Final Project

Collin Register

2024-04-23

Introduction

With the increasing prevalence of artificial intelligence in today's decision-making processes, concerns about algorithmic bias have also been on the rise. One common application of artificial intelligence is using algorithms to determine if prospective job candidates are a good fit for employment. Within this classification, algorithms often misclassify groups at different rates, maintaining a level of bias in decisions, which often leads to discrimination. This discrimination can be harmful when it is arbitrary because of the unfair classifications based on a particular feature and the unequal treatment based on that classification. In employment specifically, arbitrary discrimination can be extremely damaging to those discriminated against because it economically impacts their lives. Also, the bias within the classification is often hard to detect and can make it difficult to alter the decision-making process fairly and in a way that doesn't discriminate against certain classes.

The prevalence of bias within an algorithm is a complex issue that becomes increasingly complicated when the tradeoff between fairness and accuracy is introduced. Often, to maintain a high level of fairness, a certain degree of accuracy must be compromised to ensure that fairness. There is often a struggle to find the right balance of fairness and accuracy, making it a difficult task. One study, using an artificial intelligence software toolkit, highlights this tradeoff in terms of accuracy and fairness. In a paper detailing this study, "Bias mitigation with AIF360: A comparative study," the methods and results of how this is explored are explained in detail. The goal of this study is to determine the performance of these bias mitigation algorithms and analyze the fairness and accuracy of each of them in predicting employment status, excluding gender and race as variables that seem to cause the most arbitrary discrimination in employment. With different variations of race and gender excluded, we can see how the model arbitrarily discriminates against people of a certain status.

Dataset

The dataset used for this analysis is the 2013 U.S. census data, with original records of over 130,000 observations. Upon initial assessment and using the BinaryLabelDatasetMetric function in Python, it was found that men were associated with 9.5% more favorable labels (employed) than women, and whites were associated with 9.7% more than non-whites. Given this, the privileged groups were set to be whites and males, and the unprivileged groups to be women and non-whites. After finding this initial bias, the researchers first did pre-processing on the data, where all null values were removed for better model performance, and they were left with 56827 observations. Then, the data was sorted so that age was binned by decades, and anyone under the age of 14 was removed due to not being of working age. The race attribute was grouped into non-white and white.

Introduction to ROC and PR

After detecting the initial bias, the researchers aim to combat this bias using bias mitigation algorithms. They use Reject Option Based Classification (ROC) and Prejudice Remover (PR), two of the bias mitigation algorithms within the package AIF360. The original algorithm used to classify is logistic regression, which is an algorithm used to predict the outcome by fitting the data points to a logistic curve and then using a particular cutoff to determine which side of the decision boundary the data point lies on. In this case, using

the package sklearn, the data is standardized to a normal distribution, and the logistic regression is run with the sklearn default parameters. In-processing techniques such as Prejudice Remover alter the learning model, while post-processing techniques such as Reject Option Based Classification are applied to the predicted labels after the model has been run.

ROC operates at the decision boundary, assigning favorable outcomes to unprivileged groups where outcomes are ambiguous. A central aspect of this technique is within the critical region of a logistic regression. This is the area within the decision boundary where the output of the label is ambiguous, and the number of instances that fall within the decision boundary is adjusted by theta. A higher theta means more instances fall within the decision boundary, and the effect is less discrimination. A little more of a formulaic definition follows: Within the decision boundary, when instance X is a member of the unprivileged group Xd, classify it with the favorable label, C+. Otherwise, classify it as the undesirable label, C-. When outside of the decision boundary, classify X as a desirable label, C+, when the probability of the favorable label C+ occurring, given X, is higher than the probability of the undesirable outcome label C- occurring, given X. Otherwise, assign X an undesirable label, C-. This helps us to ensure fairness because when unprivileged group instances fall within the boundary, they are given the favorable label, but when outside of the boundary, the label depends on the greater probability.

Prejudice refers to a statistically dependent relationship between the sensitive attribute and other attributes. This can be split up into three subgroups of prejudice: direct, indirect, and latent. Direct Prejudice refers to a sensitive attribute being used in classification. In our case, this would be race or gender. Indirect prejudice refers to the case when there is a dependence between the target label and the sensitive label. Latent prejudice occurs when there is a statistical dependence in the relationship between the sensitive attribute and the non-sensitive attribute, given that y is already influenced. Prejudice removal adjusts the model's predictions by adding regularizers to ensure higher degrees of fairness in the model. The prejudice remover minimizes the regression and avoids overfitting the model as well.

Statistical Measures of Fairness

Once the algorithms are applied, they are evaluated and compared to the original logistic regression classifications without the fairness constraints, using statistical measures of fairness such as disparate impact, statistical parity difference, average odds difference, and equal opportunity difference. Disparate impact computes the measure of the intentional bias in the label assignment. It is found by dividing the proportion of favorable outcomes for privileged groups by the proportion of favorable outcomes by the unprivileged group. A value of 1 means that both groups are equally likely to have a favorable outcome, while a value above 1 would mean the privileged group is more likely to receive a favorable outcome. Another metric, statistical parity, is the difference in probabilities that the protected and unprotected groups have when being assigned the favorable label. This is found by subtracting the proportion of favorable outcomes for privileged groups from the proportion of favorable outcomes by the unprivileged group and then taking the absolute value. If the disparity is near zero, there is little to no difference in outcomes between the two groups. If the disparity has a larger, nonzero value, discrimination favoring one group over the other may be present. The benefit of this measure is that there is no need to ensure a protected vs. unprotected class because of the absolute difference between the two.

Also, the average odds difference is a combination of the difference between the true-positive rates of both groups and the false-positive rate of both groups in classification. This is calculated by finding the false positive rate of the privileged group minus the false positive rate of the unprivileged group plus the difference in the true positive rate of the privileged group minus the true positive rate of the unprivileged group and then dividing by 2. This measure provides a summary of the correct classification rates and misclassification rates. The equal opportunity difference is the difference in probability of the two groups being wrongly assigned the unfavorable label. This is found by calculating the absolute difference in the true positive rate of the privileged class and the unprivileged class. This is an important metric because it shows that members of both groups received the same favorable outcomes. In this study, all the fairness metrics looked at group fairness as opposed to individual fairness due to the lack of specificity within the dataset and to avoid having to make assumptions regarding individuals.

Analysis

Many of the methods have similar approaches and are a repetition of actions performed in previous steps. With this, much of the code use and methods do seem somewhat cyclic. All the coding is done in Python by the researchers and is available in the following repository: https://github.com/throwaway02062020/INFO381/tree/master/notebooks To begin, the researchers started with the ROC bias mitigation algorithm. For this, they shuffled and split the processed data into 70% train, 15% test, and 15% validate. This splitting ensures that the model can learn using the training data and find the optimal threshold in the validate set to be applied to the remaining test portion. Next, they set up the privileged and unprivileged classes. Using the package sklearn, the data is standardized to a normal distribution, and the logistic regression is run on the train data using the default parameters from the package. Then predictions are made based on this logistic regression for the train data and labeled with the outcome. Next, the same procedure is performed for the validation set, and the optimal threshold is found by looping through each threshold value set and finding the one with the highest balanced accuracy. Then, including the fairness constraints, the optimal classification threshold is estimated and returns a value of 0.7227 accuracy and a threshold of 0.0622. Following this, the function "ClassificationMetric" is used to calculate the accuracy, statistical parity, disparate impact, average odds difference, and equal opportunity difference are calculated for the validation set. Then, the learned ROC is applied to the validation set with fairness constraints to ensure it has not become worse. Next, the metrics are computed for the test set without fairness constraints and without ROC applied. Afterwards, they apply the ROC technique, include fairness constraints on the test set, and compute the fairness metrics.

Now the researchers use the PR bias mitigation algorithm to ensure fairness. Prejudice Remover is an in-processing algorithm that adds regularization terms to the model. They begin by setting the privileged and unprivileged groups and splitting the data into 70% training and 30% testing. The statistical fairness metrics are computed for both the original train and test data. Next, the model without a prejudice penalty is fit on the train data, and then the fit model is applied to the test data. The statistical fairness metrics are calculated using the "ClassificationMetric" function. Then the prejudice-removal model is fit to the train data and then applied to the test data once it is fit. The metrics are calculated for both. Lasty, the prejudice removal algorithm is run for all values of n between 0 and 30.

Next, the researchers look at PR and ROC together, using race as the privileged group, and then split the data into 70% train, 15% validate, and 15% train. After that, they scale the dataset and verify that it doesn't affect the group label statistics. Then, they train the model with the prejudice remover set at 10 and apply the model to the test set. Next, they standardize and fit the logistic regression and get the predictions, labels, and test statistics. The same process is repeated, but for the validation set. A loop is then run to find the best-balanced accuracy for each threshold and find the best threshold without the fairness constraints. Now those values are used to fit ROC and transform the test set with ROC. Lastly, they calculate the true positive rate and false positive rate and run the algorithm for different n values from 0 to 30 and calculate accuracy, statistical parity, average odds difference, and equal opportunity for each.

Results

Following the application of each of the bias mitigation techniques, we are left with the following results: For ROC with gender as the protected attribute, accuracy decreased by about 2 percent compared to the classification without the fairness constraint, while significant improvements were made regarding the measures of statistical parity difference, disparate impact, average odds, and equal opportunity. When looking at ROC with race as a protected attribute, there was a small decrease in accuracy (roughly half a percent), while significant improvements were made regarding the measures of statistical parity difference, disparate impact, average odds, and equal opportunity. When comparing the application of PR with gender as the protected attribute to the original, accuracy decreased by roughly four percent and average odds decreased. However, significant improvements were made regarding the measures of statistical parity difference, disparate impact, and equal opportunity. However, when comparing the application of PR with race as the protected attribute to the original, we see an improvement in all the statistical measures of fairness and a slight decrease in accuracy.

An interesting result is that ROC overshoots for both gender and race for the average odds difference. For

example, ROC with fairness constraints sways from 13.88% in favor of men to 4.01% in favor of women. All of the other fairness metrics seemed to increase drastically towards removing bias. In this instance, we can see that for most of the bias mitigation techniques, accuracy decreased by a few percent, but the statistical measures of fairness improved greatly.

From the results of the bias mitigation, we can see that ROC outperformed PR for accuracy under all conditions. With that said, the different algorithms had varying successes for each of the statistical measures of fairness for each of the protected features. Neither of the algorithms clearly outperformed the other, and we saw each of them vary between the two.

Moral Considerations

As can be seen by the employment dataset, the inclusion of sensitive features such as race and gender can invoke bias within classification, in turn leading to arbitrary discrimination. If the features are included, the predicted outcomes may and often times will disproportionally predict the privileged group to have more favorable outcomes. This case is known as arbitrary discrimination, where the distinguishing characteristic is irrelevant or prejudicial. In the algorithms, race and gender have a large impact on the classification and should not be included due to their prejudicial nature.

There are many harms associated with arbitrary discrimination, such as self-fulfilling prophecy, a lack of equal opportunity, and overgeneralizations. In the employment vs. non-employment classification, the self-fulfilling prophecy would be that those groups who were disproportionately predicted to be unemployed could potentially fulfill that calculation. The predictions could become true because of arbitrary discrimination when employers are looking to hire candidates. Similarly, the prevalence of arbitrary discrimination could lead to overgeneralizations where the negative attributes become applied to the group as a whole. While this algorithm is predicting if someone is employed or not, it still could impact hiring practices and influence status within the economy.

Another consideration is whether to use the algorithm if arbitrary discrimination is prevalent. There are various approaches to answering this question, and they consider various definitions of justice. John Rawls argues that if there is potential for arbitrary discrimination, a different method should be used. Rawls poses the example of the veil of ignorance and states that we should choose fairness based on being "blind" to characteristics such as race or gender. In this case, defining characteristics are ignored, and everyone is given opportunity based on merit or their actions. Rawls also claims that when differences exist, we should allocate resources to protect the most vulnerable. Opposing this view, Robert Noziak claims that as long as the training and testing partitions are representative of one another and legitimate methods are used, then disparities in outcomes are not necessarily problematic.

Conclusion

Generally, the inclusion of fairness metrics through the PR and ROC algorithms in a logistic classification decreases arbitrary discrimination in this case. In most instances, the misclassification decreases drastically, with only a small decrease in overall accuracy for the binary classifications. Overall, the bias mitigation is effective and helps to eliminate the arbitrary discrimination associated with the inclusion of race and gender. The bias mitigation techniques would align closely with John Rawls perspective on justice, aiming to reduce disparities and give an equal and fair outcome to all groups.

In this study, the bias mitigation worked well with regard to the privileged groups of race and sex. Another takeaway is that PR is computationally more expensive because you have to pick the N parameter, but it allows for flexibility because it is in-process. This flexibility means that the bias can be changed sooner within the model and provides more opportunity to diminish discrimination. Lastly, the AI360 bias mitigation technology is fairly new, and I will be interested to see how the continued application will promote fairness in new classifications.

Sources

Hufthammer, Knut T., et al. “Bias Mitigation with AIF360: A Comparative Study.” Bergen Open Research Archive, Norsk IKT-konferanse for forskning og utdanning, 15 Dec. 2020, bora.uib.no/bora-xmlui/handle/11250/2764230.