Gautam Samudrala and Jacob Friedland
Professor Krebhiel
CSCI 165
June 2023

A Dive into Algorithmic Fairness: Predicting Race-based and Blind Classification Methods on Income

Algorithmic fairness is a critical topic in the field of machine learning, aiming to address biases and ensure equitable outcomes when deploying algorithms in decision-making processes. It involves developing and implementing classification models that are not influenced by sensitive attributes such as race, gender, or ethnicity or in the case of their influence create predictions and classifications in the most equitable manner. The use of machine learning algorithms for income prediction, such as determining whether an individual makes more or less than a certain threshold, has gained significant attention due to its potential impacts on various applications, including loan approvals and financial decision-making.

The significance of income prediction lies in its ability to provide insights into an individual's financial capacity and stability, enabling lenders to assess the risk associated with granting loans. By leveraging algorithms to predict income levels, financial institutions can make informed decisions based on quantitative data, reducing the likelihood of biased or arbitrary judgments. Algorithmic fairness emphasizes the importance of mitigating biases and ensuring equal treatment for all individuals. When used in loan approval processes, fair algorithms promote social equality by preventing discrimination against individuals based on protected attributes such as race, gender, or age. By prioritizing fairness, these algorithms can help address historical biases and create a more inclusive financial system. Equitable machine learning libraries that implement algorithmic fairness have the potential to foster equitable opportunities and dismantle systemic inequalities. The scope of our research focuses on income prediction, but the application of our research can be projected towards a variety of different settings that utilize predictive models for different purposes.

Lending has become a leading opportunity space for AI technologies. While access to affordable credit is crucial for financial freedom and equal opportunities, lending has been marred by structural and cultural racism throughout history. In 2020, the Central Financial Protection Bureau (CFPD) estimated that twenty percent of the US adult population is underserved for credit and people from this population are more likely to be from minority groups (Weber, et al.). With one of the core mandates of the CFPD being to find the balance between lending being helpful versus harmful, machine learning can be used by such organizations to implement algorithmic fairness in the scope of loan approvals. While one could oppose that machine learning libraries are also biased, face-to-face lending inequity also exists. A 2019 study of 3.2 million mortgage applications and 10.0 million refinance applications from Fannie Mae and Freddie Mac found evidence of racial discrimination in face-to-face lending as well as in algorithmic lending (Weber, et al.). Black and Latino applicants receive higher rejection rates of 61% compared to 48% for other races. The same marginalized groups additionally suffer from race premiums on interest rates, paying up to 7.9 extra basis points on mortgage rates (Weber, et al.). This creates a 756-million-dollar annual race premium in the country. Clearly, statistics signify the importance of spending more time dedicated to finding algorithmic solutions that optimally increase fairness. The same study revealed that despite algorithmic solutions being biased, the acceptance rate parity of loan approvals improved, but race premiums were still imposed by 5.3 basis points for purchase mortgages and 2.0 basis points for refinance mortgages (Weber, et al.). These are still sizable differences that do not help bring minority groups closer to the treatment that other groups face regarding lending. However, hope exists as the bias in algorithmic lending is a lot easier to fix than removing the bias from a human individual.

Our research project aims to investigate the impact of race-blind classifiers versus non-race-blind classifiers on datasets related to income prediction. By leveraging Python, pandas, and sklearn machine learning libraries, we explore the performance of different classifiers and analyze their effects on the resulting outcomes. Specifically, the primary objective determines whether race-blind classifiers outperform non-race-blind classifiers when predicting an individual's income level. We will be using Random Forest algorithms as our classifier for this project. Our project involves feeding a US Adult income dataset from the 1994 US Census Database into different machine learning libraries, selectively omitting race-sensitive information in some instances. After preprocessing the data, the libraries predict whether an individual from the dataset earns greater or less than fifty thousand dollars a year. Consisting of anonymous information including, occupation, age, native country, race, capital gain, capital loss, education, and work class, the set is paired with a test dataset that displays the correct incomes for every individual. By comparing the accuracies and confusion matrices obtained from each combination of dataset and library, we provide valuable insights into the effectiveness of race-blind classification methods in the context of income prediction.

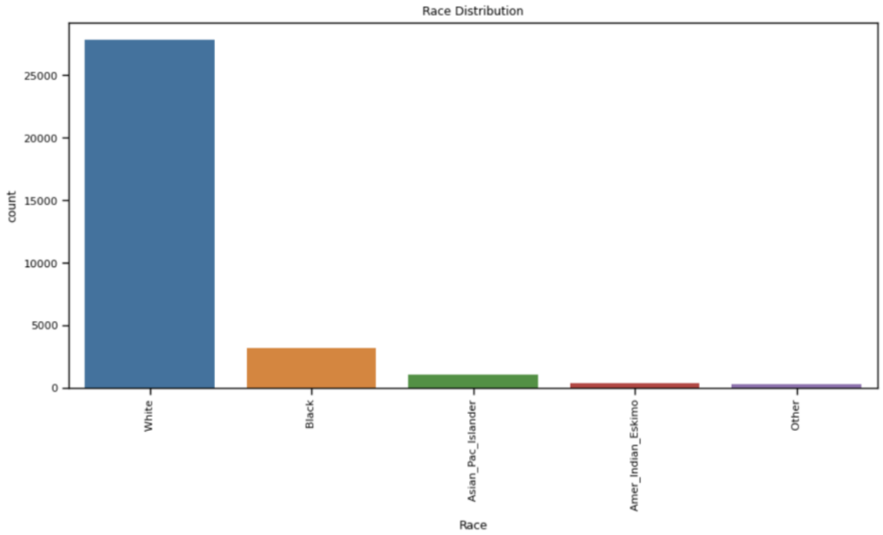
Before discussing our project, we will define and understand the premise of algorithmic fairness within the scope of our study. We will explain certain criteria that are implemented for fairness purposes and explore the random forest classifier used for our testing. Algorithmic fairness is notably composed of three components that help employ fairness tactics in prediction related settings. Separation is a strategy where developers are excluded from fairness processes to ensure that they won't intentionally or unintentionally implement bias in the system. Secondly, blindness avoids sensitive attributes when making decisions. The third method involves group fairness or statistical parity. In the context of our study, accuracy parity would involve individuals from all racial groups to have the same prediction accuracy regarding income. In terms of fairness and loan approvals, the goal for an algorithm would be to reduce type one and type 2 error as both extreme ends of the spectrum would result in certain individuals getting favored or disfavored despite meeting loan approval thresholds.

Using Random Forest classifiers is a robust and accurate method for income prediction due to its ability to handle complex relationships and diverse data types. By aggregating predictions from multiple trees, it reduces the risk of overfitting and provides more reliable results while being robust to outliers and missing data. With its versatility and scalability, we decided that RFC would be well-suited for analyzing large datasets with numerous features, making it a strong choice for the UCI dataset.

General explanation of RFC:

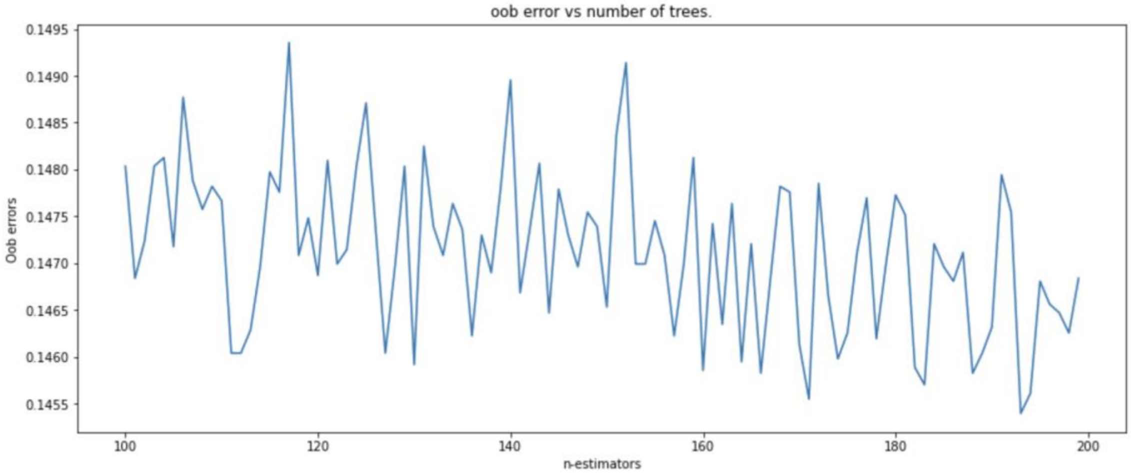
1. **Random sampling**: Given a training dataset with N samples, the algorithm randomly selects a subset of these samples (with replacement) to build each tree in the forest
2. **Random feature selection**: Each node of a decision tree, a random subset of features is considered for splitting the data
   - Introduce diversity among the trees
   - Reduces correlation between them
3. **Growing decision trees**: For each tree in the forest, the algorithm recursively grows a decision tree using the selected samples and features. At each node, the algorithm identifies the best feature and threshold to split the data
4. **Voting or averaging**: Once all trees are constructed, the Random Forest combines their predictions by majority voting in the case of classification
   - Highest # of votes wins

Initial preprocessing of the data showcases several key points to take into consideration. Shown below, there are significantly more white individuals in the dataset when compared to other minority groups.



This is a key reason for why running the classifier multiple times with different combinations of subgroups is an important part of our analysis and conclusions. When the majority of the data consists of one subgroup, the resulting difference between race-blind and race-based classification methods would be similar due to a lack of data from minority groups.

To divide the data into multiple subgroups, we used the pandas library. For preprocessing, to train the race-aware classifier we included all the feature columns and to train the race-blind classifier we removed the race columns. With these different dataframes, we also did one-hot encoding on both datasets. We used one-hot encoding to convert the categorical variable into numerical variables, as those would be easier for the classifier to analyze and take in as inputs. After preprocessing, we needed to find the optimal number of trees used to build the classifiers. We ran a loop through the numbers 100 to 200 and saw which number would return the lowest OOB error.



After finding the optimal number of trees, we trained a classifier with that number and started our testing phase. We took the original datasets and grouped them by race. The new groups were then used to test our classifiers accuracy.

All of our results can be seen in the document summarization doc uploaded with this report. From our results, we can see there is not much of a difference between each of the classifiers, including the subgroups. Overall, there were no accuracy differences between the race-aware and race blind classifier on the overall dataset. The only major differences for accuracy were seen in the Asian and Black subgroups, where the accuracy was higher by 1 or 2 percentage points on the race-aware classifier, but the accuracy was lower for Native American and other subgroups. These results generally show that Random Forest Tree is a fair classifier. Some of the reasons our results might have turned this way might be due to the fact of not enough data samples for all subgroups compared to the majority subgroup white. Another reason is that there could be multiple variables that would implicate race in the classifier, so leaving out race would not have much of an impact. For instance, the origin country column may have led to the implication being made.

While race may not play a large role in determining the income of an individual, there is a large correlation between lower income families and such families coming from minority groups. This brings in the idea of affirmative action, as used in institutions, to promote and support the potential for minority individuals to create a better future for their families. As discussed in class, determining parity between different racial groups while maintaining fairness gives various options for machine learning algorithms to decide how to make predictions. Emphasizing positive predictive value, accuracy, and true positive parity are all choices that a developer can use to tweak resulting algorithms. In addition to these options, this paper will explore another possible mathematical approach to algorithmic fairness and its application towards loan approvals.

Another suggestion towards solving group discrimination in machine learning algorithms is the notion of individual fairness. Algorithms do not have the ability to anticipate the possible subgroups from data. Resultantly, individual fairness works to ensure equal treatment between each individual rather than focusing on subgroups. The mathematical notation is given below:

For a machine learning model with mapping function h that takes inputs from X and outputs to space Y:

$d_Y (h(x_1), h(x_2)) <= Ld_X(x_1, x_2)$ for all $x_1, x_2 \in X$

In words, the distance between h(x_1) and h(x_2) in the output space d_Y is less than or equal to a Lipschitz constant. For inputs that are very similar to each other, this guarantees that the outputs are also very similar to each other handled by the constraint.

Lipschitz constant: a measure of the maximum rate of change of a function in a region over which the function is defined.

The issue with this method is that real-world practicality is very difficult as the condition is very restrictive and challenging to satisfy the condition universally in an algorithmic fairness setting. Defining a fair metric represented by d_X, is nontrivial and would require consideration of each domain and context in which a model is applied.

Due to these issues, Mikhail Yurochkin, a statistician at IBM presents the idea of distributionally robust fairness. The same machine learning model h applies, but an algorithm possesses distributionally robust fairness if:

$$\max_{P:W(P,P_n)\leq\epsilon}\int_{X\times Y}\ell(x,y,h)dP(x,y)\leq\delta$$

The maximum value of a certain expression, which involves a loss function (ℓ), should be smaller than a certain threshold ((δ). The expression considers the data's distribution (P), a small tolerance parameter (ε), and a concept called Wasserstein distance that measures the difference between distributions. Comparing this definition to a previous definition of individual fairness, we see that the loss function replaces the metric used to measure fairness in the output space. The tolerance parameter is similar to the Lipschitz constant, which ensures fairness. The important difference is that DRF considers the average violations of individual fairness, allowing for statistical analysis and establishing guarantees of fairness in broader contexts.

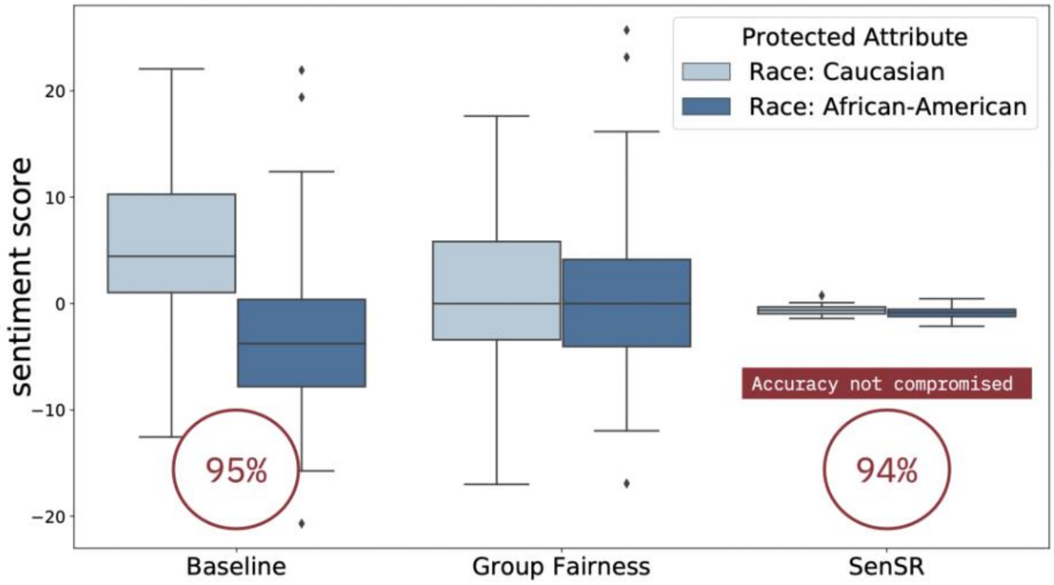
The MIT article presents the steps to create DRF in a clear manner:

1. Observe the data and learn a fair metric.
2. Generate a fictional set of all possible similar individuals to see which ones have different outcomes, as measured by the loss.
3. Take the single worst counterfactual (most unfair example) then update the parameters to account for this.
4. Repeat until the DRF condition is satisfied.

We reword the article's steps in the context of loan approval to present an application towards our topic of study:

1. Start by examining a loan approval dataset including attributes like income, credit history, and demographic information. Identify relevant features and characteristics. Then, define a fair metric that captures the attributes that should determine similarity or fairness between individuals.
2. Create a fictional dataset consisting of hypothetical individuals who are similar to each real-world loan applicant. This involves varying the demographic attributes while keeping the other features constant. By creating this set, you can analyze potential disparities in loan approval outcomes between minority and majority groups.
3. Identify the worst counterfactual. Compare the loan approval outcomes between minority and majority groups within the fictional dataset. Measure the loss (ℓ)(which quantifies the discrepancy between the predicted outcome (h) and the true outcome (y) for each individual (x). Identify the fictional individual whose outcome is most unfair or discriminatory compared to similar individuals from the majority group.
4. Adjust the parameters of the machine learning model to mitigate the unfairness detected in the worst counterfactual. This can be done by incorporating fairness constraints that enforces equal treatment amongst different demographic groups.

This image below shows how DRF is satisfied through an algorithm called SenSR. The model created sentiment for names for Caucasian and African-American individuals. While sentiment from names should be neutral, the group at IBM discovered a traditional NLP (Natural Language Processing) model showcased racial bias. By purely implementing group fairness the means would equal 0 in sentiment score, but that would not remove the massive difference between certain individuals amongst the two groups. Instead, DRF closes this bias significantly while maintaining a level of accuracy.



In the context of income prediction and its implications on loan approval, it is essential to acknowledge the complexity and diversity within our human population. The intricacies of different subgroups and their combinations cannot be simply reduced to black and white distinctions. Algorithms alone cannot erase the historical racial oppression, male privilege, or any other forms of injustice embedded in the historical data. However, they provide a step forward in our increasingly technological world that can promote equality and prevent discrimination through a socio-technical lens.

Reference:

Weber, M., Yoruchkin, M., Botros, S., & Markov, V. (2020, December 14). *Black loans matter: Fighting bias for AI fairness in lending - MIT-IBM watson AI lab*. MIT. https://mitibmwatsonailab.mit.edu/research/blog/black-loans-matter-fighting-bias-for-ai-fairness-in-lending/