Clear and Present Danger: Dataset Bias in Classification Models

Julie E. Cestaro
CSCI-GA 2565 Final Course Project [ID 20]
New York University
julie.ces@nyu.edu

Abstract

Garbage in, garbage out is an old colloquialism in computer science, and it is never more true than in machine learning. As society continues to ask machine learning systems to make increasingly high impact decisions, it becomes increasingly imperative to recognize the importance of unbiased representation in the datasets used to train and test models. This work demonstrates how easily a biased perspective can become the basis for predictions in a classification model, and offers a simple method to mitigate this bias.

1 Introduction

Human history is rife with bias and, despite the best intentions, our datasets often reflect the biases that are embedded in the construct of our society. There have been numerous examples over the past five to ten years of egregious biases being perpetrated in various high impact spaces ranging from recidivism predictions in criminal justice [1], to facial recognition models [2], to the word embeddings that shape our language tools [3]. Even if machine learning is not directly used to make a high impact decision, it is increasingly likely that a model will be integrated into a pipeline that ultimately results in a high impact decision[2].

One can imagine a bank loan allocation system where the decision to allocate a loan to a given applicant depends entirely, or largely in part, on the applicant's predicted annual income. If the model were to incorrectly predict the income of an applicant based on a protected attribute, like sex, then an individual would be likely to be denied a loan based on their membership of that protected class. This would continue to prop up historical barriers to generational wealth accumulation[4]. Therefore, it is critical to output unbiased results, even if the model is not directly making high impact decisions.

This paper makes two contributions. First, I empirically demonstrate that the predictions of classification models will explicitly perpetuate the biases of the dataset on which they were trained. Second, I show that the bias of the dataset can be decreased by increasing the representation of the underrepresented group.

This underscores the need for focus on unbiased and fair representation at the dataset level when training a new model.

2 Related Work

2.1 Early Fairness Research

The notion of bias seeping into a machine learning system is certainly not an entirely novel one, but compared to the entire discipline of machine learning, the study of bias and fairness is still relatively

new. An early foray into the dangers of biased machine learning came just five years ago in a 2018 paper which pointed out the steep fall of Microsoft's "Tay" chatbot, originally trained to respond like a normal teenaged girl, into anti-Semitism, racism, and sexism in a very short sixteen hour period[5].

In 2019, Thea Gasser performed an extensive literature review which highlighted the high likelihood that bias would become problematic in machine learning and, more broadly, in artificial intelligence (AI)[6]. Gasser also provides a comprehensive overview of possible sources of model bias, including dataset bias, although the subsequent framework for unbiased AI relies more on organizational processes than actual empirical measurement[6].

2.2 A Formal Definition for Fairness

Davies et al. introduces a formal mathematical representation for fairness using the examples of diabetes screening and college admissions[7]. In that, they define five definitions of fairness: demographic parity, equalized false positive rates, counterfactual predictive parity, counterfactual equalized odds, and conditional principal fairness[7]. The work presented in my paper will primarily focus on the demographic parity and counterfactual fairness definitions of fairness.

It should be noted that it is mathematically impossible to maintain demographic parity, predictive parity, and equalized odds at the same time; and actually only one can be satisfied by a well calibrated classifier at any given time[8]. This is the reason for the focus on demographic parity in my paper.

2.3 Recent Work in Fairness

More recent research has brought light to the dangers of dataset bias and the widespread failure of unbiased representation in datasets along the axes of age[9], race[10], and gender[11], among others. With this acknowledgement also came propositions for mitigating bias at the model-level, including processes like REPAIR[12] and RESOUND[13].

My work addresses this failure of representation by specifically investigating dataset-level bias in the training set for classifiers and the subsequent impact on predictive fairness. Rather than looking for a model-level bias mitigation strategy, I outline a simpler method to decrease bias at the dataset-level.

In considering the fairness of classifiers, though, recent research on the inherently reductive nature of classification in general[14] should not be ignored. My work unfortunately, though perhaps inevitably, falls prey to this via the binary representation of sex.

3 Method

The method for this paper begins with a hypothesis: that models trained on biased data will directly reflect the biases of the dataset in their predictions. First, I demonstrate the effects of a deliberately unbiased dataset on the predictive accuracy of classification models as a baseline for comparison. Then, I create a dataset that is highly biased towards a specific protected class to exaggerate the dangers of insufficient data for groups that have been, and still are, marginalized in our society. From there I confirm that the protected class is indeed the cause the bias in the predictions by demonstrating a breakdown in counterfactual fairness. Finally, I show that a simple way to mitigate this bias is by improving representation.

4 Dataset

This work was done using the Adult Census Income dataset[15] from the UC Irvine Machine Learning Repository. The German Credit Risk dataset[16] was also considered, but it only contains 1,000 examples. Having over 30,000 examples in the Adult Census Income dataset allowed plenty of space for pre-processing and subsampling without any unintentional underrepresentation.

The Adult Census Income dataset[15] was extracted from the 1994 Census Bureau database by

Ronny Kohavi and Barry Becker, and it outlines a classification task based on income; specifically whether an individual makes over $50,000 per year. The unprocessed dataset contains 32,600 examples with fourteen features including age, education level, marital status, occupation, race, sex, and native country, among others. Without pre-processing, 76% of the total dataset are labeled as making more than $50,000 per year, and males represent 66% of the examples.

4.1 Pre-processing

This dataset in an unprocessed state is not highly useful for training a classification model. A vast majority of the data are represented as strings, many rows are missing data for at least one feature, and there is substantial data leakage between some features.

4.1.1 Addressing Data Leakage

Each row has a feature called "fnlwgt" (final weight), which resembles a similarity score among the individuals. Similar individuals have similar values for "fnlwgt". This is an obvious point of data leakage. Also, in addition to having a feature called "sex" for each individual in the dataset, there is a feature called "relationship", which often contains a gendered representation of marital status (i.e. husband or wife), from which an individual's sex would be easy to infer.

In trying to understand how the feature value for sex changed predictions, it was necessary to ensure that an individual's sex could not be inferred from other features. The features "fnlwgt" and "relationship" were dropped in pre-processing due to data leakage concerns. "Capital.gain" and "capital.loss" were also dropped due to 95% of the values for both features being 0. The feature "education" was dropped in favor of the numerical feature "education.num", which represented the number of years the individual was educated as opposed to a string type description of their education level. Then, I filtered for rows containing individuals whose "native.country" was the United States, to ensure that there would be no confounding culture bias in the results. Finally, I dropped any of the remaining examples that had incomplete data. In all, this left 30,162 examples to work with.

4.1.2 Label Encoding

The next phase of pre-processing involved converting all of the features represented as strings to numerical representations. For each feature represented as a string, I assigned a number to each unique string that appeared as a value for that feature. Then the string representation was replaced by the new numerical representation. Sex and income were specifically given binary representations. I represent female as 0 and male as 1. Individuals with an income over $50,000 per year are given a label of 1, individuals making less are given a label of 0.

Figure 1 shows the distribution of men and women across both income groups in the original dataset after pre-processing. Figures 2 and 3 show the distributions of the feature values after converting them to numerical representations.

5 Experiments

For this project, four experiments were conducted. The first created an unbiased baseline against which to measure subsequent results. The second created an intentionally biased model to measure how predictions changed. The third demonstrated the extent to which the biased models were incapable of achieving counterfactual fairness when trained on the biased dataset. The fourth and final experiment showed how the models may be de-biased in order to achieve reasonably unbiased results. All of the experiments described below involve the training of six different classification models using the following six algorithms, respectively: Logistic Regression, Linear SVM, Decision Tree, Random Forest, Naive Bayes, and K-Nearest Neighbors. All below references to "the models" can be assumed to refer to these because, in large part, the accuracies for each model are very similar.

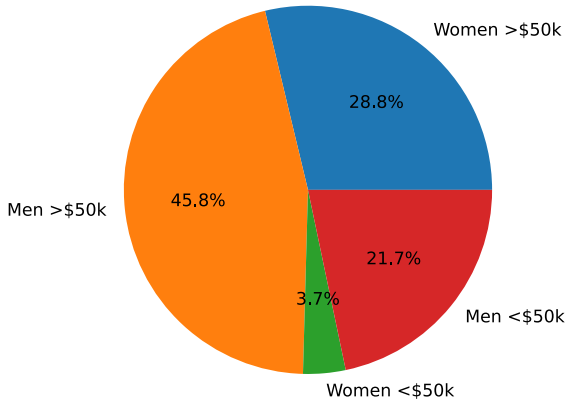

Figure 1: Distribution of men and women across income groups in original dataset

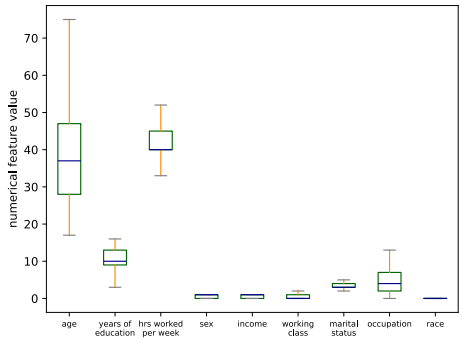

Figure 2: Distribution of feature values after pre-processing

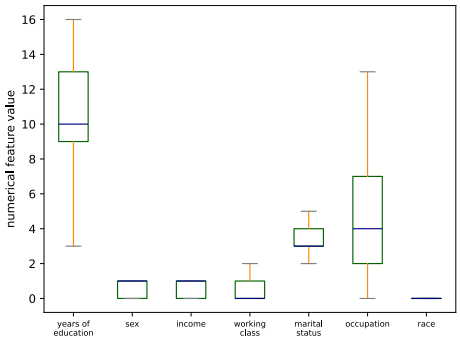

Figure 3: Distribution of feature values after pre-processing: a closer look

5.1 Making Predictions after Training on an Unbiased Dataset

In creating an unbiased baseline, it was not enough to simply use the dataset with its existing distributions of men, women, and their associated incomes. I wanted to be certain that I would not introduce any latent bias into my baseline results. I subsampled the pre-processed dataset such that the new, unbiased dataset represented 7,817 men and 7,817 women. Both the set of men and the set of women make more than $50,000 per year in equal proportions – 42.8% each – as is demonstrated in Figure 4.

I then split the dataset into a training set (70%) and a testing set (30%), as is the generally accepted practice to avoid overfitting. I trained the models using the training set and then predicted on the test set. The results can be seen in Table 1. All of the models achieved highly similar accuracy rates.

To determine a more specific baseline for each sex represented in the data, I subsampled the test set for only women, and predicted on this subsampled test set. The resulting accuracies can also be seen in Table 1. I performed a similar subsampling and prediction process for men as well.

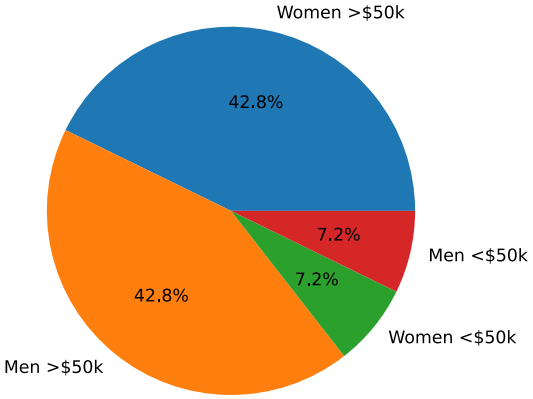

Figure 4: Distribution of men and women across income groups in unbiased dataset

	Men and women	Women	Men
logistic regression	0.87	0.88	0.87
linear SVM	0.87	0.88	0.86
decision tree	0.86	0.88	0.84
random forest	0.90	0.91	0.90
naive bayes	0.86	0.87	0.85
k-nearest neghbors	0.87	0.87	0.86

Table 1: Accuracy of predictions for different test groups after training on the unbiased dataset

Those accuracies can also be seen in Table 1. The accuracy rates for the subsamples of women and men are, as expected, not markedly different from each other. This indicates demographic parity and counterfactual fairness.

5.2 Making Predictions after Training on a Biased Dataset

When deciding how to create planned bias in my dataset, the choice to create a bias that favors women was an intentional one[17]. However, I wanted to ensure that the world that the biased dataset represented didn't completely disregard men, because using a model to predict on entirely unseen data would simply be a test of its ability to generalize. Instead, I wanted to create a dataset that represented both men and women, but only men whose income was less than $50,000 per year.

The complete, biased dataset contained 14,903 individuals in total, 40% of whom were men. See Figure 5 for a complete picture of the distribution. I split the dataset into a training set (70%) and a testing set (30%) to avoid overfitting. I then subsampled the test set for only women, and predicted on this subsampled test set. The resulting accuracies can be seen in Table 2. Note that these accuracy rates are largely unchanged from the accuracy rates for women after training on the unbiased dataset (Table 1).

When making predictions for men, I wanted to see how the biased models would perform when asked to predict on both men who make less than $50,000 per year and men who make more than $50,000 per year. In order to do this, I combined the men from the test set – again, to avoid overfitting – with the men that were kept out of the biased dataset due to the fact that they made over $50,000 per year. The distribution of income in this test set is shown in Figure 6. The accuracies of the predictions for that set are shown in Table 2.

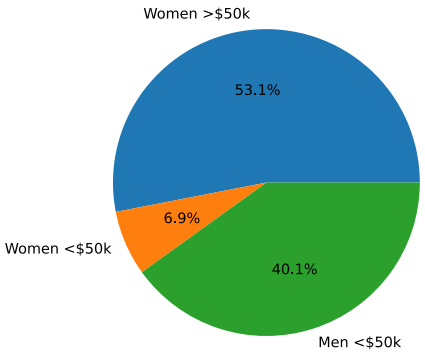

Figure 5: Distribution of men and women across income groups in biased dataset

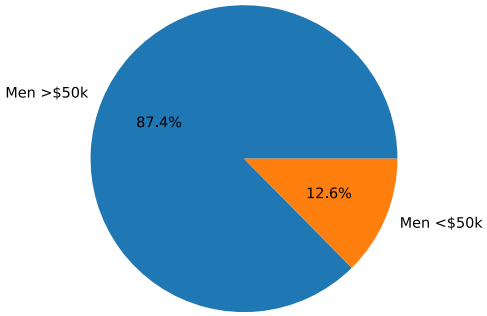

Figure 6: Distribution of income for men in the test set for the biased models

	Women	Men
logistic regression	0.88	0.13
linear SVM	0.88	0.13
decision tree	0.94	0.13
random forest	0.95	0.13
naive bayes	0.88	0.13
k-nearest neghbors	0.89	0.52

Table 2: Accuracy of predictions for different test groups after training on the biased dataset

It is important to note that 12.6% of the test set consisted of men making less than $50,000 per year. This aligns with the predictive accuracy of the biased models for the test set of men. In fact, nearly every model had an average prediction of 0 (the label for an income of less than $50,000 per year) for this test set of men. See Table 3 for specifics. This clearly demonstrates that by training a model on data that shows that no man makes more than $50,000 per year, I implicitly taught these models that there does not exist a man who makes more than $50,000 per year. The models translated that rule directly into their predictions.

5.3 The Breakdown of Counterfactual Fairness

In order to rule out the possibility that some other feature was causing the large decrease in accuracy for the models trained on a biased dataset when predicting for men, I conducted the following experiment. I created a test set consisting of only men who make more than $50,000 per year; in fact, I used the exact set of individuals who were removed from the dataset in order to create the biased dataset. This dataset consisted of 12,601 men whose income was more than $50,000 per year.

	mean prediction
logistic regression	0.0
linear SVM	0.0
decision tree	0.0
random forest	0.0
naive bayes	0.0
k-nearest neghbors	0.42

Table 3: Average predictions for the test set of men after training on the biased dataset

	Men >$50k	Counterfactual
logistic regression	12601	773
linear SVM	12601	10473
decision tree	12601	2847
random forest	12601	2390
naive bayes	12601	0
k-nearest neghbors	6731	5376

Table 4: Number of incorrect predictions made by the biased models in different test groups

	+1000	+2000	+3000	+4000	+5000	+6000	+7000
logistic regression	0.31	0.47	0.53	0.63	0.69	0.72	0.76
linear SVM	0.46	0.36	0.43	0.64	0.76	0.72	0.73
decision tree	0.49	0.56	0.65	0.69	0.74	0.78	0.81
random forest	0.45	0.59	0.67	0.73	0.78	0.81	0.84
naive bayes	0.15	0.31	0.39	0.44	0.49	0.53	0.58
k-nearest neghbors	0.52	0.58	0.63	0.68	0.72	0.74	0.76

Table 5: Accuracy of predictions after adding back samples of men whose income is >$50k

I then used the biased models to predict on this set of examples. Unsurprisingly, they performed poorly, with many of the models unable to make a single correct prediction, shown in Table 4.

Then, for each model, I took all of the individuals that received an incorrect prediction and converted them to their counterfactual representation. That is, I left all other features unchanged and switched the value for "binary.sex" from 1, representing a man, to 0, representing a woman. I then fed those counterfactual examples back into the biased models. Table 4 shows that the number of incorrect predictions dramatically decreased when the biased models perceived each individual as a woman as opposed to a man. This clearly demonstrates that the bias introduced in the experiment described in section 5.2 was based on the reported sex of each individual in the dataset. Because many of the predictions became dependent on the sex of each individual, this can be described as a breakdown in counterfactual fairness.

5.4 De-biasing the Biased Dataset

When considering the dangers of a biased dataset and a subsequently biased model, it seems natural that one would consider how to avoid or improve upon a biased model. If a model can learn bias from a biased dataset, then it should follow that decreasing the bias in the dataset should decrease the bias in the model. This turns out to be true.

There are approximately 6,000 men who make less than $50,000 per year represented in the biased dataset. I started by adding an additional 1,000 men who make more than $50,000 per year to the biased dataset, and then continued to iteratively add examples of this type in batches of 1,000.

Adding 3,000 men who make more than $50,000 per year brought the accuracies of the models to about a 50/50 guess, shown in Table 5. While not ideal, this is still a non trivial improvement for models that were initially incapable of making a single correct prediction, as shown in Table 4. Adding 7,000 men who make more than $50,000 per year brought the accuracies of most of the models within 0.10 of their accuracies on the unbiased dataset (Table 1).

It should be noted that even with this marked increase in the representation of men making over $50,000 per year in the dataset, women are still over represented in comparison, with nearly 8,000 women shown to make more than $50,000 per year in this test set. Regardless, the increased representation still improved the results for the underrepresented group, even though they remain underrepresented.

6 Conclusion

I started by generating some baseline accuracies for each model by training them on an unbiased dataset. These accuracies were very similar between models and demonstrated demographic parity. Then I introduced bias by removing all men making over $50,000 per year from the dataset and training the models on that biased dataset. I demonstrated that all but one model trained on this dataset indiscriminately predicted that every man made less than $50,000 per year. I further confirmed this bias by showing that the number of incorrect predictions made by the biased models decreased when presented with counterfactual examples instead. Finally, I demonstrated a simple bias mitigation strategy in the form of increased representation by showing increased accuracy with increased representation.

The most important conclusion to draw from this is that these classifiers will very easily adopt the bias of the dataset on which they were trained and then perpetuate that bias in their predictions. This amplifies the need for fair and unbiased representation of the full population on which the models will be expected to predict. This is especially clear when considering the effectiveness of increased representation in mitigating bias and increasing accuracy. Tangentially, this work also shows how easily a bad actor could poison a dataset to unfairly benefit or punish a specific group.

6.1 Further Work

The work described in this paper is fairly narrow in scope as it only investigates binary classifiers and only involves a protected attribute, sex, with two assigned values. Further work should be done to investigate the perpetuation of dataset bias in regression models, or of protected attributes with more than two possible values. There is also a burgeoning line of research around fairness via structural causal models[8]. It would be worthwhile to investigate if these models can be swayed by a sufficiently biased training dataset given their robust promises of fairness.

References

[1] Alexandra Chouldechova. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big Data*, 5(2):153–163, 2017. PMID: 28632438.

[2] Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In Sorelle A. Friedler and Christo Wilson, editors, *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, volume 81 of *Proceedings of Machine Learning Research*, pages 77–91. PMLR, 23–24 Feb 2018.

[3] Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016.

[4] William Gale and Benjamin Harris. Changing wealth accumulation patterns: Evidence and determinants. In *Measuring and Understanding the Distribution and Intra/Inter-Generational Mobility of Income and Wealth*. University of Chicago Press, 2020.

[5] Daniel J. Fuchs. The Dangers of Human-Like Bias in Machine-Learning Algorithms. *Missouri ST's Peer to Peer 2*.

[6] Thea Gasser. Bias – A lurking danger that can convert algorithmic systems into discriminatory entities : A framework for bias identification and mitigation. *Missouri ST's Peer to Peer 2*.

[7] Sam Corbett-Davies, Johann D. Gaebler, Hamed Nilforoshan, Ravi Shroff, and Sharad Goel. The measure and mismeasure of fairness, 2023.

[8] Kailash Karthik Saravanakumar. The impossibility theorem of machine fairness – a causal perspective, 2021.

[9] Joon Sung Park, Michael S. Bernstein, Robin N. Brewer, Ece Kamar, and Meredith Ringel Morris. Understanding the representation and representativeness of age in ai data sets. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '21, page 834–842, New York, NY, USA, 2021. Association for Computing Machinery.

[10] Michele Merler, Nalini K. Ratha, Rogério Schmidt Feris, and John R. Smith. Diversity in faces. *CoRR*, abs/1901.10436, 2019.

[11] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. Men also like shopping: Reducing gender bias amplification using corpus-level constraints, 2017.

[12] Yi Li and Nuno Vasconcelos. Repair: Removing representation bias by dataset resampling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[13] Yingwei Li, Yi Li, and Nuno Vasconcelos. Resound: Towards action recognition without representation bias. In *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018.

[14] Kaiyu Yang, Klint Qinami, Li Fei-Fei, Jia Deng, and Olga Russakovsky. Towards fairer datasets: Filtering and balancing the distribution of the people subtree in the imagenet hierarchy. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, page 547–558, New York, NY, USA, 2020. Association for Computing Machinery.

[15] Barry Becker and Ronny Kohavi. Adult. UCI Machine Learning Repository, 1996. DOI: https://doi.org/10.24432/C5XW20.

[16] Hans Hofmann. Statlog (German Credit Data). UCI Machine Learning Repository, 1994. DOI: https://doi.org/10.24432/C5NC77.

[17] Caroline Criado Perez. *Invisible Women: Data Bias in a World Designed for Men*. Abrams Press, 2019.