Alex Miale

CS4100

Final Paper

Introduction

With the rise of machine learning (ML) and artificial intelligence (AI), one or the other has woven its way into technology systems everywhere. From recommendations for Netflix to analysis for sports teams to systems that reject or deny job applications, there is a good chance that some type of AI was even used to help you decide what to have for dinner. While recommendation systems work well with ML, some other tasks have proven difficult when trying to maintain a fair system. This is due to how AI and ML work at a basic level. AI relies heavily on data and it uses this data to predict a certain type of outcome which can be boiled down into one of two categories: regression and classification. To accurately predict the outcome, lots of data is required. But, not all of it can be found in recent times so data from the past is used to help solidify the model and its predictions. This seems like a practical solution to the problem because the data collected is still "real-world data". However, depending on the task at hand, the data needs to be clean and unbiased so that the model made from the data is also unbiased. Unfortunately, this is not the case for a variety of fields. These models are supposed to make predictions about people's lives and an unfair model leads to a continuation of bias in society and unethical decisions for the person or people involved.

The article, *Bias and Unfairness in Machine Learning Models: A Systematic Review on Datasets, Tools, Fairness Metrics, and Identification and Mitigation Methods*, is a literature review going over the datasets and methods used to study bias and unfairness in machine

learning models. One of the acknowledged datasets is Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) which is meant to predict whether an inmate will re-offend within two years of release with data on sensitive attributes such as race and gender. The likelihood of such recidivism was denoted by a score of 1 to 10, called a decile score. With such a controversial topic, the authors claim that this article is "one of the most widely used datasets for bias and fairness experiments" (Pagano et al., 13). The literature review cites the ProPublica article of its analysis of the dataset which found its algorithm to overestimate the rate of recidivism by 45% for black people and only 23% for white people (Larson et al). The purpose of this project was to see if there are ways to mitigate the bias of this (or any) dataset and to see if there was a way to make a "fair" model. Using the COMPAS dataset, 3 preprocessing techniques were used before running a model to see how "fair" the outcome was.

Methods

The dataset used was taken from the ProPublica analysis of COMPAS, specifically the raw scores data in a comma-separated value (CSV) file. This dataset had the purest form of the data so the work could try to mitigate the bias at its source. The data was loaded into a Pandas DataFrame which is optimal for data inspection and works well with other libraries like NumPy and SciKit-Learn. The raw data had lots of different features, such as first, middle, and last name, marital status, and date of birth. Lots of these features were not necessary, so those columns were dropped. Other columns that were deemed important like legal status and race were converted from a categorical string datatype to numerical values in a process that is referred to as label encoding. Many of the categorical types already had an associated numerical value column, so the categorical column was removed. There were also dictionaries created with the keys being the category and the value as the associated numerical value. This was done with the idea that

possible analysis later on would require it but, due to time constraints, this did not happen. The data was then split into train and test sets using the SciKit-Learn train_test_split method, and further processed using a StandardScaler and MaxAbsScaler. StandardScaler removes the mean by setting it to 0 and scales the standard deviation to unit variance, 1. MaxAbsScaler divides every value by the maximum absolute value, scaling the values to -1 and 1. The final preprocessing technique used was from Aequitas. Aequitas is an open-sourced toolkit that is made for auditing models and making models fair. This technology was found in the literature review where it was described as one of the most common tools for addressing bias and unfairness (Pagano et al, 1-2). The method used was the DataRepairer which makes the data independent of a given sensitive attribute, race in this case. There were now four datasets: raw, standard scaled, maximum absolute scaled, and Aequitas. These were put into a SciKit-Learn pipeline with a Polynomial Features transformer and an ElasticNet algorithm. This is how SciKit-Learn does Polynomial Regression with an elastic net (L1 and L2) regularizer. The polynomial features method was used to find non-linear relationships in the data and Elastic Net was used due to possible correlation between features.

Results

After running the models, the score was printed out too. The score for an Elastic Net model is the coefficient of determination which measures, in short, how well a model predicts an outcome. For example, a coefficient of determination of 0.9 says that 90% of the variance in the dependent variable can be explained by the independent variable. This would indicate a strong relationship between the variables and a good fit from the model. Strangely, the dataset with the highest score was the raw data, with a score of 0.92. This is followed by the Aequitas set with 0.82, standard scaling at 0.52, and the maximum absolute scaling, which ran into some issues,

had a score of 0. The maximum absolute dataset will largely be ignored from here on out. It is rather strange that the unprocessed data had the highest score, in other words, the highest correlation. This will be discussed in later sections. The other metric used to analyze the data was plotting histograms of the predicted and actual data. The ProPublica article showed that the COMPAS dataset had lower decile scores for white defendants and more even distribution for black defendants, highlighting the skew that leads to bias in the models. The predicted decile score for black defendants tended to overestimate the number of people with a certain decile score by about 200-300 people. There was also over-prediction for white defendants, but it was usually about 100 higher than the actual number, which highlights the predictive bias in the unprocessed data. The standard-scaled data had a very high over-prediction for decile scores between 2 and 4 for both white and black defendants, estimating at a rate two times the actual value for the area. The scores of 5-10 were very low. The absolute maximum scaling had every prediction between 3 and 4, which can be attributed to an error in the model and can be disregarded. For the Aequitas dataset, both defendants' graph predictions had a similar shape, though it was skewed slightly higher for black defendants. For black defendants, the data peaked at 1000 predictions for a decile score between 2 and 4, had the same level at 5, and tapered off with a similar shape to the actual values. White defendant's prediction had the same peak but between 1 and 4 before tapering off. The Aequitas dataset had more similar shapes to the predictions which I would claim to be the "most fair and unbiased" dataset used for this evaluation.

Discussion

The evaluation of the different preprocessing methods highlighted various aspects of evaluating fairness and bias in ML models. To begin with, the model with the highest accuracy

had the highest bias. The unprocessed, raw scores had bias encoded into the dataset as shown by ProPublica. This means that the model created from this data also had the same bias in it. However, it finished training with a score of 92%, which was 10% better than any of the other models created. This shows a fairness/accuracy trade-off that is becoming more apparent in ML research. Maximizing the accuracy of an algorithm typically leads to the model fitting closely to the training data. As we know the data is biased, the model with higher accuracy with biased data will also be biased. This was also shown by the model with data preprocessed with Aequitas. The model had a lower score (82%) but seemed to be more fair than the model with unprocessed data. The graphs show that the model underpredicted instances of decile scores of 0 to 2 and 6 to 10 while overpredicting scores of 3 to 5 for white and black defendants. This seems to level the field for the predictions between the two groups. This can be attributed to the Aequitas DataRepairer class which looks to make the data independent of a certain attribute, and the graphs prove that it did. While the model is not perfect in this instance, it highlights that a possible way to reduce bias in datasets is to make the data independent of the sensitive attribute. Logically, this makes sense as removing some form of correlation between the sensitive attribute and the prediction should lead to fewer instances of scores being skewed because of the attribute. Some philosophers like Brian Hedden claim that unfairness towards individuals comes from decisions made “in virtue of their membership in a certain group” (213). Should this hold, it makes sense that removing dependence would remove decisions being made because of membership in a group. The project fell short in a couple of areas, namely model selection, and depth of fairness evaluation. Only using one model was not ideal, and there was no section dedicated to feature selection which could have reduced bias and/or increased accuracy. The evaluation of fairness came down to human evaluation from the graphs and could have been

much more extensive by finding the loss of the models and looking at an average decile score for each group from each dataset. For example, the article written by ProPublica used a GLM model and was able to look at false positives and false negatives in their evaluation. Using a more complex model leads to better results and the explicit use of type I and II errors also leads to a more in-depth evaluation. With that being said, the research done for this paper still showed bias in the COMPAS dataset and highlighted a possible way to reduce this bias.

Conclusion

Bias in ML models and AI has become a forefront of technology research. With AI being implemented in both everyday uses and difficult decision-making, fair models are necessary to maintain the status quo. Especially if ML will be used to help predict recidivism, fairness is essential in deploying such a technology. Based on the project, removing dependence between the data and a sensitive attribute appears to have a positive effect on the fairness of models and their datasets. Going forward, preprocessing techniques including this should be used to help make proper decisions and mitigate the bias in future AI and ML models.

Works Cited

1. Hedden, Brian. “On statistical criteria of Algorithmic Fairness.” *Philosophy & Public Affairs*, vol. 49, no. 2, Mar. 2021, pp. 209–231, https://doi.org/10.1111/papa.12189.
2. Larson, Jeff, et al. “How We Analyzed the Compas Recidivism Algorithm.” *ProPublica*, 23 May 2016, www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm.
3. Pagano, Tiago P., et al. “Bias and unfairness in Machine Learning Models: A systematic review on datasets, tools, fairness metrics, and identification and mitigation methods.” *Big Data and Cognitive Computing*, vol. 7, no. 1, 13 Jan. 2023, p. 15, https://doi.org/10.3390/bdcc7010015.

Technologies Used: NumPy, SciKit-Learn, Pandas, Aequitas