Exploring Fairness-Aware Machine Learning for Binary Classification Tasks

Selin Jessa
Computer Science and Biology
selin.jessa@mail.mcgill.ca

Adam Cavatassi
Integrated Circuits and Systems
adam.cavatassi@mail.mcgill.ca

Elsa Riachi
Electrical and Computer Engineering
elsa.riachi@mail.mcgill.ca

***Abstract*—Machine learning is being deployed to assist in decision making in many systems with enormous impact on people's lives. As the role of machine learning in society expands, as does a need for incorporating ethical concerns, including fairness. Here, we explore how discrimination can be reduced in the binary machine learning classification setting. We implement three methods for data transformation which reduce discrimination, a modified Naive Bayes classifier, and a fairness-constrained logistic regression model. We compare the performance of these models on two data sets, in terms of both accuracy and three ratio-based measures of discrimination. We find that while all the methods we explored successfully reduced discrimination, the results were notably dependent on the particular data set examined. We outline the limitations of these methods and propose future directions of research into fairness-aware machine learning. The data, code, and analysis for our work is available online at https://github.com/selinj/fair-ML Where appropriate, we have included links to our analysis within the text. A video abstract of our work can be viewed at https://youtu.be/aW-8Kygooew.**

I. Introduction

Machine learning is increasingly being incorporated into decision-making processes in a range of systems with enormous societal impact. Artificial intelligence is being used for jury selection and predictive policing, to decide who gets hired and who gets bail, and to distribute loans and make diagnoses. Yet, human decisions in these arenas reflect structural biases ingrained in human behaviour, and many researchers have sounded the alarm that when these models learn from historical data, they may also be learning and reinforcing these biases. As the role of machine learning in our social institutions grows, a critical component of the safe and responsible practice of machine learning is to develop models which are both accurate *and* fair. These issues motivate our study of leverage points for reducing discrimination in machine learning.

Research in the area of discrimination in machine learning has studied sources of discrimination, the discovery and quantification of discrimination, and methods for fair machine learning. A key finding has been that learned models can exhibit bias without any malicious intent on the part of the developer. Evidently, bias can arise when training data reflects a discriminatory decision or process [1], but more saliently, bias can arise even when sensitive features (*e.g* gender or race) are removed from training data because the underlying processes that we attempt to model often do not satisfy the assumptions implicit in the learning algorithm [2].

Consequently, researchers have explored methods of discovering discrimination in machine learning predictions and quantifying levels of discrimination. Some methods are constructed around a comparison between classification probabilities for different groups, and consider only the sensitive attribute and prediction in a data set. Žliobaitė [3] surveyed measures of discrimination applicable to the binary classification setting, with a focus on statistical tests for presence of discrimination and absolute measures which quantify discrimination. Another class of discrimination discovery techniques involve learning from the database. Bonchi et al. [4] explore a probabilistic approach for uncovering a network of causal relationships between features and show that traversing this network can allow quantification of various types of discrimination as defined in the legal literature; others have studied methods for discovering discrimination by using Bayesian networks [5], investigating classification rules learned by a model [6], and employing variants of k-nearest neighbours to measure disparate impact on similar examples [7].

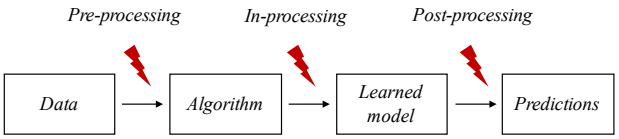

Fig. 1. The typical machine learning process gives natural points of intervention for reducing discrimination.

Machine learning generally involves a process in which *data* is input to an *algorithm* resulting in a *learned model* which can be applied to new data to make *predictions* (Figure 1). Algorithmic methods for reducing discrimination in machine learning, known as *fairness-aware data mining*, can then be naturally grouped by the step in the process at which they intervene. In pre-processing approaches, training data is transformed to reduce discrimination. In-processing approaches incorporate fairness or non-discrimination constraints into the learner. Post-processing approaches learn a model on unaltered data and then apply some modification such that the predictions on new data will be fair.

Here, we explore how discrimination can be reduced in the binary machine learning classification setting. This paper is

organized as follows: first, in Section II, we clarify definitions, define the problem, and describe the data sets we use. Next, in Section III, we describe our experimental methodology including data pre-processing, implementations of methods for discrimination reduction, and implementations of measures of discrimination. We then present the results of our analyses in Section IV, and finally, in Section V, we contextualize the results, discuss limitations, and propose future directions for research in fairness-aware machine learning.

II. Problem definition

A. Notions of fairness

First, it is critical to define discrimination and how it differs from bias. Bias is an imbalance in a data set or algorithm which causes learners to overfit the training data. Bias is a common occurrence in many machine learning applications and it affects prediction accuracy for test sets. However, discrimination can be considered a form of bias, which causes protected features to influence the prediction of a learning algorithm. Specifically, a protected feature can be designated as any feature which should not affect the expected value of the prediction. Hence, we consider a model fair if it predicts y such that $E(y|X,S) = E(y|X)$, where S is the protected characteristic (*e.g.* gender or race, which give the grounds of discrimination) and X represents all non-protected features. In order to study discrimination in machine learning, besides a protected feature, a given task needs also needs polar outcome (*i.e.* whether the decision the outcome is binary, categorical, or continuous, some outcome(s) must be preferred). We use the terms "sensitive characteristic", "protected feature", "protected attribute", *etc.*, interchangeably.

B. Investigating fairness in the binary classification setting

Our work explores whether fairness-aware machine learning techniques which are simple to implement are applicable in real-world scenarios where discrimination may be present. In particular, we address the following problem: how do various methods for reduction of discrimination in machine learning compare in terms of accuracy and fairness? In order to answer this question, we implement and characterize six discrimination-reduction methods by studying their performance and quantifying how they reduce discrimination in a machine learning task by three measures. Here we focus on the binary classification setting with one binary sensitive characteristic. This contributes to a foundation for future investigation into discrimination in more complex learning scenarios, but tasks with binary sensitive features (*e.g.* gender) and binary outcomes (*e.g.* hiring or loan decisions) are ubiquitous, and therefore important to study in their own right.

C. Description of data [data]

We apply our methods to two different data sets, referred to in our work as "Crime" and "Recidivism". The *Communities and Crime* data from the UCI Machine Learning Repository [8] aggregates socio-economic, law enforcement, and crime data (totalling 128 features) about 1994 U.S. communities for a regression task involving prediction of the number of violent crimes per capita. This data has been normalized feature-wise. The 3-year recidivism data set was obtained from the United States DATA.GOV repository by the State of Iowa. It describes information about race, sex, and initial offenses for 17061 offenders and whether or not they recidivated within three years of their release. The pre-processing of all data (distinct from discrimination-reduction methods which use pre-processing as their mechanism) is described in the next section.

III. Methodology

To briefly summarize our work, we obtain two data sets do the necessary tidying and feature selection. We implement several methods for reducing discrimination and measures of discrimination, and we compare each method applied to each data set by evaluating discrimination by each measure. This experimental setup is outlined in Figure 2. All models were evaluated using 10-fold cross-validation.

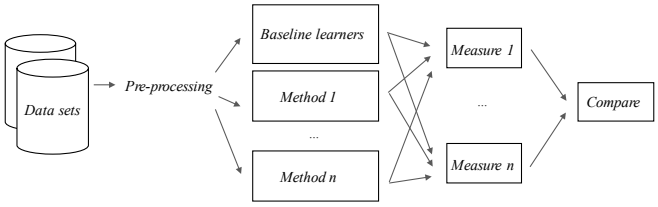

Fig. 2. Schematic outline of our experiments.

A. Data pre-processing

This section describes how we have processed raw data and selected features in order to obtain data sets suitable for training models.

1) Crime data set: For the Crime data set, we follow exactly the method described by [9] and binarize the outcome on the feature ViolentCrimesPerPop using a threshold of 0.375, and a sensitive feature black which describes whether or not the population of a community is predominantly Black using the feature racepctblack and a threshold of 0.06 [code]. We retained only numerical features.

2) Recidivism data set: The data was completely categorical and required mild modification in order to become useable for machine learning purposes [code 1, 2]. The sex category was made to be binary, while the race category was extrapolated into binary non-white and hispanic features. This allowed for any one of these three features to be analyzed as protected prisoner characteristics for fair learners. The rest of the features were adjusted to normalized numerical features. Each can take the value of a discrete fraction to represent 1 of N possible categories, where N is the number of unique categories in a given feature. This brief pre-processing of the Recidivism data set allowed us to apply our best fair machine learning methods, based on the performances on the Crime data set. The focus of our project was to explore methods of

discrimination-free machine learning, and so this basic pre-processing is considered to be adequate for the purposes of our study.

B. Measures of Discrimination

In order to determine whether given techniques for reducing discrimination are effective, we need to designate a few meaningful metrics. Although there are many ways to statistically analyze machine learning predictions, we decided to focus on three measures described by Žliobaitė [3] in their survey: impact ratio, elift ratio, and odds ratio, which we implemented in both python [code] and R [code] with parameterized scripts and functions. We have chosen these measures because ratio-based measures allow quantification of the level of discrimination, making them useful for analyzing how well particular method reduces discrimination, and because they are suitable for binary classification tasks with a binary protected feature.

Impact ratio is a measure of the probability of a positive outcome for a protected group divided by the probability of a positive outcome for the general group. In the case of race, it can outline how many positive outcomes for black people occur in relation to how many positive outcomes occur for white people. Ideally, these numbers should be close to equal if there is no discrimination. This metric is interesting because it is used by US courts to develop crime statistics and detect discrimination in criminal sentences.

$$r_{impact} = \frac{p(y^+|s^1)}{p(y^+|s^0)} \tag{1}$$

Elift ratio is a measure of the probability of a positive outcome for a protected group divided by the total probability of positive outcomes. It can portray the likelihood of a positive outcome is for a protected group in relation to all groups. Returning to the example of race, the proportion of positive outcomes for black people should be equal to the proportion of total positive outcomes if no discrimination has taken place.

$$r_{elift} = \frac{p(y^+|s^1)}{p(y^+)} \tag{2}$$

Lastly, odds ratio can be used to quantify the association between a protected group and a certain outcome. Odds ratio is a common notation which is used in logistic regression calculations for estimating relationships between features and a prediction. This propery makes it valuable since it is the most sensitive to discrimination of all three ratios. This measurement can then be employed to characterize the relationship between race and a favourable outcome.

$$r_{odds} = \frac{p(y^+|s^0)p(y^-|s^1)}{p(y^+|s^1)p(y^-|s^0)} \tag{3}$$

When no discrimination is present, all of these ratios should approach 1. As discrimination increases, these ratios begin to drop. Maximum discrimination occurs when the ratios are 0, and reverse discrimination begins to occur as the ratio exceeds 1. These ratios are only valid when the favourable outcome is known. In some data spaces, a 0 may be considered a favourable outcome. Our scripts are written such that the user can designate whether a 1 or 0 should be treated as favourable when computing these ratios.

With any machine learning, accuracy is always of highest priority. It is then critical to analyze the performance setbacks induced by incorporating discrimination reduction techniques in fair machine learning. There will often be a trade-off between maximizing accuracy and minimizing discrimination. For clarity, we note that each of these measures can be applied to measure discrimination in a labelled data set and in a data set paired with its predictions by some model.

C. Methods for reducing discrimination in machine learning

1) Pre-processing: One class of methods for fairness-aware machine learning arises from the notion that discrimination arises due to bias in historical data. These methods attempt to rectify the situation by manipulating the training data. Kamiran and Calders [9] propose several pre-processing methods, and we implement three [code].

In the *massaging* technique, we flip the class labels of M pairs of examples, where the value of M is chosen specifically to make the labelled data set non-discriminatory. The pairs of examples to flip are chosen by learning a ranker R on the training data, and then choosing M examples from the protected group with the unfavoured class to "promote" by choosing from among them the M examples ranked lowest by R, and choosing M examples from the non-protected group with the favoured class to "demote" by choosing from among them the M examples ranked highest by R. For R, we use a Gaussian Naive Bayes Classifier implemented in the scikit-learn library [10], and examples are ranked by the class probability for the favourable outcome.

The *reweighting* method adds a new feature of weights to the data. A weight is calculated for each combination of sensitive attribute and class, and is given by given by the expected probability of an example being in a certain group and class if sensitive attribute/class are independent divided by the observed probability.

Finally, in the *uniform sampling* technique, we compute the number of examples needed in each protected group/class combination which would make the data non-discriminatory, and then construct the data set by uniformly sampling with replacement from each quadrant of the data as required.

After pre-processing each data set using each of these techniques to generate "fair" data, we used them as input for three baseline learners: logistic regression with $L2$ regularization (logr), linear SVM (svm), and Gaussian Naive Bayes (gnb), all using scikit-learn implementations. Having observed that the Recidivism data set did not appear to be discriminatory (Figure 7), we did not train baseline learners on this data.

2) Constraints on linear classifiers: A common measure of discrimination supported by the U.S Equal Employment Opportunity Commision is the $p\%$ rule, which resembles the impact ratio. The $p\%$ rule sets a lower bound on the ratio of the percentage of subjects having a sensitive attribute assigned

a favourable outcome to the percentage of subjects not having that sensitive attribute also assigned the favourable outcome [11].

A decision boundary classifier is said to follow the $p\%$ rule if:

$$min(\frac{P(d_\theta(x) \geq 0|z=1)}{P(d_\theta(x) \geq 0|z=0)}, \frac{P(d_\theta(x) \geq 0|z=0)}{P(d_\theta(x) \geq 0|z=1)}) \geq \frac{p}{100} \tag{4}$$

Where $P(d_\theta(x)$ is the distance from sample point x_i to the decision boundary, and z is the sensitive feature. Equation 4 is difficult to incorporate in a convex classifier since it is non-convex with respect to the classifier weight vector θ. For this reason, a different but closely related constraint can be added to a decision boundary classifier to ensure that it follows the $p\%$ rule.

The decision boundary covariance measures the dependence of the signed distance from the decision boundary on the sensitive feature. The decision boundary covariance can be computed as $\frac{1}{N}\sum_{i=1}^{N}(z_i - \bar{z})d_\theta(x_i)$. A low covariance indicates low dependence and therefore less discrimination. Convex decision boundary classifiers can be formulated with fairness constraints as shown below:

$$min\ L(\theta) subject\ to \tag{5}$$

$$\frac{1}{N}\sum_{i=1}^{N}(z_i - \bar{z})d_\theta(x_i) \leq c \tag{6}$$

$$\frac{1}{N}\sum_{i=1}^{N}(z_i - \bar{z})d_\theta(x_i) - c \tag{7}$$

Where the variable c, is the upper bound on the decision boundary covariance. We have implemented fairness-constrained logistic regression on the crime data set with the scipy optimize package [12], using the log-likelihood as the objective function[code]. Predictions were obtained for the entire data set using 10-fold cross validation, where the model was trained on 9 partitions and predicted on the tenth partition. The upper-bound c on the covariance was varied between 0.1 and 1. The ROC curves for the different c values are shown in figure 3.

Additionally, we implemented accuracy constrained logistic regression. This is a convex optimization problem where the objective is to minimize discrimination, subject to accuracy constraints as shown below.

$$min|\frac{1}{N}\sum_{i=1}^{N}(z_i - \bar{z})d_\theta(x_i)| \tag{8}$$

$$subject\ to L(\theta) \leq (1+\gamma)L(\theta^*) \tag{9}$$

Where γ is the percentage of additional loss we are willing to incur in order to minimize the decision boundary covariance, and $L(\theta^*0$ is the log-loss of unconstrained logistic regression. Figure 4 shows the ROC curves for values of gamma between 0.1 and 1.

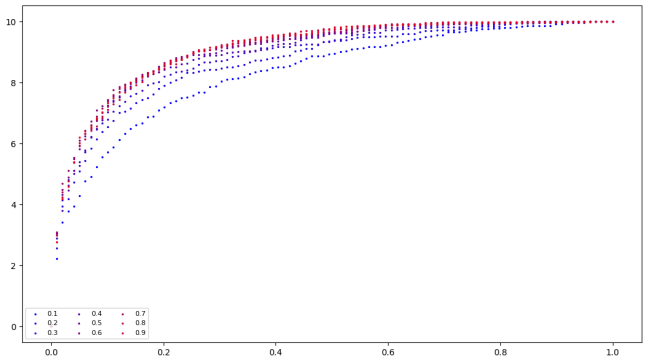

Fig. 3. ROC curve for fairness constrained logistic regression for various C values. The area under the curve decreases as the upper bound on the covariance decreases.

3) Training separate models: Calders and Verwer [13] propose three variants of Naive Bayes classifiers for reducing discrimination. Their work characterizes the performance of each method, and we implemented their best method, *2 Naive Bayes*. In this technique, which can be thought of as an ensemble learner, we partition the data set according to the value of the protected attribute, and train a separate classifier for each. Then, to predict on a new example, we simply use the classifier which corresponds to which group for the protected attribute it belongs to. We implemented this using Gaussian Naive Bayes classifier implementations from scikit-learn [code]. In effect, this prevents learning of explicit associations between the protected attribute and outcome. In other words, this feature becomes useless in training each of the two models because all the training data for each model belongs to the same group.

IV. Results

A. Constrained Logistic Regression

Since the resulting constrained optimization problems are convex, the trade-off between accuracy and fairness should be pareto-optimal. This implies that we should expect some loss in accuracy as discrimination is decreased. The ROC curves in figure 3 show that the area under the curve decreases as the upper bound on the boundary covariance decreases. Likewise, figure 4 shows that the area under the curve increases as we increase the parameter gamma for the fairness constrained problem. This demonstrates that the described constrained logistic regression problems are pareto-optimal.

Figure 5 and 6 show the tradeoff between accuracy and 3 fairness measures: Impact Ratio, Elift Ratio, and Odds Ratio. As expected, the impact ratio and elift ratio are closely related. The odds ratio, however, changes more erratically, and does not seem to be directly related to the fairness or accuracy constraints. The impact and elift ratios increase when accuracy decreases, and decrease when accuracy increases. While the impact ratio ($p\%$ rule) is not directly incorporated as a constraint, figure 5 shows that it approaches 1 as the fairness constraint is made more strict. This shows that the

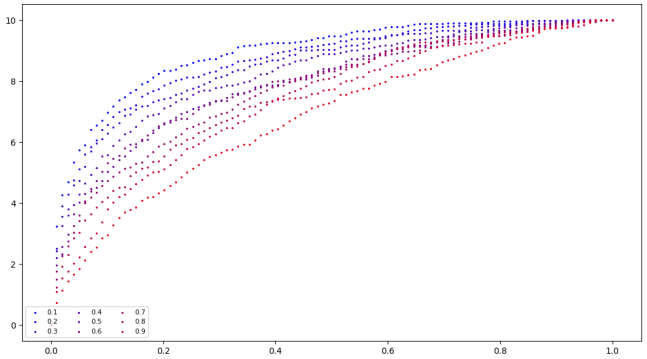

Fig. 4. ROC curve for accuracy constrained logistic regression for various gamma values. The area under the curve increases as gamma decreases.

fairness constrained logistic regression can be used to train a classifier that is supposed to obey a $p\%$ rule.

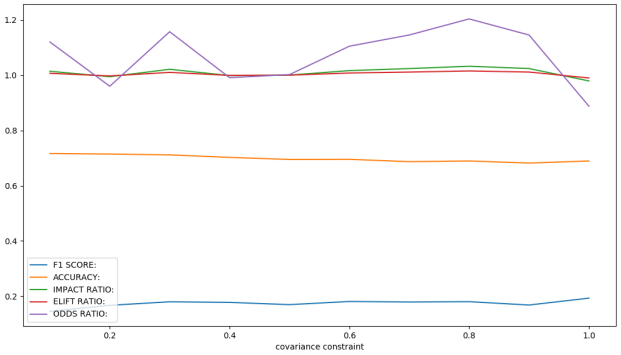

Fig. 5. Fairness and F1 Score for fairness constrained logistic regression for various values of C

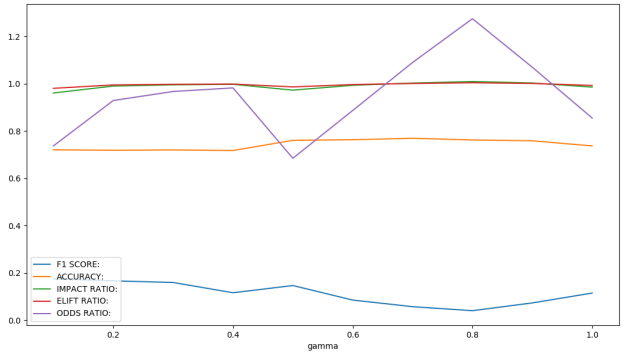

Fig. 6. Fairness and F1 Score for accuracy constrained logistic regression for various values of gamma

B. Comparison of measures of discrimination

Given that a range of measures for discrimination have been proposed in the literature, we compared the performance of each discrimination-reduction method by the three ratios. These results are summarized in Figure 7. First, we note that the labelled Crime data set appeared to be discriminatory according to each of the three measures, while the Recidivism data did not. All the methods that we implemented successfully set each ratio to 1, indicating no discrimination, however we observed that the baseline learners (*i.e.* models where no explicit anti-discrimination technique was performed) *also* reduced discrimination.

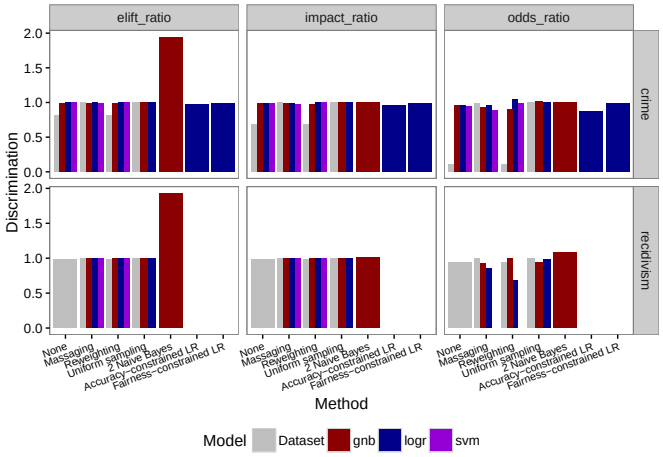

Fig. 7. Comparison of discrimination ratios for raw data and fair learners using crime and recidivism data sets

C. Performance of fair learners

Next, since it is important that fair learners be both fair *and* accurate to be suitable for application, we computed ROC curves by calculating mean FPR and TPR as the decision threshold was varied between 0 and 1, over 10-fold cross validation (Figure 8), and calculated the corresponding AUC (Table I).

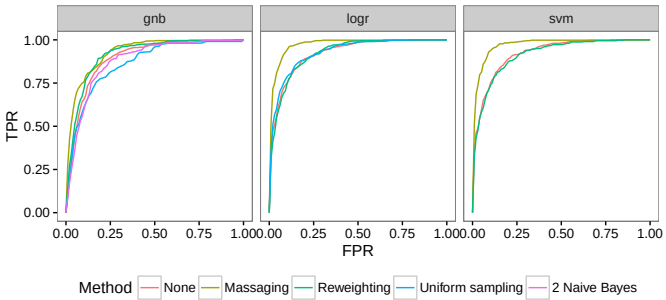

Fig. 8. ROC curves for discrimination-reduction methods, excluding constrained logistic regression. Values are mean FPR/TPR computed over 10-fold cross validation. AUC scores are given in Table I.

Figure 9 shows accuracy, F1, and precision and recall scores for each model on each data set. In general, the models learned on the data sets transformed by the *massaging* technique, logistic regression and SVM in particular, appeared to be the best performing in terms of accuracy, F1, and precision and recall. Interestingly, these models outperformed the baseline

	Method	2nb	gnb	logr	svm
1	2 Naive Bayes	0.88			
2	Massage		0.93	0.97	0.97
3	None		0.89	0.92	0.92
4	Reweighted		0.91	0.92	0.91
5	Unisample		0.86	0.93	

TABLE I
AUC FOR VARIOUS DISCRIMINATION-REDUCTION METHODS APPLIED TO THE CRIME DATA SET.

models on the Crime data set. In contrast, the results were quite different on the Recidivism data set, and here, models learned on data transformed by the *reweighting* technique performed best. The constrained logistic regression models had noticeably poor F1, precision, and recall scores; we discuss possible reasons in Section V. In order to investigate the possible trade-off between accuracy and discrimination, we plotted accuracy against discrimination for each model and data set in Figure 10. We observed that according to the impact and elift ratios, all models performed comparatively in terms of both accuracy and discrimination. The odd ratio proved to have greater variance. We did not observe a trend demonstrating that accuracy and discrimination are related by a trade-off, but only two of our methods directly incorporated fairness constraints into an optimization problem, which is the primary setting in which that trade-off holds.

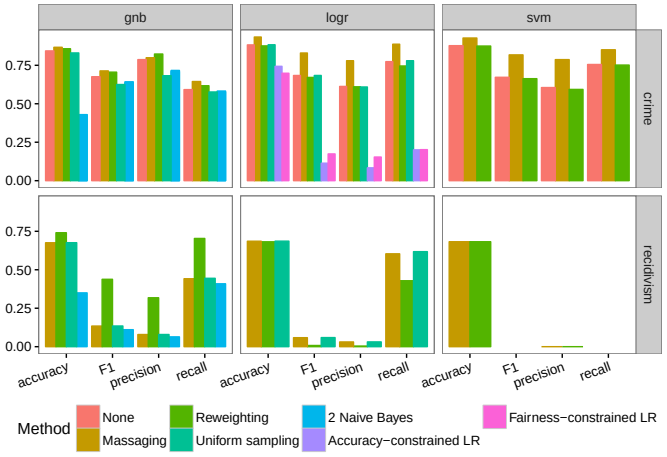

Fig. 9. Comparison of accuracy metrics for all fair learners using crime and recidivism data sets

D. Variance of discrimination with decision threshold

Zliobaite [14] carried out an interesting study of how the decision threshold and consequently what can be termed the "acceptance rate" of a model impact measures of discrimination. They confirmed what we might understand intuitively: that discrimination will be low with a very low acceptance rate or very high acceptance rate (*i.e.* when almost nobody or everybody is favoured by the outcome, respectively), and high when the acceptance rate takes on middle values between 0 and 1. In order to see if we observed the same trends on these data using these particular measures, we studied how

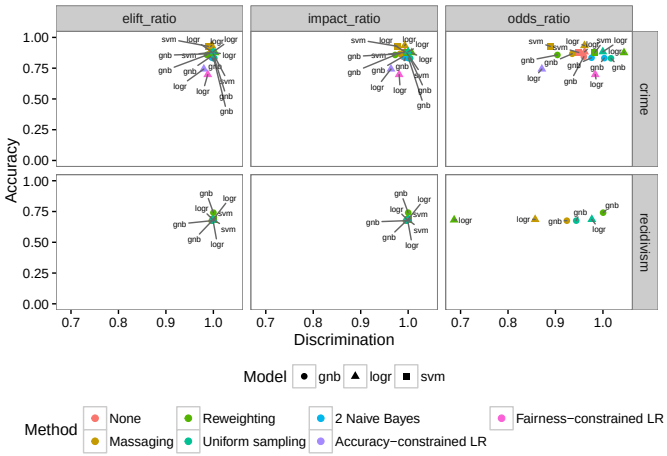

Fig. 10. Comparison of fair learners using accuracy and discrimination measurement

discrimination by each measure varied as the discrimination threshold varied between 0 and 1. Figure 11 shows the results. We found that all three ratios remained relatively constant as the decision threshold changed, however these results are not surprising since all three measures we considered are based on ratios between groups, which can be preserved as the decision boundary varies. The experiments in [14] used a measure based on differences between groups, which explains the discrepancy between our results.

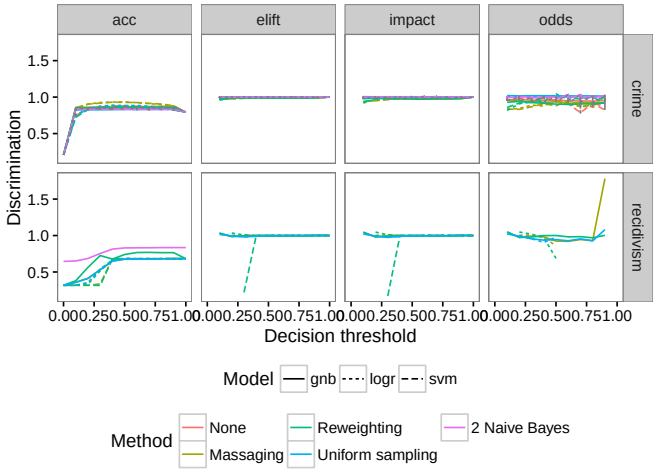

Fig. 11. Effect of decision threshold variation on discrimination in fair learners

V. DISCUSSION AND CONCLUSIONS

A. Exploring the performance of our methods in terms of accuracy and fairness

Discrimination measurement was the primary metric used to determine the effectiveness of the fair learning methods. First, the crime data set was used to test out all the methods of interest. Initially, the cleaned up data was measured to

learn what level of discrimination is inherent in neighbourhood crime likelihood. The Communities and Crime data set indicated significant discrimination with an impact ratio of 0.686 and an odds ratio of 0.103 [results]. Without any consideratoin for fairness, the level of discrimination had already reduced considerably in the three baseline learners. This result is surprising since the baseline learners are expected to capture the nuances of the training data, including any discrimination that may be present. One possible explanation for this is that the baseline learners generalize very well. If the odds ratio of the baseline predictions is very close to 1, then that means there is a minimal correlation between neighbourhoods being mostly black and a prediction of high crime in those neighbourhoods. Since all of our machine learning methods are linear, the baseline algorithms are able to compute decision boundaries using all data features without directly discriminating against the black feature. Although this result is promising, there are accuracy improvements to be made on top of the baseline.

From the array of fair machine learning methods we used, a few results stand out. The best performance was that of logistic regression on the massaged data. The massaged data alone raised the odds ratio to 0.989, allowing the logistic regression to then focus on minimizing error. This fair method resulted in an accuracy of 93.4% and F1 score of 0.898 [results], while maintaing an odds ratio of 0.961. Many other fair machine learning methods produced promising results, but there were also some poor performances. The fair logistic regression method was successful at eliminating discrimination with acceptable accuracy. The two different constrained versions of fair logistic regression presented very similar results. However, precision and recall for these accuracy- and fairness-constrained methods are low, which indicates that the adequate accuracy is an artifact of predicting mostly 1 or 0 on imbalanced data [results][results]. Moreover, we postulate that unbalanced datasets are more strongly affected by a loss in F1 score with constrained logistic regression, since the model is then forced to adjust to the distribution of the data in order to decrease discrimination while maintaining accuracy. A possible enhancement of constrained decision boundary classifiers would be to accommodate imbalanced datasets. One way to do that would be to use a weighted loss function.

Initial measurement of the recidivism dataset was not discriminatory, with an odds ratio of 0.941. This finding is unexpected, since crime data tends to be unbalanced with respect to race. Using the fair learning methods to predict recidivism produced generally weak results. Accuracy barely exceeded a random guess and the F1 score was in the 0.1 range. The only exception was when the reweighted dataset was used to train the Naive Bayes learner. The accuracy using this method presented an accuracy of 74.1% and an F1 score of 0.438 while maintaing an odds ratio of 1.0 [results][results].

B. Limitations of these methods

A common issue in most machine learning applications is that the success of an approach is data dependent. Our project investigated two unique sets of real data which produced varying degress of success using our methods. In order to perform a full survey on the effectiveness of our fair learning methods, we believe more unique data sets would be beneficial. Having more case studies allows for a more in depth understanding of our fair models and how they react with different types of data.

Pre-processing often plays an integral role in effective learning algorithms. Shaping data to better fit the type of learning being trained always improved validation accuracy. Since our project focused on exploring a multitude of alternative learners to reduce discrimination, optimizing the data sets was not a priority. Since only a basic cleaning of the two training sets was carried out, it is possible that accuracy could have been improved with a more involved approach.

Another consideration for our process is that we only investigated discrimination effects when considering one protected group. In reality, there will often be more than one protected group in a discrimination-prone decision, such as sex, age, or sexuality. Furthermore, there also exists the possibility that some non-protected features are correlated to protected ones. This phenomenon then has the potential to cause indirect discrimination towards a protected feature. Future studies would be well suited to investigate the effect of correlated non-protected features on discrimination against protected features.

Training two separate Naive Bayes classifiers seems to eliminate direct discrimination, since it decouples the prediction from the sensitive feature and allows the outcome to depend more on other features. However it does not address indirect discrimination, where the other features might be correlated with the sensitive attribute. Moreover, the separate Naive Bayes classifiers can still learn the frequency of the favorable outcome for each value of the sensitive feature, and introduce discrimination in this manner.

We noticed that the baseline linear classifiers were able to decrease discrimination even when using the sensitive feature as a decision input. We postulate that the reason for this is because most of the data points that are labeled due to discrimination with respect to the sensitive feature occur near the decision boundaries, and since the decision boundary is linear, it does not capture these contortions. This results in less discriminatory predictions for such data points, which reduces discrimination overall.

C. Conclusions and outlook on fairness-aware machine learning

As machine learning researchers committed to building helpful autonomous systems, the safe and responsible use of machine learning should be at the forefront of our work. Our analysis has highlighted that several methods for successfully reducing discrimination in the binary classification setting exist, and the methods we have explored are relatively simple to implement and represent a range of points in the machine learning process at which we can intervene to ensure fairness, which make them suitable for a range of applications.

In the course of our work, we have encountered several directions for future research in fairness-aware machine

learning. A crucial next step would be to complete a more comprehensive characterization of the methods available for discrimination reduction, and of measures of fairness. The field has not yet reached a consensus on this, and broad usage of anti-discrimination principles as well as their incorporation into legislation surrounding machine learning uses in the public sphere would require consistent definitions. We also found that many researchers exploring this area describe methods for reducing discrimination without providing clear, well-documented, user-friendly implementations. We have tried to combat these issues in our work (which is available in its entirety on GitHub), but more generally, a toolkit of discrimination reduction techniques alongside machine learning libraries would be extremely useful. Many of the methods and measures we have explored here also extend themselves to more complex learning scenarios. For example, situations with categorical protected features and outcomes or a need for a model which can incorporate streaming data also apply. In addition, more sophisticated models which can address correlation between explicitly protected features and others are sorely needed. Moreover, since neural networks are being used more frequently, further research on reducing discrimination in neural network models is needed. This is a difficult topic since neural networks present a non-convex optimization problem with a very large number of parameters.

The questions of how ethical concerns can be incorporated into machine learning are young, but ethical concerns about discrimination and fairness in general are not. While achieving equity in how various groups are treated by social institutions will require systemic change, investigating fairness in machine learning is an important and exciting opportunity for researchers to contribute to this process.

VI. Statement of Contributions

SJ performed data pre-processing, implemented the pre-processing and 2NB methods, trained baseline classifiers on the data, and analyzed results. **AC** performed data pre-processing, implemented discrimination and accuracy measurement scripts, result compilation scripts, and analyzed results. **ER** implemented constrained logistic regression models with corresponding graphs. All authors contributed equally to the writing of the report.

We hereby state that all the work presented in this report is that of the authors.

References

[1] Toshihiro Kamishima et al. “Fairness-Aware Classifier with Prejudice Remover Regularizer”. In: *Proceedings of the 2012 European Conference on Machine Learning and Knowledge Discovery in Databases - Volume Part II*. ECML PKDD’12. Springer-Verlag, 2012, pp. 35–50. URL: http://dx.doi.org/10.1007/978-3-642-33486-3_3.

[2] Toon Calders and Indrė Žliobaitė. “Why Unbiased Computational Processes Can Lead to Discriminative Decision Procedures”. In: *Discrimination and Privacy in the Information Society: Data Mining and Profiling in Large Databases*. Springer Berlin Heidelberg, 2013, pp. 43–57. URL: http://dx.doi.org/10.1007/978-3-642-30487-3_3.

[3] Indrė Žliobaitė. “A survey on measuring indirect discrimination in machine learning”. In: *CoRR* abs/1511.00148 (2015). URL: http://arxiv.org/abs/1511.00148.

[4] Francesco Bonchi et al. “Exposing the Probabilistic Causal Structure of Discrimination”. In: *CoRR* abs/1510.00552 (2015). URL: http://arxiv.org/abs/1510.00552.

[5] Koray Mancuhan and Chris Clifton. “Combating Discrimination Using Bayesian Networks”. In: *Artif. Intell. Law* 22.2 (June 2014), pp. 211–238. ISSN: 0924-8463. URL: http://dx.doi.org/10.1007/s10506-014-9156-4.

[6] Dino Pedreschi, Salvatore Ruggieri, and Franco Turini. “A Study of Top-k Measures for Discrimination Discovery”. In: *Proceedings of the 27th Annual ACM Symposium on Applied Computing*. SAC ’12. Trento, Italy: ACM, 2012, pp. 126–131. URL: http://doi.acm.org/10.1145/2245276.2245303.

[7] Binh Thanh Luong, Salvatore Ruggieri, and Franco Turini. “k-NN As an Implementation of Situation Testing for Discrimination Discovery and Prevention”. In: *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD ’11. San Diego, California, USA: ACM, 2011, pp. 502–510. ISBN: 978-1-4503-0813-7. URL: http://doi.acm.org/10.1145/2020408.2020488.

[8] D.J. Newman A. Asuncion. *UCI Machine Learning Repository*. 2007. URL: http://archive.ics.uci.edu/ml/.

[9] Faisal Kamiran and Toon Calders. “Data preprocessing techniques for classification without discrimination”. In: *Knowledge and Information Systems* 33.1 (2012), pp. 1–33. ISSN: 0219-3116. URL: http://dx.doi.org/10.1007/s10115-011-0463-8.

[10] F. Pedregosa et al. “Scikit-learn: Machine Learning in Python”. In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830.

[11] M. Bilal Zafar et al. “Fairness Constraints: Mechanisms for Fair Classification”. In: *ArXiv e-prints* (July 2015). arXiv: 1507.05259. URL: http://adsabs.harvard.edu/abs/2015arXiv150705259B.

[12] Eric Jones, Travis Oliphant, Pearu Peterson, et al. *SciPy: Open source scientific tools for Python*. 2001–. URL: http://www.scipy.org/.

[13] Toon Calders and Sicco Verwer. "Three Naive Bayes Approaches for Discrimination-free Classification". In: *Data Min. Knowl. Discov.* 21.2 (Sept. 2010), pp. 277–292. ISSN: 1384-5810. DOI: 10.1007/s10618-010-0190-x. URL: http://dx.doi.org/10.1007/s10618-010-0190-x.

[14] Indre Zliobaite. "On the relation between accuracy and fairness in binary classification". In: *CoRR* abs/1505.05723 (2015). URL: http://arxiv.org/abs/1505.05723.