
6. Calculated categorical frequency groupings to unveil trends and distributions; for example, 1,043 employees out of the 1,470 total employees rarely traveled. This type of frequency groupings is useful to unveil insights and important relationships within the data.
7. As a simplification step, the levels on the 'Education' feature were reduced from 5 to 4 by combining levels 4 and 5, since level 5 had a reduced number of employees.
8. The data set was one-hot encoded (and original features were removed) to prepare the data as input to the:
   (a) Machine learning model
   (b) Correlation matrix function
9. A correlation matrix between features is also provided for analysis purposes.
10. A Matplotlib graph chart of the correlation matrix was built for easy visualization.

5 Experiment

The anticipated outcome of this study is to demonstrate that age bias can be removed from a model based on stochastic gradient decent type models with adversarial learning. The design of our analysis is built on two models. Model one is our baseline logistic regression model trained and tested on the IBM employee data set, where the resulting accuracy of predictions are measured across the five- and 10-year age increments. Model two will use adversarial learning to mitigate bias resulting from age to predict attrition, where the desired outcome is strong accuracy and demographic parity. The parity metrics reveal some statistical measure across groups, and demographic parity is where the proportion of both protected unprotected classes receive a positive outcome at equal rates (i.e. True Positives + False Positives). Both models will split the IBM data with an 80% training set 20% test set, a common standard for machine learning. In both, the algorithms will be trained using gradient descent; this is especially critical for adversarial learning in model two when the algorithms are trained in tandem.

To design the adversarial model (model two), we turned to the work of Zhang, Lemoine, and Mitchell that demonstrates how the principles of GANs can setup two competing models with the goal to mitigate bias. The advantages of this approach are that: it can be used to enforce demographic parity, equality of odds, or equality of opportunity, which each seek to achieve a different definition of fairness; output variables and/or protected variables can be discrete or continuous; the model can be employed with both simple or complex predictions provided it is trained using gradient-based methodology; and it achieves optimality when the predictor converges to a model that satisfies the desired fairness [20].

Adversarial learning to mitigate bias starts with a model to predict Y given explanatory variables X. This part of the model is similar to the generator(G) illustrated above in the GAN overview. The negative of the gradient of the weights $\bigtriangledown_W$ is passed to the adversary model. The adversary model predicts the protected variable Z given only Ŷ. This architecture with the adversary receiving Ŷ is specifically related to demographic parity. The adversary model is similar to the discriminator(D) [20].

$$\bigtriangledown_W L_P - proj_{\bigtriangledown_W L_A} \bigtriangledown_W L_P - \alpha \bigtriangledown_W L_A \tag{2}$$

Equation 2 is how the weights should be updated according to the loss functions. It is what sets up the adversary feedback between the two models. The first term $\bigtriangledown_W L_P$ is the gradient of weights of the loss for predictor. “The middle term $proj_{\bigtriangledown_W L_A} \bigtriangledown_W L_P$ prevents the predictor from moving in a direction that helps the adversary decrease its loss” [20] The last term $\alpha \bigtriangledown_W L_A$ is much like the maximization side of the loss function of the discriminator(D). Figure 6 illustrates the model’s framework, where training occurs only with the predictor model in Step 1 and alternates between the predictor and adversarial models in Step 2.

The prediction accuracy is then measured by age groups that were created in the initial data exploration to get base reading on bias in the age groups. The adversarial and prediction algorithms can then be trained to measure the accuracy between the age groups to compare the bias between the two methods. The anticipated improvement from the baseline model to the model resulting from adversarial learning will demonstrate that the age bias in the IBM data has been mitigated.

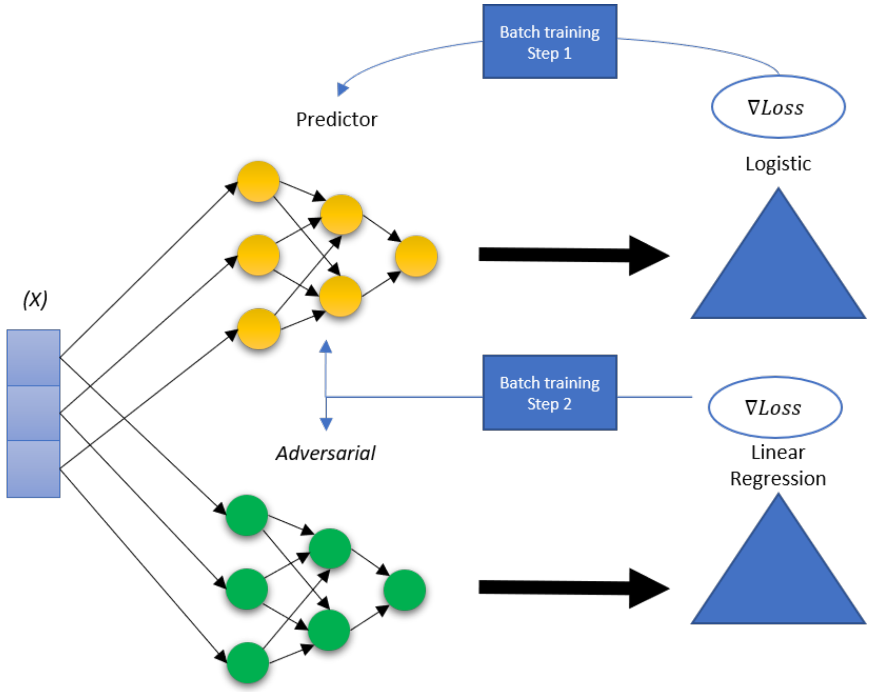


Fig. 6: Two Step Training Process for Predictor and Adversarial Models

The neural nets created for this experiment were intentionally left shallow for simplicity. The predictor has two hidden layers with 25 and 50 nodes respectively using Relu for the activation function. The output layer uses a sigmoid activation for a logistic regression to do binary classification. A binary_crossentropy loss function is selected to train the predictor and calculate the gradients to update the weights. The adversarial has one hidden layer with 25 nodes using Relu as its activation function. The output layer uses a liner activation as it is trying to predicate age. Both models are wrapped together with a singular loss function 3 that is a modified mean squared error designed to maximize the loss for the adversaries. The α is a tunable parameter that dampens or magnifies the adversarial in relation to the predictor to balance the training so, one does not over power the other. The negative turns the equation into a maximization function instead of a minimization and the rest of equation 3 is a *MSE* function.

$$\alpha * -(\frac{1}{n}\Sigma_{i=1}^{n}\left(\frac{d_i - f_i}{\sigma_i}\right)^2) \tag{3}$$

6 Results

As described above, we designed the architecture of our models to be able to measure the change in demographic parity from the baseline model to the adversarial model (after implementing a GAN-like process). From these two resulting models, we calculate standard metrics to evaluate performance, as well as several specifically related to group fairness.

First, we evaluate accuracy and would like our models to show relatively equally accuracy across all of the age groups. Initially, we anticipated lower accuracy in higher age groups given the distribution of our data. However, the results of our baseline model show the opposite of our expected results. Age groups less than 35 have a lower accuracy than age groups between 35 and 45. We then see a decrease in accuracy in our older population (>50). We attribute this trend to the larger population in the middle age ranges and the lower attrition in the older age group. However, accuracy across all the age groups was above 75% (ranging from 75% to 95%), and the overall model accuracy was nearly 86%. We feel this is a good baseline from which to work.

For comparison, we looked at accuracy by age group after running the adversarial model. It is known that limiting models with fairness constraints – such as the variables the adversary model can access – will decrease accuracy. As expected, we did see decreases in accuracy in most age categories. While the overall model did not see a significant decrease in accuracy – down to 83% - the range of accuracy across age groups shifted lower to between 70% and 91%. Table 1 summarizes the changes in accuracy.

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

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

In addition to accuracy, we looked at several group fairness metrics that require parity of some statistical measures across groups. It is important to note that most fairness metrics are considered between the "protected" and "unprotected" groups. Since our experiment does not have a "protected" group per se, we considered the values of fairness metrics in comparison across each age group. Ultimately, our goal is to have demographic parity in our final adversarial model, which is achieved when each group has equal likelihood to be assigned a positive outcome. Stated another way, demographic parity means the positive predictions in subgroups are close to each other (both True Positive and False Positive predictions of the model).

To assess demographic parity equitably between groups, we calculated it as a proportion of the size of each age group in the test set (TP + FP / total test set). Both are achieved if the proportion of positive predictions in the subgroups are close to each other. Our results related to these metrics for the baseline model range between 94% and 100%, which would indicate our baseline has good parity across the age groups. For the adversarial model, the range demographic parity is 98% to 100%, an improvement over the baseline model. See Table 2. Given these results, we believe that our objective to reduce age bias in our data set to classify attrition of employees was achieved.

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

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

We acknowledge the limitations of demographic parity since it only considers positive predictions and does not take into account all aspects the results found in a confusion matrix. It can make an otherwise accurate classifier unfair due to limiting the assessment to positive predictions, and also may treat similar individuals differently merely because they belong to different groups [16].

We did not design our adversarial model with equality of odds or equality of opportunity in mind, but we can still calculate those metrics from our baseline and adversarial model confusion matrices for each group to see what impact the adversarial architecture had with regard to other aspects of fairness. A full summary of the confusion matrices for each age group can be found in the Appendix. Equality of odds looks at both the true positive rate (TPR) and the false positive rate (FPR) for each age group in the test set as comparison [18]. Equality of odds is satisfied when the true positive rates and false positive rates are equal across groups (TPR = FPR). While the baseline model may not meet the equality of odds criteria since the false positive rate for the 40-45 age group is significantly lower, the adversarial model shows strong results across each age group. Our adversarial model comes close to satisfying equality of odds. See table 3.

EQUALITY OF ODDS		**<= 25**	**(25, 30]**	**(30, 35]**	**(35, 40]**	**(40, 45]**	**(45, 50]**	**>50**	**OVERALL**
Baseline	***TPR***	1.0000	1.0000	0.9726	0.9811	1.0000	0.9524	1.0000	0.9772
	FPR	0.8	0.8	0.8125	0.8	0.5	1.0000	1.0000	0.7377
Adversarial	***TPR***	1.0000	1.0000	1.0000	0.9811	1.0000	1.0000	1.0000	0.9967
	FPR	1.0000	1.0000	0.9375	1.0000	1.0000	1.0000	1.0000	1.0000

Table 3: Equality of Odds of Baseline and Adversarial Models by Age Group

A less stringent measure of equality of odds is equality of opportunity, which only considers the false negative rate [6]. Mathematically, a classifier with equal false negative rates will have equal true positive rates, and achieving equalized odds implies achieving equal opportunity. Both our models do well to satisfy this metric, and the adversarial model shows stronger equality of opportunity than our baseline.

7 Ethics

Ethics is a complex topic in any field, and scenarios related to bias, fairness, and ethics in both machine learning and the workplace should not be taken lightly. Our experiment to use adversarial networks to mitigate possible age bias led us through a critical thinking process about sources of data bias, aspects of fairness related to people, systems, and outcomes, and the decisions to be made as a result of the process. Many resources exist to dig into these questions in detail; our goal is to summarize our thinking process and pose critical, high-level questions to spur more critical consideration of bias and fairness.

First, it is important to understand the sources of bias in data. Data used for training machine learning processes may encode biased human decisions, encapsulate systemic issues related to different groups of people or society, or reflect outdated perspectives. Missing data or sample/selection bias may result in systems that do not represent the target population. While data analysis often will specifically exclude "sensitive" or protected variables such as age, race, or gender, this information can be encoded as "proxy" attributes which may influence the machine learning process unintentionally. In our data set, we found that "total working years," "years at company," "years in current role," and "monthly income" could be a proxy for age, particularly when considered in tandem. Finally, bias can result from algorithmic objectives that unfairly favor or disadvantage certain groups. [16] Cathy O'Neil notes "big data processes codify the past. They do not invest in the future. Doing that requires mortal imagination, and that's something that only humans can provide. We have to explicitly embed better values into our algorithms, creating Big Data models that follow our ethical lead" [15]. With this call to action, we directed our attention to bias in models and to identifying an appropriate definition of fairness we were trying to achieve.

As a starting point, it is important to understand fairness falls into two general categories: disparate treatment and disparate impact. Disparate treatment is intentional mistreatment based on a particular attribute. Considered to be direct discrimination, there are legal systems in place to prevent this kind of treatment, and US Civil Rights Act of 1964 clearly articulates attributes or classes considered protected, such as race, color, sex, age, marital status, religion, disability, national origin, among others. Leaving these protected attributes out of a modeling process will likely prevent disparate treatment or discrimination; however, unintentional disparate impact or unintentional discrimination may negatively affect a protected class, even if seemingly neutral policies are in place. We acknowledge that age is a sensitive attribute, and it is also known to influence employee attrition. Therefore, we were mindful of indirect discrimination in our process.

Next, we considered what groups the machine learning process will impact or is intended to serve. This helps determine the type of fairness trying to be achieved, as well as illuminate fairness metrics to be considered. As noted in the introduction, a process may seek to achieve fairness among groups or subgroups, or individuals, or some combination of both. The Center for Data Science and Public Policy at The University of Chicago has developed Aequitas, an open source bias audit toolkit, to assist in determining the type of fairness trying to be achieved. For us, the "Fairness Tree" [4] was helpful in determining the type of group fairness – demographic parity – that was appropriate to our problem. Overall, the decision tree encourages practitioners to consider the following questions related to fairness: should fairness be based on representation or on the errors of a system; are people to be selected by specific numbers or by proportion of population; is the machine learning process meant to intervene in a punitive or assistive capacity; is the intervention intended for most people related to the process or a small fraction; and is the impact intended for everyone without regard to actual need or outcome, or some other subset related to intervention or need? The tree then defines the fairness metrics to consider in designing a machine learning process.

There are many articles that define fairness criteria mathematically and illustrate how to design machine learning process that attempt to achieve these differing aspects of fairness. While understanding these details is critical, it is beyond the scope to summarize our critical thinking about bias and fairness in designing the adversarial machine learning process we used. There are many open source tools and toolkits (IBM's AI Fairness 360 Open Source Toolkit, the Aequitas toolkit), software packages (fairness R package), and books (The Ethical Algorithm, Fairness and Machine Learning) to serve as references. These were extremely useful to us in designing our technical machine learning architecture, but also prompted good discussions we considered how to create an unbiased and fair system.

Looking ahead, there is great opportunity to embrace fairness and unbiased processes within machine learning. [14] notes three aspects the machine learning community can continue to research and explore:

- Synthesizing a definition of fairness. Our own experience revealed many different definitions of fairness, often with slight nuances of difference between seemingly similar concepts. While we were able to achieve the definition of fairness we set out to achieve, it is unclear how this would fare in a difference use case or related to a different definition of fairness. [14] notes this as an open research problem.
- Searching for Unfairness. [14] states that "given a definition of fairness, it should be possible to identify instances of this unfairness in a particular dataset." For us, this concept is not necessarily relevant to our data set, but presents an interesting thought process. Is it possible to extract elements of unfairness in data or in a system, rather than approaching the design of a system a fairness objective? Are there processes that could remove unfairness akin to how a surgeon removes a tumor?

– From Equality to Equity. Most of the literature we reviewed to learn about bias and fairness focused on equality, which can be defined as "ensuring that each individual or group is given the same amount of resources, attention or outcome" [14]. Considering how to move from equality to equity – where individuals or groups have the resources they need to succeed – is a timely and relevant topic, both for machine learning and for society as a whole. We are hopeful that society will evolve to be more equitable, and with foresight, the machine learning community can help lead the design of systems that support and encourage this.

8 Conclusions

Our team set out to determine whether or not it was possible to use the adversarial network structure proposed in [20] to mitigate bias in a data set related to a continuous variable. Many examples of this technique used binary predictions and binary protected classes or attributes, so we wanted to explore if the technique would be successful in other applications. Overall, we were successful in mitigating some bias related to age as an influencing predictor of whether or not an employee would leave (attrition). We achieved an acceptable level of demographic parity among the age groups we defined, and did not trade-off too much prediction accuracy for improved fairness.

To expand on the work in this experiment, we see several possible avenues to pursue. While we were pleased with the results, there would be added value to achieving a more balanced distribution across age groups. This could entail pre-processing work to over- and under-sample from the employee population. Another avenue to consider would be additional tuning of the prediction models for more balanced outcomes. Additionally, technical aspects of the code could be refined to create an "early stop" in the models when the adversary has sufficiently removed bias and correlation is no longer detected in the adversarial model for $Z(X)$. Not only might this improve results, it may increase the architecture's capacity to handle larger data sets. Lastly, we could redesign the adversarial architecture to specifically achieve a different type of fairness, such as equality of opportunity or equality of odds. This would dictate different variables to be shared with the adversarial network for prediction purposes.

Ultimately, the opportunity to mitigate bias in continuous variables using adversarial network architecture shows promise. However, these opportunities cannot allow practitioners to become complacent and confident that systems are unbiased and fair. The myriad of tool kits, packages, process interventions, new techniques, or improved data collection cannot – and must not – replace the sensitivity to foresee decisions consequence, inquisitiveness, skepticism, mortal imagination, and compassion that humans bring to bear to on machine learning.

References

1. Bias, https://www.merriam-webster.com/dictionary/bias

2. Ibm hr analytics employee attrition performance
3. Webcast slides: Removing unfair bias in machine learning webcast recording slides, https://community.ibm.com/community/user/datascience/viewdocument/removing-unfair-bias-in-machine-lea
4. (Feb 2020), http://www.datasciencepublicpolicy.org/projects/aequitas/
5. Frye, A., Boomhower, C., Smith, M., Vitovsky, L., Fabricant, S.: Employee attrition: What makes an employee quit? SMU Data Science Review **1**(1), 9 (2018)
6. Garg, P., Villasenor, J., Foggo, V.: Fairness metrics: A comparative analysis (2020)
7. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: Advances in neural information processing systems. pp. 2672–2680 (2014)
8. Gu, S., Pednekar, M., Slater, R.: Improve image classification using data augmentation and neural networks. SMU Data Science Review **2**(2), 1 (2019)
9. Hom, P.W., Lee, T.W., Shaw, J.D., Hausknecht, J.P.: One hundred years of employee turnover theory and research. Journal of Applied Psychology **102**(3), 530 (2017)
10. Kearns, M., Roth, A.: The Ethical Algorithm: The Science of Socially Aware Algorithm Design. Oxford University Press (2019)
11. Lipnic, V.: The state of age discrimination and older workers in the us / 50 years after the age discrimination in employment act (adea), https://www.eeoc.gov/eeoc/history/adea50th/report.cfm
12. Lum, K., Johndrow, J.: A statistical framework for fair predictive algorithms. arXiv preprint arXiv:1610.08077 (2016)
13. McKenna, M.: Three notable examples of ai bias (Oct 2019), https://aibusiness.com/three-notable-examples-of-ai-bias/
14. Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., Galstyan, A.: A survey on bias and fairness in machine learning (2019)
15. O'neil, C.: Weapons of math destruction: How big data increases inequality and threatens democracy. Broadway Books (2016)
16. Pessach, D., Shmueli, E.: Algorithmic fairness (2020)
17. Rubenstein, A.L., Eberly, M.B., Lee, T.W., Mitchell, T.R.: Surveying the forest: A meta-analysis, moderator investigation, and future-oriented discussion of the antecedents of voluntary employee turnover. Personnel Psychology **71**(1), 23–65 (2018)
18. Verma, S., Rubin, J.: Fairness definitions explained. In: 2018 IEEE/ACM International Workshop on Software Fairness (FairWare). pp. 1–7 (2018)
19. Xu, J.: Algorithmic solutions to algorithmic bias: A technical guide (Jul 2019), https://towardsdatascience.com/algorithmic-solutions-to-algorithmic-bias-aef59eaf6565
20. Zhang, B.H., Lemoine, B., Mitchell, M.: Mitigating unwanted biases with adversarial learning. In: Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society. pp. 335–340 (2018)

Appendix A:

	<= 25		**(25, 30]**		**(30, 35]**	
	PreGan	***PostGan***	***PreGan***	***PostGan***	***PreGan***	***PostGan***
Accuracy	0.757576	0.69697	0.84	0.8	0.831461	0.831461
Demographic / Proportional Parity	0.939394	1	0.96	1	0.94382	0.988764
Positive Predicted Value	0.741935	0.69697	0.833333	0.8	0.845238	0.829545
False Discovery Rate	0.258065	0.30303	0.166667	0.2	0.154762	0.170455
Negative Predicted Value	1	n/a	1	n/a	0.6	1
False Omission Rate	0	n/a	0	n/a	0.4	0
True Positive Rate	1	1	1	1	0.972603	1
False Negative Rate	0	0	0	0	0.027397	0
True Negative Rate	0.2	0	0.2	0	0.1875	0.0625
False Positive Rate	0.8	1	0.8	1	0.8125	0.9375
Equalized Odds	1	1	1	1	0.972603	1
Predictive Parity	0.741935	0.69697	0.833333	0.8	0.845238	0.829545
Specificity	0.2	0	0.2	0	0.1875	0.0625

	(35, 40]		**(40, 45]**		**(45, 50]**	
	PreGan	***PostGan***	***PreGan***	***PostGan***	***PreGan***	***PostGan***
Accuracy	0.913793	0.896552	0.946429	0.892857	0.869565	0.913043
Demographic / Proportional Parity	0.965517	0.982759	0.946429	1	0.956522	1
Positive Predicted Value	0.928571	0.912281	0.943396	0.892857	0.909091	0.913043
False Discovery Rate	0.071429	0.087719	0.056604	0.107143	0.090909	0.086957
Negative Predicted Value	0.5	0	1	n/a	0	n/a
False Omission Rate	0.5	1	0	n/a	1	n/a
True Positive Rate	0.981132	0.981132	1	1	0.952381	1
False Negative Rate	0.018868	0.018868	0	0	0.047619	0
True Negative Rate	0.2	0	0.5	0	0	0
False Positive Rate	0.8	1	0.5	1	1	1
Equalized Odds	0.981132	0.981132	1	1	0.952381	1
Predictive Parity	0.928571	0.912281	0.943396	0.892857	0.909091	0.913043
Specificity	0.2	0	0.5	0	0	0

	>50		**OVERALL**	
	PreGan	***PostGan***	***PreGan***	***PostGan***
Accuracy	0.794118	0.794118	0.858696	0.831522
Demographic / Proportional Parity	1	1	0.9375	0.997283
Positive Predicted Value	0.794118	0.794118	0.869565	0.833787
False Discovery Rate	0.205882	0.205882	0.130435	0.166213
Negative Predicted Value	n/a	n/a	0.695652	0
False Omission Rate	n/a	n/a	0.304348	1
True Positive Rate	1	1	0.977199	0.996743
False Negative Rate	0	0	0.022801	0.003257
True Negative Rate	0	0	0.262295	0
False Positive Rate	1	1	0.737705	1
Equalized Odds	1	1	0.977199	0.996743
Predictive Parity	0.794118	0.794118	0.869565	0.833787
Specificity	0	0	0.262295	0

Table 4: Complete Fairness Metrics for Baseline and Adversarial Models