Evaluating the robustness of group fairness penalty on biased datasets

Mads Høgenhaug mkrh@itu.dk, Marcus Friis mahf@itu.dk, Mia Pugholm mipu@itu.dk
IT University of Copenhagen, May 19, 2023

1 Abstract

Fairness in machine learning has received increasing attention as modern methods have gotten better. Along with this, new methods for combating model bias in-processing has been proposed. In this paper, we build on research by Berk et al. [1], and look into the application and robustness of group fairness as a constraint. We test the method in single- and multiple group settings on *adult income* and *student dropout* data. We find that the originally proposed method does not work in our setting and propose an altered version, producing similar results as Berk et al. reports. We validate that as we increase the strength of the group fairness constraint, we sacrifice performance for fairness, both in single group settings as well as when constraining on multiple groups.
All code created and used for the project can be found at :
https://github.com/Marcus-Friis/afae-exam

2 Introduction

As machine learning has evolved, an increasing effort has been put into making algorithms fair. Algorithmic decision-making can range from seemingly trivial everyday applications to life-changing decisions, such as whether someone gets a loan, admitted to school, a job, etc. Particularly, as more algorithms are used in production systems that influence human lives, an increasing need for fairness rises. However, debiasing and making algorithms fair is no trivial task, and many methods have been presented in the pursuit of fairness. Bias and prejudice is often rooted in data, which is problematic when algorithms reflect this bias, and analysing the source of this bias is key to understanding the problem at hand [2].

Debiasing machine learning models is typically done either with pre- post- or in-processing methods. This paper focuses on the last method; how in-processing methods can be used to make models more fair. Specifically, we focus on the constrained optimization method "Group Fairness". Group fairness aims to ensure that the model's predictions are consistent across different groups of individuals, which means that the groups should have equal probability of being assigned to the positive predicted class. In this paper we seek to reproduce and extend the work of Berk et al. in their paper *'A Convex Framework for Fair Regression'* [1] by investigating the impact of adding group penalties on both single and multiple groups. Specifically, we will explore the trade-offs between accuracy

and fairness when multiple groups are considered and compare the performance of the model with a baseline and the results presented in *'A Convex Framework for Fair Regression'*.

3 Related work

When working with fairness, we first need to establish *"what is fair?"*. There are several definitions of fairness in the machine learning literature. Different fairness notions can in general be categorized as either group, subgroup, or individual fairness [3]. In this paper, we focus on a well-known instance of group fairness called equalized odds [4]. Equalized odds requires that an algorithm predicts equal true positive rates and equal false positive rates across a group. Other fairness definitions such as demographic parity [5], treatment equality [6], fairness through awareness [7] etc. could be adopted and used. However, for this project, we use equalized odds as a working definition and examine the fairness of our models through its' lens, as it has been recognized for its ability to achieve a high level of algorithmic fairness [8].

To achieve a fairness criteria, modern machine learning methods are typically divided into pre-, in- and post-processing methods [9]. Much research has gone into investigating different pre- [10, 11] and post-processing [12, 13] methods. This paper focuses on modern in-processing methods. In-processing methods usually frame the goal of fairness as an optimization issue, by using fairness penalties as an optimization constraint [1, 14–16]. These methods have been proven to achieve decent to good results, and we focus on a specific implementation of such a constraint.

3.1 Key paper

Our research is grounded in the work done by Berk et al. [1]. In this research, they propose 3 fairness penalties: an individual fairness penalty, a group fairness penalty, and a hybrid penalty. The individual fairness penalty enforces that a model should be fair for each individual observation. In contrast, the group fairness penalty allows for individual unfairness as long as the model compensates on other predictions. The hybrid method combines these two penalties by summing them. Their methods work in both regression and classification settings. The authors evaluate the effectiveness of their constraints on 6 datasets across different models. Our work aims to partly replicate and extend their research, specifically by focusing on the application and robustness of the group fairness penalty in binary classification settings.

4 Data

To understand the group fairness penalty's [1] robustness, we need data to train models on to facilitate this evaluation. For this project, we use two datasets; a *student dropout*, and an *adult income* dataset.

Both are real-life datasets, where it could be imagined an algorithm would predict an outcome on people, and thus, it needs to be fair. The criteria for choosing these datasets is that both datasets support a similar binary prediction task, contain data about humans, and contain protected attributes and biases towards them. Depending on the definition of a protected class, different features could qualify [17]. For each dataset, we will not treat all features that could be considered protected as protected, but we will choose a subset for simplicity's sake. Plenty of exploratory data analysis have been done for this project. There are two separate notebooks on the GitHub repository, one for each dataset. Moreover, a few plots have been selected from the notebooks and added to the appendix B section.

4.1 Student dropout

The first dataset is the student dropout dataset, originally called *"Predict students' dropout and academic success"* [18]. The data contains information about students, including demographic information, academic performance and whether the student is currently enrolled, graduated or dropped out of college. The goal of the classification task is to predict whether a student will graduate or dropout. As such, we drop all rows of current enrolled students, as they are not interesting in our classification task. After dropping enrolled students, the data consists of 3630 rows and 35 columns, containing information such as "Marital status", "Mother's occupation", "Previous qualifications" etc. We treat "Age at enrollment", "Nationality" and "Gender" as protected, which have to be considered when making models fair.

4.2 Adult income

The second dataset is the *adult income* dataset, also commonly referred to as *"adult"* [19]. This dataset contains information about descriptors of people across the United States, and whether they have a yearly salary above or below $\$50,000$. It has 48842 rows and 15 columns. Similar to the student dropout dataset, this data contains various demographic information such as "Marital status", "Workclass", "Education", etc. For this data, we work with the protected features "race", "gender", and "age".

5 Methodology

The goal of our research is to analyse and evaluate the robustness of the group fairness constraint [1]. Specifically, we constrain logistic regression classifiers to mitigate biases in otherwise biased models. Berk et al. [1] also do this, but they do not report the uncertainty of their results, and they only apply it in single group settings. While this method yields good results for them, it is unrealistic to strictly

apply it to one group, since in many real world applications, there are multiple protected attributes to ensure fairness for. For instance, in the *adult income* dataset, we have information about gender, age and race, and the dataset contains biases towards these groups. Our goal is to recreate their results by comparing a constrained model to a baseline, and to test the method in a more realistic setting by penalizing multiple groups at a time. In other words, we create and evaluate 3 models for each dataset: (i) a baseline binary logistic regression model without any fairness constraints. (ii) A gender-fair logistic regression model, constrained on the group fairness for gender. (iii) A triple constrained logistic regression model, constrained on gender, race/nationality and age. The focus is thoroughly testing the group fairness constraint.

5.1 Original group fairness constraint

From [1], we formulate the group fairness constraint as follows:

$$f_2(\mathbf{w}, S) = \left(\frac{1}{n_1 n_2} \sum_{\substack{(\mathbf{x}_i, y_i) \in S_1 \\ (\mathbf{x}_j, y_j) \in S_2}} d\left(y_i, y_j\right) \left(\mathbf{w} \cdot \mathbf{x}_i - \mathbf{w} \cdot \mathbf{x}_j\right) \right)^2$$

where S_1 and S_2 are all observations from group 1 and 2 respectively, $d(y_i, y_j)$ is a distance function, $\mathbf{w}$ is the model weights, and $n_1 = |S_1|$ and $n_2 = |S_2|$. The distance function, in this binary setting, is defined as

$$d(y_i, y_j) = \begin{Bmatrix} y_i = y_j & 1 \\ y_i \neq y_j & 0 \end{Bmatrix}$$

Informally, the constraint sums over all cross pairs from two groups, and punishes a model when the predictions between groups are dissimilar. It does not punish individual instances, since if the model predicts in favor of one group on a cross pair, it can "correct" this bias by predicting oppositely on a different cross pair, thus achieving group fairness. It is implemented like any other regularization function, and the creators use it with l2-regularization. As such, the optimization goal becomes

$$\min_{\mathbf{w}} \ \ell + \gamma f_2(\mathbf{w}, S) + \lambda ||\mathbf{w}||_2$$

When we faithfully implement this constraint and apply it on an optimization task, we experienced that it did not work for us. We experienced 2 issues: it was extremely slow to compute, and the resulting fair losses were exceptionally unstable. In an initial analysis, we did a parameter gridsearch for finding the optimal regularization strength γ on the *adult income* data, and the resulting figure 1 showed that no matter the regularization strength, the model converged to always classifying negative. Upon further inspection, we saw that the batch losses were wildly unstable as seen in figure 2.

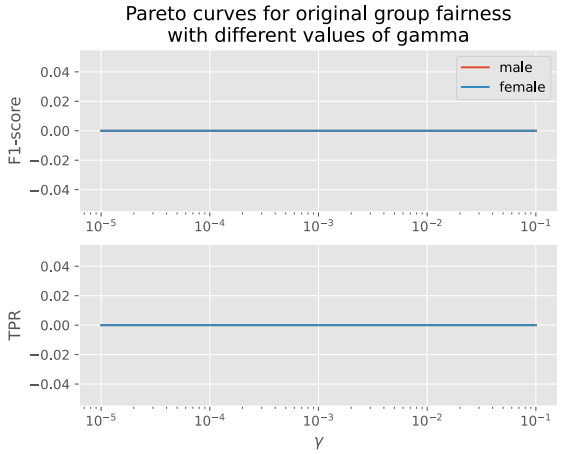

Figure 1: Plot of how the f1-score and true positive ratio changes for different regularization strengths for the original group fairness penalty definition. It does not work for us, therefore, we see the classifier strictly assign all observations as negative.

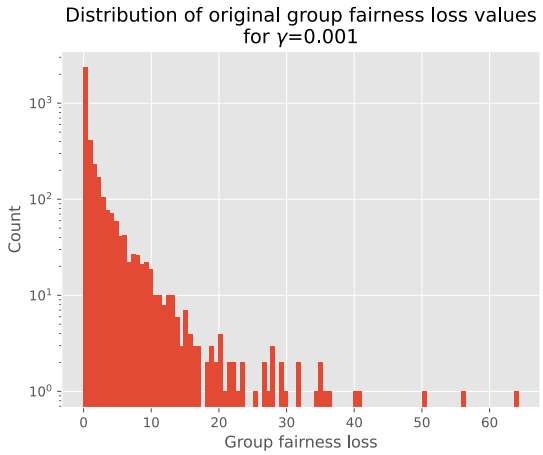

Figure 2: Histogram of fair penalty batch losses for the original group fairness formulation. Note how, even though most losses are lower than 1, sometimes, the penalty gets exceedingly large. This results in an unstable training process.

5.2 Modified group fairness constraint

Since the original group fairness constraint did not work, we modified it. We made two changes; firstly, instead of using the logits of the logistic regressor $\mathbf{w}x$ for punishing dissimilarity of two predictions, we use the logistic regressions' predicted probability $\frac{1}{1+e^{-(\mathbf{w}x)}}$. Additionally, the cross pair normalization term $\frac{1}{n_1 n_2}$ was moved outside of the parenthesis to achieve better loss values and align more closely to other normalized cost functions such as mean squared error. With the implemented changes, we let $g_{\mathbf{w}}(x) = P(Y = 1|x)$ denote the predicted positive probability, and the formulation of the penalty becomes

$$f_2(\mathbf{w}, S) = \frac{1}{n_1 n_2} \left(\sum_{\substack{(\mathbf{x}_i, y_i) \in S_1 \\ (\mathbf{x}_j, y_j) \in S_2}} d\left(y_i, y_j\right) \left(g_{\mathbf{w}}(\mathbf{x}_i) - g_{\mathbf{w}}(\mathbf{x}_j)\right) \right)^2$$

This formulation yields significantly more stable losses (see figure 3) and provides a massive performance speedup. As such, we use this definition of the group fairness penalty for the rest of this project.

5.3 Experiment setup

With the group fairness penalty defined, we construct an experiment for evaluating the robustness and applicability of said method. We preprocess our data as follows: we onehot encode all categorical features and standard scale numerical features. We do not split our data into a train-test split, but instead use k-fold cross validation for all models such that we test the robustness of the evaluated

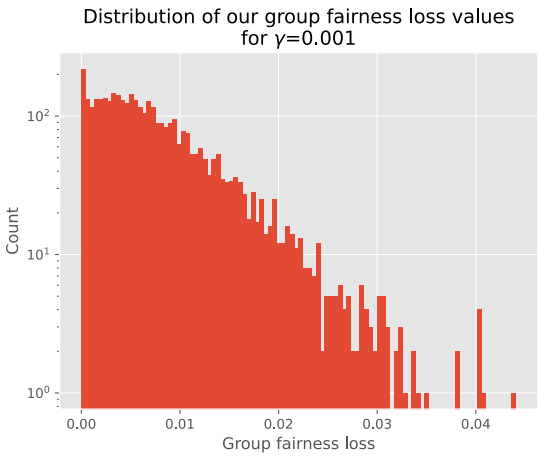

Figure 3: Histogram of fair penalty batch losses for the new group fairness penalty. Note how the values are significantly less scattered compared to figure 2.

method. On all models, we evaluate the performance with accuracy, balanced accuracy, f1-score, and ROC AUC. We evaluate the fairness under equalized odds by looking at true positive and false positive ratios across groups. Lastly, we inspect the feature importance in each model. We do this by looking at the odds ratios [20]. We achieve this by raising our model's weights to the power of e, which in logistic regression translates to feature importances represented as odds ratios.

As for the actual experiment, we start with **(i) defining a baseline model**. Our baseline we build our experiment upon is a simple binary logistic regression model. As per the instructions of Berk et al. [1], we use l2-regularization on this model. We pick an arbitrary regularization strength $\lambda = 0.01$, and do not experiment with any other values as to not over-complicate the task.
We then **(ii) apply group fairness penalty** on a logistic regression model. For this part, we only enforce the fairness criteria on a single group. Specifically, we apply it on gender on both datasets.
Afterwards, we **(iii) scale the experiment**. We apply the group fairness penalty on gender, race/nationality, and age, and evaluate the impact on the model. Especially since Berk et al. never tested the use of multiple group fairness penalties for debiasing multiple groups.

6 Results

The results we report are the performance, true positive and false positive ratios (related to equalized odds), pareto curves for optimal γ value, and feature importances of the trained models.

Our final three models on both the *adult income* and *student dropout* data are evaluated using our chosen performance metrics in figure 4 and 5. We see that, as expected, increasingly constraining models on fairness leads to a performance drop, illustrating the performance-fairness tradeoff.

For finding the optimal regularization strengths for the fairness constraint, we plot our resulting pareto curves in figure 6 and 7 for the gender constrained model, and 8 and 9 for the triple constrained

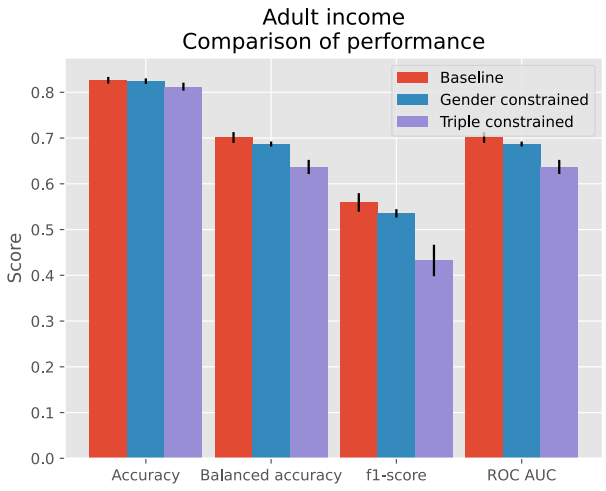

Figure 4: Barplot showcasing the performance over various metrics of the three created models for the *adult income* data. Note how, as expected, we see that as models become more costrained on fairness, we sacrifice performance.

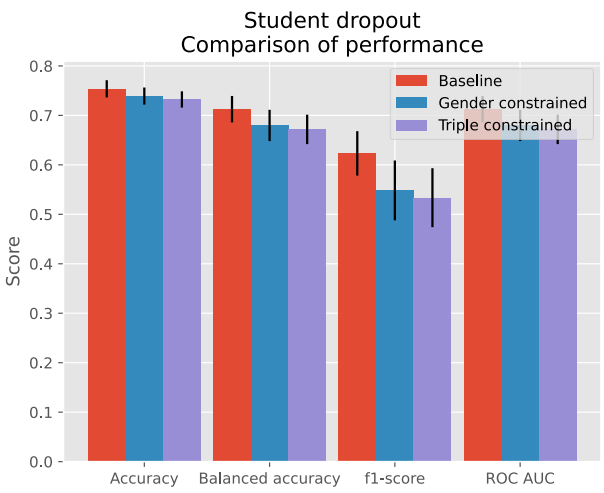

Figure 5: Barplot showcasing the performance over various metrics of the three created models for the *student dropout* data. Similarly to figure 4, we see the expected trade-off between performance and fairness.

model. The *adult income* models generally showed γ values where the model reaches a good trade-off between fairness and performance in contrast to the *student dropout* data.

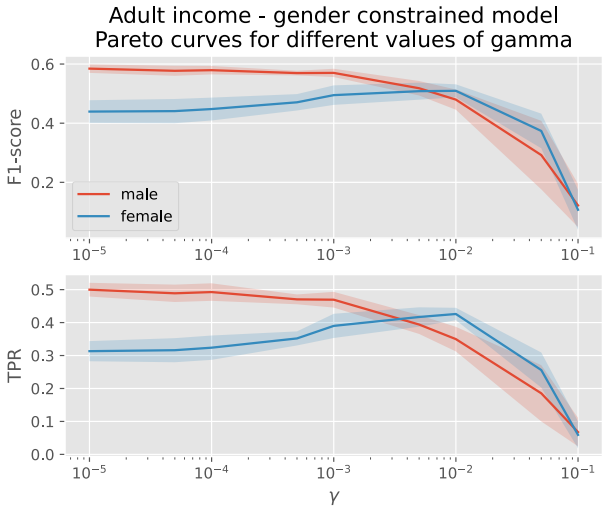

Figure 6: As the γ value increases, the Pareto curves show a decrease in both the F1-score and true positive rate for males. The pareto curve for female increases and decrease towards the end.

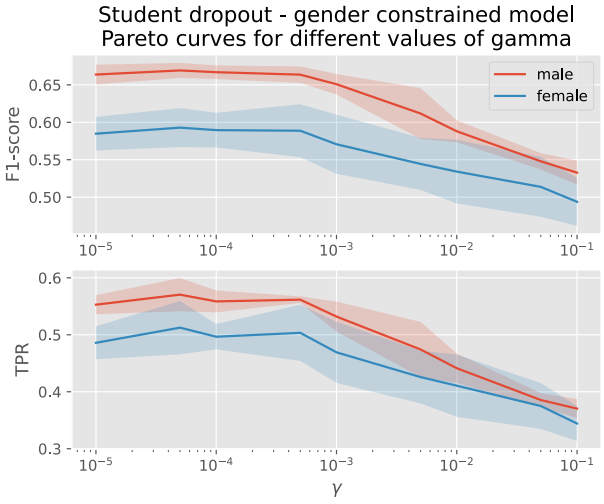

Figure 7: As the γ value increases, the model slowly converges towards equalized opportunity at the cost of performance dropping as well. Unlike for the *adult income* models in figure 6, there is no clear convergence point nor optimal γ value.

We plot the TPR and FPR for each group for each model to show how closely the models fulfill equalized odds. For the *adult income* models, the trend is models become fairer as we enforce more fairness. We also see that as models become more fair, the general positive rates decrease. The *student dropout* results are not as good. The nationality group is more uncertain (likely due to a imbalanced

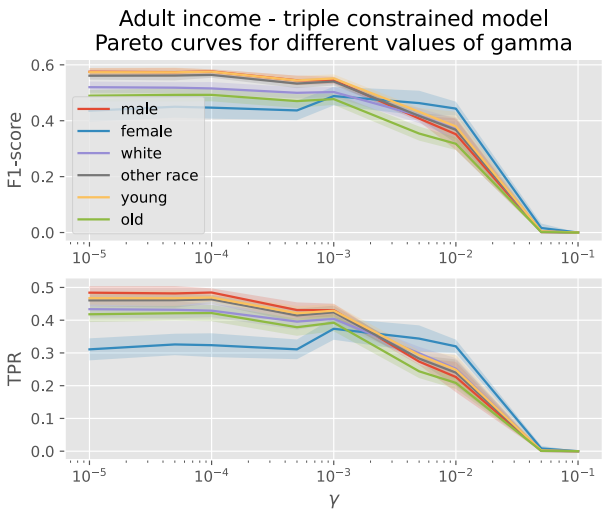

Figure 8: As γ increases, the model converges towards the most fair solution; classifying everything as negative. Along the way, the gap between the TPR of groups gets narrower at the cost of decreased performance.

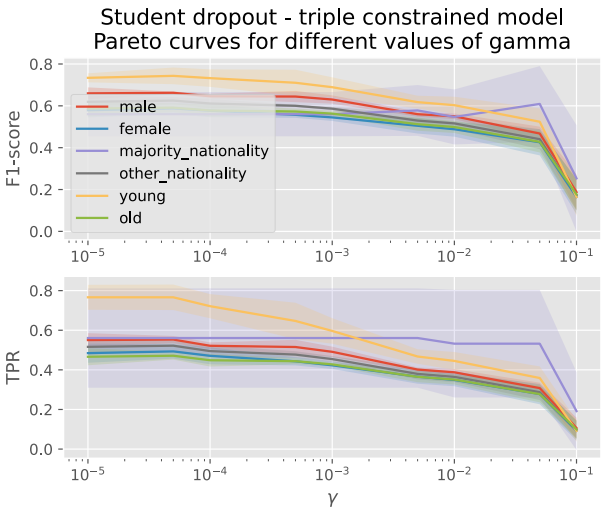

Figure 9: As γ increases, the model converges towards the most fair solution; classifying everything as negative. Along the way, the gap between the TPR of groups gets narrower at the cost of decreased performance. Additionally, we notice a extremely large uncertainty for the nationalities. This is most like due to the fact that nationalities have become binary, with 3544 instances in group 1 and 86 instances in group 2

distribution and less data), but the general trend from before is still there. As we increase fairness, we get closer to equalized odds across all 3 groups.

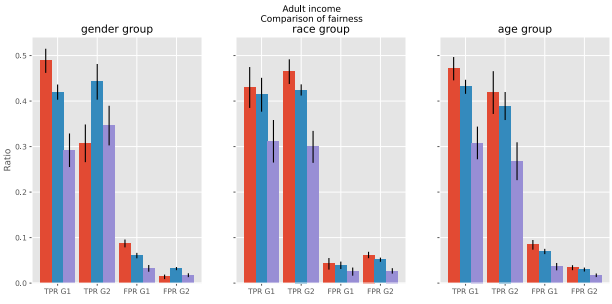

Figure 10: We show the various performance metrics across the 3 different models, for each of the protected attributes on the *adult income* dataset. The color scheme follows the one in figure 4 & 5.

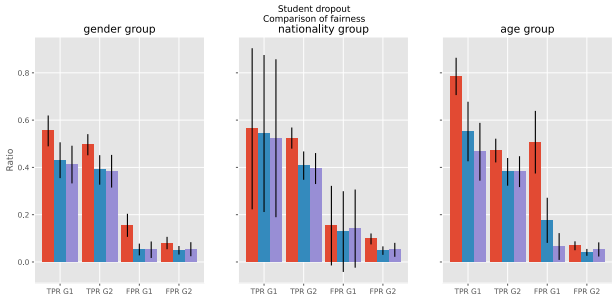

Figure 11: We show the various performance metrics across the 3 different models, for each of the protected attributes on the *student dropout* dataset. The color scheme follows the one in figure 4 & 5.

Lastly, we plot the feature importance of each model to show how the decision process changes as we enforce more fairness. As seen in appendix figure 12, for the *adult income* data, as we punish models on gender bias, the model reweighs the importance of highly gender correlated features (see appendix figure 14) such as "marital status - wife" and "marital status - husband". The results are not as clear for *student dropout* models in figure 13, and it is harder to build an intuition on what it is

changing.

7 Discussion

The datasets we have investigated in this paper can resemble real-world problems that can be addressed through algorithmic solutions.

With the *student dropout* dataset, we saw that by using various features such as prior academic performance, demographic attributes, and other relevant factors, we can predict with some accuracy the likelihood of a student dropping out or achieving academic success. This could be used to identify students who may require additional assistance or who should be granted scholarships or the like. The *adult income* dataset with features such as education, occupation, and other relevant variables, made it possible to predict whether individuals have a yearly salary above or below $\$50,000$. This could be used to assess the creditworthiness of individuals and make decisions regarding loan approvals.

These real-world problems show high complexity, involving relationships among numerous variables. In this study we have seen how leveraging the power of machine learning algorithms, can extract complex patterns in the data and make accurate predictions. These algorithms learn from the data and if the training data is biased the resulting models can unintentionally perpetuate and amplify those biases. We have seen how enforcing fairness prevent this. However, had we chosen a more complex machine learning model such as neural network, the complexity of the model would make it difficult to interpret how the predictions were made.

This challenge contrasts with the perspective of traditional symbolic artificial intelligence, often referred to as Good Old-Fashioned Artificial Intelligence (GOFAI) [21]. In GOFAI, systems are built using explicit rules and logic, making it possible to trace the decision-making process. These systems are transparent, as their inner workings can be examined and reasoned about. However while GOFAI may provide more transparency and interpretability, it still requires human designers who have expertise in the dataset to explicitly encode fairness considerations, which introduces a risk of bias due to the subjective interpretation.

In our case, considering the case of predicting student dropout, the datasets consists of numerous features. Additionally, there is to our knowledge no apparent reason within the data that can reliably determine whether a student will drop out or not, making it challenging to establish explicit rules as required by GOFAI. This specific problem benefits from being solved with modern machine learning techniques as it can capture patterns in the dataset beyond what GOFAI is capable of.

Another thing to consider when making algorithmic solutions for real-world problems is to consider ethical approaches and how the model should inherit the moral values they represent. Following

the ethics of utilitarianism we should seek to make decisions that results in the greatest good for the greatest number of people while minimizing suffering [22]. With utilitarianism we consider the negative impacts of biases on marginalized groups and seek fairness in order to maximizing wellfare. However, pursuit of overall welfare may in some cases lead to the exclusion of certain groups instead of inclusion if this benefits the overall wellfare more. Considering the ethic approach of deontology we establish a set of rules that we have a moral duty to follow [23]. Deontological ethics recognizes the moral duty to refrain from discrimination using protected personal traits, as mandated by the law [17]. Unlike utilitarianism, which considers the outcome when determining moral rightness, deontology prioritizes moral duties regardless of the consequences of the decision. In the context of our study, when considering the trade-off between fairness and model accuracy, fairness should be prioritized over accuracy in order to follow the moral law. The ethical approach of virtue ethics would similarly not look at the consequences of an action, but instead of solely relying on legal obligations, virtue ethics encourages individuals to embody these virtues in their actions and decisions [24].

Finding a balance between fairness, accuracy and other ethical considerations remains a complex challenge that necessitates careful consideration and ongoing refinement of machine learning models and their deployment.

8 Conclusion

Investigating the robustness and application of group fairness, we found that the original definition as proposed by Berk et al. [1] did not work. Instead, we propose a slightly altered method, that shifts from using the logits of the model to using the output probabilities, and evaluate this changed method instead. On our two datasets, we find different levels of effectiveness. On one dataset, we see robust results and validate the classic trade-off between fairness and performance, but manage to successfully constrain models in both single group as well as multi-group settings. The change in decision process aligns with our expectations based on the preliminary exploratory data analysis, and the models change their behaviour on features correlated with the constrained group. We do not achieve the same results for the other dataset, and instead achieve more ambiguous and less robust results, likely due to lower data quantities and more sparsely distributed features.

References

[1] Richard Berk, Hoda Heidari, Shahin Jabbari, Matthew Joseph, Michael Kearns, Jamie Morgenstern, Seth Neel, and Aaron Roth. A convex framework for fair regression. *arXiv preprint arXiv:1706.02409*, 2017. URL https://arxiv.org/abs/1706.02409.

[2] Harini Suresh and John V Guttag. A framework for understanding unintended consequences of machine learning. *arXiv preprint arXiv:1901.10002*, 2(8), 2019.

[3] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)*, 54(6):1–35, 2021.

[4] Moritz Hardt, Eric Price, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016. URL https://proceedings.neurips.cc/paper_files/paper/2016/file/9d2682367c3935defcb1f9e247a97c0d-Paper.pdf.

[5] Alekh Agarwal, Miroslav Dudik, and Zhiwei Steven Wu. Fair regression: Quantitative definitions and reduction-based algorithms. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 120–129. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/agarwal19d.html.

[6] Alexandra Chouldechova. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments, 2016.

[7] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pages 214–226, 2012.

[8] Mike Teodorescu. Exploring fairness in machine learning for international development, 2020. URL https://shorturl.at/gmstY.

[9] Emmanouil Krasanakis, Eleftherios Spyromitros-Xioufis, Symeon Papadopoulos, and Yiannis Kompatsiaris. Adaptive sensitive reweighting to mitigate bias in fairness-aware classification. In *Proceedings of the 2018 world wide web conference*, pages 853–862, 2018.

[10] Zhe Zhang, Shenhang Wang, and Gong Meng. A review on pre-processing methods for fairness in machine learning. *Advances in Natural Computation, Fuzzy Systems and Knowledge Discovery: Proceedings of the ICNC-FSKD 2022*, pages 1185–1191, 2023.

[11] He Lyu, Ningyu Sha, Shuyang Qin, Ming Yan, Yuying Xie, and Rongrong Wang. Advances in neural information processing systems. *Advances in neural information processing systems*, 32, 2019.

[12] Pranay K Lohia, Karthikeyan Natesan Ramamurthy, Manish Bhide, Diptikalyan Saha, Kush R Varshney, and Ruchir Puri. Bias mitigation post-processing for individual and group fairness. In *Icassp 2019-2019 ieee international conference on acoustics, speech and signal processing (icassp)*, pages 2847–2851. IEEE, 2019.

[13] Alan Mishler, Edward H Kennedy, and Alexandra Chouldechova. Fairness in risk assessment instruments: Post-processing to achieve counterfactual equalized odds. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 386–400, 2021.

[14] Toshihiro Kamishima, Shotaro Akaho, Hideki Asoh, and Jun Sakuma. Fairness-aware classifier with prejudice remover regularizer. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2012, Bristol, UK, September 24-28, 2012. Proceedings, Part II 23*, pages 35–50. Springer, 2012.

[15] Bhanu Jain, Manfred Huber, and Ramez Elmasri. Increasing fairness in predictions using bias parity score based loss function regularization. *arXiv preprint arXiv:2111.03638*, 2021.

[16] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rogriguez, and Krishna P Gummadi. Fairness constraints: Mechanisms for fair classification. In *Artificial intelligence and statistics*, pages 962–970. PMLR, 2017.

[17] Cornell Lawschool Legal Information Institute. Protected characteristic. URL https://www.law.cornell.edu/wex/protected_characteristic.

[18] Valentim Realinho, Jorge Machado, Luís Baptista, and Mónica V. Martins. Predict students' dropout and academic success, December 2021. URL https://doi.org/10.5281/zenodo.5777340.

[19] Dheeru Dua and Casey Graff. UCI machine learning repository, 2017. URL http://archive.ics.uci.edu/ml.

[20] J Martin Bland and Douglas G Altman. The odds ratio. *Bmj*, 320(7247):1468, 2000.

[21] John Haugeland. *Artificial Intelligence: The Very Idea*. Cambridge: MIT Press, 1985.

[22] Henry R. West. Utilitarianism. URL https://www.britannica.com/topic/utilitarianism-philosophy.

[23] Larry Alexander and Michael Moore. Deontological Ethics. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Winter 2021 edition, 2021.

[24] Rosalind Hursthouse and Glen Pettigrove. Virtue Ethics. In Edward N. Zalta and Uri Nodelman, editors, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Winter 2022 edition, 2022.

A Feature importance

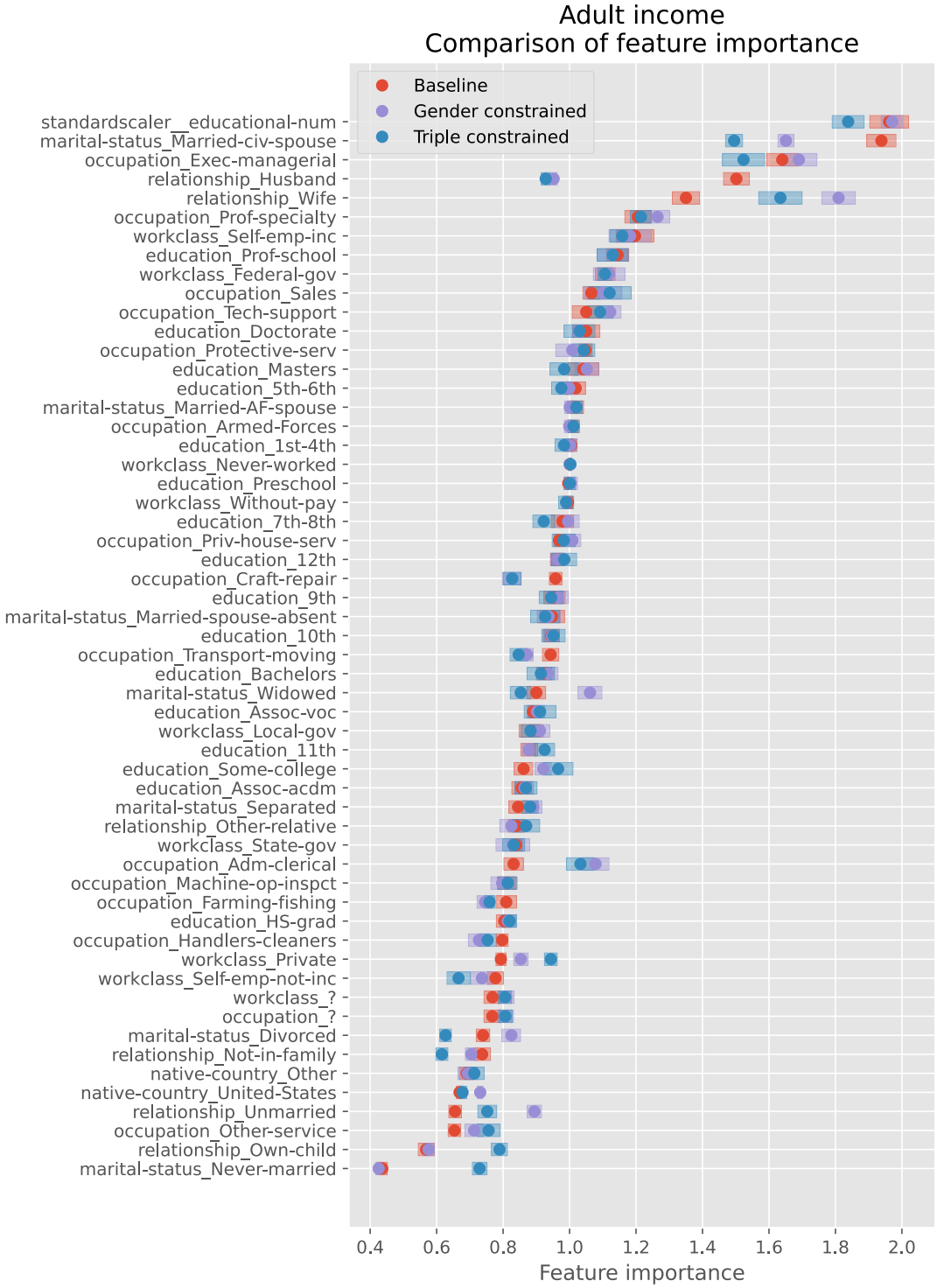

Figure 12: Plot showing the feature importance of 3 logistic regression models, trained on the *adult income* data, constrained increasingly on fairness. The trend is not as clear as we had initially hoped, but in general, we see that once we constrain our models, features correlated with the constrained feature shift their importance. For instance, both the gender constrained and triple constrained models drastically change the importance of husband and wife once we enforce fairness for gender, which we would expect as these are highly correlated with gender.

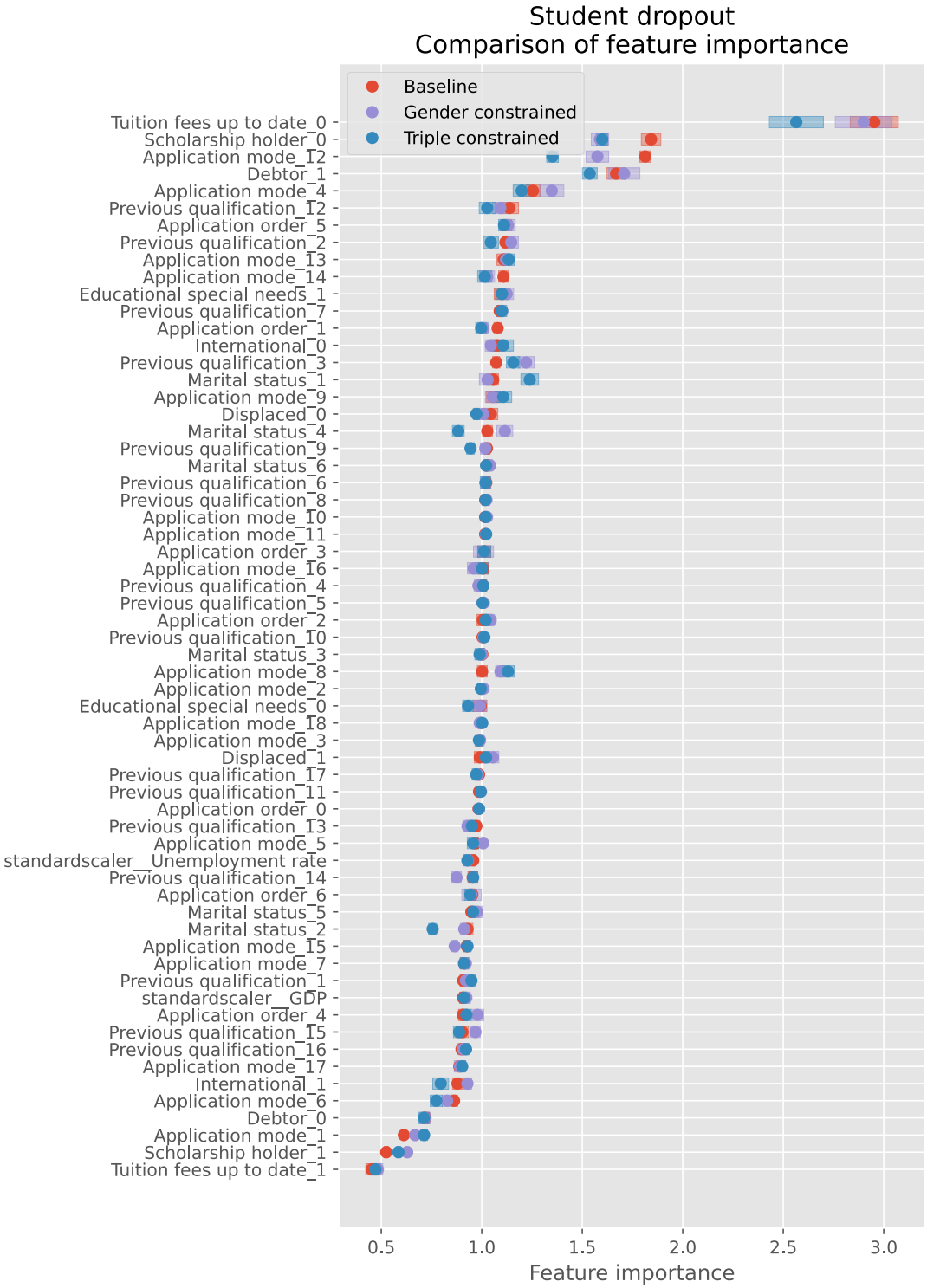

Figure 13: Plot showing the feature importance of 3 logistic regression models, trained on the *student dropout* data, constrained increasingly on fairness. Unlike in figure 12, the pattern is not as easily understandable.

B Highlights of exploratoy data analysis

B.1 Adult income dataset

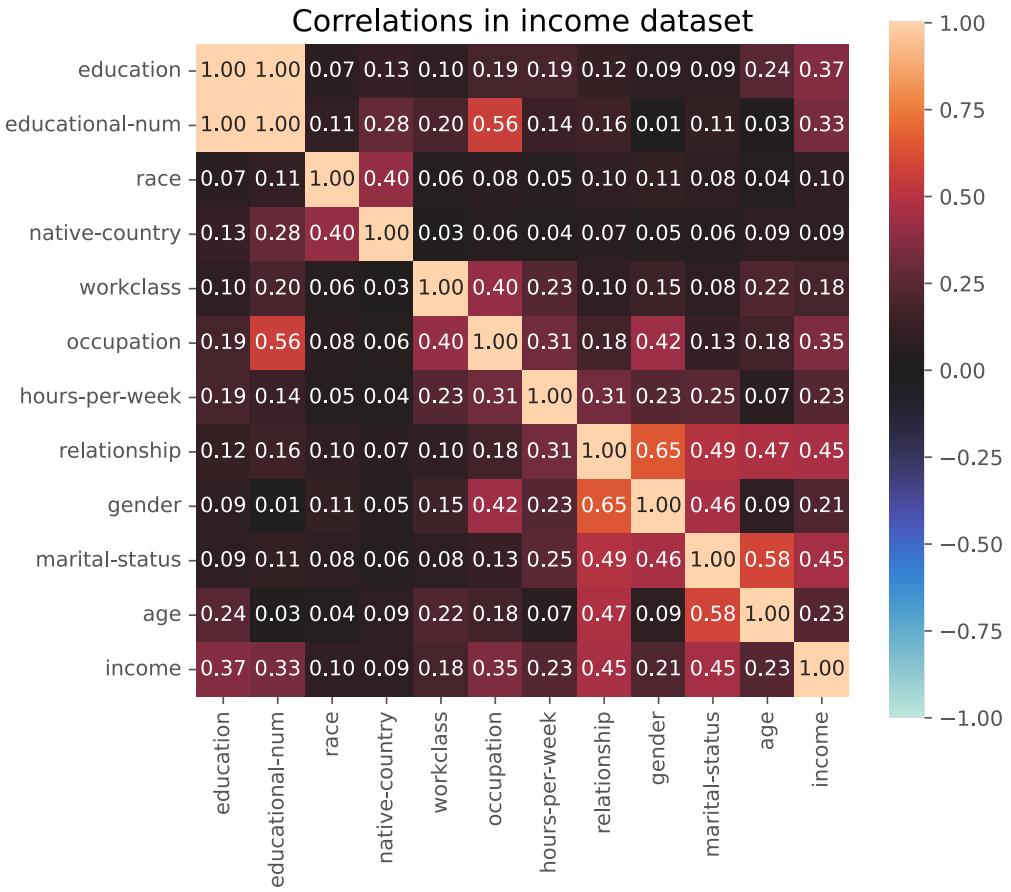

Figure 14: Correlation matrix for the adult income dataset. High correlations between our protected attributes and other features, is an indicator that some of the non-protected attributes are proxies for the protected ones. This means that, even if we constrain one the protected attributes, a proxy for the same attribute could still potentially lead to biased decisions

By investigating the correlation matrix in figure 14 we get an intuition of which features are proxies for the protected attributes. This is important to know, because it can allow us to better understand the underlying patterns and biases in the data that could potentially lead to unfair or discriminatory decisions if not properly managed. A closer look of the matrix shows that the protected attribute 'age' has a correlation of 0.58 & 0.47 with 'marital-status' & 'relationship' respectively. Thus despite removing 'age' as a predictor, the model might still implicitly use age-related information through these two highly correlated features.

Figure 15: Income in relation to age. The general tendency seems to be that people making above \$50.000 is normally distributed with $\mu = 44.27$ years old

Figure 16: Income in relation to gender. This plot shows that males generally have a higher ratio of earning above \$50.000, introducing a bias for the gender attribute.

B.2 Student dropout dataset

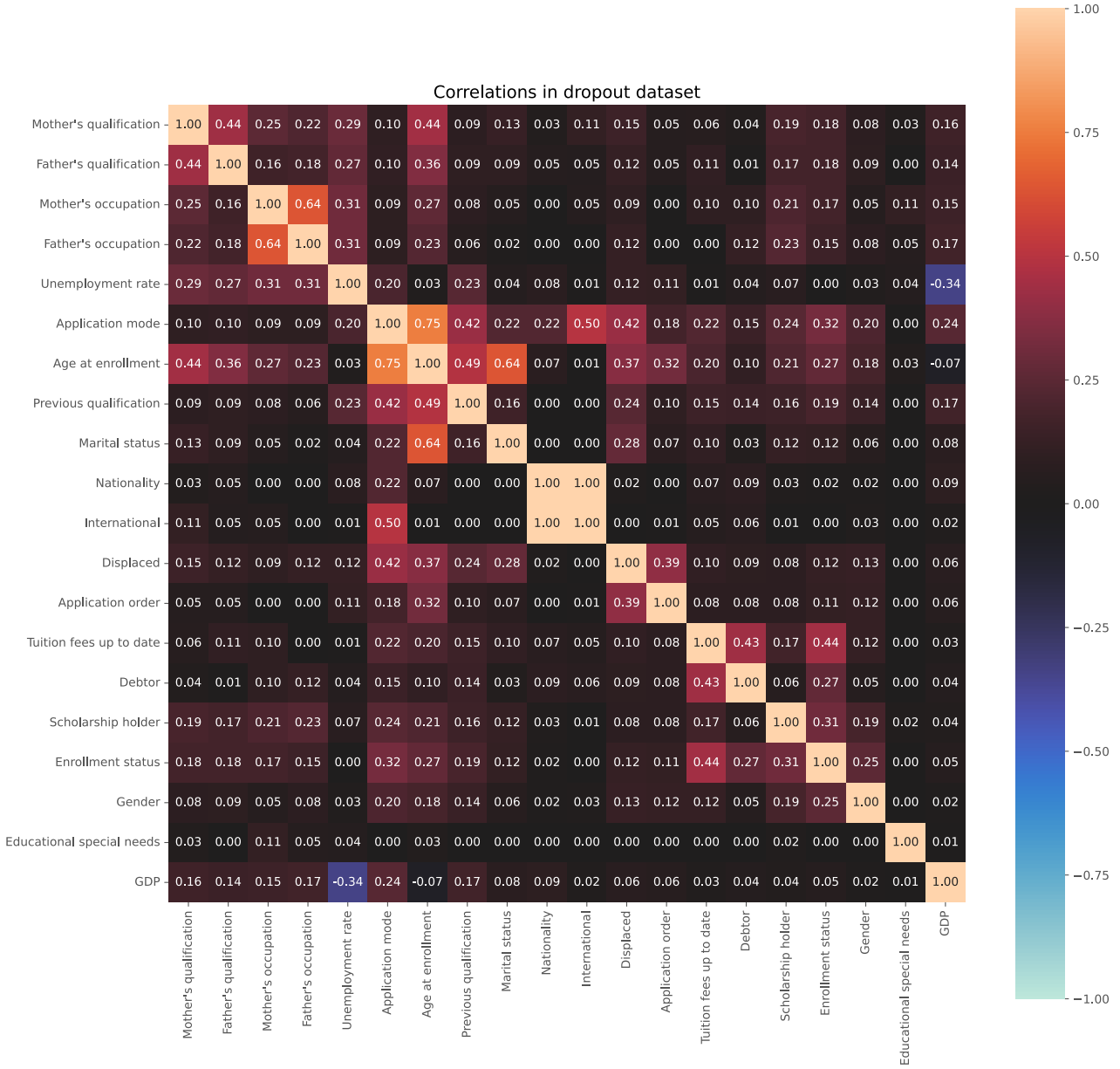

Figure 17: Correlation matrix for the student dropout dataset. High correlations between our protected attributes and other features, is an indicator that some of the non-protected attributes are proxies for the protected ones. This means that, even if we constrain one the protected attributes, a proxy for the same attribute could still potentially lead to biased decisions

The pattern identified in figure 18 shows that the risk of dropping out is not uniformly distributed across genders, a significant finding considering our fairness-oriented approach. According to the plot, men and females have a similar number of dropouts in terms of absolutes. However, the females' graduation rate is significantly larger than the men's. In a prediction domain, this imbalance could lead to predicting men dropping out more frequently, simple due to the dataset bias. In a fair setting the predicted dropout rates should not be based on your gender, but on other, more relevant factors that directly impact academic performance and continuation.

As seen in figure 19 there is an upward trend in the dropout percentage as the age at enrollment surpasses 25 years. Similarly to gender, dropout rates should not be determined solely based on age in a fair setting. When developing a model that predict dropout probabilities, this plot shows that it is crucial to be mindful of the potential biases associated with age.

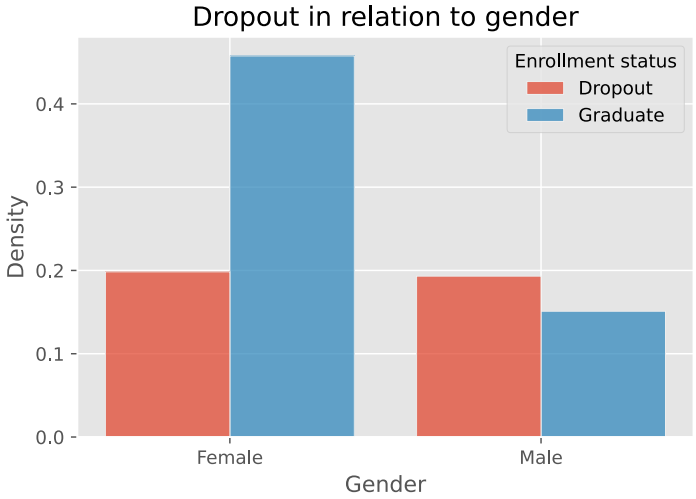

Figure 18: Looking at the distribution of gender we see a clear imbalance. Females are over-represented, both in raw numbers (2381 females vs 1249 males) as well as in proportion of graduates to dropouts. This imbalance introduces a clear bias towards men in the dataset, where using gender - without any preprocessing steps to even out the distributions - could be prominent feature for predicting the dropout rates.

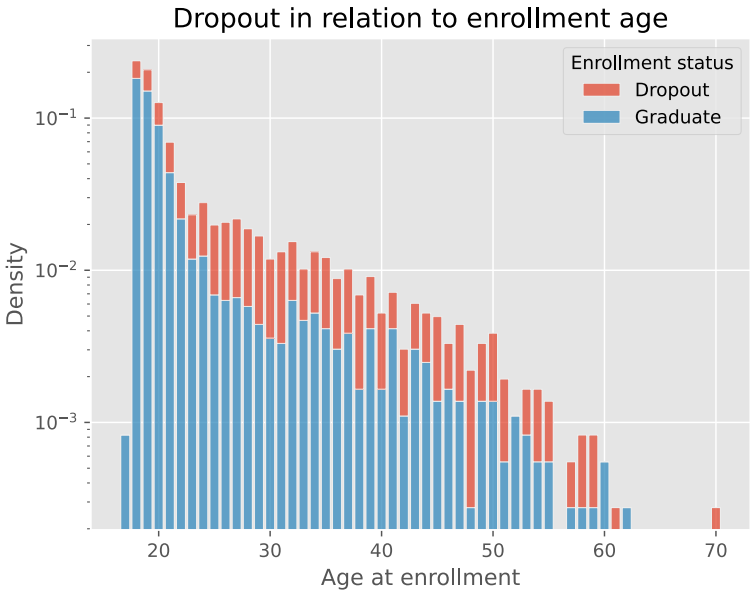

Figure 19: The plot illustrates that as the age at enrollment increases, the dropout percentage tends to rise, indicating a potential correlation between older age at enrollment and increased likelihood of dropout. This correlation can give rise to fairness concerns, as it can maintain biased outcomes and disadvantage individuals based on factors beyond the students' control.

In the field of income prediction, biases may be introduced due to disproportionate representation of these protected features. For instance, income tends to increase as an individual grows older, necessitating a representative sample across different age groups. Without a diverse representation of age groups, or other protected features, the predictive model may inherit these biases which can lead to unfair outcomes. The same reasoning applies to race and gender. To ensure accurate and fair results, it is crucial that the distribution of these protected attributes in the dataset mirrors their real-world distribution across the United States[1].

[1]The data is sampled from the US