Preliminary Evaluation of Fair-SMOTE Given Doctored Data

Ellis Wilson
North Carolina State University
ejwilso2@ncsu.edu

ABSTRACT

There has been a relatively large amount of research recently being done into how to avoid innate bias when running machine learning algorithms. Fair-SMOTE [1] has made great strides toward improving fairness while maintaining performance in the face of unintentional bias within datasets, but it has not been tested against *intentional* bias.

This work tests Fair-SMOTE on datasets which have been doctored in a simple way to various degrees. While much of the resulting data is inconclusive, a few key observations stand out: Many metrics of both performance and fairness actually show improvement in the face of data doctoring, some metrics show different trends depending on the dataset being analyzed, and Fair-SMOTE is not able to entirely overcome the effects of doctored data.

KEYWORDS

Fairness, Machine-Learning, Fair-SMOTE

ACM Reference Format:
Ellis Wilson. 2021. Preliminary Evaluation of Fair-SMOTE Given Doctored Data. In *Proceedings of ACM Conference (Conference'17).* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

The question of "can this be trusted" with regard to algorithms, programs, and even data has been a constant thorn in the side of anyone looking to use them. It is an even more pressing concern to people who do not have access to or control over these tools. Stories flood the news of social media showing people what will engage them the most, with no regard to its truth or other side effects. Websites which track shopping habits recommend useful suggestions, but this can also cause problems when unwanted or when shared with others.

But this question is most pressing when applied to software which is set up to make significant, immediate changes to people's lives. Software which decides whether it would be too risky to give an applicant a loan, whether a patient has a good enough chance to survive an operation, whether a prison occupant is likely to commit another crime if they are released from jail.

When closely inspected, there is more than one meaning of the word "trust" with regard to these types of important decisions. The simplest definition is with regard to accuracy: how well does this piece of software predict what it is intended to? If 95% of people offered a loan manage to pay it off while 95% of people rejected are later confirmed to have been truly unworthy of that loan, by whatever standard, then one might say that this piece of software is trustworthy.

Another meaning of the word "trust", however, has to do with the biases within the software. If the same hypothetical loan offering software were to assign those loans such that the 5% who were offered a loan but did not manage to pay it off were white while the 5% of those who were rejected but were worthy of the loan were of some other ethnicity, this software would suddenly look a lot less trustworthy. These two types of trustworthiness are referred to in this work as **Performance** and **Fairness**.

In the context of predictive software, if something has high **performance** it only needs to do a good job at predicting the outcome. All that matters is that, given a particular case, the prediction is as likely to be correct as possible. For something to be **fair** it needs to ensure that protected groups are treated the same way. These protected groups are problem specific, but are often categories like age, sex, and race.

Many works have attempted to identify, explain, and reduce unfairness within software [1–6], but in many cases have found that doing so has an unfortunate side effect of also reducing performance. This is so prevalent that Berk et al. [7] claimed "it is impossible to achieve fairness and high performance simultaneously (except in trivial cases)".

One recent attempt to subvert this claim came from Chakraborty, Majumder, and Menzies in the form of Fair-SMOTE [1], a software package which works almost as a preprocessing step, inspecting the data and strategically eliminating biased labels and reordering internal distribution in order to reduce bias without heavily impacting accuracy. Fair-SMOTE was shown to have good results with several important datasets.

While Fair-SMOTE has been shown to perform well on datasets with slight systemic bias and innate bias from data collection, it has not yet been tested on datasets which have been deliberately tampered with to increase bias.

This type of data is referred to as doctored data. Doctored data is a relevent issue today, as individuals and companies attempt to circumvent laws or push an agenda. As the amount of data and the methods to analyze it increase, it becomes more and more important to have ways to ensure that the methods we use to analyze data are able to withstand some amount of data doctoring.

2 PROBLEM STATEMENT AND OBJECTIVES

This work seeks to begin an evaluation of Fair-SMOTE in the presence of doctored data. This work will focus on one particular type of doctored data, which has been chosen due to the nature of Fair-SMOTE. The datasets which Fair-SMOTE is set up to work with have a few key characteristics:

Conference'17, July 2017, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

(1) These datasets have a single "target" attribute with a favorable and unfavorable outcome. These are typically something similar to "should this person be offered a loan".
(2) These datasets have one or more "protected" attributes, for which Fair-SMOTE tries to reduce bias. These are typically variables such as "race", "age", and "sex", and typically have easy characterizations such as "privileged" and "unprivileged".

The type of doctoring we will be examining works as follows: A hypothetical adversary wishes to push an agenda that a privileged group should have favorable outcomes. They take the dataset in question and inspect it line by line. For each line of data such that the protected attribute is privileged and the target attribute is unfavorable, there is some probability that they change the target attribute to favorable.

The question which we seek to answer then is this: **how robust is Fair-SMOTE with respect to this type of data doctoring?**

To answer this question, we plan to analyze the trends which different metrics of both performance and fairness take when Fair-SMOTE is exposed to higher and higher probabilities of data being doctored.

We will also run our datasets through a naive randomized obfuscator, in order to inspect how Fair-SMOTE compares to a simpler method of bias mitigation under these circumstances.

3 BACKGROUND

3.1 Fair-SMOTE

To understand our experiments, it is important to understand Fair-SMOTE [1]. Fair-SMOTE is a piece of software based on SMOTE [6] and developed by Chakraborty et al. It is used to reduce the bias of machine-learning software working with biased datasets without greatly impacting their performance. Fair-SMOTE acts primarily as a pre-processing step on the dataset itself in order to solve data imbalance.

Fair-SMOTE separates the dataset into groups based on protected and target attributes. If the protected attributes can be defined as privileged or unprivileged and the target attribute can be defined as favorable or unfavorable, Fair-SMOTE separates the dataset into four groups: favorable and privileged, favorable and unprivileged, unfavorable and privileged, and unfavorable and unprivileged. Fair-SMOTE then works to equalize the size of these groups by synthesizing new entries in the smaller three groups until all groups have the same number. It synthesizes these new entries in such a way that they follow the same trends as existing entries. These new entries belong to the same distributions as the previous ones.

We also use a naive random obfuscator here, taken from the work done for xFAIR [5]. This is (and is intended to be) a much less refined piece of software, which merely randomizes the protected attribute in the dataset before giving it to the learner.

3.2 Metrics of Fairness and Performance

There is no one metric which can be used to identify the output of a learner as good.

Many different metrics of performance exist. We evaluate based on four of them here: Accuracy, Precision, Recall, and F1. These are more fully explained in Table 1.

We also evaluate our results for different metrics of fairness. The question of what is "fair" is a large and nontrivial one. When we refer to fairness in this work it is typically to some abstract concept where if the target attribute in the output of a learner is affected by the protected attribute, it is unfair. But this abstract definition is difficult to quantify and examine. Many different metrics for fairness have been proposed and studied [2–4, 8], but we ultimately chose the metrics AOD, EOD, SPD, DI, and FR—which are explained in Table 2—in an attempt to follow the format set by Peng et al. [5]

4 IMPLEMENTATION

4.1 Data

For this experiment, we are going to be looking at four datasets which are publicly available. All of these datasets are easily used for binary classification, all of them have at least one attribute which should be protected, and all of them have been used in studies of fairness before now [1, 3, 5].

These datasets are as follows:

(1) **Adult Census**: 1994 census data used to predict personal income. This dataset has the protected attributes of race and sex, and the target attribute of income. A favorable output would be income greater than $50,000.
(2) **Compas**: Criminal history data used to predict the likelyhood of re-offence. This dataset has the protected attributes of race and sex, and the target attribute of re-offence. A favorable output would be not re-offending.
(3) **German Credit**: Many pieces of personal information used to predict credit rating. This dataset has the protected attribute of sex, and the target attribute of credit. A favorable output would be good credit.
(4) **Bank Marketing**: Marketing data used to predict whether a client will subscribe to a term deposit. This dataset has the protected attribute of age, and a target attribute of subscription. A favorable output would be subscribing.

4.2 Method of doctoring data

This data was doctored probabilistically. Each line where the protected attribute was "privileged" and the target attribute was "unfavorable" had its target attribute changed to "favorable" with some probability P. Each dataset was corrupted eleven times, with $P \in \{0.01, 0.05, 0.1, 0.15, 0.2, 0.25, 0.3, 0.35, 0.4, 0.45, 0.5\}$.

This method of doctoring ensures that some existing connections between other attributes and the target attribute remain in place, though weakened, in a way that would not be true for a dataset which had been entirely fabricated. It also ensures that, if any other connections are changed, they are overpowered by the connection between the protected and target attributes in a way which would not be true if a dataset had been changed according to some larger criteria which encompassed multiple attributes. This method of doctoring is very basic, and therefore a good place to begin testing the robustness of Fair-SMOTE.

4.3 Bias mitigation and learner

This data was subject to two different methods of bias mitigation: Fair-SMOTE and a simple random obfuscator.

Metric	Mathematical Definition	Description	Ideal Value
Accuracy	(TP + TN)/(TP + TN + FP + FN)	Ratio of correct results to total results	1
Precision	TP/(TP + FP)	Ratio of true positive results with total positive results	1
Recall	TP/(TP + FN)	Ratio of true positives to true positives and false negatives, also called the true positive rate.	1
F1	2*(Precision*Recall)/(Precision + Recall)	Harmonic mean of Precision and Recall	1

Table 1: Description of performance metrics used in these experiments. Metrics rely on True Positive (TP), False Positive (FP), True Negative (TN), and False Negative (FN).

Metric	Mathematical Definition	Description	Ideal Value
Average Odds Difference (AOD)	$((FPR_U - FPR_P) + (TPR_U - TPR_P))/2$	Average of differences of true and false positive rates for privileged and unprivileged groups	0
Equal Opportunity Difference (EOD)	$TPR_U - TPR_P$	Difference of true positive rates for privileged and unpriveleged groups	0
Statistical Parity Difference (SPD)	$P[Y = 1\|PA = 0] - P[Y = 1\|PA = 1]$	Difference of probability that unprivileged group has a favorable prediction and privileged group has a favorable prediction	0
Disparate Impact (DI)	$P[Y = 1\|PA = 0]/P[Y = 1\|PA = 1]$	Ratio of probability that unprivileged group has a favorable prediction and privileged group has a favorable prediction	1
Flip Rate (FR)		Ratio of outcomes where the prediction will change if the protected attribute is flipped from privileged to unprivileged or vice versa	0

Table 2: Description of performance metrics used in these experiments. Metrics rely on True Positive Rates (or Recall) for Privileged (TPR_P) and Unprivileged (TPR_U) groups and False Positive Rates (FP/(FP + TN)) for Privileged (FPR_P) and Unprivileged (FPR_U) groups. Some metrics also rely on probabilities related to the privileged group (PA=1), unprivileged group (PA=0), favorable prediction (Y=1) and unfavorable prediction (Y=0).

Fair-SMOTE attempts to balance groups as described in section 3.1 before sending it along to the learner. Because this method of data doctoring increases the size of the privileged-favorable group by decreasing the size of the privileged-unfavorable group, high probabilities of corruption lead to the creation of many more entries into the dataset. Fair-SMOTE attempts to keep the same distributions for each group as it adds entries, so the only group where the distribution is different from the uncorrupted dataset is the privileged-favorable group. This change is likely to make predictions more difficult as it disrupts the connections between other attributes and the target attribute.

The random obfuscator reassigns a random value to the protected attribute before sending it along to the learner. This method of bias mitigation would be excellently placed to counteract this type of data corruption if there were no connections between the protected attribute and the other attributes, or if the protected attribute had been changed in the corruption rather than the target attribute.

After going through a mitigator, the new datasets are sent sent to a simple random forest classifier.

4.4 Other parameters

For the training and testing, each dataset is randomly separated into two subsets: a training set, which contains 80% of the data, and a testing set, which contains the other 20%. The learner is trained on the training set and tested on the testing set, where its output target attributes are compared to the original target attributes. This comparison testing gives us are true positives, false positives, true negatives, and false negatives.

Each combination of dataset, corruption probability, and mitigator is run 20 times, with a different testing set chosen each time. It is important to note that the corruption probability *is not run each time.* **The dataset is corrupted once, before being separated into the testing and training sets.** This means that in the end, the output will be compared to corrupted results, making it harder to tell whether the results are the way they would be with new subjects. Running the experiments in this way, then, simulates the experience of testing the software on already doctored datasets, rather than training the model on a corrupted dataset and using it in the real world.

The datasets Adult Census and Compas each have two protected attributes, race and sex. In both of these cases we only doctored the data with respect to sex, but we tested mitigating the datasets for each protected attribute separately. This allows us to also note the effect of data corruption in situations where it is not directly being mitigated in any way.

We took our basic structure directly from Peng et al. [5].

5 RESULTS

5.1 Fair-SMOTE Direct Mitigation

We first inspect the results for our experiments using Fair-SMOTE and mitigating based on the corrupted attribute.

Immediately upon inspecting our data, an unexpected trend appears. We had expected every metric to either get worse, as the dataset is compromised by the doctored data creating false connections, or to show no particular trends. The one thing which we were not expecting was for some of the metrics to actually improve with an increasing corruption probability, especially those metrics related to fairness.

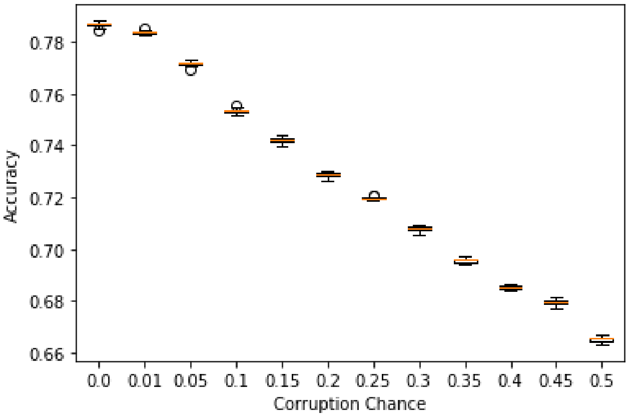

Figure 1: Adult Census accuracy mitigating for sex as a function of corruption probability. Accuracy is ideally 1, so this shows that it gets worse as corruption increases.

Looking at each metric individually:

(1) **Accuracy:** Accuracy is based entirely on the number of correct predictions with relation to the total number of predictions. It acts more or less acted as expected, getting worse with data corruption. Figure 1 shows the trend for the Adult Census dataset, with the other datasets either following this trend or not significantly changing.

(2) **Precision:** Precision is based only on true positives and false positives. Precision acts nearly the opposite of what was expected; as shown for the Adult Census dataset in Figure 2, it may increase as the corruption increases. The other datasets either followed this trend or did not significantly change.
One possible reason why this might be the case is that there are simply more positive results in the testing dataset. Because precision is based on positive results, as shown in Table 1, if more entries are have a positive target value, there is a greater chance for having a true positive result and a smaller chance of having a false positive result.

(3) **Recall:** Recall, which looks at true positives and false negatives, is one of the least intuitive results here. Recall sometimes shows trends within datasets but not between datasets, as shown in Figures 3 and 4. In each of these figures there is a clear trend, with recall improving as corruption increases for the Compas dataset, but worsening for the Adult Census dataset. Furthermore, for the Bank Marketing dataset, recall initially worsens as corruption increases until the chance of corruption reaches about .25, after which it improves. This can be seen in figure 5.

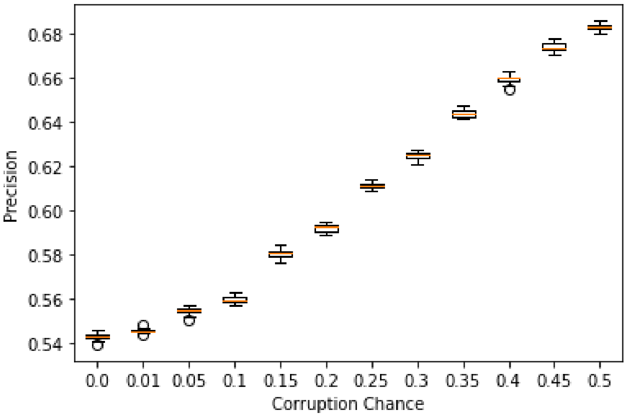

Figure 2: Adult Census precision mitigating for sex as a function of corruption probability. Precision is ideally 1, so this shows that it gets better as corruption increases.

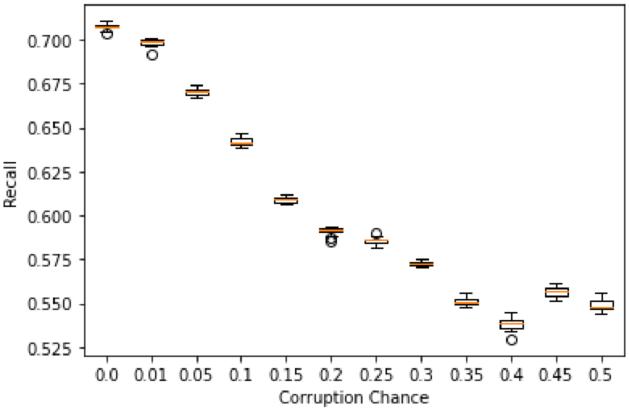

Figure 3: Adult Census recall mitigating for sex as a function of corruption probability. Recall is ideally 1, so this shows that it gets worse as corruption increases.

It is still unclear why recall behaves in this way. The number of true positives increases or stays the roughly the same, as seen in the inspection of precision, so why does the number of false negatives act so differently between datasets?

(4) **F1:** F1, which is a function of precision and recall, shows no significant trends in any of the datasets except the Bank Marketing dataset, where it follows the same trend seen in its recall in Figure 5. This is not unexpected; precision and recall for the Adult Census dataset go in opposite directions, and at least one of the two for all other datasets shows no significant trend.

(5) **Average Odds Difference:** AOD, based on true and false positive rates for privileged and unprivileged groups respectively, shows another unexpected result. Two of the datasets, Compas and German Credit, show no significant

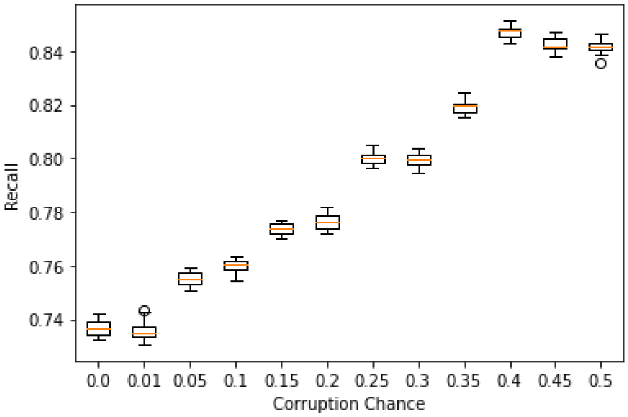

Figure 4: Compas recall mitigating for sex as a function of corruption probability. Recall is ideally 1, so this shows that it gets better as corruption increases.

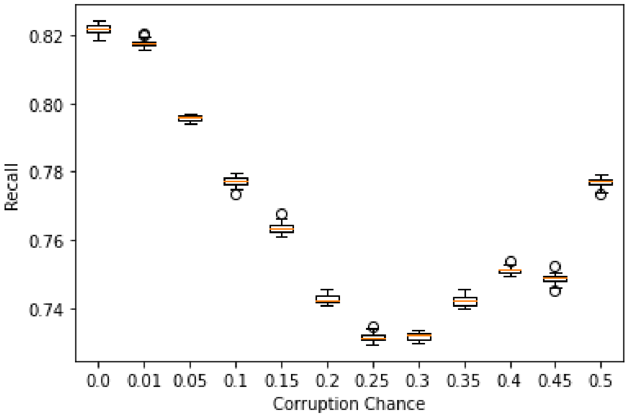

Figure 5: Bank Marketing recall mitigating for age as a function of corruption probability. Recall is ideally 1, so this shows that it gets worse and then better as corruption increases.

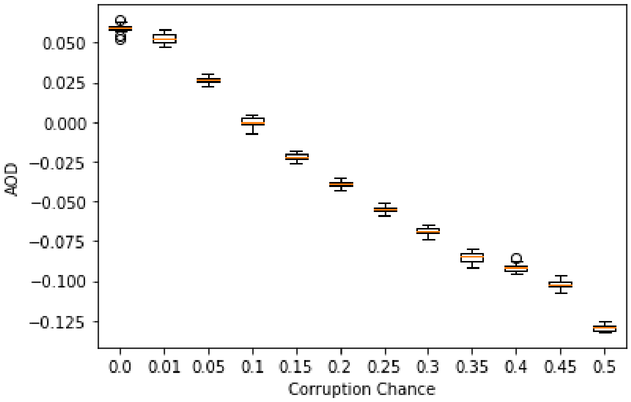

Figure 6: Adult Census AOD mitigating for sex as a function of corruption probability. AOD is ideally 0, so this shows that it gets worse as corruption increases.

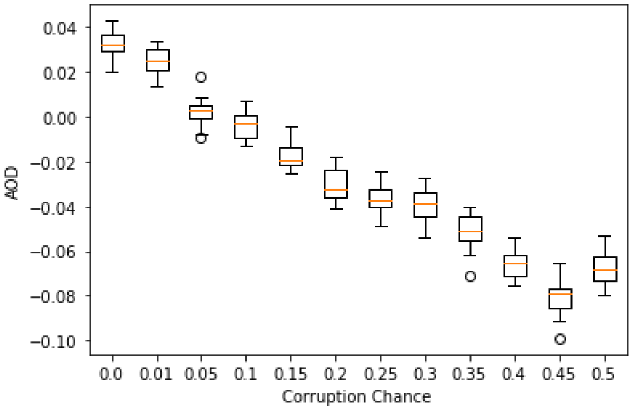

Figure 7: Bank Marketing AOD mitigating for age as a function of corruption probability. AOD is ideally 0, so this shows that it gets better and then worse as corruption increases.

trends for AOD with respect to corruption chance. The other two, Bank Marketing and Adult Census, show strange behavior. In both cases, the Adult Census dataset shown in Figure 6 and the Bank Marketing Dataset shown in Figure 7, the AOD first improves until it reaches an ideal level, after which it worsens once more. For both datasets, this inversion happens at a low probability. For the Adult Census data, this cross happens at about a .1 probability of corruption, while with the Bank Marketing dataset it happens at about .05.

As yet, we have no hypothesis for why this would happen. We are especially interested in why the Bank Marketing dataset recall and AOD act in this manner.

(6) **Equal Opportunity Difference:** EOD, once again based on the true positive rate, or recall, of the different groups, shows similar trends to AOD. Interestingly, while EOD decreases for both the Adult Census dataset and the Bank Marketing dataset, the EOD for the Adult Census dataset begins as a positive number and therefore improves as corruption increases, while the EOD for the Bank Marketing dataset starts negative and therefore worsens as corruption increases.

This implies that either the true positive rate for the privileged group specifically increases with corruption or that the true positive rate for the unprivileged group specifically decreases when corruption increases.

Perhaps this is explained with the same possible explanation as that for precision: there are simply more positive outcomes for the privileged group in the testing set. If this is the case, it is interesting that Recall, which is the true positive rate for the entire dataset, does not reflect it.

(7) **Statistical Parity Difference:** SPD is based on the probabilities of favorable predictions for the different groups. In the absence of bias mitigation this is expected to decrease as corruption increases, as the probability for the privileged group to receive a favorable prediction is expected to increase.

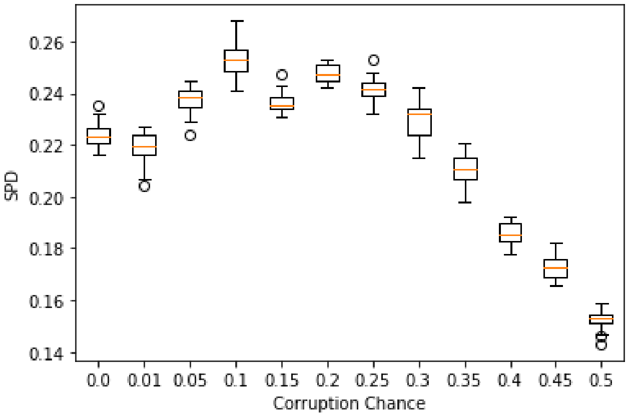

Figure 8: Bank Marketing SPD mitigating for age as a function of corruption probability. SPD is ideally 0, so this shows that it gets better and then worse as corruption increases.

Side-note: unless I'm missing something here, shouldn't a positive SPD mean that the unprivileged group is more likely to have a good outcome than the privileged group? In both the Fair-SMOTE paper and the xFAIR paper SPD is presented without any absolute value operation, but is always positive. Yet, isn't the whole meaning of a group being 'privileged' that they are more likely to receive a favorable outcome? Maybe the groups are mislabeled? But DI in those papers acts how it would if SPD were swapped. Other sources I've found have the different groups swapped or an absolute value there, but it seems like quite a large difference whether it's the other way around or whether it's absolute value. I haven't found a single negative SPD value, so I think it's absolute value, but until I understand the metric, there's not a whole lot of analysis I feel like I should do.

In the Adult Census dataset SPD rises steadily. In the Bank Marketing dataset, the trend switches once again, this time at a corruption probability of about .2, as shown in Figure 8.

(8) **Disparate Impact:** DI is very similar to SPD, but is a ratio rather than a difference. we also noticed a problem with our data here: with the Adult Census dataset, the SPD was positive, but the DI was less than 1. This has caused us to doubt either the veracity of the calculations or of our understanding of these two metrics.

That said, once more there is a disagreement between Adult Census and Bank Marketing Data, with the other two datasets once again showing no trends. The Adult Census dataset, as shown in Figure 9, improves as more corruption is introduced, while the Bank Marketing dataset, shown in Figure 10, follows an expected trend of worsening as the corruption gets especially bad.

It is very unclear here why this corruption should increase the probability of the privileged group getting a favorable outcome in one dataset, while decreasing that probability in another.

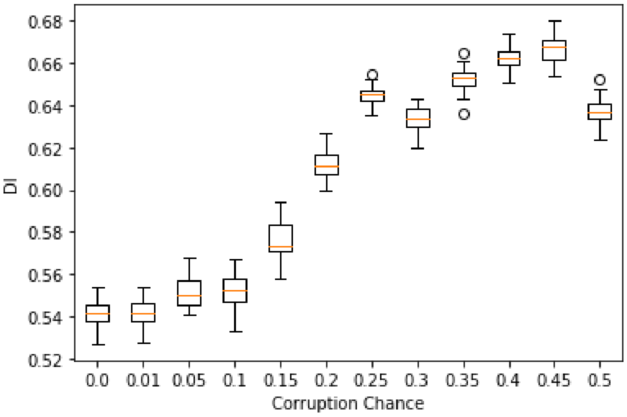

Figure 9: Adult Census DI mitigating for sex as a function of corruption probability. DI is ideally 1, so this shows that it gets better as corruption increases.

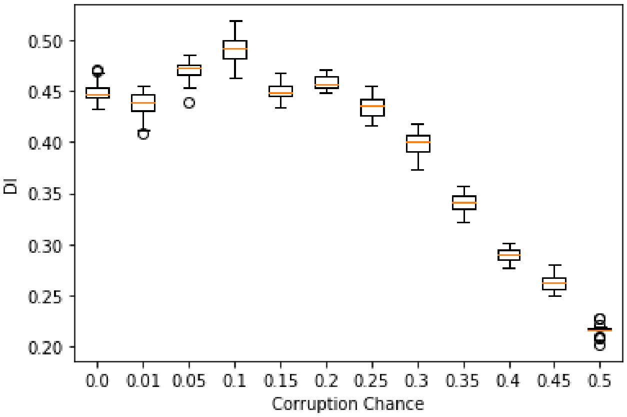

Figure 10: Bank Marketing DI mitigating for age as a function of corruption probability. DI is ideally 1, so this wavers for a bit before worsening dramatically.

(9) **Flip Rate:** FR is a metric unlike any of the others. FR is, quite simply, measuring whether changing the protected attribute of an entry would change its prediction, and therefore is agnostic toward the target attributes of the testing set. If the FR increases with corruption, then, that means that the models are becoming more biased.

FR shows the trends which we expected; it either worsens or stays about the same as corruption increases. This is a clear indication that, in one way at least, the models become less fair as the corruption increases. **Fair-SMOTE is not able to entirely make up for the corruption caused by doctoring data in this way.**

5.2 Fair-SMOTE Indirect Mitigation

Next we investigate what happens when we attempt to mitigate for a protected attribute that is not the one which was not doctored. We only looked at two datasets for this, the Adult Census dataset and the Compas dataset. In both cases the datasets were doctored with respect to sex, but we attempted to mitigate for race.

Under these conditions, we expected to see performance dip as the datasets began to be corrupted, but then increase rapidly as the datasets were more completely corrupted. We expected the connection of the target attribute with sex to increase, at first interfering with but later drowning out the other attributes. We did not expect this to have a great effect on the fairness metrics, as we were judging fairness only with respect to race.

The Compas dataset laregly showed no trends, and where it did the differences in value were too small to be significant.

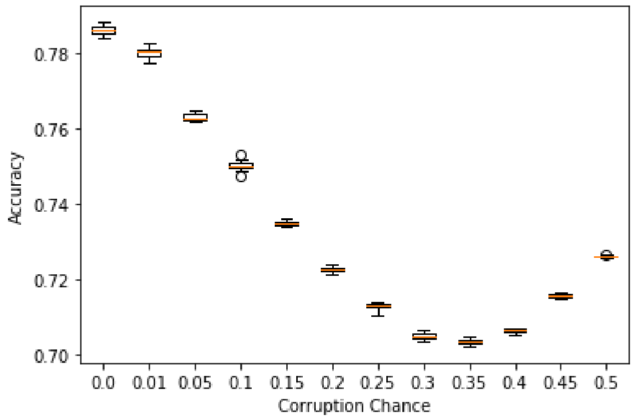

Figure 11: Adult Census accuracy mitigating for race as a function of corruption probability. Note that the corruption was based on sex rather than race. Accuracy is ideally 1, so this shows that it initially gets worse, but starts to get better as corruption increases.

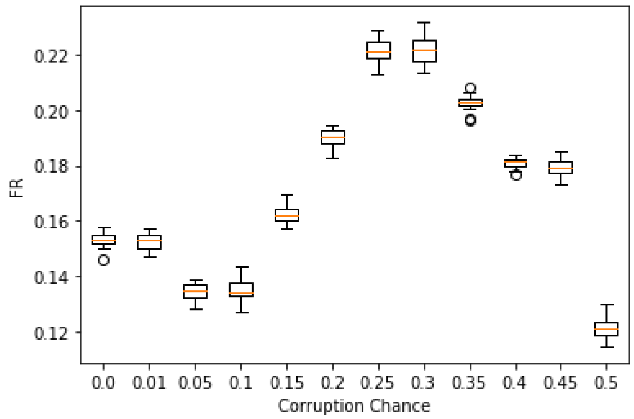

Figure 12: Adult Census FR mitigating for race as a function of corruption probability. Note that the corruption was based on sex rather than race. FR is ideally 0. This shows no clear trend, but still suggests some kind of structure.

The Adult Census dataset largely followed our expectations, with many of the performance metrics initially dipping but all eventually trending positively with the corruption, and the fairness metrics all showing no clear trends.

Figures 11 and 12 are good examples of this. A few of the fairness metrics, such as FR shown here, suggest that there might be some kind underlying structure. It may be worth investigating more closely.

5.3 Comparisons with Random Obfuscator

We ran our datasets through a naive random obfuscator as a method of bias mitigation as well as through Fair-SMOTE. The random obfuscator, hereafter referred to as Random, randomly assigns each entry a new value for its protected attribute in an attempt to decrease bias. In many instances the trends seen by the two different mitigators agree, with perhaps a slight difference in rate, but there were some notable differences as well.

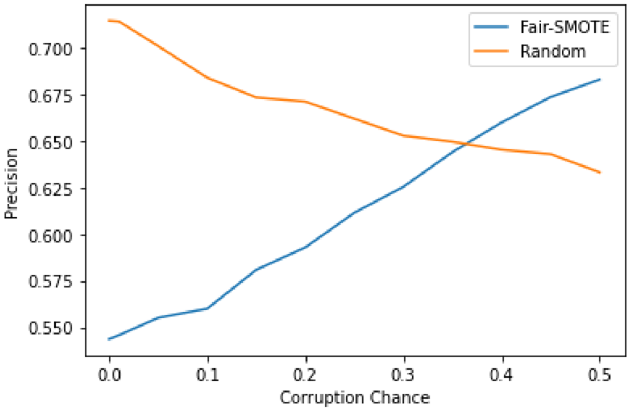

Figure 13: Adult Census precision mitigating for sex as a function of corruption probability. Lines show the median values over the 20 trials. Precision is ideally 1. This shows that Fair-SMOTE and Random react differently to corruption.

The most notable difference is in the precision ratings for the Adult Census dataset, shown in Figure 13. In this Figure, it can be seen that the precision after Random decreases slightly as the corruption increases, moving in the opposite direction from the same dataset after being treated by Fair-SMOTE. While both are very small changes in precision, the presence of additional favorable predictions in the testing set seems to harm the performance after being treated by Random, while aiding that of Fair-SMOTE. Similar trends were seen in other datasets, but the changes were even less significant than that of the Adult Census dataset.

Another disagreement, shown in Figure 14, is that of DI. Here, Fair-SMOTE causes DI to improve with increasing Corruption, while Random causes it to worsen.

For both precision and DI, the learners fed the datasets from Random followed our original expected trend, of both worsening fairness and performance, while those fed the datasets from Fair-SMOTE managed to improve with corruption. This suggests that, in some ways, Fair-SMOTE is better able to deal with corruption than Random.

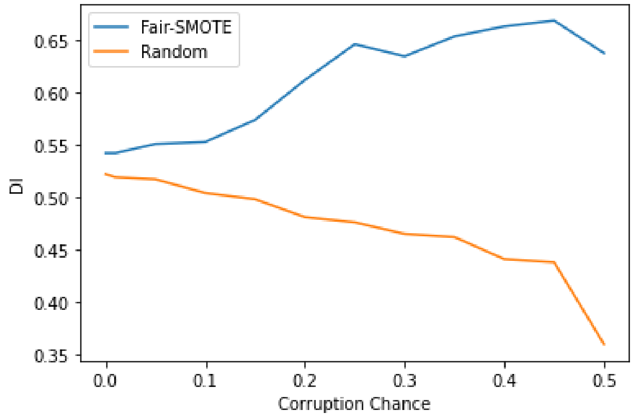

Figure 14: Adult Census DI mitigating for sex as a function of corruption probability. Lines show the median values over the 20 trials. DI is ideally 1. This shows that Fair-SMOTE and Random react differently to corruption.

6 LIMITATIONS

6.1 Future Work

The largest hole in our data is that of comparing methods of bias mitigation with the learner alone. In the future we would like to run these tests again through the random forest learner without first attempting to mitigate the problems, in order to compare and evaluate how Fair-SMOTE does when compared to nothing.

In addition, we would like to try corrupting the data after separating out the testing set from the original dataset. This would require a more thorough understanding of Fair-SMOTE and the workflow which we adapted from Peng et al. [5] This would allow us to better evaluate the effects which data doctoring would have should its results ever be used to teach learners in the real world. We expect worse performance and fairness across the board under these conditions.

We would also like to experiment with different datasets and different learners. These experiments were quite narrow, only using four datasets, two of which showed almost no trends, and a single learner. We expect it would be enlightening if we were able to expand our experiments in this way.

We would also like to experiment with different methods of bias mitigation. Fair-SMOTE has shown excellent results under normal conditions, but it is possible that different mitigators would have different resistances to different methods of corruption.

We would also love to experiment more with different methods of corruption. Especially if we were comparing different mitigators, this might help shed some light on how one might design an adversary-resistant bias mitigator.

6.2 Threats to Validity

There were a few concerns that arose during this study.

One obvious problem was the scope of the study; looking at so few datasets and a single learner opens up the possibility of sampling bias.

There were other, less concrete signs of possible problems in this study. One was in our understanding of SPD and DI, which was explained in section 5.1. Another sign was that in the uncorrupted runs of our study, Fair-SMOTE had a lower score on most fairness metrics than Random did. While this is also present in the xFAIR study [5], which uses nearly the same workflow and setup, it is not nearly as pronounced there as it is here. A third sign was that, in adapting the workflow of xFAIR for use here, several problems arose which prevented us from using some datasets and mitigators used there. It is possible that there was something wrong with our setup which may have impacted the data we were able to collect.

7 CONCLUSION

In general, this work asks as many questions as it answers. When corrupting data by changing varying numbers of privileged-unfavorable entries into privileged-favorable entries we see many trends across different types of metrics, but few of them are conclusive. Some of the trends, while clear within datasets, do not hold accross different datasets. Some of these even show conflicting trends. Some of the trends act in ways directly contrary to our expectations, notably precision, which improved as the datasets were altered. This method of data doctoring, while simple to perform, has complicated effects on the end results of bias mitigation and machine learning.

While many of these issues do not have answers, one solid conclusion can be made: As shown by flip rate, Fair-SMOTE cannot entirely overcome the effects of this type of doctored data.

REFERENCES

[1] J. Chakraborty, S. Majumder, and T. Menzies, "Bias in machine learning software: Why? how? what to do?" in *Proceedings of the 29th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering*, ser. ESEC/FSE 2021. New York, NY, USA: Association for Computing Machinery, 2021, p. 429–440. [Online]. Available: https://doi.org/10.1145/3468264.3468537

[2] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, "A survey on bias and fairness in machine learning," 2019.

[3] S. Caton and C. Haas, "Fairness in machine learning: A survey," 2020.

[4] D. Xu, S. Yuan, L. Zhang, and X. Wu, "Fairgan: Fairness-aware generative adversarial networks," 2018.

[5] K. Peng, J. Chakraborty, and T. Menzies, "xfair: Better fairness via model-based rebalancing of protected attributes," 2021.

[6] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: Synthetic minority over-sampling technique," *Journal of Artificial Intelligence Research*, vol. 16, p. 321–357, Jun 2002. [Online]. Available: http://dx.doi.org/10.1613/jair.953

[7] R. Berk, H. Heidari, S. Jabbari, M. Kearns, and A. Roth, "Fairness in criminal justice risk assessments: The state of the art," *Sociological Methods & Research*, vol. 50, no. 1, pp. 3–44, 2018. [Online]. Available: https://doi.org/10.1177/0049124118782533

[8] "Ai fairness 360: an extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias." [Online]. Available: https://github.com/IBM/AIF360