Geographic Unfairness Mitigation of School Evaluation Model

Juan David Nieto García, Giovani de Almeida Valdrighi

1 Introduction

In our last report on modeling, we developed a predictive model to identify schools that needed support, adjusting the hyperparameters and using feature selection techniques (such as Mutual Information and Causality based on Markov Blanket). Among the important findings, a high predictive correlation associated with the socioeconomic level of the schools was identified, supported by the significant importance of the socioeconomic dimension in the modeling. The model presented good performance. However, it is important to mention that the data is being repurposed and can have inner biases. The target variable is also a proxy of the intended label. Despite that, the model could still be useful with a critique use of it.

To improve our initial model, it is important to mitigate biases and make them fairer—meaning to identify schools that require support programs without providing an advantage to those that do not. Fairness algorithms address biases present before and after training the model.

In this report, we first studied the biases present in our model and identified a bias toward the geographic regions of the school. This was later mitigated by using a diverse set of fairness techniques. Ultimately, we present an extra experiment that tries to create new labels without bias by using a clustering step.

2 Related Works

Bias mitigation in Machine Learning (ML) can be addressed at different stages of the process, such as data pre-processing, processing, and post-processing [1]. Pre-processing seeks to identify and correct inherent biases in the data before the model is trained. Pre-processing techniques try to transform the data so the underlying discrimination is removed [2]. If the algorithm is allowed to modify the training data, then pre-processing can be used.

During processing, algorithms and techniques are implemented that seek to minimize biases while the model is trained. This involves adjusting model weights and parameters to reduce the influence of biased features.

By last, post-processing is performed after training by accessing a hold out set that was not involved during the training of the model. If the algorithm can only treat the learned model as a black box with out any ability to modify the training data or learning algorithm, then only post-processing can be used in which the labels assigned by the black-box model initially get reassigned based on a function during the post-processing phase [3].

Regarding impartiality, three key aspects stand out: separation, sufficiency and independence. The separation implies that model decisions should be based on relevant factors and not on characteristics that are related to protected groups. Sufficiency refers to ensuring that the model has enough information to make accurate decisions without over-reliance on biased features. Finally, independence seeks to prevent the inclusion or exclusion of a characteristic from unduly influencing the model's predictions, thus ensuring equitable treatment. These criteria are essential to achieve equity in the development and application of machine learning models [4].

Regarding works related to fairness in educational applications [5], various standard measures have been employed to assess the fairness of educational systems across different groups, such as speakers of different languages or examinees with disabilities [6], [7], [8]. The two most common analyses involve standardized mean score differences and overall model performance for different groups based on human scores (predictive ability). Recently, other measures have been employed, including differential item functioning to analyze the performance of individual features or variance in the model's mean residual across groups [9], [10]. The predominant approach in many previous studies on the fairness of automated scoring has been that significant differences between groups in both human and automated scores on any of these measures may indicate potential equity issues in the system. Therefore, it is necessary to intervene in the systems with the aim of improving the decisions.

3 Methodology

To provide context and rationale for the intervention applied to our trained model, the following sections outline the sensitive attributes, these attributes play a pivotal role in evaluating and addressing potential disparities or unfair discriminations that may arise in the model's predictions. And subsequent sections detail the metrics utilized for evaluation and the fairness models implemented in this study.

3.1 Sensitive attributes

Group fairness is the most suited approach to our study, as we can separate the schools into groups regarding some

sensitive attributes of the school itself or the individuals related to it. Group fairness is also more suited for public policy concerns, as the regulations in different countries are defined. Our data has a large set of features available, and in the previous study, we selected a subset of features using causality tests. Despite that, the features that were not used by the final model should also be considered for evaluating fairness.

In our available data, we could consider three possible sensitive attributes: the director's race, region, and socioeconomic index. The race of the director, despite being a sensitive attribute, should not be considered as we are looking at evaluating the schools: it is not expected that the distribution of the race of the director and the school features have a high correlation. The remaining two attributes are going to be studied. In Fig. 1, we present the distribution of the label for both the original data (Y) and our model ($\hat{Y}$). By first looking at the distribution of $P(Y = 1)$, we study the distribution of the original data, i.e., how the SAEB score is distributed along the sensitive attributes. It is possible to see that schools from the north and northeast have a higher probability of $Y = 1$, i.e., schools from the north and northeast have lower SAEB scores. This is similar to the socioeconomic index, i.e., schools with a lower socioeconomic index have a higher probability of lower SAEB scores. The model exaggerates this trend, as we can identify by looking at the distribution of $P(\hat{Y} = 1)$.

This analysis permits us to conclude a question that was presented in our initial proposal: are the SAEB scores biased regarding the socioeconomic condition of the school? This plot shows that it, in fact is, and it is also biased towards the geographic region of the school. Considering the SAEB school was an indicator to identify schools in need of support, this shows that the schools that need it are the ones with lower socioeconomic index. The social condition of classes, teachers, and students is highly correlated to how students perform in math and Portuguese tests. This may also be an indicator that the SAEB score is not very useful in analyzing how educational projects affect students, as the results of the test will be impacted by social conditions. In Brazil, the geographic regions have some correlation with the socioeconomic condition; however, this does not fully explain the bias toward south regions on the SAEB score. Students from the South are not expected to know more about math and Portuguese; this bias can be related to how these tests are formulated.

This analysis also demonstrated the importance of using fairness techniques, as we should not use a model that does not correct (and even increase) the discriminatory bias in the data. The following sections present the development of these techniques.

3.2 Metrics

There are different methods to quantify unfairness in machine learning models, and an important initial step in the study of our models is to consider which of the fairness metrics are going to be utilized. Our positive outcome equals $Y = 1$, i.e., the school will be selected to receive special support. One important fairness metric to consider is equality of opportunity, the difference between the true positive rate between privileged and unprivileged groups. It will measure if the model has a tendency to miss positive outcomes for any of the groups. Another important metric is demographic parity, this metric considers that different groups should receive the same amount of resources, i.e., both groups should have the same probability of a prediction $Y = 1$.

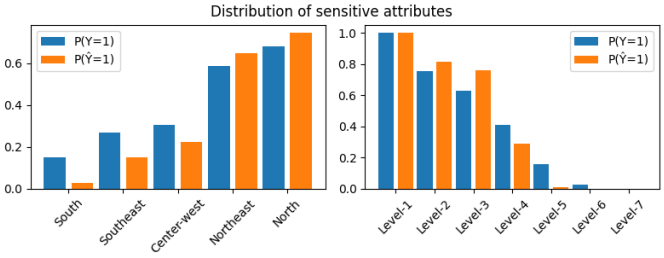

Fig. 1: Distribution of positive label (need of support) conditioned on sensitive attributes (left: region and right: socioeconomic index) for the original data Y and the model $\hat{Y}$.

Metric		Region	Socio-economic index
Equality of opportunity		0.520	0.731
Demographic Parity	Dataset	0.366	0.427
	Model	0.546	0.564

TABLE 1: Fairness metrics calculated considering both the sensitive attributes.

At Tab. 1, we present the metrics calculated considering both the sensitive attributes discussed in the previous section. Note that demographic parity can be calculated independently for the dataset and the model, while equality of opportunity is only defined by using both the dataset and model. The metrics should be close to 0 to guarantee fairness. To facilitate the formulation, we simplified the sensitive attribute into two groups. The north and northeast regions are replaced by the value 1, and the remaining regions are replaced by 0. Schools with socioeconomic indexes lower than 4 are replaced by 1 and the remaining by 0. The results showed that we actually have a high level of unfairness with all combinations of metric and sensitive attributes. As we discussed in the previous section, the model increased the unfairness presented in the dataset.

It is important to make a decision about which sensitive attribute to consider and which fairness metric to focus on. We will focus on **minimizing the demographic parity regarding the geographic region**. From a governance point of view, we would want to give the same amount of resources to each of the regions, and we also should not expect a difference in their performance, differently from the socioeconomic index, which is already known to impact school performance.

3.3 Methods

3.3.1 Pre-processing

Pre-processing is the bias mitigation method applied at the initial stage. Some advantages of pre-processing a dataset

are that the pre-processed data can be used for any downstream task and it is not necessary to modify the model. However, this method is inferior in performance in terms of precision and fairness. Furthermore, pre-processing can only be used to optimize a limited number of fairness metrics, as we do not have the label information at this stage.

Reweighing: Proposed by Calders et al. [11], [12], consists of adding weights to each training data sample. Therefore, it can only be applied by models that can incorporate weights in their learning, which is generally done by using them in terms of the loss function. Designed for classification tasks and categorical sensitive attributes, the weights are defined to enforce that the data labels are independent of the sensitive attribute. If there is independence, we have that $P(Y = y \wedge Z = z) = P(Y = y) \times P(Z = z)$; however, most of the datasets will not satisfy this property. The weights are defined for each group z and label y as $W_{y,z} = \dfrac{P(Y = y) \times P(Z = z)}{P(Y = y \wedge Z = z)}$ and, in practice, the probabilities are calculated from the empirical distribution of the observations:

$$W_{y,z} = \frac{\left(\frac{1}{n}\sum_{i=1}^{n} I_{[Y_i=y]}\right) \times \left(\frac{1}{n}\sum_{i=1}^{n} I_{[Z_i=z]}\right)}{\frac{1}{n}\sum_{i=1}^{n} I_{[Y_i=y \wedge Z_i=z]}}$$

Where I_A is an indicator variable that is 1 if A is true and 0 otherwise. The reweighing strategy updates the weights so that the not-so-well-observed scenarios (positive outcomes for the unprivileged group) are also significant in the loss value.

Fair-SMOTE: SMOTE [13] is a method of resampling the data for scenarios of imbalanced learning. Considering that there is a minority class present in the data, it tries to create new samples (that are feasible) from this class to achieve a better proportion of samples from each class. Fair-SMOTE [14] is an adaptation of this methodology to consider missing observations of a specific value of the sensitive attribute. It combines the information of the sensitive attribute Z and the label Y, considering that they are binary, to produce four classes. For example, one class is $Z = 1 \wedge Y = 1$ (underprivileged attribute with positive outcome). Then, new samples are generated for every class but the majority class.

3.3.2 In-processing

Another category of bias mitigation methods is in-processing, which encompasses methods that directly modify the training algorithms. In other words, the decision-making process is modified by changing the feedback used to assess whether the decision was reasonable. Those methods have the advantage of allowing the flexibility of choosing which trade-off between fairness and precision is the most suited for the use case and also tend to result in better precision and fairness than the methods from the other categories.

Demographic Parity Classifier: Logistic Regression is a simple yet very reliable model as we discussed in our previous report, and can be adjusted to incorporate fairness. Zafar et al. [15] designed an in-processing mitigation method that changes the loss function of convex margin-based methods, logistic regression in particular. This approach minimizes the prediction error with a constraint of max correlation between the sensitive attribute and the prediction, i.e., ensuring independence. Let θ be the vector of parameters of the logistic model, the constraint is formulated as:

$$\left|\frac{1}{n}\sum_{i=1}^{n}(Z_i - \overline{Z})\theta^T X_i\right| \leq c$$

Where $\overline{Z}$ is the mean of the sensitive attribute of samples, this constraint can be formulated to consider demographic parity, as presented above. This technique is generally used by training with different values of c and selecting a model from the results based on the trade-off between performance and fairness.

FairGBM: Gradient boosting is among the state of the art for tabular data. Recently, different mitigation methods have already been used to incorporate fairness in this method; in particular, FairGBM [16] is a modification of the gradient-boosting algorithm that enforces fairness with constrained learning. A fairness metric is also minimized jointly with the prediction error at each iteration. One important detail is that, as the boosting algorithm needs to use a differentiable function, it was necessary to use a differentiable proxy function of the fairness metric, which they used demographic parity and equality of opportunity.

3.3.3 Post-processing

Such methods are widely used when facing a "black box" model. These methods alter the model output to satisfy a particular fairness constraint, such as changing a classifier's threshold of positive and negative prediction. Despite the versatility, they do not have the flexibility to choose the trade-off between accuracy and fairness and can produce worse improvements than other approaches.

Threshold Optimizer: Proposed by Hardt et al. [17], this approach works with the score value, i.e., the model output, before applying a threshold to decide for the positive prediction. The ROC curve is commonly created by calculating the false positive and true positive rates for different threshold values. But, in this mitigation method, the curve is calculated for each group, i.e., the rates are calculated for the subset of samples of each group. If the original score is fair, the curves will be similar, and the threshold could be the same for the groups, however, this is not generally the case. Then, a point of minimal prediction loss is searched at the intersection of the convex hull of ROC curves (the area between the curve and the diagonal). The prediction loss can be minimized inside this region without unfairness because the searched area is of fair solutions, and the selected point will have different threshold values for each group.

3.4 Training details

The discussed models were trained using the selected 50 features from the causality tests. Using the same approach

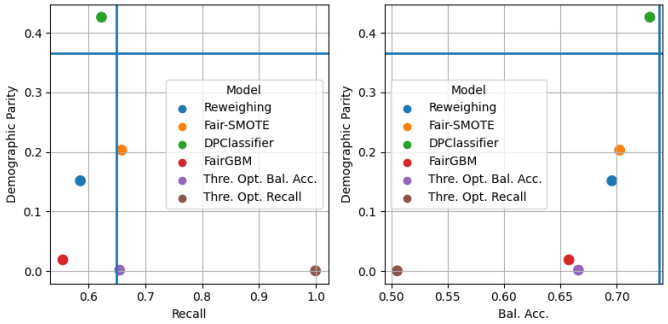

Fig. 2: Resulting metrics of fairness and performance for the fairness methods. The vertical blue line denotes the original performance metric, and the horizontal blue line is the demographic parity of the initial model.

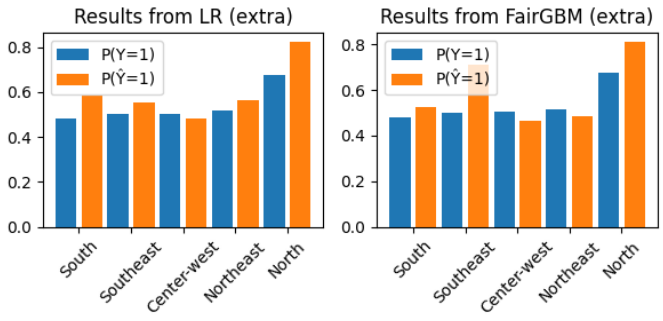

Fig. 3: Distribution of positive label (need of support) conditioned on sensitive attributes (left: region and right: socioeconomic index) for the new obtained Y and two models trained with it.

from previous studies, the categoric features were one hot encoded, and the numeric ones were standardized. For the pre-processing methods, the fairness method was used with fixed parameters and the parameter C of logistic regression was tuned to maximize recall. Both complexity parameters and fairness weight were optimized for the in-processing techniques to maximize recall. The post-processing technique was used with two scenarios, to maximize recall and balanced accuracy.

4 Results

The metrics of demographic parity versus recall and balanced accuracy for the six models used are presented in Figure 2. Regarding preprocessing methods, we can see that in both metrics (Recall and Bal. Acc), the SMOTE method showed higher values than the Reweighing method, with greater demographic parity for SMOTE. This means that the SMOTE method has achieved a better balance in the distribution of positive outcomes among different demographic groups. However, the demographic parity is still 0.2, distant from the objective to be 0.

Regarding the in-processing methods, it was observed that the Demographic Parity Classifier has the worst results for demographic parity compared to the FairGBM method and all other methods, even worse than the original demographic parity value. It could occur for many reasons, but it is possible that by removing the correlation between the sensitive attribute and the prediction, other features contain the information of the sensitive attribute.

The post-processing method was able to obtain the perfect demographic parity, equal to 0, however with a high cost in the balanced accuracy. By maximizing recall with the demographic parity constraint, the method obtained a recall equal to 1 by predicting the positive outcome to almost any sample. This is not a desirable model, despite the fact that demographic parity and recall are the best values.

At Fig. 4 we present the distribution of the prediction of all fairness methods. We can see, for example, that the threshold optimizer with the recall criteria resulted in all predictions being equal to 1 for all geographic regions. We can also see that FairGBM obtained a similar probability of $\hat{Y} = 1$ for all geographic regions and was the best model.

5 Extra Experiment

In the previous study, we decided to use SAEB scores to identify schools that need support and selected the threshold of 250 to binarize the problem. However, the SAEB score correlates to the schools' sensitive attributes, i.e., they depend on the socioeconomic condition, geographic region, etc. If we want to identify which schools are underperforming due to pedagogic decisions, we should remove the effect of their socioeconomic condition. With that in mind, we perform an extra experiment to create a new label Y. The idea is to cluster the schools based on their socioeconomic condition; each cluster k will have a mean value of the SAEB score $\overline{Y}_k$, and then for each school, we subtract from its SAEB score the mean value of its cluster. To binarize the problem again, we transform to 1 if this new label is lower than 0, i.e., if the school had a SAEB score lower than the mean of its group. In more detail, the clustering was performed with K-means using the socioeconomic index, the geographic region, the school location, and if it is in the capital. The schools were clustered into five groups. In Fig. 3, we present the distribution of $Y = 1$ for this new dataset. It is possible to see that it has a lower bias than our original dataset, i.e., the probability of $Y = 1$ is similar for all geographic regions. We also trained two models with this needed dataset: Logistic Regression and FairGBM. The distribution of model predictions is present in Fig. 3. By using less biased data, the models also resulted in less biased predictions.

However, this was a test of an unconventional approach. A more detailed study is necessary to verify whether this new label represents school performance. It would be important to validate the modeling decisions with domain experts, such as which features to use in the clustering. An interesting analysis would be to compare which scenarios the original label and this new label agree or disagree.

References

[1] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, "A survey on bias and fairness in machine learning," *ACM computing surveys (CSUR)*, vol. 54, no. 6, pp. 1–35, 2021.

[2] B. d'Alessandro, C. O'Neil, and T. LaGatta, "Conscientious classification: A data scientist's guide to discrimination-aware classification," *Big data*, vol. 5, no. 2, pp. 120–134, 2017.

[3] R. K. Bellamy, K. Dey, M. Hind, S. C. Hoffman, S. Houde, K. Kannan, P. Lohia, J. Martino, S. Mehta, A. Mojsilovic *et al.*, "Ai fairness 360: An extensible toolkit for detecting," *Understanding, and Mitigating Unwanted Algorithmic Bias*, 2018.

[4] C. Hertweck and T. Räz, "Gradual (in) compatibility of fairness criteria," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 11, 2022, pp. 11 926–11 934.

[5] A. Loukina, N. Madnani, and K. Zechner, "The many dimensions of algorithmic fairness in educational applications," in *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, 2019, pp. 1–10.

[6] B. Bridgeman, C. Trapani, and Y. Attali, "Comparison of human and machine scoring of essays: Differences by gender, ethnicity, and country," *Applied Measurement in Education*, vol. 25, no. 1, pp. 27–40, 2012.

[7] Z. Wang, K. Zechner, and Y. Sun, "Monitoring the performance of human and automated scores for spoken responses," *Language Testing*, vol. 35, no. 1, pp. 101–120, 2018.

[8] J. Burstein and M. Chodorow, "Automated essay scoring for non-native english speakers," in *Computer mediated language assessment and evaluation in natural language processing*, 1999.

[9] M. Zhang, N. Dorans, C. Li, and A. Rupp, "Differential feature functioning in automated essay scoring," *Test fairness in the new generation of large-scale assessment*, pp. 185–208, 2017.

[10] S. Malmasi, K. Evanini, A. Cahill, J. Tetreault, R. Pugh, C. Hamill, D. Napolitano, and Y. Qian, "A report on the 2017 native language identification shared task," in *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications*, 2017, pp. 62–75.

[11] F. Kamiran and T. Calders, "Data preprocessing techniques for classification without discrimination," *Knowledge and information systems*, vol. 33, no. 1, pp. 1–33, 2012.

[12] T. Calders, F. Kamiran, and M. Pechenizkiy, "Building classifiers with independency constraints," in *2009 IEEE international conference on data mining workshops*. IEEE, 2009, pp. 13–18.

[13] 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, pp. 321–357, 2002.

[14] 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*, 2021, pp. 429–440.

[15] M. B. Zafar, I. Valera, M. Gomez Rodriguez, and K. P. Gummadi, "Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment," in *Proceedings of the 26th international conference on world wide web*, 2017, pp. 1171–1180.

[16] A. F. Cruz, C. Belém, J. Bravo, P. Saleiro, and P. Bizarro, "Fairgbm: Gradient boosting with fairness constraints," *arXiv preprint arXiv:2209.07850*, 2022.

[17] M. Hardt, E. Price, and N. Srebro, "Equality of opportunity in supervised learning," *Advances in neural information processing systems*, vol. 29, 2016.

APPENDIX

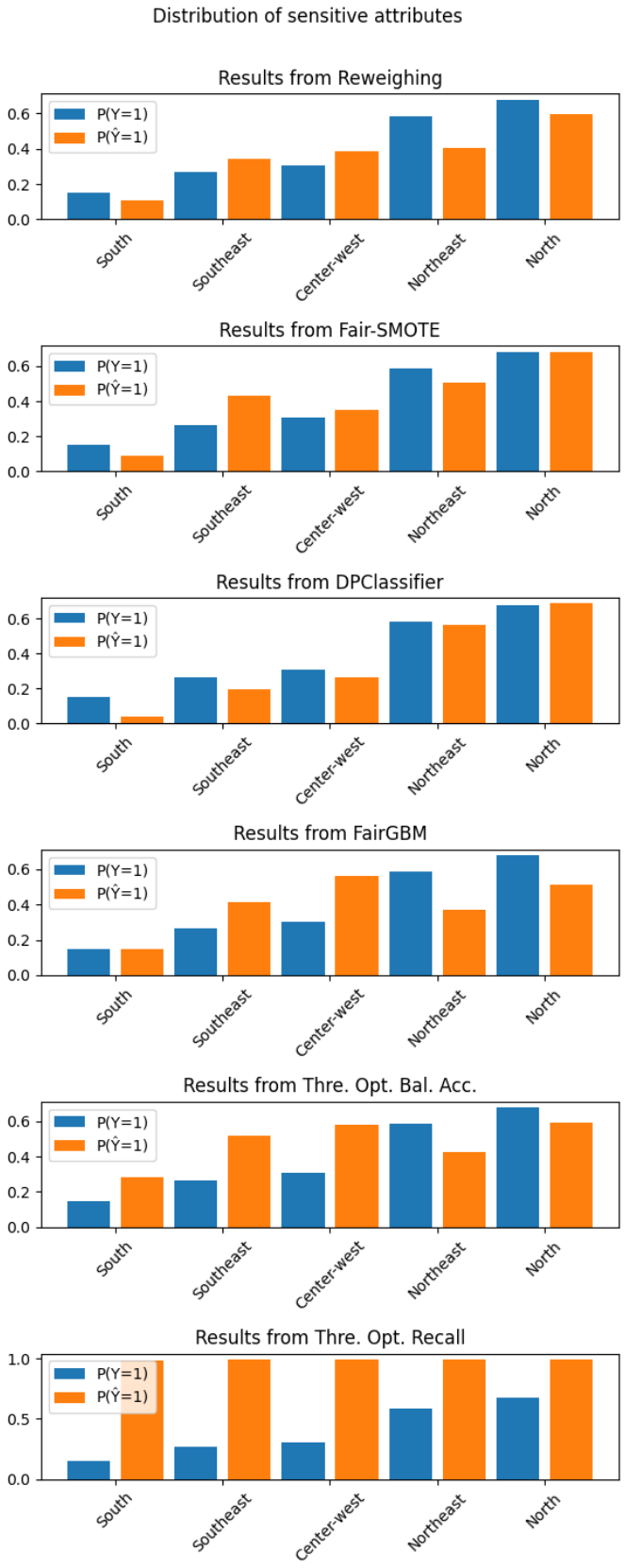

Fig. 4: Distribution of prediction $\hat{Y}$ conditioned on the sensitive attribute for all the fairness methods. The more fair models are the ones that reduced the probability of $P(\hat{Y} = 1)$ for north and northeast and increased in the other regions.