AI and Prejudice: Analyzing the Issue of Implicit Bias in Machine Learning Algorithms

RP445

Abstract

Innovations in Artificial Intelligence and Machine Learning have vastly changed the way technological systems operate. However, systems that are capable of accomplishing tasks that would require human intelligence are also fallible to issues such as bias. We investigated the extent of the effects of bias on machine learning systems through an analysis of past scandals involving machine learning systems. In order to observe bias in a machine learning system, we created our own machine learning model that classifies if a person has received a higher education (Bachelor's Degree or higher) or not. We utilized data from the 2013 American Community Survey conducted by the US Census Bureau to train and test our model. With this data, we created three models: one with blatant racial bias, one with implicit racial bias, and one without racial bias. After evaluating our models, we concluded that by removing biased variables, the model's bias decreased overall. However, the removal of these variables, and thus the removal of unfair bias, negatively affected the performance of our model. This trade-off, known as the "fairness-accuracy" trade-off, is one of the greatest ethical challenges present in Machine Learning today. We examine the implications of this dilemma through a review of preexisting literature and an analysis of our own.

Background Literature

Machine learning is the subfield within artificial intelligence that involves instructing computers to complete tasks without explicit programming. Unlike traditional computer programs which follow a sequence of instructions that are programmed, machine learning algorithms are programmed to learn the types of operations to be performed by feeding the algorithm training examples so that it can optimize the operations for the task. A classic example of a machine learning task is spam email filtering. Most email services like Gmail and Yahoo have spam email filters, in which they scan through the emails in a person's inbox to determine if the email should be labeled as spam or not. A task like this could not be easily done without using a machine learning algorithm because it would require casework to account for every possible spam email. Machine learning algorithms, on the other hand, can learn to detect spam email by training it to recognize patterns and features of spam emails.

Machine learning has been revolutionary in terms of enhancing user experience. Previously, many services would have every user go through a similar experience, but the advent of machine learning caused many services to become personalized. For example, Netflix users receive recommendations for movies and television shows after finishing one. This recommendation system is powered by machine learning, since the system learns what types of shows and movies the user enjoys in order to suggest a new show or movie. Another example is Face ID, which enables users to log into their phone simply by pointing their phone cameras at their faces. This facial recognition is also an example of machine learning, where the algorithm must learn to recognize key facial points and be able to differentiate a user's face from those of others.

Overall, Machine Learning has numerous applications in the real world and is currently utilized in many different industries including retail stores, entertainment, transportation, health care, energy, and the government in order to automate processes and enhance the user experience. However, just like human systems, Machine Learning systems are susceptible to issues such as bias and injustice, which can have massive repercussions on certain individuals.

Fairness and Bias

A major concern for the ever-ubiquitous applications of machine learning algorithms is the possibility of bias that discriminates against individuals in an unfair manner, such as considering "protected attributes" (race, gender, color, age, etc.) in making decisions where the use of such data is inappropriate. Notorious examples of this phenomenon occurring in machine learning algorithms include COMPAS, a recidivism algorithm that predicted a defendant's risk of reoffending. In 2016, Larson et al. of ProPublica discovered that the algorithm was racially biased: it disproportionately flagged black individuals as high-risk at a rate more than twice that of white individuals (Angwin). Another notable case involved an advertisement service by Google, which was found to have shown men advertisements for higher-paying jobs at a rate 6 times that of women, thus exacerbating the issue of disproportionate salaries between genders (Carpenter). Perhaps the most embarrassing incident, however, occurred when two African friends witnessed Google Photos tag their selfie as containing two gorillas instead of people (Guynn). The ubiquity of machine learning algorithms in everyday life means that unless unfair biases are eliminated, countless numbers of people will be unfairly victimized in a subtle and seemingly undetectable manner. The Future of Privacy Forum indicates that there are four major harms of algorithmic bias: loss of opportunity, economic loss, social detriment, and loss of

liberty (Smith). Clearly, the impacts of unfair bias are severe and this issue must be addressed immediately. Though it often appears such bias is the result of intentional malice, the truth is that these are almost always cases of unintentional oversight that lead to statistical bias.

Bias refers to the tendency to overestimate or underestimate the value of a statistical parameter. Not all biases in algorithms are necessarily bad. Machine learning models must make assumptions in order to properly make predictions about the data; this is known as productive bias. The issue is when this bias becomes unfair by taking into consideration the aforementioned "protected attributes." This bias is caused by flaws in their training data, which are used to develop and tune the model. Amazon's recruiting algorithm contained gender bias because it was trained on the company's previous recruiting data, which was biased towards men due to the industry's male-dominant history (Lauret). The training data must be properly processed to ensure that minimal unfair bias is present. However, this is much more difficult than it seems. For instance, simply removing demographic information from recruiting data often times is insufficient to truly eliminate unfair bias. The algorithm can still learn to be racially biased through other data, such as the home address of the individual, since race is often correlated with neighborhood location. Fortunately, there are methods that assist with eliminating unfair bias in machine learning algorithms.

Certain systems contain biases that are implicit and are harder to detect. Hence, machine learning systems need to follow certain standards to prevent reduce any bias. In general, bias in machine learning can be mitigated through greater transparency and awareness of the systems in place. In 2018, the European Union established the General Data Protection Regulation, which included key principals in protecting data property rights and setting the foundation for the

ethical handling of people's data (European Commision). These laws establish the limits of the usage of a person's data, treating it more like property. Thus, these limits allow for more fair machine learning systems to be built since the data given to systems will be properly filtered and used for its intended purpose. Though this is a correct step in the right direction, stricter regulation of data usage and making machine learning systems more transparent will become necessary to reduce bias.

Methodology

In order to better understand the effects of machine bias on real-world applications, we decided to create our own machine learning model that would determine the education level of an individual based on various characteristics, such as profession, age, salary, etc. The model would determine whether an individual, given his or her information, possessed an education level equivalent to or higher than a bachelor's degree. We developed various iterations of this model in order to test varying degrees of bias. Each iteration of our model would be scored for unfair bias using a custom index we developed that will be further described in the "Model Evaluation" section.

Materials

In order to process the data and build the models, we used Python and some of its various libraries, which were Pandas, Scikit-Learn, XGBoost, Pickle, and Matplotlib. We also used Google Sheets for data visualization and Google Colab as our computing platform.

Data

The data used to train and test the models was sourced from the 2013 American Community Survey, a service provided by the U.S. Census Bureau. The survey provides comprehensive data about 3.5 million households regarding a wide variety of topics, including "ancestry, education, work, transportation, internet use, and residency." During data processing, we removed all variables that were not correlated with education level, in order to reduce noise and improve model performance. The remaining variables were age, poverty level, wage, total earnings, social security income, supplementary social security income, healthcare, type of healthcare (public or private), occupation, number of hours worked per week, number of weeks worked per year, mode of commute to work, and minutes spent commuting to work. We chose to keep these variables because such variables, with the exception of age, are indicative of socioeconomic status, which is correlated with education level (Aikens 2008). Age was kept because it also provides some indication into the education level of the individual; for example, young adults in their early twenties likely have less education than their older counterparts. We also introduced some variables into our model that could possibly cause racial bias, namely the race of the individual, language spoken at home, and how well the individual could speak English (as rated by him or herself).

We continued with our data processing by filtering out individuals who were unemployed, or were not African American or Caucasian. Unemployed individuals were excluded because many of our variables could only provide meaningful information if the person of interest was employed (e.g. occupation, number of hours worked per week, etc.). We chose to exclude individuals of other racial groups in order to better examine and compare the effects of

machine bias on one historically underprivileged group and on one historically privileged group. Finally, we filtered out individuals who had missing data values to preserve the completeness of our data. After completing data processing, we had 1,302,171 data points in total, to be used for either training or testing our model.

Model

We selected our algorithm out of three that we felt were appropriate for our task at hand. These algorithms were Logistic Regression, XGBoost, and Support Vector Machine. Logistic Regression is incredibly efficient to train and offers a simple implementation. XGBoost is a highly versatile algorithm that has inherent safeguards against overfitting, an error which hampers a model's ability to generalize to new data. Support Vector Machine performs well with high dimensional data, which is important for our task as the data used contains over 600 dimensions. We ended up choosing Logistic Regression over the other algorithms because it is far more difficult to fine tune a more complex model like XGBoost and Support Vector Machine. We decided to choose Logistic Regression over XGBoost and Support Vector Machine because they offered similar performances in our preliminary analysis, but Logistic Regression is much faster to train than the other two algorithms. In addition, since Logistic Regression is a simpler model and provides equal performance, it offers a more informative interpretation of results and provides a better generalization to new data.

In order to train our model, we first split up our census data into two datasets: the training dataset and testing dataset. The training dataset is used to feed the Logistic Regression model examples of people with higher education and lower education in order to learn to differentiate the two. The test dataset is used to test our model's performance on data that it has not seen

before, thus allowing us to evaluate its actual performance. We also utilize Stratified K-fold Cross Validation during training in order to better assess the model's performance and monitor for overfitting. Once the model is trained, we store the model in a pickle file (.pkl) and then test it on the test dataset.

We created three iterations of our model: one containing blatant racial bias, one containing implicit racial bias, and one without racial bias. The model with blatant racial bias, referred to as Model A, contains all of the variables mentioned in the "Data" section, including race, home language, etc.We consider Model A to contain blatant racial bias because a variable is present in the data which explicitly provides information to the model about the race of the individual, making it easy for racial bias to be incorporated into the model. Model B is considered to contain implicit bias because while it does not explicitly factor in race, it still uses variables that can act as a proxy for race: namely, the language spoken at home and proficiency in speaking English. The language spoken at home can indicate the heritage of an individual, which is strongly associated with race, and underprivileged groups often face adversity in academics, which can be indicated by a lack of proficiency in English. While less obvious, racial bias can still be incorporated due to the presence of such variables. Finally, Model C is without unfair bias, or at least to the extent of our abilities. For this model, we excluded variables at risk of both explicit and implicit bias. It should be noted, however, that oftentimes variables at risk of implicit bias may be useful variables to consider when constructing a model. We noted that proficiency in English, while certainly correlated with race, is also a strong indicator for education level, as those with higher education should also have stronger proficiency in English. In the end, we decided that Model C should not take into account English proficiency in order to

construct a model as unbiased as possible, although this issue will be elaborated on further in the "Model Evaluation" section.

Model Evaluation

After running the three models on our testing dataset, the prediction data was gathered and compiled into the data table below:

	Ground Truth	Model A	Model B	Model C
# of Upper Education Black	8643	6255	7020	7143
# of Upper Education White	108379	93270	92691	93388
# of Lower Education Black	216690	29049	28284	28161
# of Lower Education White	201002	216111	216690	215993

We evaluated model performance using two metrics: F-Score and accuracy score. The accuracy score tells us how many times the model accurately predicts a person's education level out of all the people it analyzes. The F-Score (or F_1 score) is a weighted average between the precision and recall of the model. The precision of a model is the ratio of correctly predicted upper education observations to all upper education observations. This tells us how many actually have a higher education out of all the people labeled as higher education. The recall is the ratio of correctly predicted higher education observations to all of the observations. This tells us how many were labeled with a higher education correctly out of all the people who have a higher education. The F-Score takes into account both precision and recall and is a better metric than accuracy alone because it accounts for uneven distributions in the data.

Each of the model's metric scores are displayed in the table below:

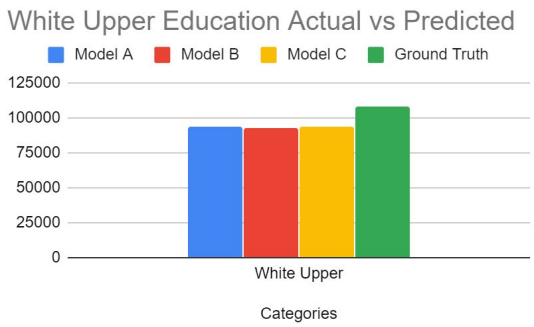
	Model A	Model B	Model C
F-Score	0.724013	0.723686	0.722298
Accuracy	0.825062	0.824867	0.824010

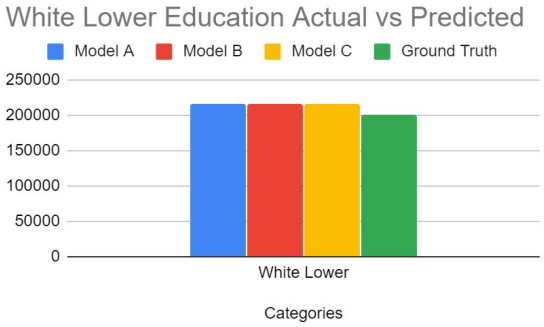
One important observation to make is that the F-Score and Accuracy decrease as we remove biased variables from the datasets.

In order to actually quantify the amount of racial bias present in each model, we use the Wasserstein Distance to evaluate the statistical distance between the frequency distribution of race by the model predictions and the frequency distribution of the ground truth (the actual frequencies). A model is more biased if the statistical distance is greater (and therefore less reflective of the actual distribution of race). The statistical distances are displayed below:

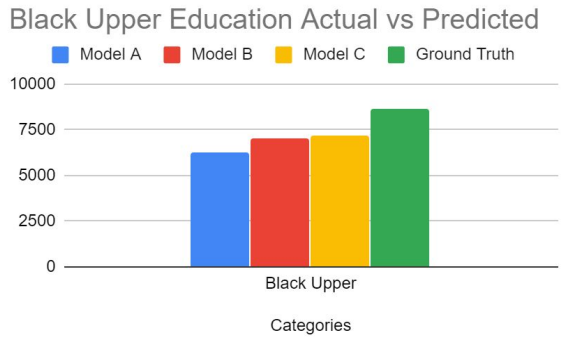
Model A	Model B	Model C
8748.5	8655.5	8245.5

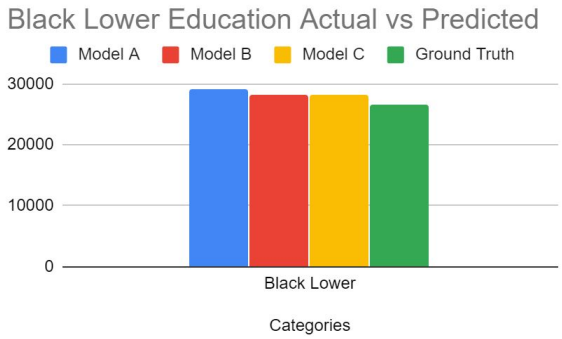
In line with our predictions, Model A is the most biased with the highest Wasserstein distance, Model C is the least biased with the lowest distance, and Model B is in between with a distance greater than that of Model C but lower than that of Model A. These differences in bias can also be visualized through the frequency distribution plots of education level by race. Each of the model's graphs for upper and lower education are displayed below:





Although the difference is small, as we go from Model A to C there is a slight upward trend that indicates as we remove variables that may cause bias, the model gets less accurate in determining the number of people with a higher education. To test this observation, we also graph the model's performance for African Americans with a higher education.





Our model's fairness (as calculated by the Wasserstein Distance) on predicting the educational level of African Americans dramatically increased as we removed variables susceptible to bias. When we removed race when considering education level, the amount it predicted increased by a lot more than when we removed both race and language spoken. This indicates that our model is biased towards race and automatically considered it as a strong indicator of education despite the fact that race does not determine education level. Rather, the reason is because underprivileged minorities have historically faced adversity in education. Once we remove race as a considerable factor for our model, it made less mistakes with classifying

African Americans than when the model considered race. This demonstrates that the model itself is not responsible for the bias, but rather the data that it was trained on.

Conclusion

Our analysis of the three models we built demonstrates that unfair bias in machine learning algorithms is easy to incorporate, but difficult to remove. Even upon removal of explicit variables, bias can still be found indirectly through other implicit variables. Furthermore, complete removal of these biased variables presents another problem. Although biased, such variables still provide valuable information to the model can be used legitimately in performing its task. This "fairness-accuracy" tradeoff is clearly exhibited through the decline in the performance metric scores of our models as bias was further reduced, but is also present in the real-world as well.

These examples all demonstrate the difficulty of perhaps the hardest problem in the field of Machine Learning. Data scientists must build their models so that they attain as high of a performance as possible, yet still ensure that their models are as unbiased and fair as possible. This can only be done with thorough knowledge of the task at hand, in order to determine how the data should be processed in order to eliminate as much unfair bias as possible. This near-impossible feat is why the field of Machine Learning is tremendously difficult, both from an ethical and scientific standpoint.

Works Cited

Aikens, N. L., & Barbarin, O. (2008). Socioeconomic differences in reading trajectories: The contribution of family, neighborhood, and school contexts. Journal of Educational Psychology, 100, 235-251. http://dx.doi.org/10.1037/0022-0663.100.2.235

Angwin, Julia, et al. “Machine Bias.” ProPublica, 9 Mar. 2019, https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

.Carpenter, Julia. “Google's Algorithm Shows Prestigious Job Ads to Men, but Not to Women. Here's Why That Should Worry You.” The Washington Post, WP Company, 28 Apr. 2019, https://www.washingtonpost.com/news/the-intersect/wp/2015/07/06/googles-algorithm-shows-prestigious-job-ads-to-men-but-not-to-women-heres-why-that-should-worry-you/.

Dressel, Julia, and Hany Farid. “The Accuracy, Fairness, and Limits of Predicting Recidivism.” Science Advances, American Association for the Advancement of Science, 1 Jan. 2018, https://advances.sciencemag.org/content/4/1/eaao5580

European Commision, “Data Protection.”European Commission, 6 Aug. 2019, https://ec.europa.eu/info/law/law-topic/data-protection_en.

Guynn, Jessica, and USA TODAY. “Google Photos Labeled Black People ‘Gorillas.’” USA Today. Accessed 12 Jan. 2020.

Lauret, Julien. “Amazon's Sexist AI Recruiting Tool: How Did It Go so Wrong?” Medium, Becoming Human: Artificial Intelligence Magazine, 16 Aug. 2019, https://becominghuman.ai/amazons-sexist-ai-recruiting-tool-how-did-it-go-so-wrong-e3d14816d98e.

Smith. “Future of Privacy Forum.” Future of Privacy Forum ICal, 11 Dec. 2017, https://fpf.org/2017/12/11/unfairness-by-algorithm-distilling-the-harms-of-automated-decision-making/.

Spielkamp, Matthias. “Inspecting Algorithms for Bias.” MIT Technology Review, vol. 120, no. 4, July 2017, pp. 96–98. EBSCOhost, search.ebscohost.com/login.aspx?direct=true&db=a9h&AN=123605629&site=ehost-live.

US Census Bureau. “American Community Survey (ACS).” The United States Census Bureau, 11 Dec. 2019, https://www.census.gov/programs-surveys/acs/.