Evolving Fair Models: Fair and Accurate Classification with Multi-Objective Evolutionary Computing

Julian Zucker, zucker.j@gmail.com
John Rachlin, j.rachlin@northeastern.edu

Abstract

Machine learning models must be trained for both fairness and accuracy. We use multi-objective evolutionary computing to accomplish that goal, and as a tool to understand the tradeoffs between fairness and accuracy or different metrics of fairness. We explore four different fairness metrics, and show only a small cost to accuracy when greatly increasing fairness. Multi-objective evolutionary computing produces a Pareto frontier, which is used to identify how much accuracy must be sacrificed to achieve a specific level of fairness on a given fairness metric. We present a methodology for identifying whether two fairness metrics are correlated or anti-correlated, reducing the number of metrics which must be examined to understand the overall fairness of a model.

1. Introduction

Machine learning models play a role in decisions that alter the life prospects of their subjects. At the same time, these models are often biased [2], and because making decisions based on biased methods is unethical and illegal [4], it is important that machine learning practitioners take action to reduce bias in their models [6]. Bias in algorithmic decision-making has many sources: a model can end up with bias if it is trained on historical data from humans who made biased decisions [7], a misrepresentative subset of data that paints a biased picture [20], or data with more datapoints for certain subgroups of the population [21].

Practitioners can identify biases in their models by applying metrics of fairness to their models and datasets. However, it can be difficult for practitioners to choose which metrics to use – the AIF360 project has over 70 [5]. Indeed, many of these metrics are contradictory, that is, an increase in one may necessitate a decrease in another [12]. For example, a definition of fairness could measure how consistent the model's accuracy is for different subgroups. Fairness and accuracy conflict when one subgroup is easier to make predictions for than another. If the easy subgroup is large, overall performance metrics will obscure weaknesses on smaller subgroups. In addition to fairness, practitioners require that their models be accurate, i.e., that they correctly classify

examples outside of the training set at a high rate, with respect to the ground truth labels in the dataset.

Practitioners can mitigate biases by preprocessing the input datasets to remove bias, post-processing the model's predictions to ensure fairness, or training models using algorithms that take fairness as well as accuracy in to account [2]. This paper will focus on the last method, introducing a new way to train models that optimizes simultaneously for accuracy and fairness.

Training a fair and accurate model is a multi-objective optimization problem, because one must optimize simultaneously for accuracy and some definition of fairness. In this paper, we use multi-objective evolutionary computing to train fair and accurate models. First, we show that evolutionary computing, optimizing for accuracy alone, can produce models with similar accuracy to benchmarks. We then show that when optimizing for fairness as well as accuracy, the fair models are only slightly less accurate than the unfair models. Using the generated Pareto frontiers, we show the tradeoffs required between each pair of fairness metrics, giving insight into which are correlated or anti-correlated in practice. The full source code for this analysis is available online [1].

2. Fairness Metrics

A fairness metric is a function of a model and a dataset that produces a real-valued output representing the degree of unfairness of the given model on the given dataset. For simplicity, we limit our models to produce binary classifications: "positive outcome" or "negative outcome." We also limit our datasets to only have two groups: "privileged" and "unprivileged." Fairness metrics, then, are measures of how fairly the positive and negative outcomes are distributed across the privileged and unprivileged classes [6]. We will explore four fairness metrics:

1. *Disparate impact,* the ratio between the ratio of positive to negative predictions for the privileged group and the ratio of positive to negative predictions for the unprivileged group. While the ratio is normally conceived of as being greater than one if the privileged group has a higher true positive rate, and less than one is the unprivileged group does, we chose to represent the ratio as the higher ratio divided by the lower one. With this adjustment, the minimum score is 1, and inequality favoring the priviliged or unpriviliged groups causes the score to be greater than 1. This step ensures that minimizing the adjusted score will lead to unbiased models, instead of models biased towards the unpriviliged group.
2. *False negative rate ratio,* the ratio between the false negative rate for the privileged group and the unprivileged group. The same ratio adjustment as for disparate impact was applied.

[1] https://github.com/julian-zucker/evolving-fair-models

3. *True positive rate ratio,* the ratio between the true positive rate for the privileged group and the unprivileged group. The same ratio adjustment as for disparate impact was applied.
4. *Between-group generalized entropy,* the amount of information required to represent the inequality in the distribution of benefits between groups by the model. This is the only one of our chosen metrics where a score of 0 is perfect (representing equal distribution of outcomes). We choose specifically the between-group Theil index, a special case of between-group generalized entropy where α is 1 (for a more in-depth description, see [19]).

3. Evolving Fair Models

Training a machine learning model that is both accurate and fair is a multi-objective optimization problem. Because there are multiple objectives, some of which are not differentiable, standard methods of fitting models such as gradient descent and convex optimization—even the multi-objective version of gradient descent proposed by [9]—which rely on a differentiable loss function will not work [22]. Furthermore, we are not trying to find the optimal point in Euclidean space by some function, we are trying to find an optimal model. We must find the optimal model in model-space, not the optimal point in Euclidean space. Evolutionary optimization approaches have proven effective under these constraints, as in [23] where decision trees were evolved with false negative and false positive rate as objectives. Evolutionary computing can also support optimizing models for multiple fairness metrics at once, an advantage over parameterized fairness/accuracy calculations like the ones performed in [11].

We used the open-source Evvo framework[2] for evolutionary computing to train our models. The core of evolutionary computation implemented by Evvo is:

1. Generate a starting population of random solutions.
2. Starts threads which asynchronously copy and modify solutions in the population. The modified solutions are added to the population.
3. Start theads which asynchronously select random samples from the population and delete the solutions that are bad. We chose to delete dominated solutions from the sample. Dominated solutions are those which score worse than another solution on every objective, for example, a model that is less accurate and less fair than another model would be dominated.
4. When some amount of time has elapsed, stop the system.

To evolve a decision tree with Evvo, one needs an initial population, modification operators, and criteria for when to stop. Our initial population consists of

[2] https://github.com/evvo-labs/evvo

randomly generated decision trees, trained by splitting greedily to minimize entropy. We implemented the modification operators introduced in [13]. These operators change the feature that a node examines, change the threshold of a node, swap the class that a given leaf predicts, change a leaf to a node, or change a node to a leaf. Following [16], we employ a "crossover" operator, which takes two decision trees and swaps a random subtree of one with a random subtree of another. After running for a specified amount of time, the system stops and returns the Pareto frontier. In contrast with running mutation operators a fixed number of times, time-based criteria ensure that increasing the computational complexity of our mutation operators is penalized, as those operators will take longer to run [10].

Machine learning practitioners often evaluate models on accuracy alone, as increasing the accuracy of a model can directly produce value for their employer [15]. If we can show that large increases of fairness are obtainable without sacrificing much accuracy, perhaps we can increase the adoption of fair machine learning models.

We evaluated our model training method on two datasets: the German credit dataset [8] and the AIF360 pre-processed COMPAS dataset [14]. Both datasets contain numeric attributes and a binary prediction task. For the German dataset, the task is to predict whether the person described by the data point will default on a loan or not. For the COMPAS dataset, the task is to predict whether a criminal will recidivate. For both of these tasks, there are multiple privileged classes with an intersectional effect. We have chosen to measure bias only across gender in the German dataset and race in the COMPAS dataset, for economy of presentation and the simplicity of the resulting fairness metrics. Gender and race, respectively, are not included as inputs to the models being trained.

On each dataset, we will have to define the success of a machine learning model at the prediction task. We do so by examining the model's accuracy and fairness on data it hasn't seen before. We first split each dataset into a training set, consisting of a random sample of 70% of the data, and a test set, consisting of the remaining 30% of the data. During evolution, we optimize for accuracy and fairness on the training set. At the end, when we evaluate the model's performance, we measure its accuracy and fairness on the test set. To quantify how difficult each dataset is to model, we present in Table 1 accuracies achieved by evolving decision trees for accuracy alone, as compared to decision trees produced by SciKitLearn [17]. Table 1 shows that the models produced by evolutionary computing optimizing just for accuracy perform only slightly worse than those produced by alternative training regimes. The aim of this paper is to show that fair models produced by evolutionary computation perform almost as well as unfair models; there are likely possible improvements to the training methods that will increase accuracy to be more competitive.

Dataset	Our Acc	Benchmark Acc	Our DI	Benchmark DI
Adult Income	.808	.859	2.22	2.64
German Credit	.752	.766	1.01	1.08
COMPAS	.645	.685	1.70	2.05
Taiwan Credit	.817	.827	1.22	1.33

Table 1: Accuracy and disparate impact obtained by evolving models using false positive rate and false negative rate as objectives, compared to a SciKitLearn DecisionTreeClassifier trained with a RandomizedSearchCV. Our reported accuracy is the accuracy on the test set of the decision tree with the highest accuracy on the training set. While our accuracies are lower than the SciKitLearn model, our fairness is higher.

4. Results

Traditional methods of training models only produce one model, while evolutionary computing produces a Pareto frontier. Instead of requiring manual human exploration of the potential models, based on an incomplete understanding of the fairness/accuracy tradeoff space, the human in charge of developing the model can see the Pareto frontier of many possible solutions. When examining the Pareto frontier, it is usually clear how much an increase in accuracy or fairness would cost in the other metric. The entire search space is presented to the practitioner choosing which model to deploy, allowing for decisions to be made based on more than just a few hand-picked data points.

On the COMPAS dataset, it is clear that improving the fairness of models decreases the model's accuracy. The points that are in the bottom-left-most parts of each figure, corresponding to the highest accuracy, are the ones with the most unfairness. However, models with near-perfect fairness only slightly underperform the best models. The results were qualitatively similar for the German dataset, where fair models only slightly lagged unfair models in accuracy.

Figures 1 and 2 look quite similar, because disparate impact and true positive rate ratio are both measures of inequality in the allocation of the positive label. For both measures, an increase in the overall amount of positive predictions allows for a higher value of inequality. This effect is shown by the darker areas in the bottom right of the graph, where there are more positive predictions (and thus a higher false positive rate). Figure 3 shows the opposite effect, because the false negative rate ratio is bounded by the number of total negative predictions.

The value of the overall Theil index must be lower than the maximum entropy for three classes ($\log_2(3) \approx 1.585$) as it is equivalent to the entropy of a series of observations with only three values. And, as noted in [19], the between-group entropy is a small amount (often less than one percent) of the overall entropy when you have few groups, so the small values in Figure 4 are to be

expected.

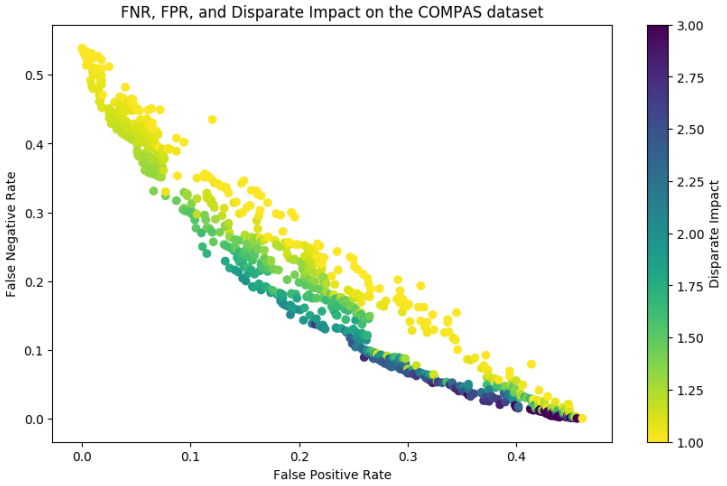

Figure 1: False Negative Rate, False Positive Rate, and Disparate Impact on the COMPAS dataset. Higher false positive rates are correlated with higher disparate impact scores, and higher-accuracy models have worse disparate impact scores. Disparate Impact has a much wider spread than TPR Ratio and FNR Ratio, because there are more positive label predictions than true positives or false negatives.

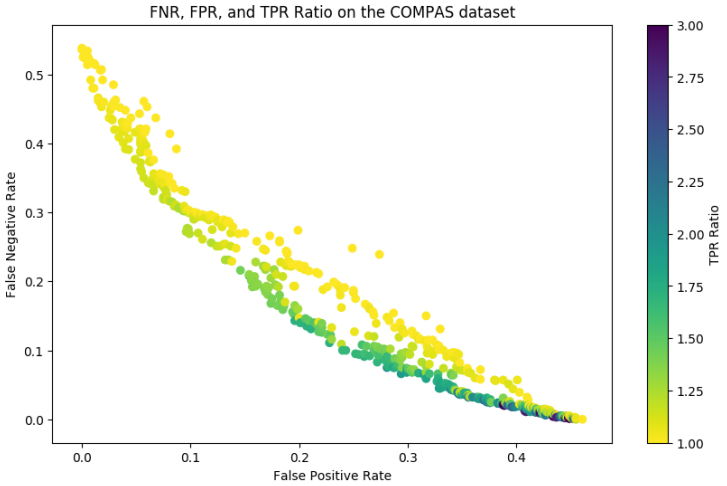

Figure 2: False Negative Rate, False Positive Rate, and True Positive Rate Ratio on the COMPAS dataset. As above, the the true positive rate ratio is highest when the false positive rate is highest.

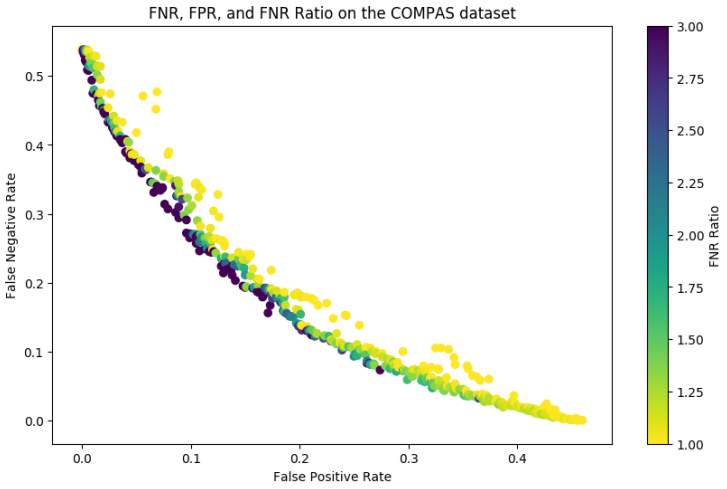

Figure 3: False Negative Rate, False Positive Rate, and False Negative Rate Ratio on the COMPAS dataset. This figure shows the opposite trend as Figures 1 and 2, as false negative rate ratio is naturally correlated with false negative rate and not false positive rate.

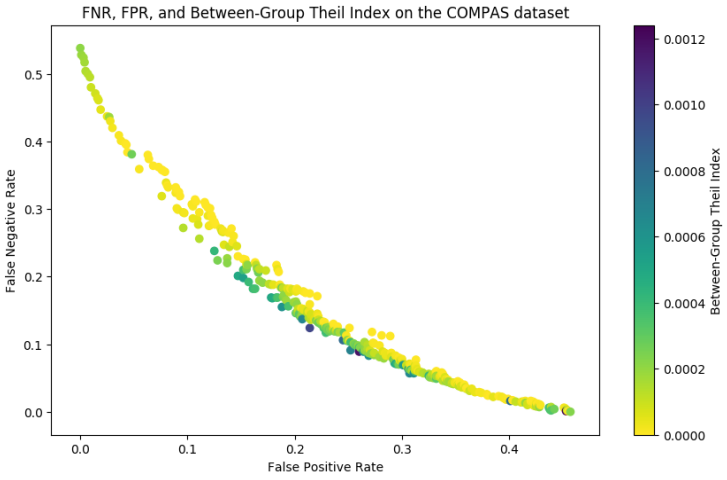

Figure 4: False Negative Rate, False Positive Rate, and Between-Group Theil Index on the COMPAS dataset. Between-Group Theil Index is the only fairness metric that isn't correlated with overall false positive and negative rates.

In Figures 5 and 6 we present the Pareto frontier for two pairs of objectives. Visualizations of the tradeoffs for the other four pairs are available in the Appendix. These results were obtained using false positive rate, false negative rate, and a given pair of metrics as the training objectives. Examination of these Pareto frontiers reveals which metrics conflict with each other (for

example, Between-Group Theil Index and Disparate Impact), and which are obtainable at the same time (for example, TPR Ratio and Disparate Impact). Between-Group Theil Index and Disparate Impact conflict because the Theil index requires that benefits be exactly evenly apportioned between groups, regardless of the labels on the data points in the test dataset, while the disparate impact measure requires the proportion of positive predictions for each group to scale with the proportion of positive labels in the test dataset for that group. True Positive Rate Ratio and Disparate Impact, on the other hand, are both measures of fairness of the distribution of positive labels, so they both can be satisfied at the same time.

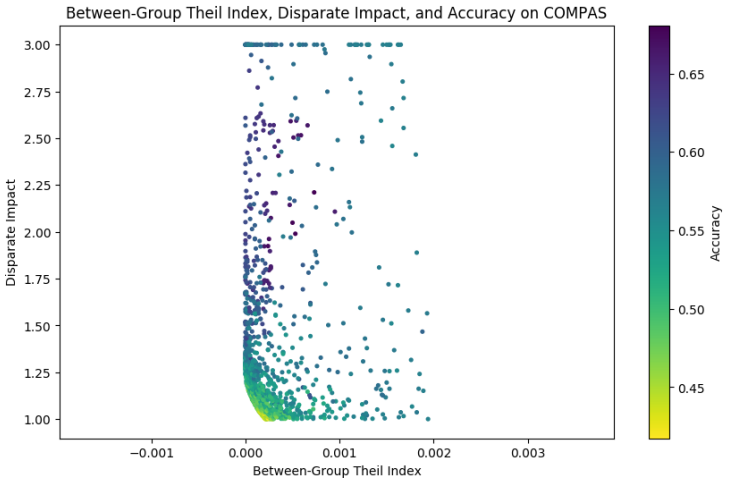

Figure 5: Theil Index and Disparate Impact are anti-correlated, as evidenced by the lack of points in the bottom-left corner. It is also clear from this figure that accuracy suffers as these fairness metrics increase.

Fairness Threshold	TPR Ratio	FNR Ratio	Disparate Impact
1.00	0.580	0.633	0.539
1.05	0.653	0.660	0.568
1.10	0.653	0.660	0.634
1.20	0.653	0.660	0.634
1.30	0.653	0.660	0.634
1.40	0.654	0.660	0.634
1.50	0.658	0.660	0.635
2.00	0.661	0.662	0.653
3.00	0.661	0.667	0.657

Table 2: Peak accuracy for models that achieved different levels of each fairness metric on the COMPAS dataset. Only models with fairness less than or equal to the fairness threshold are included in the accuracy calculation.

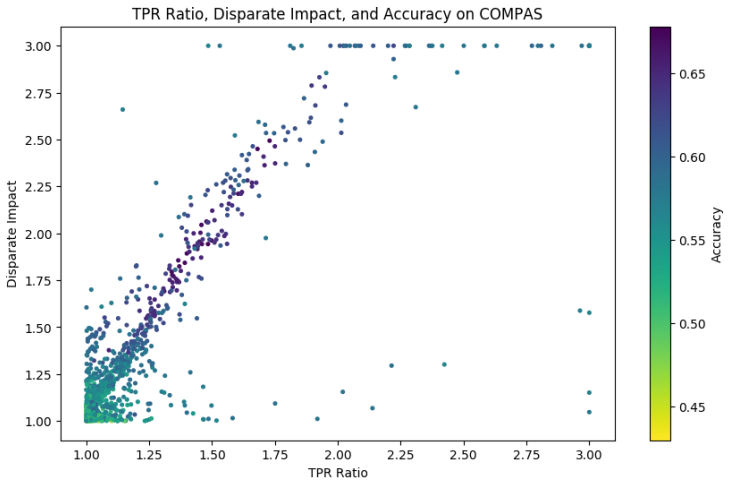

Figure 6: Disparate Impact and TPR Ratio are correlated, which makes intuitive sense as both measure inequality in distribution of positive labels. As with Figure 5, accuracy is lower when the models are more fair, but the effect is lesser here.

Finally, we will examine the maximum accuracy achieved by fair models trained in each of these cases. Table 2 shows the accuracy of our models on the COMPAS dataset, when bounded by some threshold of a fitness metric. Recall that the unfair model, trained solely for accuracy, achieved an accuracy of 0.684. There are clearly diminishing returns to decreasing fairness, as most

of the accuracy benefits can be obtained at a relatively high level of fairness (namely, 1.1). In fact, a disparate impact of 1.1 falls within the "four-fifths rule" established by Title VII [4], while having an accuracy only 2% lower than the highest accuracy of any model we trained. So, ensuring the legality of making decisions based on machine learning models may require only a 2% decrease in accuracy.

5. Further Work

As mentioned when introducing the datasets, each dataset has multiple protected subgroups, and we debiased against only one. Adding objectives for each protected feature may present a different picture. Another straightforward extension would be the application of this framework to different fairness metrics. In addition, the evolution of different types of models could be explored, to see if types other than decision trees may produce fairer, more accurate models. The AIF360 project provides pre- and post-processing steps for models, to reduce their bias. Evolution of models combine with pre- and post-processing might provide an additional way to reduce the cost of fairness.

6. Conclusion

We have shown that fair models can be trained with only a minor loss of accuracy. This result holds across multiple definitions of fairness and multiple datasets. In the face of ethical and legal concerns, this minimal tradeoff may motivate even purely profit-seeking companies to train and deploy fair models. Furthermore, we have used evolutionary computing not only as an optimization technique, but also to produce Pareto frontiers for the exploration of different tradeoffs in fairness. We have explored the three-dimensional Pareto frontiers between false negative, false positive, and different fairness metrics, as well as between pairs of fairness metrics, allowing an exploration of the entire accuracy/fairness tradeoff space.

References

[1] Abellán, J. and Castellano, J. G. (2017). A comparative study on base classifiers in ensemble methods for credit scoring. *Expert Systems with Applications*, 73:1 – 10.

[2] Angwin, J., Larson, J., Mattu, S., and Kirchner, L. (2016). Machine Bias: There's software used across the country to predict future criminals. And it's biased against blacks.

[3] Bao, W., Lianju, N., and Yue, K. (2019). Integration of unsupervised and supervised machine learning algorithms for credit risk assessment. *Expert Systems with Applications*, 128:301 – 315.

[4] Barocas, S. and Selbst, A. D. (2016). Big Data's Disparate Impact. *California Law Review*.

[5] Bellamy, R. K. E., Dey, K., Hind, M., Hoffman, S. C., Houde, S., Kannan, K., Lohia, P., Martino, J., Mehta, S., Mojsilovic, A., Nagar, S., Ramamurthy, K. N., Richards, J. T., Saha, D., Sattigeri, P., Singh, M., Varshney, K. R., and Zhang, Y. (2018). AI fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias. *CoRR*, abs/1810.01943.

[6] Binns, R. (2017). Fairness in machine learning: Lessons from political philosophy. *CoRR*, abs/1712.03586.

[7] Calders, T. and Zliobaite, I. (2013). *Why Unbiased Computational Processes Can Lead to Discriminative Decision Procedures*, pages 43–57.

[8] Dua, D. and Graff, C. (2017). UCI machine learning repository.

[9] Désidéri, J.-A. (2012). Multiple-gradient descent algorithm (mgda) for multiobjective optimization. *Comptes Rendus Mathematique*, 350(5):313 – 318.

[10] Eiben, A. E. and Smith, J. E. (2015). *Introduction to Evolutionary Computing*. Springer Publishing Company, Incorporated, 2nd edition.

[11] Friedler, S. A., Scheidegger, C., Venkatasubramanian, S., Choudhary, S., Hamilton, E. P., and Roth, D. (2019). A comparative study of fairness-enhancing interventions in machine learning. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, pages 329–338, New York, NY, USA. ACM.

[12] Kleinberg, J. M., Mullainathan, S., and Raghavan, M. (2016). Inherent trade-offs in the fair determination of risk scores. *CoRR*, abs/1609.05807.

[13] Kretowski, M. and Grzes, M. (2005). Global learning of decision trees by an evolutionary algorithm. *Inform Process Security Syst*, pages 401–410.

[14] Larson, J., Mattu, S., Kirchner, L., and Angwin, J. (2016). How We Analyzed the COMPAS Recidivism Algorithm. Technical report, ProPublica.

[15] Packer, B., Mitchell, M., Guajardo-Céspedes, M., and Halpern, Y. (2018). Text embeddings contain bias. here's why that matters. Technical report, Google.

[16] Papagelis, A. and Kalles, D. (2000). Ga tree: genetically evolved decision trees. In *Proceedings 12th IEEE Internationals Conference on Tools with Artificial Intelligence. ICTAI 2000*, pages 203–206.

[17] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

[18] Shen, F., Zhao, X., Li, Z., Li, K., and Meng, Z. (2019). A novel ensemble classification model based on neural networks and a classifier optimisation technique for imbalanced credit risk evaluation. *Physica A: Statistical Mechanics and its Applications*, 526:121073.

[19] Speicher, T., Heidari, H., Grgic-Hlaca, N., Gummadi, K. P., Singla, A., Weller, A., and Zafar, M. B. (2018). A unified approach to quantifying algorithmic unfairness: Measuring individual & group unfairness via inequality indices. *CoRR*, abs/1807.00787.

[20] Suresh, H. and Guttag, J. V. (2019). A framework for understanding unintended consequences of machine learning. *CoRR*, abs/1901.10002.

[21] Torralba, A. and Efros, A. A. (2011). Unbiased look at dataset bias. *Proc. of IEEE Computer Vision and Pattern Recognition, 2011*, pages 1521–1528.

[22] Zerbinati, A., Desideri, J.-A., and Duvigneau, R. (2011). Comparison between MGDA and PAES for Multi-Objective Optimization. Research Report RR-7667, INRIA.

[23] Zhao, H. (2007). A multi-objective genetic programming approach to developing pareto optimal decision trees. *Decision Support Systems*, 43(3):809 – 826. Integrated Decision Support.

7. Appendix

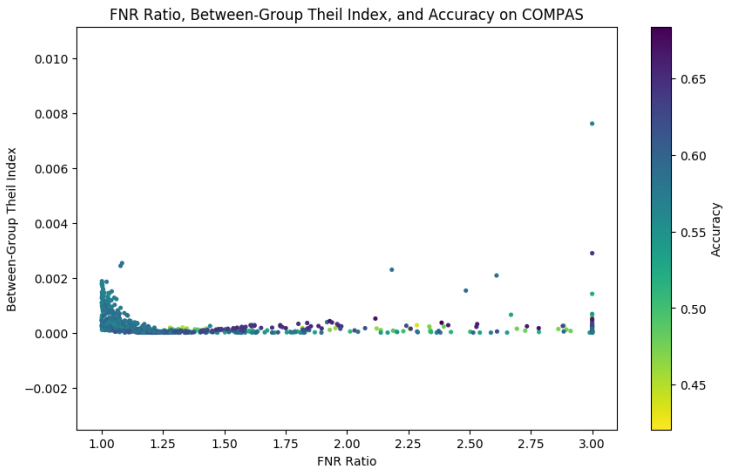

Figure 7: Between-Group Theil Index and False Negative Rate Ratio are can be optimized jointly, with little cost to predictive accuracy.

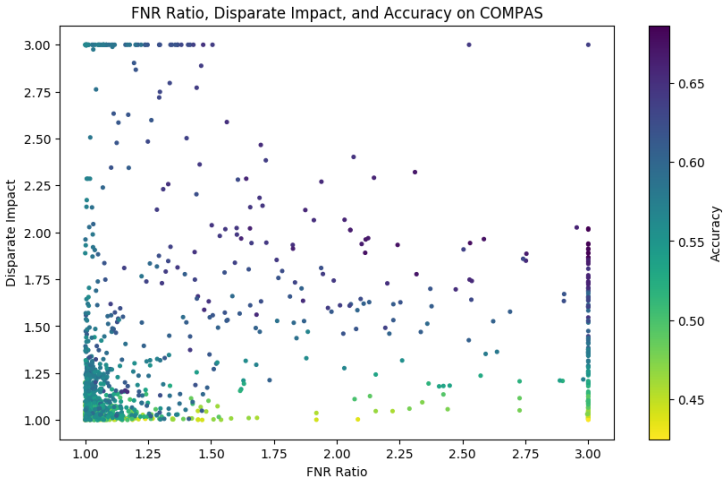

Figure 8: Decreasing Disparate Impact has a large influence on accuracy, while False Negative Rate Ratio has less of an affect.

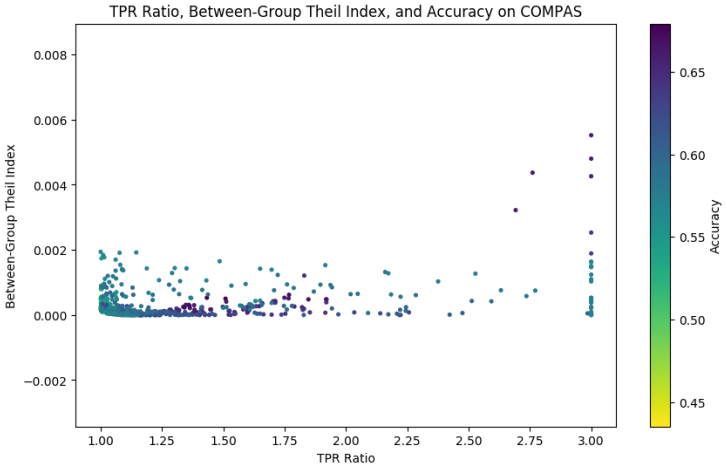

Figure 9: As with False Negative Rate Ratio in Figure 7, accuracy is relatively high, and models can be fair by both metrics.

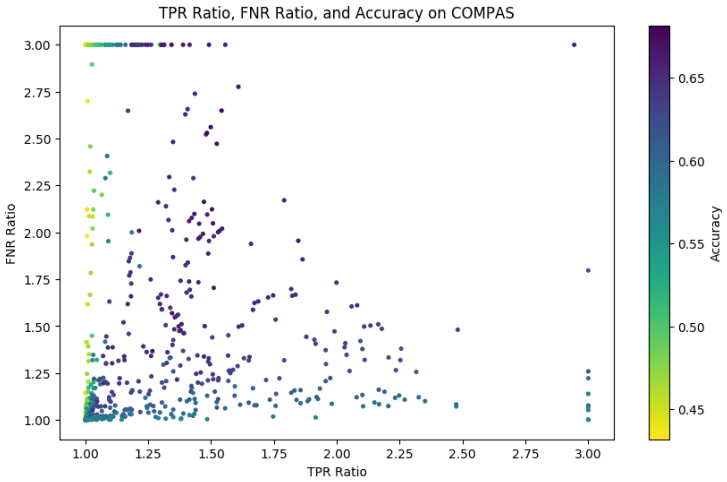

Figure 10: Here, the cost of True Positive Rate Ratio in accuracy is very high, although some points quite near (1,1) represent models that are incredibly fair, and still relatively accurate.