Erasmus University Rotterdam

Erasmus School of Economics

On the Effects of Fairness Constraints on Interpretable Decision Rules for Binary Classification

Bachelor Thesis BSc^2 Econometrics & Economics

Name student: Robert Praas

Student ID number: 468166

Supervisor: dr. M.H. Akyuz

Second Assessor: U. Karaca MSc

Date final version: July 3, 2021

Abstract

In order to use algorithms in high-stake decision making, they must be interpretable and fair. In this paper, we investigate Boolean rule sets as a classification tool using a fair column generating approach inspired by Lawless and Günlük (2021). The following fairness metrics are separately implemented as constraints for the algorithm: Equality of opportunity, Equalized odds and Demographic parity. Experiments with four datasets, including the additional German credit dataset, indicate the highest accuracy of the model is obtained by tuning for Equality of opportunity. However, using the Equalized odds metric produces competitive accuracy with a stricter fairness definition. Demographic parity proves to be a surprisingly challenging metric as benchmark algorithms CART and Logistic regression achieve better performance than the Fair Column Generating algorithm. In addition, a set of recommendations is proposed to elevate the quality of fair and interpretable classification methods.

Keywords— machine learning, fairness, interpretability, boolean rule sets, integer optimization

Contents

1 Introduction

Many decision-making processes are increasingly taken care of by machines. Algorithms are becoming more competitive against humans due to rapid technological advancements. Machines have the potential to be quicker, cheaper and more precise than humans. Nevertheless, as fast as the machine learning field is developing, so are the concerns (Lemm, Blankertz, Dickhaus, & Müller, 2011; Cabitza, Rasoini, & Gensini, 2017; Makridakis, Spiliotis, & Assimakopoulos, 2018). Humans can usually be questioned about the reasoning behind their decisions. In contrast, a large body of algorithmic functions is defined as black boxes, that cannot be understood by humans. Machine learning algorithms can hold so much predictive power partly because they can detect hidden structures (Elman & Zipser, 1988). Nonetheless, there is a demand for transparency. The European Union for instance, already obligates decisions made by algorithms to hold explanatory power about the reasoning behind that decision (Goodman & Flaxman, 2017).

A second issue with algorithms concerns that the outcomes of several algorithms were found to be biased against groups with protected features such as gender, race or age. Next to the apparent undesirableness of discrimination, algorithmic bias might threaten the societal trust in algorithmic decision-making. Consequently, a large body of research on fairness for machine learning has emerged during the past few years, exemplified by the twenty-one different fairness definitions considered by Verma and Rubin (2018). However, the combination of fair and interpretable models is still limited and only started to appear more frequently in the literature during the past two years (Berrendorf, Faerman, Vermue, & Tresp, 2020; Geden & Andrews, 2021). This could be partly due to the complexity of the fairness-accuracy trade-off (Menon & Williamson, 2018). To investigate this trade-off further, the results of the Fair Column Generation (CG) algorithm (Lawless & Günlük, 2021) are replicated and extended by applying the algorithm to a new dataset. This algorithm applies the fairness metric Equality of opportunity to Boolean rules in disjunctive normal form (DNF, "OR-of-ANDs") (Lawless & Günlük, 2021). In addition, two extra fairness constraints are implemented to the model, namely Equalized odds and Demographic parity. Furthermore, we investigate whether inherent control for one fairness metric simultaneously produces promising results for the other fairness metrics. As success is measured by an algorithm that is accurate and fair, the research question is:

What is the effect of different fairness metrics on the accuracy of an interpretable machine learning model?

One distinction between fairness metrics is made by distinguishing individual versus group fairness, also called statistical parity (Kearns, Neel, Roth, & Wu, 2018). Individual fairness aims for similar classification for similar individuals, whereas group fairness aims for similar classification across different groups. It prevents different classification between groups based on a protected attribute. This research focuses on three measures of group fairness. The first, Equality of opportunity, entails similar false negative rates across groups. Secondly, Equalized odds considers similar false negative rates as well as false positive rates between groups (Corbett-Davies & Goel, 2018). Lastly, we implement the additional

metric of Demographic parity, which measures the rate of positive predictions among groups. Due to the constraints, the differences between groups cannot be larger than a certain threshold. In order to answer the research question, two sub-questions are considered:

- Does the Fair CG algorithm perform similarly when it is applied to different datasets?
- Which fairness metric leads to the highest accuracy in prediction?

This further develops the limited research on fair and interpretable machine learning models by extending empirical research on fairness metrics for Boolean decision rule sets. The added value of this research lies in:

- Extending the research to the German *credit* dataset
- Extending the fairness metric set to include *Demographic parity*
- Assessing the effects of different fairness metrics on accuracy when the model is tuned for another fairness metric

This is relevant in real-world applications as fairness is a growing pre-requisitive for the use of algorithms, while the model needs to be accurate enough to be useful in practice, which highlights the societal relevance. Moreover, the use of the additional Demographic parity fairness measure in a Boolean rule set context and comparing fairness metrics when the algorithm is tuned for another metric form the scientific relevance. The rest of the paper is structured as follows: Section 2 assesses the existing literature concerning interpretability and fairness. In Section 3 the methodology of Lawless and Günlük (2021) is summarized and extended to include Demographic parity. Section 4 contains data, implementation details and the results of experiments and Section 5 and 6 consist of the discussion and conclusion of the paper.

2 Literature review

The use of algorithms to classify the recidivism risk of criminals is hardly surprising as human judges were found to widely classify crimes differently among each other (Austin & Williams III, 1977). However, using software in such situations only improves social issues such as discrimination if the model itself is not prone to biases. For instance, a commonly used criminal risk assessment tool provided not only unreliable, but also racially biased predictions (Angwin, Larson, Mattu, & Kirchner, 2016). The developers, however, showed their predictions were fair when considering a different fairness metric. The *compas* dataset they used is also used for the analysis in this paper. Criminal risk assessment is only one of the possible applications, and similar situations may occur in settings such as providing loans, school admissions and default risk scores.

Interpretability is suggested to be a prerequisite for trusting algorithms (Ribeiro, Singh, & Guestrin, 2016a). Moreover, the potential of interpretable models is justified by the Rashomon set argument: Assume the data can be explained by a large set of reasonably accurate predictive models. Since this set

is large, there should be at least one interpretable model. Then that model is interpretable and accurate (Rudin, 2019). To put this this theory into practice, Boolean rule sets in DNF form are investigated. They are unordered (Lawless & Günlük, 2021) and are relatively easy to comprehend for humans (Freitas, 2014). An alternative to Boolean rule sets is the use of decision trees, which can be rewritten to a decision rule list (Quinlan, 1987). However, this may be harder to understand due to the more complex interpretation of ordered rules (Lawless & Günlük, 2021).

The Boolean rule sets are researched in a fairness setting. An example of a commonly used fairness metric is fairness under unawareness (Chen, Kallus, Mao, Svacha, & Udell, 2019), where protected attributes such as race or gender are simply excluded from the model. However, biases can still prevail after excluding attributes due to hidden correlations (Dwork, Hardt, Pitassi, Reingold, & Zemel, 2012). Caution is in any case required while using fairness metrics as they could lead to inaccurate predictions (Dwork et al., 2012). Other metrics used are the aforementioned individual fairness or counterfactual fairness (Kusner, Loftus, Russell, & Silva, 2017). Nevertheless, in this paper the focus will be on the group fairness metrics of Equality of opportunity, Equalized odds and Demographic parity. The first two are replicated based on Lawless and Günlük (2021) and Demographic parity (also referred to as Statistical parity) (Yao & Huang, 2017) is an additional metric based on similar positive rates among groups.

Research into fair algorithms or interpretable algorithms separately is not uncommon (Lawless & Günlük, 2021). However, the combination of fair and interpretable models is still limited and only started to appear more frequently in the literature during the past two years (Berrendorf et al., 2020; Geden & Andrews, 2021; Kehrenberg, Bartlett, Thomas, & Quadrianto, 2020). Since both notions naturally restrict accuracy, the combination of the two is expected to decrease the accuracy even more. Algorithmic use in social situations usually demands interpretability as well as fairness (Wang, Han, Patel, Mohideen, & Rudin, 2020). In their paper, these authors acknowledge the location dependency of algorithmic success and the need for timely updates. Current explorations of fair and interpretable machine learning models include decision trees (Kamiran, Calders, & Pechenizkiy, 2010; Aghaei, Azizi, & Vayanos, 2019), regression (Berk et al., 2017), rule-based and association rule-based classifiers (Pedreshi, Ruggieri, & Turini, 2008) and the recently added boolean rule sets (Lawless & Günlük, 2021). Boolean rule sets are used to find associations with binary data (Mannila & Toivonen, 1996). The main methodology (Lawless & Günlük, 2021) is based on Dash, Günlük, and Wei (2018) and makes the trade-off between simplicity of the rule sets and accuracy. More complex rule sets can increase accuracy, but decrease interpretability. Mita, Papotti, Filippone, and Michiardi (2020) propose a boolean rule set generator that is flexible with imbalanced datasets.

Equality of opportunity is stated to be oblivious as it otherwise requires subjective interpretation or assumptions (Hardt, Price, & Srebro, 2016). This means two individuals with similar talent and ambition are entitled to similar success prospects in competition (Arneson, 1999). Furthermore, Hardt et al. (2016) found the weaker notion of Equality of opportunity to provide more utility than Equalized odds. Equality of opportunity is one of two elements of Equalized odds, together with similar false positive rates. The false positive rate has a different meaning depending on the dataset. For assignment problems with limited capacities, such as graduate admissions, excessive false positives for one group directly take the

place of those from other groups. For non-competing classification problems such as recidivism, this does not have to be the case, although differences between groups are likely to be undesirable. The following research into Equalized odds is known by us: Pleiss, Raghavan, Wu, Kleinberg, and Weinberger (2017) use relaxed Equalized odds with calibration to find it to be equivalent to randomizing a subset of predictions. Romano, Bates, and Candès (2020) make use of resampling sensitive attributes to obtain Equalized odds. Lastly, Zhang and Bareinboim (2018) investigate Equalized odds and Equality of opportunity in a causal setting.

The last fairness metric, Demographic parity, is widely apparent in the literature (Hardt et al., 2016). For instance, Zemel, Wu, Swersky, Pitassi, and Dwork (2013) aim to achieve it by learning a representation of the data independent of the protected attribute while losing as little information about the other features as possible. However, Dwork et al. (2012) show conceptual limitations to demographic parity, which is later verified by Hardt et al. (2016). The different fairness metrics have their pros and cons in terms of strictness and practical use. Rajkomar, Hardt, Howell, Corrado, and Chin (2018) argue either of these methods may be most relevant depending on the setting.

3 Methodology based on Lawless and Günlük (2021)

3.1 Fairness metrics

A supervised binary classification problem is defined by a set of training samples with labels $y_i \in \{0, 1\}$ and features $X_i \in \{0, 1\}^p$ for $i \in I = \{1, \ldots, n\}$ to generate the most accurate decision rule $d : \{0, 1\}^p \to \{0, 1\}$ (Lawless & Günlük, 2021). In the context where fairness is considered, each data point is part of a group denoted by the sensitive variable, also known as protected feature. In this research we consider three different fairness metrics. The first metric is Equality of opportunity, which requires the false negative rate to be equal between groups (Lawless & Günlük, 2021). It is portrayed by the expression

$$\mathbb{P}(d(X) = 0 \mid Y = 1, G = g) = \mathbb{P}\left(d(X) = 0 \mid Y = 1, G = g'\right) \quad \forall g, g' \in \mathcal{G} \tag{1}$$

with $d(X)$ as the prediction, Y as the actual outcome (label) and $g_i \in \mathcal{G}$ shows the protected feature of which the data point is a part of. As the probability of false negatives and true positives given a positive label sum to one, consequently the true positive rate will also be similar among groups. If we add a similar constraint to Equation 1, while interchanging the 0's and 1's, we obtain the second and stricter Equalized odds condition,

$$\mathbb{P}(d(X) = 1 \mid Y = 0, G = g) = \mathbb{P}\left(d(X) = 1 \mid Y = 0, G = g'\right) \quad \forall g, g' \in \mathcal{G} \tag{2}$$

which ensures equal false positive rates and therefore also equal true negative rates. In addition to these fairness metrics, which take the labels of the data into account, the third fairness metric Demographic parity is independent of data labels and requires equal acceptance rates among groups. The formulation is given by

$$\mathbb{P}(d(X) = 1 \mid G = g) = \mathbb{P}\left(d(X) = 1 \mid G = g'\right) \quad \forall g, g' \in \mathcal{G} \tag{3}$$

Demographic parity leads to similar proportions of positive classifications among groups. Given $\mathbb{P}(d(X) = 1) + \mathbb{P}(d(X) = 0) = 1$ for any group G, Equation 3 holds for both $d(X)$ equal to 0 as well.

Furthermore, any perfect fairness metric could lead to a very limited set of possible classification models. Therefore, maximum disparity among groups is used in practice as a proxy for unfairness and will be denoted by d. The notation of maximum disparity for Equality of opportunity,

$$\Delta(d) = \max_{g,g' \in \mathcal{G}} |\mathbb{P}(d(X) = 0 \mid Y = 1, G = g) - \mathbb{P}\left(d(X) = 0 \mid Y = 1, G = g'\right)| \tag{4}$$

has a similar form as for the other fairness metrics. It represents the maximal distance between the probabilities for different groups. The absolute difference for a fairness metric among groups is bounded by a chosen ϵ. Therefore, $\Delta(d) < \epsilon$ is added as a constraint to inherently control for unfairness.

3.2 Boolean decision rule sets

The goal is to construct an optimal DNF rule-set, which classifies a data point as 1 if it adheres to a complete rule and 0 otherwise, while taking into account fairness constraints. For p binary features, only $(2^p - 2)$ decision rules can possibly be made (Lawless & Günlük, 2021). By enumerating all rules, solving the Integer Programming (IP) problem should lead to an optimal subset of rules which minimizes classification error. In practice, this is intractable, and a Linear Programming (LP) relaxation is solved using a Column Generation framework. The objective of the LP is to minimize *Hamming loss* (Dash et al., 2018), which represents classification error as it is calculated by counting the number of rules that should be changed for correct classification.

3.3 Integer program formulation

Similar to Lawless and Günlük (2021) we define $\mathcal{K}$ as the set of possible rules. Then, $\mathcal{K}_i \subset \mathcal{K}$ is rule set met by data point $i \in I$. Moreover, c_k corresponds to the complexity of rule $k \in \mathcal{K}$. This is calculated by a fixed cost of one and increases by one for every condition in the rule. In the supervised learning context we assume the data are split into two partitions based on their labels $\mathcal{P} = \{i \in I : y_i = 1\}$, and $\mathcal{Z} = \{i \in I : y_i = 0\}$. For every group $g \in \mathcal{G}$, data points are denoted to have the sensitive attribute g with $\mathcal{G}_g = \{i \in I : g_i = g\}$. Let $\mathcal{P}_g = \mathcal{P} \cap \mathcal{G}_g$ and $\mathcal{Z}_g = \mathcal{Z} \cap \mathcal{G}_g$. For now two groups $\mathcal{G} = \{1, 2\}$ are assumed, but the number of groups can be easily extended upon. In addition, let $w_k \in \{0, 1\}$ define whether rule $k \in \mathcal{K}$ is selected. Let $\zeta_i \in \{0, 1\}$ define whether data point $i \in \mathcal{P}$ is misclassified and lastly, let C indicate the highest complexity allowed. Based on this notation, finding the optimal rule set considering the fairness constraint of Equality of opportunity is defined by the following problem:

$$z_{\text{mip}} = \min \sum_{i \in \mathcal{P}} \zeta_i + \sum_{i \in \mathcal{Z}} \sum_{k \in \mathcal{K}_i} w_k \tag{5}$$

$$\text{s.t.} \quad \zeta_i + \sum_{k \in \mathcal{K}_i} w_k \geq 1, \quad i \in \mathcal{P} \tag{6}$$

$$C\zeta_i + \sum_{k \in \mathcal{K}_i} 2w_k \leq C, \quad i \in \mathcal{P} \tag{7}$$

$$\sum_{k \in \mathcal{K}} c_k w_k \leq C \tag{8}$$

$$w \in \{0,1\}^{|\mathcal{K}|}, \zeta \in \{0,1\}^{|\mathcal{P}|} \tag{9}$$

$$\frac{1}{|\mathcal{P}_1|} \sum_{i \in \mathcal{P}_1} \zeta_i - \frac{1}{|\mathcal{P}_2|} \sum_{i \in \mathcal{P}_2} \zeta_i \leq \epsilon_1 \tag{10}$$

$$\frac{1}{|\mathcal{P}_2|} \sum_{i \in \mathcal{P}_2} \zeta_i - \frac{1}{|\mathcal{P}_1|} \sum_{i \in \mathcal{P}_1} \zeta_i \leq \epsilon_1 \tag{11}$$

The first Constraint (6) captures false negatives if no rule is selected. Then ζ_i takes on value 1 if no rules satisfied by $i \in \mathcal{P}$ are selected. The second Constraint (7) ensures ζ_i takes on value 0 if any rule is selected. The point is then correctly classified. It is taken into account that $c_k \geq 2$ as any rule contains at least one condition. The possible amount of rules are bound by complexity C. The third Constraint (8) extends this bound on complexity to the final rule set. The final two Constraints (10-11) safeguard the level of tolerated unfairness, depending on ϵ_1. Two constraints are necessary to prevent either of the groups to have a higher false negative rate.

For Equalized odds, two constraints concerning the false positives are added to the above framework

$$\frac{1}{|\mathcal{Z}_1|} \sum_{i \in \mathcal{Z}_1} \sum_{k \in \mathcal{K}_i} w_k - \frac{1}{|\mathcal{Z}_2|} \sum_{i \in \mathcal{Z}_2} \sum_{k \in \mathcal{K}_i} w_k \leq \epsilon_2 \tag{12}$$

$$\frac{1}{|\mathcal{Z}_2|} \sum_{i \in \mathcal{Z}_2} \sum_{k \in \mathcal{K}_i} w_k - \frac{1}{|\mathcal{Z}_1|} \sum_{i \in \mathcal{Z}_1} \sum_{k \in \mathcal{K}_i} w_k \leq \epsilon_2 \tag{13}$$

also here does ϵ_2 represent the maximum distance allowed between groups. Since w_k represents whether a rule set is selected, parity between selected rule sets for data labeled as a 0 outcome is a good indicator for false positives. If for one data point labeled as 0 still multiple rule sets are selected, then the prediction is further from the truth than when only one rule set was selected. These situations are penalized harder in this formulation.

Thirdly, for Demographic parity a new variable $\tau_i \in \{0,1\}$ is added, indicating whether data point $i \in \mathcal{Z}$ is misclassified. Two constraints to ensure equal positive classifications among groups are added,

$$\frac{1}{|\mathcal{P}_1| + |\mathcal{Z}_1|} \left(\sum_{i \in \mathcal{P}_1} (1 - \zeta_i) + \sum_{i \in \mathcal{Z}_1} \tau_i\right) - \frac{1}{|\mathcal{P}_2| + |\mathcal{Z}_2|} \left(\sum_{i \in \mathcal{P}_2} (1 - \zeta_i) + \sum_{i \in \mathcal{Z}_2} \tau_i\right) \leq \epsilon_3 \tag{14}$$

$$\frac{1}{|\mathcal{P}_2| + |\mathcal{Z}_2|} \left(\sum_{i \in \mathcal{P}_2} (1 - \zeta_i) + \sum_{i \in \mathcal{Z}_2} \tau_i\right) - \frac{1}{|\mathcal{P}_1| + |\mathcal{Z}_1|} \left(\sum_{i \in \mathcal{P}_1} (1 - \zeta_i) + \sum_{i \in \mathcal{Z}_1} \tau_i\right) \leq \epsilon_3 \tag{15}$$

in this way, the fractions of correctly classified data points $i \in \mathcal{P}$ and misclassified points $i \in \mathcal{Z}$ combined are the fractions of data points with a positive prediction. The positive classifications are divided by all data points in groups P and Z, because no data label assumptions are made. Here ϵ_3 works similarly to ϵ_1 and ϵ_2. Since the indicator τ_i is not linear by itself, the following constraints are added

$$(1 - \tau_i) + \sum_{k \in \mathcal{K}_i} w_k \geq 1, \quad i \in \mathcal{Z} \tag{16}$$

$$C(1 - \tau_i) + \sum_{k \in \mathcal{K}_i} 2w_k \leq C, \quad i \in \mathcal{Z} \tag{17}$$

$$\tau_i \in \{0,1\}^{|\mathcal{Z}|} \tag{18}$$

similar, but in the opposite direction to constraints (6-7), Constraint (16) ensures τ_i takes value 0 if no rule is selected for a data point with label 0. Constraint (17) ensures τ_i takes on value 1 if at least one rule set is selected.

3.4 Column Generation framework

Again, we follow the methodology by lawlessfair. A column generation framework (Conforti, Cornuéjols, Zambelli, et al., 2014) is used to solve the LP relaxation of the Master Integer Program (MIP) above. First, a subset of possible rules is taken, and the LP is solved for only these rules. Then, its optimal dual solution is used to find missing variables with negative reduced cost (Bazaraa, Jarvis, & Sherali, 2008). This is done by solving a separate integer program, also referred to as the Pricing problem. A variable with negative reduced cost is added to the subset of the LP relaxation. This problem is repeatedly solved until no variables with negative reduced cost can be found anymore. We solve the restricted LP problem for a restricted subset of rules $\hat{\mathcal{K}} \subset \mathcal{K}$. Let $\left(\mu, \alpha, \lambda, \gamma^1, \gamma^2\right)$ be an optimal dual solution. The γ's correspond to the fairness constraints whereas μ, α, λ, are associated with constraints (6)-(8) respectively. The integer program attempts to find a $k \in \mathcal{K}$ with the minimum reduced cost $\hat{\rho}_k$. It includes variable $z_j \in \{0,1\}$ for $j \in J$ to denote if a data point i has all features selected by the rule. Let variable $\delta_i \in \{0,1\}$ for $i \in I$ define whether a rule misclassifies data point i. With these variables we compute the complexity rule to put into the objective function: $\left(1 + \sum_{j \in J} z_j\right)$. The full pricing problem for Equality of opportunity becomes:

$$z_{cg} = \min \sum_{i \in \mathcal{Z}} \delta_i + \sum_{i \in \mathcal{P}} \left(2\alpha_i - \mu_i\right) \delta_i + \lambda \left(1 + \sum_{j \in J} z_j\right) \tag{19}$$

$$\text{s.t.} \quad D\delta_i + \sum_{j \in S_i} z_j \leq D \quad i \in I^- \tag{20}$$

$$\delta_i + \sum_{j \in S_i} z_j \geq 1 \quad i \in I^+ \tag{21}$$

$$\sum_{j \in J} z_j \leq D \tag{22}$$

$$z \in \{0,1\}^{|J|}, \delta \in \{0,1\}^{|\mathcal{P}|+|\mathcal{Z}|} \tag{23}$$

Set $I^- \subseteq I$ consists of the indices of variables which have a negative coefficient for δ_i in the objective function, and $I^+ = I \backslash I^-$ works similarly for positive coefficients. The objective function does not include γ^1 or γ^2 since variable w_k is not part of the fairness constraints in the MIP problem. The complexity bound D is independent of C in the MIP problem.

The fairness constraints contain w_k for Equalized odds. Therefore, γ^3 and γ^4 need to be included in the objective of the pricing problem. The following elements are added,

$$\sum_{i \in \mathcal{Z}_1} \frac{\gamma_3}{|\mathcal{Z}_1|} \mathbb{1}_{\{k \in \mathcal{K}_i\}} - \sum_{i \in \mathcal{Z}_1} \frac{\gamma_4}{|\mathcal{Z}_1|} \mathbb{1}_{\{k \in \mathcal{K}_i\}} - \sum_{i \in \mathcal{Z}_2} \frac{\gamma_3}{|\mathcal{Z}_2|} \mathbb{1}_{\{k \in \mathcal{K}_i\}} + \sum_{i \in \mathcal{Z}_2} \frac{\gamma_4}{|\mathcal{Z}_2|} \mathbb{1}_{\{k \in \mathcal{K}_i\}} \tag{24}$$

which leads to the following pricing problem:

$$z_{cg} = \min \left(1 + \frac{\gamma_3 - \gamma_4}{|\mathcal{Z}_1|}\right) \sum_{i \in \mathcal{Z}_1} \delta_i + \left(1 + \frac{\gamma_4 - \gamma_3}{|\mathcal{Z}_2|}\right) \sum_{i \in \mathcal{Z}_2} \delta_i + \sum_{i \in \mathcal{P}} (2\alpha_i - \mu_i)\, \delta_i + \lambda \left(1 + \sum_{j \in J} z_j\right) \tag{25}$$

Just as for Equalized odds, the objective of the pricing problem needs to be adapted for Demographic parity. Similar to Lawless and Günlük (2021), let $\left(\mu, \alpha, \lambda, \psi, \theta, \gamma^1, \gamma^2\right)$ be an optimal dual solution to the Restricted Master Linear Program, where ψ and θ are associated with Constraints (16) and (17), respectively. Then the coefficient for the term $\sum_{i \in \mathcal{Z}} \delta_i$ in Constraint (19) takes a similar form as the part of the objective function for $i \in P$. The new objective function for the pricing problem becomes:

$$z_{cg} = \min \sum_{i \in \mathcal{Z}} (2\theta_i - \psi_i + 1)\delta_i + \sum_{i \in \mathcal{P}} (2\alpha_i - \mu_i)\, \delta_i + \lambda \left(1 + \sum_{j \in J} z_j\right) \tag{26}$$

Hence, the three different fairness metrics share the same Constraints (20)-(23) for the pricing problem, yet they have different objective functions.

4 Results

4.1 Data

The main requirement for classification with fairness constraints is that the dataset contains a sensitive attribute, which is a variable indicating different groups, which could be gender, race or income level, for example. The sensitive variable (also known as protected attribute) is suspected to be an indicator of unfairness. The algorithm creating the boolean rule sets using column generation, abbreviated by Fair CG, is tested on three machine learning datasets by Lawless and Günlük (2021):

- *default* (Dua & Graff, 2017) is a 2005 dataset of Taiwanese credit card customers, with the objective of predicting default payments. The sensitive variable is gender.
- *adult* (Dua & Graff, 2017) is a 1994 US Census Income dataset, with the objective to predict if a person makes over $50.000. The sensitive variable is gender.
- *compas* (Angwin et al., 2016) is a dataset with the aim of classifying the risk of recidivism for convicted defendants. The sensitive variable is race.

The analysis is extended to the following dataset:

- German *credit* (Dua & Graff, 2017) is a 1975 German dataset used to classify customers as good or bad risks. This dataset is in the same domain as the *default* set, with the addition that both gender and being foreign can be used as the sensitive variable here. It has 20 features and a relatively low size (1000 rows versus 30.000 for the *default* dataset). As *credit* can be easily confused by *default*, we refer to the German *credit* dataset as German in the tables and figures.

The *compas* dataset can be retrieved from Kaggle (Ofer, 2017). The other three datasets can be retrieved from the UCI machine learning repository (Dua & Graff, 2017). Similar to Lawless and Günlük

(2021) for the *adult* dataset the training data is used, for *default* the entire dataset. The COMPAS dataset is cleaned to be restricted to only consider African American and Caucasian respondents. A binary column race indicates whether a respondent was African American. Similar to Dash et al. (2018), the data is made binary through one-hot encoding. Both the encoding and its negation are considered. Numerical values are compared against sample deciles, which are all the tenth percentiles. The negations of those comparisons are also included.

4.2 Implementation details

The implementation is based on Lawless and Günlük (2021). The Gurobi Python interface (Gurobi Optimization, 2021) is used to solve the programs.

The Master Linear Program (MLP) is solved using a barrier interior point method with the default crossover parameter. Here, the integer program is converted to a series of unconstrained programs, where a high cost is added to infeasibility or approaching the boundary from the interior (Ravikumar, 2017). For the Master Integer Program (MIP) the rule set with the highest accuracy for the training set is used. Due to the intensity of computations, standard time limits are implemented. In addition, when the MLP is (nearly) optimal, at most 1000 rules with the lowest reduced cost are used to solve the MIP, and it returns all feasible solutions found within the time limit. The Column Generation framework is approximated for large datasets using a sub-sample selected uniformly at random of the original dataset. Such a sub-sample consists of less than or equal to 2000 rows due to the problem size. A greedy heuristic is first employed four times before switching to the IP formulation, as this was found to produce the best results (Lawless & Günlük, 2021).

Due to time limitations, we use 5-fold cross-validation. The rule set is built using a two-step process. First, the Column Generation algorithm runs on a training dataset with hyper-parameters. Then the generated candidate rules are used to solve the master integer program for more potential unfairness bounds (epsilons) and complexity limits. These complexity limits are generated from finding the best accuracy when no fairness constraints are included and testing neighboring values. Solving the MLP and MIP is done with a 5-minute time limit, and each pricing problem iteration has a 45 second time limit (Lawless & Günlük, 2021).

4.3 Experimental set-up

The code is written in Python 3.7 and all results are found using a 2,3 GHz Dual-Core Intel Core i5 CPU. The python environment was configured with Anaconda and included the following package dependencies:

Table 1: Overview of package dependencies

Package	Version
Python	3.7.9
Gurobi	9.1.2
Numpy	1.20.1
Pandas	1.2.4

For Equality of opportunity, Equalized odds and Demographic parity, the ϵ used in the first phase of rule generation $(0.2, 1)$ and in the second phase of rule generation $(0, 0.01, 0.03, 0.05, 0.1, 0.2, 0.5, 1)$ are equal among the different datasets. For Equalized odds we set $\epsilon_1 = \epsilon_2$. Rule complexity parameter C is set similarly to Lawless and Günlük (2021). Values for C ranged in the experiments from 5 to 110 (Lawless & Günlük, 2021) such that the number of rules and conditions remain interpretable. For the German *credit* dataset we use complexities $(10, 17, 30)$ for the first phase and $(10, 15, 20)$ for the second phase as suggested by Lawless and Günlük (2021).

4.4 Measuring unfairness

First, let us define the fairness definition used. The constraints measure the degree of unfairness. More fairness means low differences between groups, and thus, lower differences are better. For instance, the degree of unfairness being 0 is the best fairness to be found.

The analysis structure is as follows: For the three fairness metrics used we provide the accuracy and unfairness metrics per dataset when optimizing for accuracy or fairness. In addition, graphs are shown about the relations between fairness, complexity and accuracy. For the extension dataset German *credit* we provide comparisons to other interpretable benchmarks with decision trees (CART) and Logistic Regression methods. Similar to Lawless and Günlük (2021), both are implemented using scikit-learn in python. As Demographic parity is newly implemented, benchmarks will be provided for all datasets involved with this fairness metric. Furthermore, in Section 5.7 we show results for the other fairness metrics than for which the model was originally tuned for.

4.5 Equality of opportunity

We start with the results for the Equality of opportunity measure. The goal is to replicate the results of the Fair CG algorithm by Lawless and Günlük (2021). All tables show accuracies and disparities of the fairness metrics in percentages, as well as standard deviations in brackets. Table 2 shows the optimal accuracy and unfairness results per dataset. Standard deviations are shown in parentheses. The mean accuracy and unfairness for the models tuned for accuracy and unfairness are similar to Lawless and Günlük (2021). We comment on the tunings separately. For the *adult* set, the optimal accuracy is only 0.1 lower than found in Lawless and Günlük (2021), most probably due to the use of 5-fold cross-validation. The corresponding unfairness is two points lower, exemplifying the possibility of sacrificing little accuracy for more fairness. The results for *compas* are similar, yet here a slightly lower fairness disparity is found

for similar optimal accuracy. For *default* the optimal accuracy is just higher and accompanied unfairness is just lower.

Table 2: Mean Accuracy and Fairness Results for Equality of Opportunity

	Adult		Compas		Default		German	
	Accuracy	Fairness	Accuracy	Fairness	Accuracy	Fairness	Accuracy	Fairness
Tuned for Acc	82.8(0.4)	7.0(5.2)	68.2(1.8)	22.2(5.5)	82.1(0.4)	0.2(1.5)	73.2(3.5)	2.4(6.4)
Tuned for Fair	82.3(0.4)	1.3(0.8)	53.0(1.6)	0(0)	77.9(0.4)	0(0.3)	70.0(2.9)	0(0)

Next, we tune for fairness. For the *adult* set the optimal unfairness found is 1.3 higher than Lawless and Günlük (2021) found, which was 0.3. However, here the accuracy of 82.3 outperforms theirs (78.4). This is a situation where we might consider sacrificing some fairness for more accuracy. For the *compas* set, when tuned for fairness, the accuracy is equal to the percentage of positive labeled data. This means the majority class was automatically predicted and no effective rules were used.

In addition, the algorithm is also run on the *german* dataset with gender as the sensitive attribute. As Table 2 shows, Equality of opportunity gaps are found, but yet the accuracy is at most 73.2%. Given 70% of the dataset is labeled as negative, the boolean rule sets do not seem to handle this imbalance effectively.

In Figure 1 three graphs for the *compas* dataset are used as examples. The other datasets are displayed in Appendix A.1 Figures 5 and 6. It shows that the difference in false negative rates increases if the fairness constraint is relaxed. Figures 1a and 1b show similar trends to what was found before. For the *compas* dataset the increase in the Equality of opportunity gap affects one group (G1) in a more negative way than the other (G2). Figure 1b shows that increasing complexity leads to lower false negative rates, exemplifying the interpretability-fairness trade-off. Moreover, figure 1c shows that decreasing fairness leads to more accuracy. The results are further interpreted in the Discussion in Section 6.

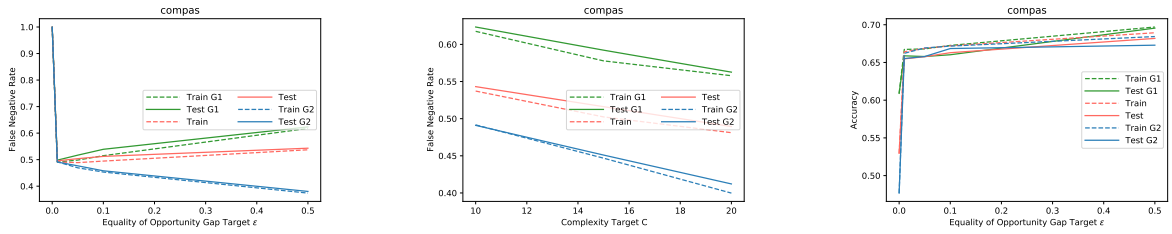

Figure 1: Impact of the fairness constraint (a), complexity constraint (b) on the false negative rate of the *compas* dataset. Impact of the fairness constraint on the accuracy (c).

For the German *credit* dataset, the Fair CG algorithm is compared to benchmark algorithms of decision trees (CART) and Logistic Regression (LR). Here, also 5-fold cross-validation is used in combination with a variety of hyperparameters. Table 3 shows the results.

Table 3: Benchmark Accuracy and Fairness Results for Equality of opportunity

	German	
	Accuracy	Fairness
Fair CG Tuned for Acc	73.2(3.5)	2.4(6.4)
Tuned for Fair	70.0(2.9)	0.0(0.0)
CART Tuned for Acc	71.3(2.5)	13.3(6.8)
Tuned for Fair	70.0(3.0)	0.0(0.0)
LR Tuned for Acc	74.4(2.4)	15.1(11.4)
Tuned for Fair	71.8(2.7)	9.3(7.7)

LR obtains the highest accuracy when tuned for it, although it has a cost of a 15.1 gap in unfairness. When tuned for fairness, Fair CG and CART resort to predicting the majority class, and in doing so they find no unfairness. Unfortunately, this result is not very useful in practice as this would mean that the algorithm would classify any person as low risk. It may be costly for the bank to also provide loans to a great share of high risks, of which there were 30% in this sample. However, the optimal accuracy (when tuned for accuracy) is not much higher, and this indicates none of these interpretable methods can classify risks with high accuracy.

Figure 2 shows the effect of the Equality of opportunity gap target (unfairness target) on the false negative rate, Equality of opportunity gap and accuracy. Notably, tested accuracy does not keep up with trained accuracy, highlighting the potential benefit of a larger dataset.

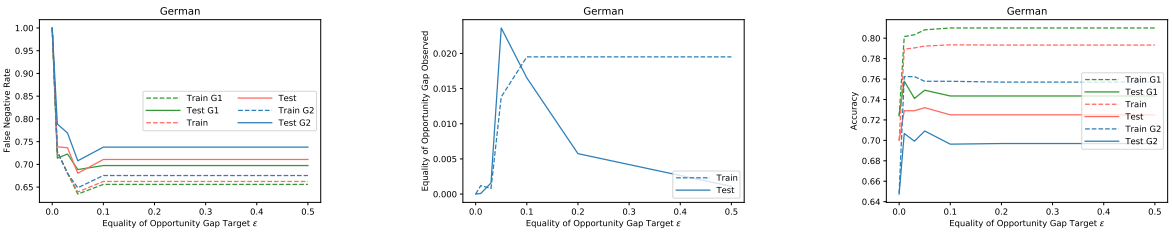

Figure 2: Impact of the fairness target on false negative rate (a) and fairness measure (b) and accuracy (c) for the German *credit* dataset.

Moreover, Figure 3 shows the relation between the complexity target and the accuracy and false negative rates. Whereas a higher complexity target leads to lower false negative rates, it does not lead to higher test accuracy. The effect of complexity seems ambiguous. It may lead to more rules which increases the probability of a positive classification. However, longer rules might make positive classifications less likely.

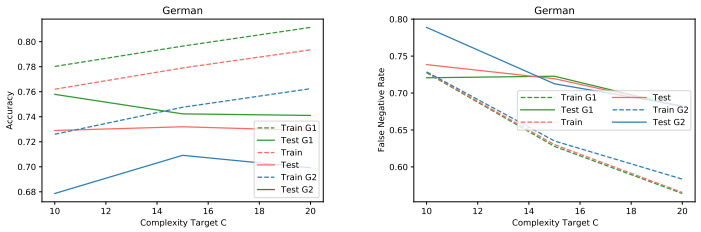

Figure 3: Impact of the complexity target on accuracy and false negative rate

4.6 Equalized odds

The second fairness measure considered is Equalized odds. Next to similar false negative rates, groups now also need to have similar false positive rates for their predictions. Similar to Lawless and Günlük (2021) the results for the Equalized odds measure are split between the (+) term, the loss from data points with a positive label and the (-) term for loss from data points with a negative label. These are also the false negative and false positive rates respectively. As Equalized odds is a stricter fairness measure, it leads to slightly lower optimal accuracy than Equality of opportunity. Table 4 shows the results for the Equalized odds (-) term. Apart from the optimal accuracy for the *adult* dataset (80.7 here vs 83.1 there), most results are very similar to Lawless and Günlük (2021).

Table 4: Mean Accuracy and Fairness Results for Equalized odds, fairness defined by the gap in Hamming loss (-) between groups

	Adult		Compas		Default		German	
	Accuracy	Fairness	Accuracy	Fairness	Accuracy	Fairness	Accuracy	Fairness
Tuned for Acc	80.7(1.2)	12.5(5.4)	67.1 (1.6)	12.2(5.4)	82.0(0.2)	1.2(1.0)	72.4(3.0)	9.5(0.7)
Tuned for Fair	75.9(0.7)	0(1.3)	53.0(1.0)	0.0(5.0)	82.0(0.2)	0.0(1.0)	70(2.8)	0.0(0.7)

Next, we consider the results for the positive gap in Equalized odds in Table 5. This is defined by positively labeled data points which did not select any rule and are thus classified as negative. Hence, the values tuned for accuracy are naturally equal to those for negative hamming loss. The results for the *default* dataset display similar accuracies, yet different degrees of unfairness due to rounding. Tuned for fairness, all datasets can be classified with a gap of 0 in unfairness. Interestingly, for *adult*, *default* and the German *credit* datasets, optimal fairness can be obtained with a decrease of not more than 5% points inaccuracy in comparison to when they are tuned for accuracy. Only for the *compas* dataset does optimal fairness lead to only 53% accuracy, which is equivalent to always predicting the majority class.

Table 5: Mean Accuracy and Fairness Results for Equalized odds, fairness defined by the gap in Hamming loss (+) between groups

	Adult		Compas		Default		German	
	Accuracy	Fairness	Accuracy	Fairness	Accuracy	Fairness	Accuracy	Fairness
Tuned for Acc	80.7(1.2)	2.1(5.0)	67.1 (1.6)	21.1 (8.9)	82.0(0.2)	0(0.2)	72.4(3.0)	5.9(2.2)
Tuned for Fair	75.9(0.7)	0(5.0)	53.0(1.0)	0(8.9)	82.0(0.2)	0(0.2)	70(2.8)	0(2.2)

We continue with benchmark results for the German *credit* dataset. Here, the benchmarks show the challenge of classifying the German *credit* dataset with Equalized odds. The sum of unfairness for HEO(+) and HEO(-) exceeds ten if tuned for optimal fairness. In terms of optimal accuracy CART and LR both find a higher accuracy than Fair CG.

Table 6: Mean Accuracy and Fairness Results for the German *credit* dataset with Equalized odds

German	Accuracy	HEO(-)	HEO(+)
Fair CG Tuned for Acc	72.4(3.0)	9.5(0.7)	5.9(2.2)
Tuned for Fair	70(2.8)	0(0.7)	0(2.2)
CART Tuned for Acc	71.3(2.5)	2.2(1.5)	13.3(6.8)
Tuned for Fair	70.0(3.0)	0.0(0.0)	0.0(0.0)
LR Tuned for Acc	74.4(2.4)	7.1(5.6)	15.1(11.4)
Tuned for Fair	71.8(2.7)	1.7(1.4)	9.3(7.7)

Lastly, we observe the figures for Equalized odds in the Appendix A.2 Figures 7-10 and see similar patterns to Equality of opportunity. Accuracy is on average 0.05 lower for test sets in comparison to training sets. Naturally, a higher complexity target leads to higher HEO(-) observed and lower HEO(+) observed. This shows for the *credit* dataset that higher complexity leads to more fairness between groups considering they are positively labeled. The German *credit* dataset was also evaluated for both Equality of opportunity and Equalized odds using foreign as the sensitive attribute, although only 3% of the dataset has this feature. The Fair CG as well as CART and LR were able to find accuracy, but no fairness, most probably due to the large imbalance. Resampling techniques could be considered as a remedy in the future.

4.7 Demographic parity

Lastly, we consider inherent control for Demographic parity, where the positive classification rates ought to be equal among groups. This seems to be a more straightforward fairness metric than the others, as it does not take labels into account. Nonetheless, implementation proves to be much harder. For all datasets, accuracy results are similar regardless of fairness target. For the *adult* dataset especially only 24.1% accuracy was acquired, which is more than twice as bad as random predictions. The results are

published for reference in Table 7 here:

Table 7: Mean Accuracy and Fairness Results for Demographic parity

	Adult		Compas		Default		German	
	Accuracy	Fairness	Accuracy	Fairness	Accuracy	Fairness	Accuracy	Fairness
Fair CG Tuned for Acc	24.1(0.8)	19.6(0.6)	47.0 (1.2)	13.3 (2.0)	78.0(0.7)	0(0.0)	70.0(2.4)	0(0.0)
Tuned for Fair	24.1(0.7)	19.6(0.6)	47.0(1.0)	13.2(3.1)	78.0(0.7)	0.0(0.0)	70.0(2.4)	0.0(0.0)
CART Tuned for Acc	85.4(0.5)	18.0(0.9)	68.1 (1.9)	27.5 (3.0)	82.1(1.5)	2.4(1.4)	71.3(2.5)	3.5(2.1)
Tuned for Fair	85.4(0.5)	18.0(0.9)	65.9(2.3)	23.8(3.9)	82.0(1.4)	1.8(1.3)	70.0(3.0)	0.0(0.0)
LR Tuned for Acc	79.9(0.3)	6.7(4.6)	68.1 (1.6)	27.7 (3.3)	78.0(1.7)	0(0.0)	74.4(2.4)	11.5(5.7)
Tuned for Fair	79.9(0.2)	4.5(0.5)	67.5(1.8)	26.9(2.9)	78.0(1.7)	0.0(0.0)	71.8(2.7)	2.3(1.0)

The benchmarks show CART and LR can find higher optimal accuracies when tuned for accuracy. Only for *default* does LR find similar accuracy and fairness in comparison with Fair CG. When tuned for fairness, both benchmarks find much better accuracy and lower unfairness for *adult*. For *compas* their optimal fairness is considerably higher than for Fair CG, but this includes on average a 20%-point higher accuracy when tuned for fairness. For *default* and *credit* the results are more comparable. The Fair CG algorithm with Demographic parity mainly classifies similar accuracy for different complexity and fairness, so graphs did not prove to be insightful.

4.8 Comparison across fairness measures

Finally, we compare different fairness results when tuned for one of them. This means when tuned for Equality of opportunity, those related constraints were added to the integer program of the fair CG algorithm. The fairness metrics of Equalized odds and Demographic parity are analyzed in that case.

4.8.1 Tuned for Equality of opportunity

The different fairness metrics satisfy different fairness definitions. We continue by investigating whether the Fair CG model optimized for Equality of opportunity also holds promising Equalized odds and Demographic parity results. Equalized odds has identical constraints to Equality of opportunity for false negatives. Therefore, we only look at the Hamming equalized odds (-), which identifies false positives. For illustration the values of Equalized odds and Demographic parity are displayed in Table 8. There Abbreviations EqOp (Equality of opportunity), HEO (Hamming equalized odds) and DemPar (Demographic Parity) are used. It shows when tuning Equality of opportunity for fairness, the Hamming equalized odds (-) is low. Demographic parity shows a similar pattern to Hamming equalized odds (-). Hamming equalized odds (-) is even lower when tuned for accuracy. This is generally also the case, although it just as Equality of opportunity is relatively high for the *compas* dataset.

Table 8: Mean Accuracy and Fairness Results when tuned for Equality of Opportunity

Adult		Accuracy	EqOp	HEO(-)	DemPar
Adult	Tuned for Acc	82.8(0.4)	7.0(1.7)	6.3(0.8)	9.2(0.4)
	Tuned for Fair	82.3(0.4)	1.3(0.8)	4.9(0.8)	9.4(0.6)
Compas	Tuned for Acc	68.2(1.8)	22.2(5.5)	16.9(6.5)	16.9(3.1)
	Tuned for Fair	53.0(1.6)	0.0(0.0)	0.0(6.5)	0.0(0.0)
Default	Tuned for Acc	82.1(0.4)	0.2(1.5)	1.0(0.3)	1.3(0.6)
	Tuned for Fair	77.9(0.4)	0(0.3)	0.2(0.3)	0.0(0.0)
German	Tuned for Acc	73.2(3.5)	2.4(6.4)	0.5(0.7)	1.8(1.8)
	Tuned for Fair	70.0(2.9)	0.0(0.0)	0.0(0.7)	0.0(0.0)

Note : EqOp is equivalent to HEO(+)

Three sets of graphs are made per dataset. In Figure 4 and Appendix A.3 Figures 11-13, the Hamming loss (-) term with respect to the Equality of opportunity gap target always finds its optimal point between the values 0 and 0.2 for epsilon. This suggests a value can be found here where Equalized odds is relatively fair. In practice, this case is not more useful than optimizing for Equalized odds immediately. In terms of the false positive rate with respect to the Equality of opportunity gap target (as in Figure 4c), the scale of the y-axis is important to take into account. Coherent to the tables, the largest gaps between groups are to be found for the *compas* dataset, then *adult* and followed by much smaller differences for German *credit* and *default*. Coherent to earlier findings, a higher complexity target leads to higher false positive rates for all datasets in Appendix A.3 figures 11-13. Only in Figure 4b we observe a slight decrease. Graphs of the effects of the Equality of opportunity target and complexity on the Demographic parity gap are displayed in Appendix A.4 Figures 14-16.

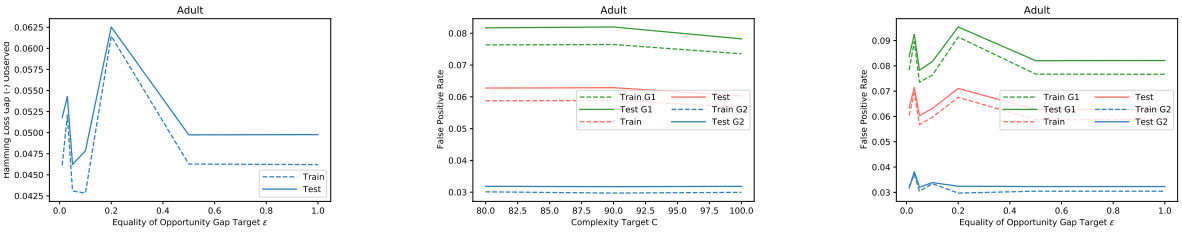

Figure 4: Impact of the fairness constraint (a) on hamming loss (-) and the false positive rate (c) and the impact of the complexity target on the false positive rate (b).

4.8.2 Tuned for Equalized odds

Since EqOp and HEO(-) in Table 9 are tuned for when considering Equalized odds, we focus on Demographic parity. When tuned for accuracy, again the disparity between groups is greater when considering

datasets *adult* and *compas*. When tuned for fairness all disparities are zero. This is further elaborated on in the Discussion section. Graphs of the effects of the Equalized odds target and complexity on the Demographic parity gap are displayed in Appendix A.5 Figures 17-19.

Table 9: Mean Accuracy and Fairness Results when constrained with Equalized odds

	Accuracy	EqOp	HEO(-)	DemPar
Adult Tuned for Acc	80.7(1.2)	2.1(5.0)	12.5(5.4)	7.2(1.2)
Tuned for Fair	75.9(0.7)	0.0(5.0)	0.0(1.3)	0.0(0.0)
Compas Tuned for Acc	67.1 (1.6)	21.1 (8.9)	12.2(5.4)	15.1(2.0)
Tuned for Fair	53.0(1.0)	0.0(8.9)	0.0(5.0)	0.0(0.0)
Default Tuned for Acc	82.0(0.2)	0.0(0.2)	1.2(1.0)	1.1(0.5)
Tuned for Fair	82.0(0.2)	0.0(0.2)	0.0(1.0)	0.0(0.0)
German Tuned for Acc	72.4(3.0)	5.9(2.2)	9.5(0.7)	1.7(1.6)
Tuned for Fair	70(2.8)	0.0(2.2)	0.0(0.7)	0.0(0.0)

Note: EqOp is equivalent to HEO(+)

4.8.3 Tuned for Demographic parity

Next, we consider the scores for Equality of opportunity and Equalized odds when constraining Fair CG for Demographic parity. Equality of opportunity is equal to the (+) term for Equalized odds. The results are displayed in Table 10.

Table 10: Mean Accuracy and Fairness Results for Demographic Parity

	Accuracy	DemPar	EqOp	HEO(-)
Default	78.0(0.7)	0.0(0.0)	0.0(0.0)	0.0 (0.0)
Adult	24.1(0.8)	19.6(0.6)	0.0(0.0)	13.0(4.0)
Compas	47.0(0.7)	13.2(4.6)	0.0(0.0)	16.6(3.5)
German	70.0(2.4)	0.0(0.0)	0.0(0.0)	0.0(2.2)

Note : The values for tuning for accuracy and fairness coincide for this table.

For *default* and *german* fair classifications are found. Only for *adult* and *compas* a relatively high score for unfairness in Demographic parity is accompanied by a relatively high score for the gap of the false positive rate (HEO(-)). As the results are similar to when tuned for fairness or accuracy, only one value is displayed. When tuned for Demographic parity, the unfairness measured by Equality of opportunity is minimal.

5 Discussion

In the discussion we consider the questions asked in the introduction: Does the Fair CG algorithm perform similarly when applied to different datasets? Which fairness metric leads to the highest accuracy in prediction? In addition, special attention is given to taking the fairness metrics simultaneously into account. We structure the discussion by separately treating the different fairness metrics and at last discussing the comparison between them.

5.1 Equality of opportunity

Similar to Lawless and Günlük (2021) we observe Equality of opportunity as a relatively straightforward fairness measure producing promising results. Three out of four datasets have relatively low fairness for optimal accuracy. Furthermore, Equality of opportunity gaps of close to 0 can be obtained with accuracies that are still competitive. There is a clear trade-off between relaxing the fairness constraint in training and the realized false negative rates. The rapid differences in Equality of opportunity when its target is relaxed shows neither data or algorithm is likely to inherently capture fairness if not specifically being programmed to do so. Likewise, a clear fairness-accuracy trade-off is seen in Figure 1c. However, when fairness restrictions are relaxed to increase accuracy, gains are very skewed to the early and relatively low decreases in fairness. After this first burst of accuracy-increase, only a slight increase in accuracy is witnessed while relaxing the fairness constraint further. Therefore, this shows satisfactory fairness can be accompanied by decent accuracy.

If the user of the algorithm were solely maximizing accuracy, then the fraction of correct classifications is expected to be highest. In the *compas* example this would be 70%. Nonetheless, if the Equality of opportunity gap were to be kept at 0.1, then around 4% of correct classifications would be sacrificed to ensure the difference in false negative rates between groups is low. Ceteris paribus this would imply the opportunity for the disadvantaged group would increase (or the opportunity for the group with more abundant positive classifications might decrease) with at least an overall cost of more false classifications. This leads to two conclusions. First, the implications of using fairness metrics must be carefully researched. In other words, what are the effects of the fairness measure on the classification performance for different groups? Secondly, in a policy setting it might help to set fairness standards and optimize accuracy constrained to these standards.

5.2 Equalized odds

As Equalized odds includes Equality of opportunity it is valuable to compare Equality of Opportunity when optimizing for either of the two. When tuned for accuracy, there is more fairness in terms of Equality of opportunity if Equalized odds is used as a constraint for three out of four datasets. Only for the German *credit* dataset does fairness decrease when optimizing for Equalized odds. The unfairness measures are for both very low when tuned for fairness, there is no notable difference. As Equalized odds is the stricter fairness measure, it takes a higher toll on accuracy. When tuned for accuracy, the difference

between the accuracy found by using Equality of opportunity is only 1 to 2-percent points higher than for Equalized odds. The additional cost of establishing similar false positive rates among groups seems low for the additional benefit concerning fairness. In deciding between Equality of Opportunity and Equalized odds, the context of the problem should be taken into account. Classifying the risk of recidivism (*compas*) might be different than classifying a customer as a bad risk (*credit*).

The Fair CG algorithm achieves 2-percent points lower optimal accuracy for the German *credit* dataset than the competing benchmarks. However, the accompanying sum of positive and negative term parts of equalized odds is considerably lower. The German *credit* dataset was evaluated using gender as the sensitive attribute. The dataset also contained foreign as a potential sensitive attribute, but given that the Fair CG algorithm is not able to provide sensible classifications for this setting, we find this to be a clear example for the need of fairness. When a group with potentially lower opportunity or odds is too small to be classified, it is hard for an algorithm to come up for their rights.

5.3 Demographic parity

The third fairness metric was not implemented in the Fair CG setting before. The measure itself, namely similar positive prediction rates among groups, seems straightforward. However, the implementation proved more challenging. For instance, special attention needs to be given to the misclassification of negatively labeled data and the pricing problem. The expectation of a relatively more straightforward fairness metric leading to potentially high accuracy and/or high fairness is not met. The Fair CG algorithm performed poorly on *adult* and *compas* datasets to the extent that random guessing in expectation leads to twice the accuracy of the Fair CG algorithm. This problem did not occur for the *default* and German *credit* datasets. With this formulation for Fair CG, Demographic parity is not competitive to Equality of opportunity and Equalized odds. This is also found by analyzing the benchmark results. The accuracies of CART and LR unanimously outperform those from Fair CG, yet the gaps in Demographic parity are remarkably high even when the results are tuned for fairness. For the *adult* dataset LR finds the fairest classifier, with a cost of 5.5-percent points in accuracy in comparison to CART. This result is surprising, given the actual positive rates in the datasets were not particularly unequally distributed. This finding tempers the optimism suggested by Lawless and Günlük (2021) after finding promising results for Equality of opportunity and Equalized odds. Namely, for other fairness metrics there is still enough work to be done.

5.4 Comparison among fairness metrics

As inherently implementing Demographic parity has proven to be difficult, we resort to finding the measure while optimizing for Equality of opportunity or Equalized odds. The reason for this is two-fold. In addition to the earlier mentioned reason, satisfying different fairness metrics is regarded as difficult by the literature. Nevertheless, if an algorithm is fair by one definition and unfair by another, this may cause confusion about the common understanding of fairness. Lack of research into this area urges us to keep looking for optimal fairness across definitions.

Equalized odds is a stricter fairness condition than Equality of opportunity in terms of false classifications. The question of interest is whether Demographic parity differs when tuned for one or the other. These differences are low. When tuned for accuracy the Demographic disparity is lower when tuned for Equalized odds for all datasets, although the difference is only larger than 11% for the *adult* dataset. The results when tuned for fairness only differ for the *adult* dataset. There the Demographic disparity when tuned for Equality of opportunity disappears when tuning for Equalized odds.

We consider the values for Equality of opportunity and Equalized odds when constraining the algorithm for Demographic parity. We see that for *adult* and *compas* the unfairness with respect to Demographic parity is shared with unfairness for the false positive rates (HEO(-)). As the positive rates differ (due to the unfairness in the Demographic parity measure), this gets expressed in a difference in false positive rates. In a worst-case scenario, one group receives more positive classifications, which are predominantly false, resulting in bad scores for both fairness metrics. This shows another side of the coin. Given correlated fairness metrics, if one is unfair, other metrics follow.

The aim of this paper is ambitious. Namely, finding a classifier that is accurate, interpretable and fair. Depending on the fairness constraint, interpretability and fairness in the boolean rule set setting do not generally influence each other greatly. However, both are shown to have a negative relationship with accuracy. One suggestion is to sacrifice interpretability for fairness. However, in delicate matters such as fairness, understanding the reasoning behind decision-making is key. This obviously is harder with a more complicated model. In the conclusion several suggestions will be laid out to take further steps in balancing the respective trade-offs between interpretability and fairness and their effects on accuracy.

6 Conclusion

This paper extended the research into Boolean rule sets constrained with Equality of opportunity and Equalized odds to the German *credit* dataset and Demographic parity. Furthermore, the various fairness metrics are evaluated when the model is constrained to a different fairness metric. We started by asking the question:

What is the effect of different fairness metrics on the accuracy of an interpretable machine learning model?

To answer this question, we first investigated if the Fair CG algorithm performs comparatively when it is applied to different datasets. The addition of the German *credit* dataset taught us two things: First, the results are not very different from the *default* dataset in terms of fairness, which is in the same domain. Second, with a 1000 rows it was clearly smaller than the other datasets, and this was noticed in its performance. Overall it is clear that FairCG had more difficulty with two datasets (*compas* and *credit*) to create a combination of high accuracy and fairness, in comparison to the other datasets where this was more doable. Next, we considered which fairness metric led to the highest accuracy in prediction. It was known that Equality of opportunity as a less strict measure could attain higher accuracy than Equalized odds. Beforehand, it was expected that Demographic parity, which does not make assumptions

about labels, might lead to higher accuracy than when accounting for other fairness measures. However, this was not the case. Generally, when we tune for fairness, all three fairness metrics attain very low levels. This showed instead of only considering fairness metrics separately, attempts to satisfy multiple definitions of fairness at once have potential.

In conclusion, among these fairness metrics, Equality of opportunity leads to the highest accuracy. However, more difficult is finding the optimal balance between fairness and accuracy. In experiments we have shown the tendency for a combination of still-low unfairness and already high accuracy. Whether that is optimal depends on the objective as well as on research outside of economics and computer science. The fairness we require should not be in the hands of the machine as it is a question for human society.

This research intended to take a next step to fair and interpretable classification. In this new field it is sensible to start with fairness definitions that are straightforward to grasp, which led to investigating Demographic parity. However, there is plenty of room for improvement. The time limitation led to only using 5-fold cross-validation. Although the effect on the larger datasets was small, higher cross-validation might lead to better results for the German *credit* dataset. Also, the Fair CG algorithm subject to the Demographic parity constraint was not able to provide competitive classification for the *compas* and *adult* datasets. Modifying the formulation as an attempt to fix this problem could greatly elevate the value of this research. Lastly, here follows a range of suggestions for further research:

- In this research the focus was on binary prediction, which can be intuitively evaluated. Nevertheless, uncertainty was not taken into account. If the model could express its degree of certainty about a prediction, this could greatly increase its practical use. Developing this might be a bigger challenge for boolean rule sets than for other types of algorithms. For neural networks for instance, there exist methods for this (Ribeiro, Singh, & Guestrin, 2016b).

- Econometric methods (here referred to as algorithms) were mostly used for research and finance. With the rapid advancements in machine learning the intersection between algorithms and economics has quickly become more important, while it seems that the moral and ethical considerations for fairness fall behind. The use of machines gives us a chance to break with unfair practices. A strong question is whether the machines can fix this for us, or if we first need to elevate human's moral standards.

- We advise to investigate the details of produced rule sets, something which should be done in cooperation with domain experts, such as lawyers, psychologists and sociologists. In the case of *compas*; Properly compare the outcomes of the algorithms with human judgment.

- The use of the German *credit* dataset has shown what is needed to make fairness standard practice in machine learning. Large data samples from different places and times, and proper checks of the labels provided. This high cost might well be worth the added predictive power of algorithms.

- Consider solving the problem in an optimization context where tuning for fairness and accuracy is done through weighting both aspects to find optimal values for the fairness-accuracy trade-off.

References

Aghaei, S., Azizi, M. J., & Vayanos, P. (2019). Learning optimal and fair decision trees for non-discriminative decision-making. In *Proceedings of the AAAI conference on artificial intelligence* (Vol. 33, pp. 1418–1426).

Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016). *Machine bias: there's software used across the country to predict future criminals. and it's biased against blacks. propublica 2016.*

Arneson, R. J. (1999). Against rawlsian equality of opportunity. *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition*, *93*(1), 77–112.

Austin, W., & Williams III, T. A. (1977). A survey of judges' responses to simulated legal cases: Research note on sentencing disparity. *J. Crim. L. & Criminology*, *68*, 306.

Bazaraa, M. S., Jarvis, J. J., & Sherali, H. D. (2008). *Linear programming and network flows.* John Wiley & Sons.

Berk, R., Heidari, H., Jabbari, S., Joseph, M., Kearns, M., Morgenstern, J., ... Roth, A. (2017). A convex framework for fair regression. *arXiv preprint arXiv:1706.02409*.

Berrendorf, M., Faerman, E., Vermue, L., & Tresp, V. (2020). Interpretable and fair comparison of link prediction or entity alignment methods with adjusted mean rank. *arXiv preprint arXiv:2002.06914*.

Cabitza, F., Rasoini, R., & Gensini, G. F. (2017). Unintended consequences of machine learning in medicine. *Jama*, *318*(6), 517–518.

Chen, J., Kallus, N., Mao, X., Svacha, G., & Udell, M. (2019). Fairness under unawareness: Assessing disparity when protected class is unobserved. In *Proceedings of the conference on fairness, accountability, and transparency* (pp. 339–348).

Conforti, M., Cornuéjols, G., Zambelli, G., et al. (2014). *Integer programming* (Vol. 271). Springer.

Corbett-Davies, S., & Goel, S. (2018). The measure and mismeasure of fairness: A critical review of fair machine learning. *arXiv preprint arXiv:1808.00023*.

Dash, S., Günlük, O., & Wei, D. (2018). Boolean decision rules via column generation. *arXiv preprint arXiv:1805.09901*.

Dua, D., & Graff, C. (2017). *UCI Machine Learning Repository.* Retrieved from `http://archive.ics.uci.edu/ml`

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., & Zemel, R. (2012). Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference* (pp. 214–226).

Elman, J. L., & Zipser, D. (1988). Learning the hidden structure of speech. *The Journal of the Acoustical Society of America*, *83*(4), 1615–1626.

Freitas, A. A. (2014). Comprehensible classification models: a position paper. *ACM SIGKDD explorations newsletter*, *15*(1), 1–10.

Geden, M., & Andrews, J. (2021). Fair and interpretable algorithmic hiring using evolutionary many-objective optimization.

Goodman, B., & Flaxman, S. (2017). European union regulations on algorithmic decision-making and a "right to explanation". *AI magazine*, *38*(3), 50–57.

Gurobi Optimization, I. (2021). *Gurobi optimizer reference manual.* `http://www.gurobi.com`.

Hardt, M., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning. *arXiv preprint arXiv:1610.02413*.

Kamiran, F., Calders, T., & Pechenizkiy, M. (2010). Discrimination aware decision tree learning. In *2010 IEEE international conference on data mining* (pp. 869–874).

Kearns, M., Neel, S., Roth, A., & Wu, Z. S. (2018). Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. In *International conference on machine learning* (pp. 2564–2572).

Kehrenberg, T., Bartlett, M., Thomas, O., & Quadrianto, N. (2020). Null-sampling for interpretable and fair representations. In *European conference on computer vision* (pp. 565–580).

Kusner, M. J., Loftus, J. R., Russell, C., & Silva, R. (2017). Counterfactual fairness. *arXiv preprint arXiv:1703.06856*.

Lawless, C., & Günlük, O. (2021). Fair and interpretable decision rules for binary classification. In *Ijcai 2021 workshop on ai for social good.*

Lemm, S., Blankertz, B., Dickhaus, T., & Müller, K.-R. (2011). Introduction to machine learning for brain imaging. *Neuroimage*, *56*(2), 387–399.

Makridakis, S., Spiliotis, E., & Assimakopoulos, V. (2018). Statistical and machine learning forecasting methods: Concerns and ways forward. *PloS one*, *13*(3), e0194889.

Mannila, H., & Toivonen, H. (1996). Multiple uses of frequent sets and condensed representations: Extended abstract. In *Proc. of the 2nd international conference on knowledge discovery and data mining (kdd'96)* (pp. 189–194).

Menon, A. K., & Williamson, R. C. (2018). The cost of fairness in binary classification. In *Conference on fairness, accountability and transparency* (pp. 107–118).

Mita, G., Papotti, P., Filippone, M., & Michiardi, P. (2020). Libre: Learning interpretable boolean rule ensembles. In *International conference on artificial intelligence and statistics* (pp. 245–255).

Ofer, D. (2017). *Compas recidivism racial bias.* `https://www.kaggle.com/danofer/compass`.

Pedreshi, D., Ruggieri, S., & Turini, F. (2008). Discrimination-aware data mining. In *Proceedings of the 14th acm sigkdd international conference on knowledge discovery and data mining* (pp. 560–568).

Pleiss, G., Raghavan, M., Wu, F., Kleinberg, J., & Weinberger, K. Q. (2017). On fairness and calibration. *arXiv preprint arXiv:1709.02012*.

Quinlan, J. R. (1987). Simplifying decision trees. *International journal of man-machine studies*, *27*(3), 221–234.

Rajkomar, A., Hardt, M., Howell, M. D., Corrado, G., & Chin, M. H. (2018). Ensuring fairness in machine learning to advance health equity. *Annals of internal medicine*, *169*(12), 866–872.

Ravikumar, P. (2017). *Barrier methods.* `cs.cmu.edu/ pradeepr/convexopt/Lecture_Slides/barrier-methods.pdf`.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016a). "Why should I trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 1135–1144).

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016b). Model-agnostic interpretability of machine learning. *arXiv preprint arXiv:1606.05386*.

Romano, Y., Bates, S., & Candès, E. J. (2020). Achieving equalized odds by resampling sensitive attributes. *arXiv preprint arXiv:2006.04292*.

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, *1*(5), 206–215.

Verma, S., & Rubin, J. (2018). Fairness definitions explained. In *2018 ieee/acm international workshop on software fairness (fairware)* (pp. 1–7).

Wang, C., Han, B., Patel, B., Mohideen, F., & Rudin, C. (2020). In pursuit of interpretable, fair and accurate machine learning for criminal recidivism prediction. *arXiv preprint arXiv:2005.04176*.

Yao, S., & Huang, B. (2017). Beyond parity: Fairness objectives for collaborative filtering. *arXiv preprint arXiv:1705.08804*.

Zemel, R., Wu, Y., Swersky, K., Pitassi, T., & Dwork, C. (2013). Learning fair representations. In *International conference on machine learning* (pp. 325–333).

Zhang, J., & Bareinboim, E. (2018). Equality of opportunity in classification: A causal approach. In *Proceedings of the 32nd international conference on neural information processing systems* (pp. 3675–3685).

A Appendix

A.1 Equality of opportunity

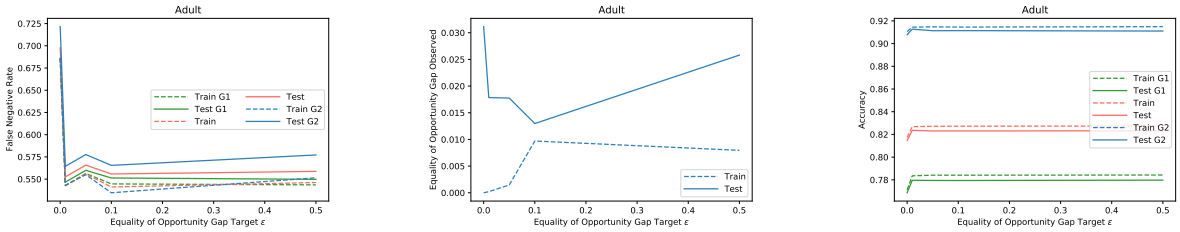

Figure 5: Impact of the fairness target on the false negative rate(a), acquired fairness (b) and accuracy (c) for the *adult* dataset.

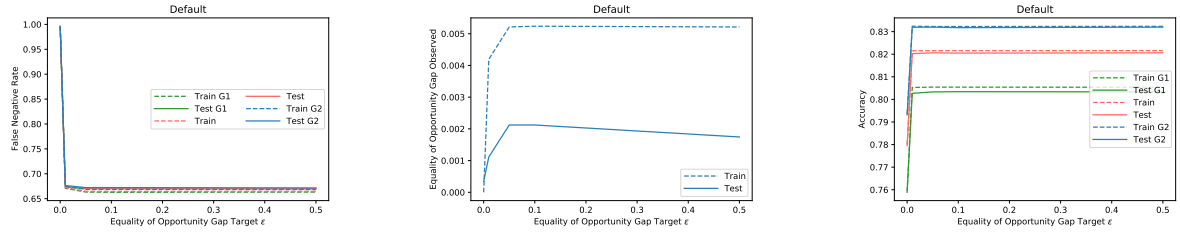

Figure 6: Impact of the fairness target on the false negative rate(a), acquired fairness (b) and accuracy (c) for the *default* dataset.

A.2 German *credit* Equalized odds

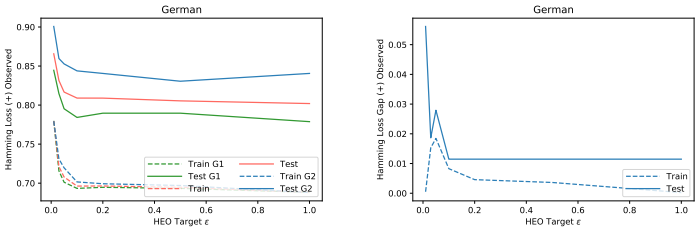

Figure 7: Impact of the Equalized odds target on Hamming loss (+) observed.

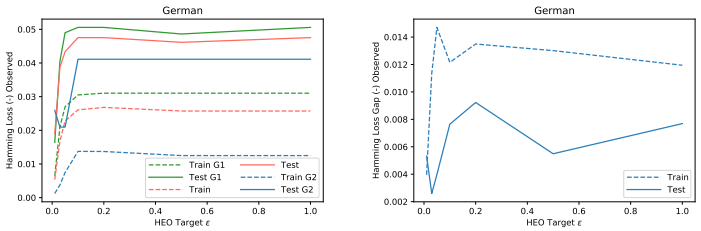

Figure 8: Impact of the Equalized odds target on Hamming loss (-) observed.

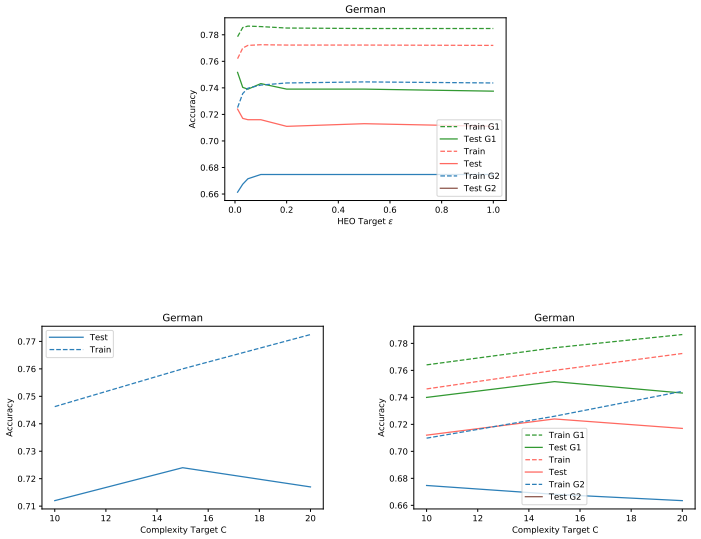

Figure 9: Impact of the complexity targets and Equalized odds target on accuracy for the German *credit* dataset.

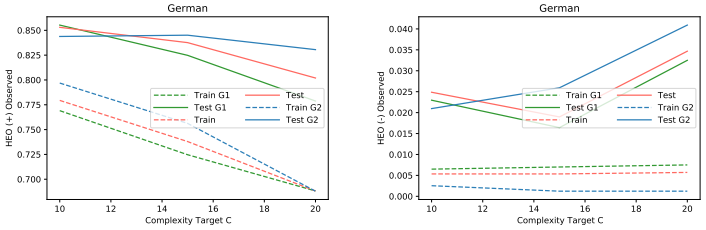

Figure 10: Impact of the complexity target on Equalized odds (+) and (-) observed for the German *credit* dataset.

A.3 Equalized odds results when tuning for Equality of opportunity

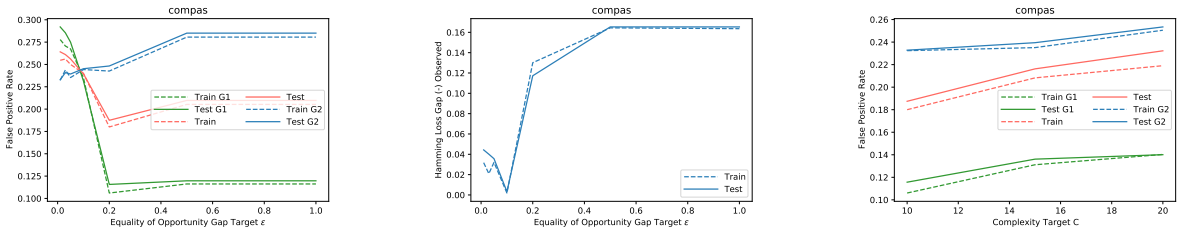

Figure 11: Impact of the fairness target on the false positive rate(a) and hamming negative loss gap (-) (b) and the impact of the complexity target on the false positive rate (c) for the *compas* dataset.

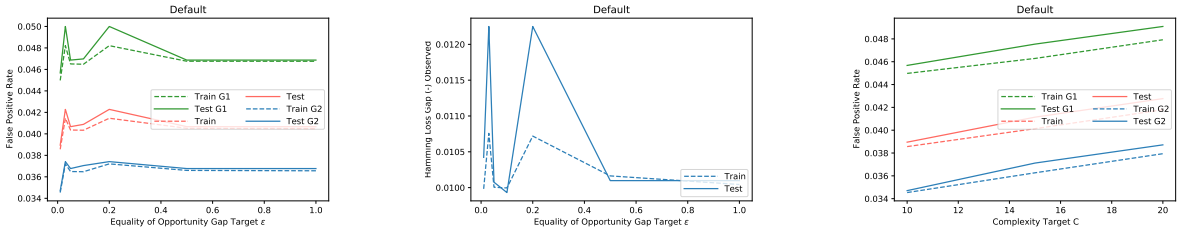

Figure 12: Impact of the fairness target on the false positive rate (a) and hamming negative loss gap (-) (b) and the impact of the complexity target on the false positive rate (c) for the *default* dataset.

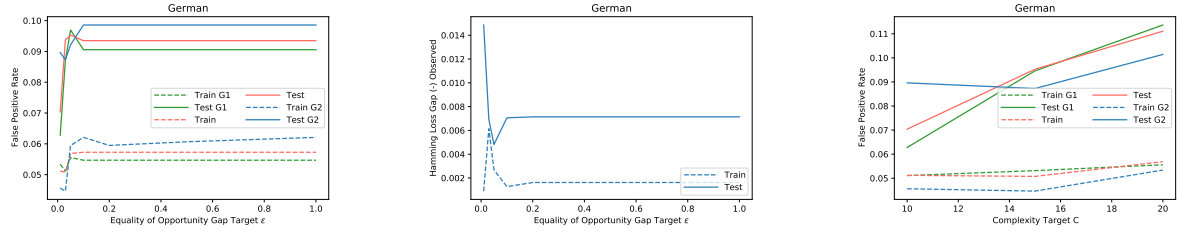

Figure 13: Impact of the fairness target on the false positive rate (a) and hamming negative loss gap (-) (b) and the impact of the complexity target on the false positive rate (c) for the German *credit* dataset.

A.4 Demographic parity results when tuning for Equality of opportunity

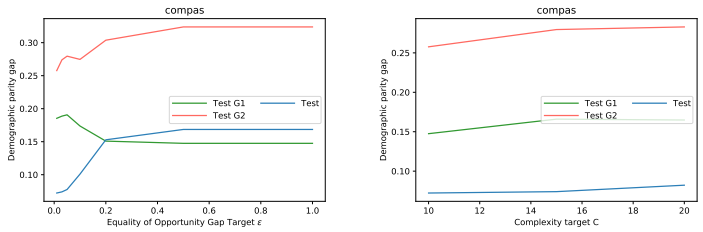

Figure 14: Impact of the Equality of opportunity target and complexity targets on the Demographic parity gap for the *compas* dataset.

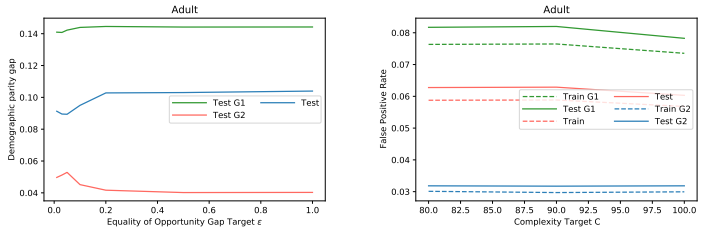

Figure 15: Impact of the Equality of opportunity target and complexity targets on the Demographic parity gap for the *adult* dataset.

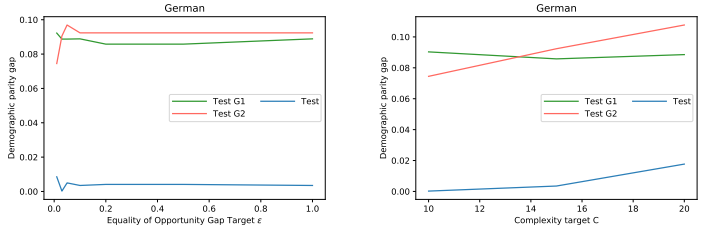

Figure 16: Impact of the Equality of opportunity target and complexity targets on the Demographic parity gap for the German *credit* dataset.

A.5 Demographic parity results when tuning for Equalized odds

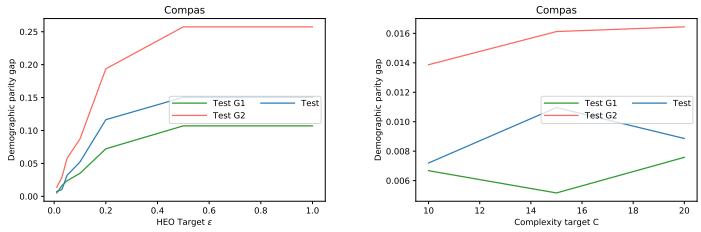

Figure 17: Impact of the Equalized odds target and complexity targets on the Demographic parity gap for the *compas* dataset.

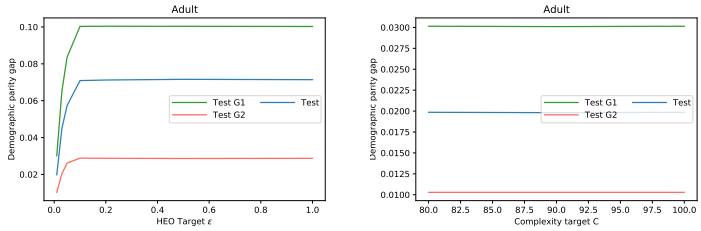

Figure 18: Impact of the Equalized odds target and complexity targets on the Demographic parity gap for the *adult* dataset.

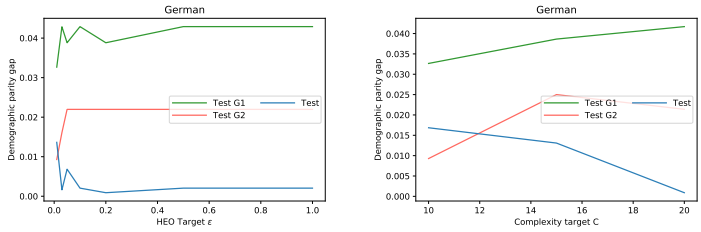

Figure 19: Impact of the Equalized odds target and complexity targets on the Demographic parity gap for the German *credit* dataset.

A.6 README of the code

A ZIP file of the code is handed in with this thesis and can be distributed upon request. Attached is the README of the code. Folders are in bold, files in italic. The code is split up in two main folders: **BRS** (Boolean Rule Sets) and **BRS_demographic**.

- **BRS** consists of the code for fairness metrics Equality of Opportunity and Equalized odds
- **BRS_demographic** consists of the code for the Demographic parity metric

Both folders have a similar structure:

- The different datasets can be found in the **data** folder
- The **Graphing** folder consists of three folders considering datasets with accuracy and fairness performance (**Results**), these are used to create graphs (**Notebooks**) which are saved to the **Graphs** folder
- The subfolders in the **Results** folder in the **Graphing** folder all contain a notebook to merge the result text files such that they can be used to compute statistics or make graphs
- The results from the **Graphing** folder are generated from the **Results** folder These results are generated in *EqOp Tests.ipynb* for Equality of opportunity and *Hamming EO Tests.ipynb* for Equalized odds. In the **BRS_demographic** folder this is the *DemPar Tests.ipynb* file
- The Benchmark algorithms are computed in the **Benchmark** folder

The algorithm works due to a couple of python files:

- The *test_helpers.py* file is called as a file which calls the other files
- The **fairness_modules** folders consist of the different representations for the different fairness metrics
- The *master_model.py* file, the **rule_generator** folder and the *CompactDoubleSidedMaster.py*, *Classifier.py*, *binerizer.py* and *DNFRuleModel.py* files are called by the *test_helpers.py* file

When considering Demographic parity a couple of changes are made:

- In the **fairness_modules** folder the *DemographicParity.py* file is created
- *The test_helpers.py* file is adapted to compute the demographic parity gaps between the groups
- The master model *CompactDoubleSidedMaster.py* is adapted for the column generation constraints and the pricing problem
- The *GeneralRuleGenerator.py* file includes the coefficient for Demographic parity in the pricing problem objective