# Achieving fair personalization policies in marketing

**Bachelor Thesis Business Analytics and Quantative Marketing**

**Otto Haanappel (497902)**

**ERASMUS UNIVERSITY ROTTERDAM**

**Erasmus School of Economics**

Supervisor: Deng, H

Second assessor: Frasincar, F

July 3, 2022

**Abstract**

As personalization might cause unintended discrimination in targeting policies, several methods have been proposed that mitigate algorithmic bias. Ascarza and Israeli (2022) propose the Bias-Eliminating Adapted Trees (BEAT) method that ensures fair personalization policies, while leveraging differences between individuals. Extensive analysis on the performance of BEAT is relevant for decision makers designing fair targeting policies and contributes to current academic literature on practical effectiveness of bias mitigating methods. This paper extends the research of Ascarza and Israeli (2022) with analysis on the Equalised Odds metric and implementation of Equalised Odds Post Processing method, introduced by Hardt et al. (2016), to provide more insights on the performance of BEAT. Experiment data of a Domino's Pizza promotion and Portuguese Bank marketing campaign is used to estimate targetability scores and design targeting policies accordingly. The methods used for estimation build on the Generalised Random Forest framework, designed to efficiently estimate heterogeneity. In addition, the Equalised Odds Post Processing (EOPP) method is implemented, which adjusts derived targeting policies to achieve Equalised Odds. This results in the comparison of several methods on performance and fairness. Following this analysis it can be confirmed that BEAT is an effective method in removing bias in practice while leveraging differences between individuals. BEAT combined with EOPP results in additional bias mitigation and performs well in both marketing applications. Therefore, BEAT-EOPP is a strong bias mitigating method that is effective in practice and can be used by policy makers to design fair marketing personalization policies.

# Contents

# 1 Introduction

The increased availability and collection of data in a wide variety of domains and for many applications has initiated the beginning of a new trend in decision making. Namely, decision makers nowadays use the seemingly endless source of information obtained from data for personalization of their policies with the aim to optimally target affected individuals. We are confronted with these targeted interventions on a daily basis through for instance pricing (tickets, insurance, renting), advertising (online ads, emails, personal promotions), medical treatments (organ allocation or Covid-19 Intensive Care availability), admissions (college, university, jobs) or news publication (social media). In arguably many cases these personalised policies offer the benefit of confronting an individual with their specific preferences, improving user experience and, if applicable, possibly profits. On the contrary, this algorithmic innovation appears to cause unintended discrimination, based on so called protective attributes (e.g., race, gender or age), in the allocation of resources through hidden correlation or relations between predictive variables. Numerous methods have been developed that aim to mitigate this bias. Recently, Ascarza and Israeli (2022) proposed the Bias-Eliminating Adapted Trees (BEAT) method that removes bias while preserving the benefits from leveraging differences in individual preferences. Their results show that BEAT performs well in removing bias according to the fairness metrics used in the paper, but this effectiveness might not be consistent for different metrics or applications. The Imbalance metric for group fairness in Ascarza and Israeli (2022) illustrates differences in distribution of protected attributes over treated and non-treated groups, leaving space for bias in the prediction errors across protected groups. Moreover, the performance of BEAT is not significantly better than the method where protected variables are removed from the data.

In this paper the research on BEAT is extended with analysis on the Equalised Odds (EO) fairness metric. EO is a well known metric for group fairness proposed by Hardt et al. (2016). EO requires equality in prediction errors and through this extension provides a new perspective on its performance. In addition, this paper implements the Equalised Odds Post Processing (EOPP) method that is introduced by Hardt et al. (2016) as well, which aims to optimise a targeting policy with respect to Equalised Odds. Moreover, the effectiveness of BEAT is validated with an additional marketing experiment. This extension is relevant to examine the general effectiveness of BEAT, providing more insights on its applicability. Providing additional evidence for a well performing practical bias mitigation method is of great relevance for decision makers that design policies with different objectives with regard to achieving fair personalization. This research also contributes to the academic field of algorithmic bias. The majority of previous literature focused on the theoretical background of algorithmic bias, whereas this paper

investigates practical use of bias mitigating methods. The aim of this research is thus to reveal algorithmic bias in practice and solve discrimination issues in targeting policies with practical solutions in a marketing context.

Therefore, the research question of this paper is: "Is the BEAT method effective in mitigating bias defined by the Equalised Odds metric?". To answer this question, experiment data of a Domino's Pizza promotion and Portuguese Bank marketing campaign, both including sensitive and non-sensitive attributes, is used to derive targeting policies using both causal inference and prediction forests. For the Domino's Pizza experiment causal inference is used to estimate treatment effects (CATE) of individuals, conditional on a set of explanatory variables. For the Portuguese Bank campaign prediction forests estimate targetability scores based on the outcome variable directly. The implemented methods are based on Generalised Random Forests (GRF), which are designed to efficiently estimate heterogeneous outcomes. Using different methods in these marketing applications, targeting policies are derived according to estimated targetability scores. The designed targeting policies for each method are then used to analyse and compare methods on efficiency and fairness, including the Equalised Odds metric. The EOPP method adjusts derived targeting policies from other methods by solving a linear program to achieve Equalised Odds.

The BEAT-EOPP method, that estimates targetability scores with BEAT to derive targeting policies and adjusts these with EOPP to achieve EO, performs best of all methods. In addition, the results show that BEAT is not only powerful in reducing imbalance while preserving a competitive performance, but reduces disparity in Equalised Odds as well. This finding confirms the capability of BEAT to remove algorithmic bias defined by different metrics and provides conclusive evidence on superior performance in practice compared with the other methods.

In Section 2 previous literature on algorithmic bias, mitigation methods and fairness is summarised. Section 3 describes the data for both marketing experiments. Next, the methodology of personalization methods and evaluation metrics is explained in Section 4. The results are given in Section 5 and finally in Section 6 and 7 the conclusion and discussion are given.

## 2 Literature review

**Algorithmic Bias** Bias in algorithms can be interpreted as systematic disproportionate targeting of individuals from different groups in a protected variable. Algorithmic bias has been extensively examined in previous literature. Earlier researches aimed to detect biases in data or algorithms and examine its consequences, whereas more recent literature also developed methods that eliminate algorithmic bias or unfairness. In order to remove discriminating bias, an

intuitive adaption by companies and organisations is removing protected variables from their data, assuming that discrimination by their algorithm is resolved. The problem that arises after the elimination of these protected variables is that it is very likely and frequently observed that the protected variables affect the predicted value through other unprotected variables that were not eliminated, in essence thus creating omitted variable bias (Pope and Sydnor, 2011). As long as the variables correlated with the protected attributes are not removed as well, the bias can not be (completely) eliminated. On the other hand, if one decides to remove these variables the resulting model would lose predictive power and becomes inefficient. This trade off between predictive accuracy and removing bias is crucial in the development of methods that eliminate bias while also capturing value from the predictions that lead to targeted interventions.

**Bias mitigation methods** In general, there are three stages in the process of personalization where methods attempt to mitigate algorithmic bias. These stages are related to the preparation, prediction and allocation phases of designing personalization policies. First, data can be subject to pre-processing methods such that biased data is transformed before being used. For instance, Feldman et al. (2015) propose such a method with the aim to change only the unprotected variables so that the ability to classify can be preserved as much as possible and Johndrow and Lum (2019) similarly describe a generally applicable method for creating an adjusted set of covariates that are independent of protected characteristics. Lohia et al. (2019) argue that these methods are generally only successful in removing bias defined according to group fairness rather than both group and individual fairness. Secondly, rather than removing bias within the data, methods have been proposed that eliminate bias in the allocation stage using constraints in optimisation problems. Goh et al. (2016) and Agarwal et al. (2018) describe how constraint optimisation removes bias defined according to different notions of fairness. These methods however do not seem to work well in the case of classification algorithms where multiple protected variables are being used, i.e. high dimensionality (Ascarza and Israeli, 2022). A third category of bias mitigating methods focuses on the prediction stage, rather than pre- (data transformation) or post-processing (allocation constraints). Kamishima et al. (2012) propose a technique to reduce indirect prejudice, meaning statistical dependency between sensitive and non sensitive variables, that is applicable to various prediction algorithms with probabilistic discriminating models. Recently, Ascarza and Israeli (2022) propose the Bias-Eliminating Adapted Trees (BEAT) method that removes bias, by ensuring a balanced allocation of resources across individuals while preserving the benefits from heterogeneity in individuals.

The methods used and proposed in Ascarza and Israeli (2022) build on the General Random Forest (GRF) framework. The technicalities of the GRF are carefully explained in detail by Athey

et al. (2019) and its adaption for causal inference, Causal Forests, by Wager and Athey (2018). In addition to the Equalised Odds criterion, Hardt et al. (2016) propose a Post Processing method that achieves Equalised Odds by solving a linear program. This method is implemented in this research to achieve fair targeting policies in marketing applications, in addition to the methods from Ascarza and Israeli (2022).

**Fairness definitions** Defining bias and unfairness in more detail is of crucial importance as different fairness metrics could imply different results and might be used in specific applications, depending on the objective of a decision maker. Group and individual fairness have therefore been thoroughly discussed in literature resulting in many different, but also somewhat identical fairness metrics. First of all, within the context of group fairness the defined metrics can be organised in three categories: Independence, separation and sufficiency, as given by Castelnovo et al. (2022). The criterion of independence states that the decisions should be independent of any protected attributes (Barocas et al., 2019). This group fairness definition is also known as demographic or statistical parity. Adjustments to this criterion can be made by conditioning on additional information, obtaining conditional demographic parity. These independence criteria do not utilise any information on the true target, but solely use the distribution of features and decisions. Therefore, Barocas et al. (2019) describe the concept of separation to define group fairness where the probability is also conditioned on the true target. Hardt et al. (2016) introduced equalised odds, where fairness is measured with differences in prediction errors between protected groups. In addition, a relaxation of this equalised odds is the equal opportunity definition that only requires equal true positive rates. Furthermore, sufficiency is described as third category by Castelnovo et al. (2022) and takes the perspective of people that are given the same model decision. More specifically, it requires parity among the treated groups irrespective of sensitive features (Barocas et al., 2019).

However, Dwork et al. (2012) explain that statistical parity could indicate fairness, while from the point of view of an individual, the outcome is seen as unfair. Castelnovo et al. (2022) add to this and states that since group fairness requires to satisfy conditions only on average among groups, it leaves room to bias discrimination inside the groups. Therefore, defining individual fairness has been addressed in literature as well. For example, Dwork et al. (2012) define individual fairness as two individuals who are similar with respect to the unprotected variables at hand should be classified similarly. Analysis on mitigating bias is thus dependent on the definitions of fairness used.

# 3 Data

For this research two data sets related to marketing experiments are analysed. The first data set that is used is obtained from a Domino's Pizza experiment conducted in 2022 by Ascarza and Israeli (2022) via MTurk. The experiment consisted of an A/B Test where individuals in the control condition were offered a choice between $5 gift cards to Panera Bread and to Domino's Pizza. In the treatment condition, participants were given a similar choice, but the Domino's Pizza gift card had a value of $10, representing the idea of a $5 coupon promotion. The experimental data of this treatment is publicly available on the Harvard Business School Dataverse (Ascarza and Israeli). The data describes 3,146 individuals with 206 variables. The data set contains two experiment variables, corresponding to the treatment allocation and final responses to the promotion, and 204 explanatory variables. These explanatory variables consist of 190 unprotected variables and 14 variables on three protected attributes: Age, race and gender. Descriptive statistics (mean and standard deviation) of these sensitive attributes and

Table 1: Descriptive statistics of experiment variables and protected attributes from the Domino's Pizza experiment data.

| | Variable | Mean | SD |
|---|---|---|---|
| Experiment variables | Gift card choice | 0.563 | 0.496 |
| | Treatment group | 0.499 | 0.500 |
| Protected attributes | Age | 41.49 | 13.00 |
| Race | American Indian / Alaska Native | 0.012 | 0.112 |
| | Asian | 0.090 | 0.286 |
| | Black / African American | 0.097 | 0.297 |
| | Hispanic / Latino | 0.056 | 0.233 |
| | Native Hawaiian / Pacific Islander | 0.004 | 0.060 |
| | Middle Eastern | 0.007 | 0.082 |
| | North African | 0.001 | 0.030 |
| | White | 0.774 | 0.421 |
| | Other | 0.007 | 0.085 |
| Gender | Male | 0.435 | 0.496 |
| | Female | 0.556 | 0.497 |
| | Nonbinary | 0.006 | 0.080 |
| | Other | 0.002 | 0.047 |

experiment variables are shown in Table 1. As the shown features are dummy variables, the mean values represent the fraction of individuals belonging to the corresponding group. Moreover, the unprotected variables describe brand preferences, consumption behaviour and personal information of the participated individuals. For example, participants were asked to rate their preferences on brands in various categories such as sportswear, apparel, restaurants and technology/social media on a three-point scale (dislike, indifferent or like). In addition, individual behaviour is measured on a six-point scale, ranging from 'never' to 'on a daily basis'. This consumer behaviour of the participant describes activity related variables such as social media activity, food delivery frequency, exercise frequency and coupon usage. Furthermore, personal information such as education, income, spending and environmental consciousness is included in the unprotected variables.

The second data set that is analysed in this research is related to a telephone marketing campaign of a Portuguese banking institution. The campaign consisted of phone calls to a list of clients to sell a term deposit between 2008 en 2010. Hence, the outcome of the campaign is a binary list indicating unsuccessful or successful contacts, with a success rate of approximately 11% as can be seen from Table 2. The data of this experiment is made publicly available by Moro et al. (2014). The full data set provided consist of 41,188 observations, but for computational purposes a stratified subset is used in this paper, resulting in 4,119 observations with 45 variables. The 44 explanatory variables consist of 5 protected and 39 unprotected variables. The protected variables in this data describe age and marital status. Descriptive statistics of the outcome variable and protected variables are shown in Table 2. The unprotected variables describe each individual with categorical variables type of job, education level, existing credit in default, housing/personal loans and numeric variables employment variation rate, consumer

Table 2: Descriptive statistics of experiment variables and protected attributes from the Portuguese Bank experiment data.

| | Variable | Mean | SD |
|---|---|---|---|
| Experiment variable | Term deposit decision | 0.11 | 0.31 |
| Protected attributes | Age | 40.11 | 10.31 |
| Marital Status | Married | 0.609 | 0.488 |
| | Divorced | 0.108 | 0.311 |
| | Single | 0.280 | 0.449 |
| | Other | 0.003 | 0.052 |

price index, consumer confidence index, number of employees and information on outcomes of previous campaigns.

For both experiments, the protected variable Age is split into two equal quantiles. This is necessary for the determination of privileged and unprivileged subgroups in the Equalised Odds computation explained in Section 4. This split results in two age categories for both data sets, depending on the quantile for Age in each data. For the Domino's Pizza experiment, causal inference is used to estimate treatment effects based on the available data and targeting policies are designed according to a decision rule. In the Portuguese Bank experiment, prediction forests directly estimate targetability scores based on the outcome variable, similarly resulting in targeting policies based on a specified targeting rule. This is explained in more detail in the next Section.

# 4 Methodology

In this Section, the methods used for deriving targeting policies and corresponding evaluation metrics are explained. First, the methods used in Ascarza and Israeli (2022) and the Equalised Odds Post Processing method implemented in this paper are described. Thereafter, the evaluation metrics from Ascarza and Israeli (2022) and the Equalised Odds metric are explained. More specifically, as illustrated in Ascarza and Israeli (2022), three types of models are used being Causal Forest, Debiasing method and the proposed BEAT method. All methods are evaluated with three metrics that reflect model performance, group fairness and individual fairness. This analysis is thus extended with the Equalised Odds metric and Equalised Odds Post Processing method.

For convenience, first the notation is introduced that is used throughout the paper. Suppose we have $n$ observations, corresponding to the number of experiment participants, given by $i = 1, .., n$. Each observation consists of a feature vector, where $\boldsymbol{x}_i$ denotes the vector of unprotected features and $\boldsymbol{z}_i$ the vector of protected features of individual $i$, respectively. $Y_i \in \{0, 1\}$ denotes the outcome variable and $W_i \in \{0, 1\}$ is the treatment indicator from the actual data for individual $i$. Throughout the paper the designed targeting policy indicator, obtained by one of the methods, for individual $i$ is denoted with $\widehat{W}_i$. Note that this is not a treatment prediction, but targeting decision based on a decision rule using predicted targetability scores which is explained next.

### 4.1 Personalization methods

Personalization is obtained by designing targeting policies, i.e. the allocation of treatments over a set of individuals, that maximise the goal or outcome of a policy (e.g., maximising profits or response rates). In order to maximise that outcome, the main objective when designing targeting policies is to identify heterogeneity in the object of interest, which is treatment effect for causal inference and the outcome variable for prediction forests. This paper derives targeting policies using both causal inference (Domino's Pizza experiment) and prediction forests (Portuguese Bank experiment). Identifying the individuals on whom a treatment would have the most predicted effect could result in better performing targeting policies, where performance depends on the policy objective. For the application of causal inference, the object of interest is thus treatment effect, which is also known as the Conditional Average Treatment Effect (CATE). For individual $i$ this is given by

$$\tau_i(x) = \mathrm{E}\left[Y_i^{(1)} - Y_i^{(0)} | \boldsymbol{x}_i = x\right]. \tag{1}$$

Here $Y_i^{(1)}$ and $Y_i^{(0)}$ denote the potential outcomes individual $i$ would have experienced with and without the treatment, conditional on covariates $\boldsymbol{x}_i$ that can include both protected and unprotected variables. Heterogeneous treatment effects can be estimated with Causal Forests (CF), which build on the Generalised Random Forest (GRF) framework, discussed by Athey et al. (2019). The concept of Causal Forest is explained in detail by Wager and Athey (2018) and a brief explanation is included in the Appendix A. Based on the identified heterogeneity in treatment effect across individuals, the method estimates a targetability score $\hat{\tau}_i(x)$ for each individual, indicating the potential effect of assigning a treatment to the corresponding individual. These scores can be used to design the targeting policy, by assigning treatments to the individuals with the highest predicted targetability scores. If for instance the target rate of a personalization policy is 50%, one would assign treatments to the individuals corresponding with the 50% highest scores. In the paper of Ascarza and Israeli (2022) three different methods are implemented to obtain these targetability scores. These method respectively are Causal Forest, Debiasing and the proposed BEAT method.

For the application of prediction forests, where the object of interest is the outcome variable, the expected outcome variable $\mu_i$ is predicted rather than the treatment effect $\tau_i$. The estimation is done with a regression forest instead of a causal forest, resulting in targetability scores based on the outcome variable. For the Portuguese Bank data, where no treatment allocation in the data is applicable, this outcome variable is the individual's decision of subscribing a term deposit. With the resulting vector of expected outcome predictions $\boldsymbol{\mu}$, targeting policies can be designed

in the same manner as for the causal inference. Given a target rate $t$, treatments can be assigned to the individuals corresponding to the $t\%$ highest targetability scores.

**Targeting methods** In this research different methods are used to derive targeting policies from targetability scores. For the analysis on the Domino's Pizza experiment causal inference is used and hence the methods estimate targetability scores with Causal Forests (CF). For the analysis on the Portuguese Bank experiment prediction forests are used and hence the methods estimate targetability scores with Regression Forests (RF). Both the CF and RF are applications of the GRF and hence RF operate equivalently to CF, but estimate targetability scores based on the outcome variable and not on treatment effects. In other words, no data on actual treatment allocation is available for prediction forests and hence treatment effects can not be estimated. The various methods used in this research are explained next.

First, a CF with the full data (CF-FD), including both protected and unprotected variables, is used to predict the targetability scores for the Domino's Pizza experiment. In addition, CF without protected variables (CF-NP), thus including unprotected variables only, is used as well. These two methods hence differ only in the data used for estimation. CF-FD is the benchmark model where no intervention with respect to bias mitigation is implemented and CF-NP represents the naive approach of removing protected variables from the data as a solution for bias mitigation. For the Portuguese Bank experiment this is equivalent to implementing both a regression forest with full data (RF-FD) and without protected variables (RF-NP), where the difference with CF is that no treatment vector $\boldsymbol{w}$ is available for estimation and is hence omitted in the input data.

Secondly, for both experiments a Debiasing method is implemented which aims to debias the unprotected variables with a random forest where the protected variables $\boldsymbol{Z}$ are the features and the unprotected variables $\boldsymbol{X}$ the outcome. The predicted values from the random forest are subtracted from the original $\boldsymbol{X}$ variables. In other words, the predictive information of $Z$ on $X$ is removed from $X$ with the aim to make it unrelated with respect to $Z$. Mathematically this is given by $\boldsymbol{X}_{Debiased} = \boldsymbol{X}_{Original} - f_x(\boldsymbol{Z})$, where $\boldsymbol{X}$ and $\boldsymbol{X}_{Debiased}$ denote the matrices of original and debiased unprotected variables respectively and $\boldsymbol{Z}$ denotes the matrix of protected variables. Next, the debiased unprotected variables are used in a CF along with the original outcome and treatment vectors $\boldsymbol{y}$ and $\boldsymbol{w}$ resulting in the CF-DB method. Again, the debiasing method using regression forests (RF-DB) for the Portuguese Bank experiment operates similarly, but estimates targetability scores based on $\boldsymbol{X}_{Debiased}$ and $\boldsymbol{y}$ and hence no treatment vector $\boldsymbol{w}$.

Finally, the BEAT method proposed by Ascarza and Israeli (2022) is implemented. The goal of this method is to, similar to the GRF, identify heterogeneity across individuals. However,

BEAT aims to estimate heterogeneity that is unrelated to the protected attributes. For that reason, BEAT build on the GRF framework but differs from GRF in the way splits in the forest are determined. GRF determines the split based on the divergence in the outcome of interest and BEAT adjusts this splitting criterion by adding a penalty term that penalises differences in the protected attributes between the split nodes. The resulting split criterion is called Balanced Divergence (BD) and maximising it when searching for splits ensures that all resulting trees are balanced with respect to the protected attributes (Ascarza and Israeli, 2022). The resulting split criterion is given by

$$BD(C_1, C_2) = \underbrace{\widehat{\Delta}(C_1, C_2)}_{\text{GRF split criterion}} \underbrace{- \gamma \, \text{Dist}\left( \overrightarrow{\overrightarrow{Z}}_{C_1} \middle| X_{k,s}, \overrightarrow{\overrightarrow{Z}}_{C_2} \middle| X_{k,s} \right)}_{\text{BEAT added penalty}}, \tag{2}$$

where $C_1$ and $C_2$ denote the child nodes, $\gamma$ is the penalty parameter and $X_{k,s}$ is the splitting point for $C_1$ and $C_2$ at dimension $k \in K$. As equation (2) shows, maximising the BD criterion implies minimising the penalty term and thus ensures a balance with respected to protected variables in the child nodes, given by $Z_{C_1}$ and $Z_{C_2}$. The value for the $\gamma$ parameter can be adjusted and determines the weight given to the balance in the protected attributes. Increasing $\gamma$ reduces the imbalance, but at the cost of efficiency. BEAT is implemented in this research with three different values for $\gamma$: $\gamma \in \{3, 5, 8\}$. Once this maximisation procedure has been completed for all splitting points and trees are fully grown, the complete forest is obtained. BEAT proceeds the same as GRF once the full forest is obtained. Hence, targetability scores are predicted for each individual. Given the balanced divergence criterion these scores are referred to as Conditional Balanced Targetability (CBT) scores (Ascarza and Israeli, 2022). CBT thus measures the adjusted treatment effect predictions, conditional on a set of unprotected variables such that there is balance with regard to the protected attributes. Finally, after BEAT estimates CBT scores for each individual, these can be used to determine the optimal targeting allocation by selecting the individuals belonging to the top $t\%$ of CBT scores, given a target rate of $t$. Note that BEAT similarly predicts heterogeneity unrelated to protected variables in outcome variable $\mu_i$ rather than $\tau_i$, for the Portuguese Bank experiment.

**Equalised Odds Post Processing** This paper extends the analysis on BEAT with the bias mitigation method Equalised Odds Post Processing (EOPP), introduced by Hardt et al. (2016). The Equalised Odds (EO) metric is defined in more detail in the next Section, but for better understanding of the method the EO criterion is defined as:

$$\Pr[\widehat{W} = 1 | Z = a, W = y] = \Pr[\widehat{W} = 1 | Z = b, W = y]. \tag{3}$$

Note that $\widehat{W}$ is a derived targeting policy obtained by the implemented methods, $a, b \in Z$ are the classes or groups in a protected variable and $y = \{0, 1\}$ is the true treatment allocation. For $y = 1$, the criterion requires that $W$ has equal True Positive Rates (TPR) with respect to the protected attribute $Z$ and $y = 0$ corresponds with equal False Positive Rates (FPR). Equalised Odds Post Processing solves a linear program, based on TPR and FPR, to find probabilities with which to change output labels from a previous obtained (binary) targeting policy to optimise Equalised Odds (Bellamy et al., 2019).

For the marketing experiments studied in this research, this amounts to changing the the obtained labels $\widehat{W}$, i.e. the derived targeting policy. The resulting new targeting policy $\widetilde{W}_p$ is given by a set of parameters, corresponding with probabilities of changing labels, given by $p = (p_{00}, p_{01}, p_{10}, p_{11})$, where $p_{wa} = \Pr[\widetilde{W} = 1 | \widehat{W} = w, A = a]$ with $\widetilde{W}$ and $\widehat{W}$ the new and original targeting vectors. Now, define

$$\gamma_0(\widetilde{W}) = \Big(\Pr\{\widetilde{W} = 1 | A = 0, W = 0\}, \Pr\{\widehat{W} = 1 | A = 0, W = 1\}\Big) \qquad \text{and}$$

$$\gamma_1(\widetilde{W}) = \Big(\Pr\{\widetilde{W} = 1 | A = 1, W = 0\}, \Pr\{\widehat{W} = 1 | A = 1, W = 1\}\Big) \qquad \text{and}$$

Then, the Equalised Odds criterion is defined to be satisfied if $\gamma_0(\widetilde{W}) = \gamma_1(\widetilde{W})$.

The objective function of the linear program is a loss function $\mathbb{E}\left[\ell(\widetilde{W}_p, W)\right]$, which takes a pair of labels and returns a value indicating the loss of predicting $\widetilde{W}$ when the correct label is $W$. An optimal targeting policy satisfying the Equalised Odds criterion is now derived by the following optimisation problem:

$$\begin{aligned} \min_p \quad & \mathbb{E}\left[\ell(\widetilde{W}_p, W)\right] \\ \text{s.t.} \quad & \gamma_0\left(\widetilde{W}_p\right) = \gamma_1\left(\widetilde{W}_p\right) \\ & \forall_{w,a} 0 \leqslant p_{wa} \leqslant 1. \end{aligned} \tag{4}$$

Finally, as an unprocessed targeting policy $\widehat{W}$ one can choose any method (predicting targetability scores) that derive a targeting policy. In this paper, the resulting targeting policies from GRF without protected variables (CF-NP/RF-NP) and BEAT ($\gamma = 8$) are post processed by the EOPP method, given the effectiveness of the two methods shown by Ascarza and Israeli (2022).

The Equalised Odds Post Processing method is implemented in Python with the package from Bellamy et al. (2018).

### 4.2 Evaluation metrics

The explained methods are evaluated with evaluation metrics that reflect performance and fairness, including the Efficiency, Imbalance and Delta-Policy metrics from Ascarza and Israeli (2022) and the Equalised Odds metric implemented as an extension in this paper.

**Efficiency** For the analysis on model performance, the efficiency metric is used. Efficiency is defined as the proportion of users choosing the discounted product (i.e., market share) if the firm was to target 50% of the population (Ascarza and Israeli, 2022). Hence, this metric depends on the outcome $Y_i$, a binary indicator of whether participants chose the Domino's gift card, the actual treatment value $W_i$ and the derived targeting assignment $\widehat{W}_i$ (based on $\tau_i$ prediction) from a method. To compute efficiency, for each individual the Inverse Probability Score (IPS) is computed. This score is a well-known metric in the field of causal inference as unbiased estimates of CATE can be obtained with IPS using propensity scores (Austin and Stuart, 2015). Normally, this IPS can be computed as

$$IPS_i = \begin{cases} \frac{1}{\hat{e}(x} & \text{if} \quad W_i = 1, \\ \frac{1}{1-\hat{e}(x)} & \text{if} \quad W_i = 0. \end{cases}$$

However, for the comparison of the various methods that are implemented in the analysis, the Efficiency metric is normalised such that 0% correspond to the case where no targeting interventions are made. Hence, by including an additional term in the IPS computation, which excludes individuals that are targeted (non-targeted) while in fact are not (are) in the treatment group, the Efficiency metric denotes the percentage increase in the final outcome $Y_i$ as a consequence of implementing the targeting policy. With this adjustment, the IPS score for individual $i$ can be computed as

$$IPS_i = \begin{cases} \frac{1}{\hat{e}(x} & \text{if} \quad W_i = \widehat{W}_i, W_i = 1, \\ \frac{1}{1-\hat{e}(x)} & \text{if} \quad W_i = \widehat{W}_i, W_i = 0. \end{cases}$$

Propensity score $\hat{e}(x)$ is the proportion of targeted individuals in the actual data, computed as $\hat{e}(x) = \frac{1}{N}\sum_{i=1}^{N} \mathrm{E}\left[W_i | \boldsymbol{x}_i = x\right]$ and hence the probability of receiving a treatment. Using the notation introduced in the beginning of this Section, the formula for efficiency can finally be written as

$$\textit{Efficiency} = \frac{1}{N}\sum_{i=1}^{N} Y_i * IPS_i.$$

**Imbalance** Imbalance is used in Ascarza and Israeli (2022) as a metric for group fairness and is defined as the average distance between standardised protected attributes of targeted and non-targeted individuals in the test data. To allow for comparison of the imbalance between different protected variables, standardised values are used in the computation. For protected variable $Z_k$ this standardisation is given by $\widetilde{Z}_{k,i} = \frac{Z_{k,i} - \overline{Z}_k}{S_{Z_k}}$. Here, $Z_{k,i}, \overline{Z}_k$ and $S_{Z_k}$ denote the value of observation $i$, the mean and standard deviation of protected variable $k \in K$, the set of protected variables. Imbalance is computed as

$$Imbalance = \sum_{k=1}^{K} \left( \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}_{\{Targeted\}} \widetilde{Z}_{k,i} - \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}_{\{Non-targeted\}} \widetilde{Z}_{k,i} \right)^2,$$

where $\mathbb{1}_{\{Targeted\}}$ is an indicator function taking the value 1 when individual $i$ is assigned in the targeting policy and 0 otherwise. Similarly, $\mathbb{1}_{\{Non-targeted\}}$ is an indicator function taking the value 1 when individual $i$ is not targeted and 0 otherwise.

**Delta-Policy** In Ascarza and Israeli (2022) individual fairness is indicated with a metric called Delta-Policy. This metric is measured as the percentage of individuals for whom the outcome would change if their most important protected attribute was different. This change depends on the type of protected attribute. For binary attributes, this change in $Z_k$ is accomplished by using $1 - Z_k$. For continuous attributes, the data is standardised and then moved by $\pm 1$ SD, according to Ascarza and Israeli (2022). In the experiment setting this paper replicates, age is the most important feature and due to continuity is thus changed by $\pm 1$ SD for each customer. Let the protected variable for age be denoted with $Z_1$. More specifically, 1 SD is added if the value in the vector of standardised values for $Z_1$ is negative and 1 SD is subtracted if that value is non-negative. Mathematically, the age adjustment for individual $i$ is given by

$$Z_{1,i,New} = \begin{cases} \frac{Z_{1,i} - \overline{Z_1}}{S_{Z_1}} - S_{Z_1} & \text{if} \quad \frac{Z_{1,i} - \overline{Z_1}}{S_{Z_1}} \geq 0 \\ \frac{Z_{1,i} - \overline{Z_1}}{S_Z} + S_{Z_1} & \text{if} \quad \frac{Z_{1,i} - \overline{Z_1}}{S_{Z_1}} < 0 \end{cases}$$

where $Z_{1,i}$ is the age observation for individual $i$, $Z_{1,i,New}$ the corresponding adjusted standardised age attribute and $S_{Z_1}$ the standard deviation of the protected variable Age. Next, for each individual the given adjustment is made and these changes are collected in a new protected variable vector $Z_{1,New}$, which replaces the old vector in the test data. With this adjusted set of protected variables $Z$ the targetability scores $\boldsymbol{\tau}$ or $\boldsymbol{\mu}$ are estimated again. Using these adjusted vector of targetability scores, a new targeting policy is designed and finally the value for Delta-Policy can be obtained by computing the percentage of individuals for whom the outcome changes due to the adjusted protected variable.

**Equalised Odds** The concept of Equalised Odds is introduced by Hardt et al. (2016) and requires the positive outcome to be independent of a protected class, conditional on the actual outcome. This is mathematically given by

$$\Pr[\widehat{W} = 1 | Z_k = a, W = y] = \Pr[\widehat{W} = 1 | Z_k = b, W = y]. \tag{5}$$

Here $\widehat{W}$ is the derived targeting vector as explained in the previous Section. The classes or groups in protected variable $Z_k$ are given by $a$ and $b$, but the criterion can be generalised for $\forall \{a, b\} \in K$. Finally, $W = \{0, 1\}$ is the true treatment allocation from the experiment data. For $y = 1$, the criterion requires that $W$ has equal True Positive Rates (TPR) with respect to the the protected attribute $Z_k$. For $y = 0$, the criterion equals False Positive Rates (FPR). In

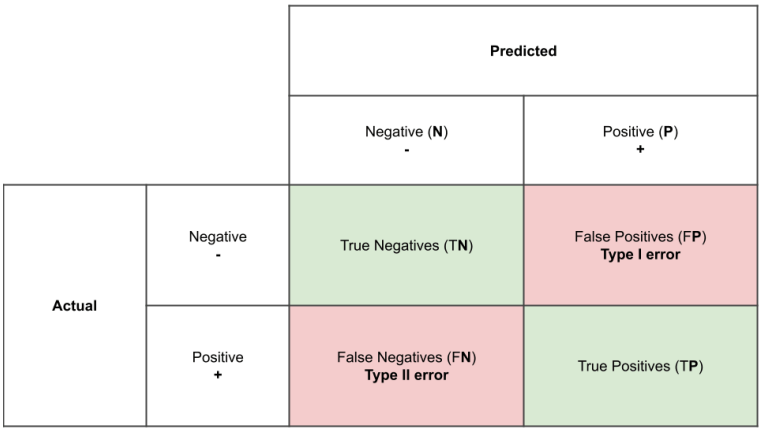


Figure 1: Confusion Matrix

words, TPR is the ratio of correctly positive predictions (TP) and actual positives, which is the sum of false negatives (FN) and true positives (TP). In a similar way, FPR is defined as the ratio of falsely positive predictions (FP) and actual negatives, the sum of true negatives (TN) and false positives (FP). TPR and FPR are based on the so called Confusion Matrix shown in Figure 1. The TPR and FPR with respect to a protected variable $Z_k$ can be computed as

$$TPR_k = \frac{TP_k}{FN_k + TP_k} \qquad \text{and} \qquad FPR_k = \frac{FP_k}{TN_k + FP_k},$$

where all individual statistics are obtained with respect to protected variable $k \in K$, with $K$ the set of protected variables.

To determine the Equalised Odds criterion specified in equation (5) it is hence required to compute the TPR and FPR rates with respect to the different protected variables for each method. Let the Equalised Odds (EO) denote the metric implemented in this paper. The EO metric measures the average of absolute differences in TPR and FPR for unprivileged and privileged groups, across the protected variables in the data.

For simplicity, this research examines Equalised Odds between two subgroups in a protected variable, hence one subgroup is defined as the privileged group and one as the unprivileged group. More specifically, for each protected variable $k \in K$, the TPR and FPR is computed for every subgroup (i.e., the privileged and unprivileged groups). The EO metric then takes the average of the absolute differences between the TPR and FPR of both subgroups. Next, the average of this absolute difference for all protected variables is taken. This procedure is repeated 100 times and the average value of these repetitions results in the final EO metric. Hence, the final EO metric can be computed as

$$EO = \frac{1}{K} \sum_{1}^{K} \frac{|FPR_{k,unprivileged} - FPR_{k,privileged}| + |TPR_{k,unprivileged} - TPR_{k,privileged}|}{2}. \tag{6}$$

Table 3: Replication results of Table 2 from Ascarza and Israeli (2022), where results are average values of 1,000 iterations.

| Method | Efficiency | Imbalance | Delta-Policy (in %) |
|---|---|---|---|
| CF-FD | 0.578 (0.050) | 0.155 (0.063) | 16.7 (3.9) |
| CF-NP | 0.576 (0.048) | 0.056 (0.022) | 0 (-) |
| Debiased | 0.568 (0.049) | 0.242 (0.088) | 42.3 (3.7) |
| BEAT ($\gamma = 3$) | 0.572 (0.050) | 0.041 (0.017) | 0 (-) |
| BEAT ($\gamma = 5$) | 0.574 (0.051) | 0.042 (0.018) | 0 (-) |
| BEAT ($\gamma = 8$) | 0.561 (0.048) | 0.040 (0.017) | 0 (-) |

The perfect scenario where all subgroups in all protected variables have equal TPR and FPR would imply that this EO metric is zero (or is close to zero if some small difference is allowed). For the Domino's Pizza experiment the privileged groups in race, gender and age are white, male and young ($Age < 40$) and hence the unprivileged groups are non-white, non-male and old ($Age >= 40$). For the Portuguese Bank experiment the privileged groups in marital status and age are married and young ($Age < 39$) and hence the unprivileged groups are non-married and old ($Age >= 39$).

The Equalised Odds metric for the privileged and unprivileged subgroups in a protected variable are computed in R by implementing the package from Kozodoi and V. Varga (2021).

# 5 Results

To illustrate the performance of BEAT in removing algorithmic bias, first the obtained results for the Domino's Pizza experiment in Ascarza and Israeli (2022) are replicated and shown in Table 3. The results are average values of 1,000 iterations. In each iteration, the data is randomly split into a train and test sample. Implementing a 90/10 splitting ratio results in train and test samples of 2,831 and 315 observations, respectively. In particular, the estimated mean and standard deviation (in parentheses) of Efficiency, Imbalance and Delta-Policy are given. The rows in Table 3 correspond with the different methods used for estimation. Columns indicate the computed metrics Efficiency, Imbalance and Delta-Policy. The values shown for Delta-Policy are given as percentages and by construction, CF-NP and the BEAT models have zero Delta-Policy, given the elimination of Age.

From the results follow that BEAT is capable of removing a significant part of the Imbalance. However, the causal forest without protected attributes (CF-NP) does not perform significantly

worse in mitigating the bias compared to BEAT. Moreover, as can be seen from Table 3, comparing the methods on the achieved Efficiency does not yield much differences, providing evidence for effectiveness of these bias eliminating methods in practice. However, the values obtained for Efficiency raise the question whether imposing a personalised policy, i.e. any of the models from the table, is preferred over a random or uniform allocation, given the insignificant or little benefit from personalization. This follows from the results shown in Table S4 from the Appendix of Ascarza and Israeli (2022), where the random and uniform policy achieve an Efficiency of 0.564 and 0.661 with Imbalance 0.039 and 0.000, respectively. Note that for a uniform targeting policies also increased costs are involved.

In order to assess the flexibility of BEAT in ensuring fair targeting policies with respect to different metrics, the analysis for Table 3 is extended with the Equalised Odds (EO) fairness metric and corresponding Equalised Odds Post Processing (EOPP) method. The results obtained by implementing these extension in the analysis are shown in Table 4, where the results are average values of 100 iterations. In each iteration, the data is randomly split into a train and test sample. Implementing a 70/30 splitting ratio results in a train and test sample of 2,202 and 944 observations. The first column in Table 4 correspond with the different methods used for estimation, including the Equalised Odds Post Processing methods. The estimated mean and standard deviation (in parentheses) of Efficiency, Imbalance and Equalised Odds are given in the next three columns. The fifth column shows the percentages increase or decrease in the EO metric using the Causal Forest with full data (CF-FD) as the benchmark model, for which Equalised Odds is set to 100% for that purpose. Hence, the value in Table 4 in row 2, column 5 should be read: CF-NP generates only 44.2% of the disparity in Equalised Odds across protected

Table 4: Extension results on Domino's Pizza experiment data including the Equalised Odds (EO) metric and Equalised Odds Post Processing methods (EOPP), where results are average values of 100 iterations.

| Method | Efficiency | Imbalance | EO | EO (relative, in %) |
|---|---|---|---|---|
| CF-FD | 0.576 (0.025) | 0.118 (0.078) | 0.129 (0.044) | 100.0 |
| CF-NP | 0.570 (0.024) | 0.033 (0.015) | 0.057 (0.015) | 44.2 |
| CF-DB | 0.567 (0.028) | 0.226 (0.114) | 0.194 (0.056) | 150.7 |
| BEAT ($\gamma = 3$) | 0.575 (0.025) | 0.020 (0.008) | 0.046 (0.012) | 36.0 |
| BEAT ($\gamma = 5$) | 0.573 (0.025) | 0.019 (0.008) | 0.047 (0.013) | 36.4 |
| BEAT ($\gamma = 8$) | 0.566 (0.025) | 0.019 (0.009) | 0.046 (0.014) | 36.0 |
| CF-EOPP | 0.571 (0.027) | 0.020 (0.009) | 0.041 (0.013) | 31.6 |
| BEAT-EOPP ($\gamma = 8$) | 0.574 (0.024) | 0.014 (0.006) | 0.032 (0.012) | 25.0 |

groups obtained when using CF-FD.

Similar to the decrease in Imbalance, also the disparity in Equalised Odds can be decreased with both a Causal Forest without protected variables and BEAT. The results of the EO metric confirm capability of these methods in removing bias, by ensuring not only less Imbalance but better Equalised Odds across protected groups as well, compared to the CF-FD. Given the values for Efficiency, which are not significantly different across the methods, Equalised Odds thus gives additional evidence for the effectiveness of BEAT in capturing value from personalization, while ensuring a more balanced treatment allocation. Note that again the Debiasing method CF-DB is not capable of removing bias and significantly increases EO disparity as well as Imbalance, which is in line with the findings in Table 3. The results from the inclusion of EO in the analysis provides managers with the evidence that BEAT is able to mitigate bias with respect to multiple objectives and therefore effective in practice.

Furthermore, the extended research on BEAT also includes two methods, based on the Equalised Odds Post Processing. The obtained results from these methods are shown in the seventh and eight rows in Table 4. The results obtained by implementing the Equalised Odds Post Processing methods on CF-NP and BEAT show that the proposed method by Hardt et al. (2016) is extremely efficient in removing bias. Both Imbalance is largely removed and better EO is achieved with the EOPP extension. Both the CF-EOPP and BEAT-EOPP methods yield better results compared to their regular method without EOPP. However, BEAT-EOPP performs best by reducing disparity in EO with 10% compared to regular BEAT, achieving the lowest Imbalance and relatively high Efficiency. Thus, for the Domino's Pizza experiment BEAT performs well according to the evaluation metrics including EO, but the hybrid method of BEAT and EOPP achieves the best targeting policy in terms of Efficiency and fairness.

To validate the obtained results for the Domino's Pizza experiment, the analysis is repeated on a marketing campaign of a Portuguese Bank where targetability scores are estimated with respect to the outcome variable and hence Regression Forests (RF) are used rather than CF as in the Domino's Pizza experiment. These results are shown in Table 5, where the results are again average values of 100 iterations. The data is randomly split in each iteration, resulting in train and test samples of 2,883 and 1,236 observations.

From the results in Table 5 it can be concluded that BEAT is more effective in removing bias than other implemented methods, including the RF-NP that performed nearly as good as BEAT in the Domino's Pizza analysis. In addition, these results confirm effectiveness of the EOPP methods in adjusting targeting policies obtained from RF-NP and BEAT to achieve better EO, compared to the methods without EOPP. The fairness improvement of BEAT and EOPP does

Table 5: Extension results on Portuguese Bank Marketing data including the Equalised Odds (EO) metric and Equalised Odds Post Processing methods (EOPP), where results are average values of 100 iterations.

| Method | Efficiency | Imbalance | EO | EO (relative, in %) |
|---|---|---|---|---|
| RF-FD | 0.778 (0.026) | 0.052 (0.031) | 0.082 (0.041) | 100.0 |
| RF-NP | 0.778 (0.027) | 0.015 (0.009) | 0.061 (0.022) | 74.6 |
| RF-DB | 0.780 (0.027) | 0.105 (0.045) | 0.108 (0.033) | 132.0 |
| BEAT ($\gamma = 3$) | 0.748 (0.028) | 0.007 (0.004) | 0.040 (0.018) | 49.6 |
| BEAT ($\gamma = 5$) | 0.727 (0.031) | 0.006 (0.004) | 0.042 (0.020) | 51.3 |
| BEAT ($\gamma = 8$) | 0.728 (0.035) | 0.005 (0.003) | 0.042 (0.018) | 51.9 |
| RF-EOPP | 0.752 (0.042) | 0.005 (0.003) | 0.029 (0.012) | 35.9 |
| BEAT-EOPP ($\gamma = 8$) | 0.702 (0.048) | 0.004 (0.003) | 0.027 (0.011) | 33.0 |

appear to come at the expense of some loss in Efficiency, although these losses are not significantly large. Hence, the Portuguese Bank experiment validates the effectiveness of BEAT in different applications and provides additional evidence for the performance of EOPP extension in designing fair targeting policies. Depending on the objective of a manager, these results show how two effective bias mitigating methods can be applied in practice to ensure fair personalization.

# 6 Conclusion

As the source of data used to design personalization policies keeps expanding, the hazard of unintended discrimination in these targeting policies also increases. Therefore, research into developing bias mitigation methods that are effective in practice are of great relevance. The research in this paper further investigates the finding from Ascarza and Israeli (2022) that the Bias-Eliminating Adapted Trees (BEAT) method performs well in mitigating algorithmic bias in practical experiments. In order to provide additional insights and evidence on the effectiveness of BEAT from different perspectives, this paper answers the following research question: "Is the BEAT model effective in mitigating bias defined according to the Equalised Odds metric?". Extending the analysis on BEAT with the Equalised Odds metric, requiring equal true positive rates (TPR) and false positive rates (FPR) across protected groups, and corresponding Equalised Odds Post Processing method provides another perspective on the performance of BEAT. More specifically, analysis of marketing experiments from Domino's Pizza and Portuguese Bank show that BEAT performs significantly better compared to other existing methods in achieving fair

targeting policies with respect to Imbalance and Equalised Odds. Moreover, the BEAT-EOPP method proposed in this paper outperforms the other methods, including BEAT, and results in less Imbalance and better Equalised Odds with no severe Efficiency loss. Hence, BEAT-EOPP not only ensures equality of treatment among groups and individual parity, but in addition improves equal distribution of prediction errors across protected groups, both while capturing value from personalised targeting policies. This hybrid method, that uses BEAT for estimating balanced targetability scores to design targeting policies and solves a linear program to adjust this targeting policy with the aim to achieve Equalised Odds, results in the best targeting policies with respect to fairness. Hence, the findings in this research validate the effectiveness of BEAT with the Equalised Odds criterion and propose an extension of BEAT, BEAT-EOPP, that allows decision makers to further exploit fairness in their policies. In short, by revealing algorithmic bias in marketing experiments and resolving these discrimination issues with practical methods this research has provided decision makers with an effective bias mitigating method that is proven to perform well in different practical marketing experiments.

# 7 Discussion

Mitigation of algorithmic bias is closely related with defining fairness in the first place. Since no general applicable definition of fairness is available yet, generalisation of the obtained results to fairness measures outside of the scope in this paper is not possible. Also, this research strictly applied the implemented methods on marketing experiments with binary outcome variables. The findings therefore might not be consistent with applications that have different objectives. While being an extension of the research by Ascarza and Israeli (2022), it should be noted that also the research in this paper provides suggestions for further research. The analysis on Equalised Odds (EO) implied only one privileged and unprivileged group within a protected variable. Future research could examine effectiveness of BEAT and Equalised Odds Post Processing (EOPP) in achieving Equalised Odds when multiple subgroups are considered. In addition, this research focused on the application of a binary predictor and hence the Equalised Odds Post Processing (EOPP) methods were designed accordingly. Further examination of this method could provide additional insights on the applicability and effectiveness in different applications. Furthermore, BEAT(-EOPP) has been strictly compared to different methods that are based on the framework of Generalised Random Forests. Research into different methods, for example Neural Networks, would provide an interesting comparison of different machine learning methods.

## References


Alekh Agarwal, Alina Beygelzimer, Miroslav Dudík, John Langford, and Hanna Wallach. A reductions approach to fair classification. pages 60–69, 2018.

Eva Ascarza and Ayelet Israeli. Replication Data for 'Eliminating Unintended Bias in Personalized Policies Using Bias Eliminating Adapted Trees (BEAT)'. URL `https://doi.org/10.7910/DVN/EWEB0W`.

Eva Ascarza and Ayelet Israeli. Eliminating unintended bias in personalized policies using bias-eliminating adapted trees (BEAT). *Proceedings of the National Academy of Sciences*, 119(11), 2022.

Susan Athey, Julie Tibshirani, and Stefan Wager. Generalized random forests. *The Annals of Statistics*, 47(2):1148–1178, 2019.

Peter C Austin and Elizabeth A Stuart. Moving towards best practice when using inverse probability of treatment weighting (iptw) using the propensity score to estimate causal treatment effects in observational studies. *Statistics in medicine*, 34(28):3661–3679, 2015.

Solon Barocas, Moritz Hardt, and Arvind Narayanan. *Fairness in machine learning*. 2019.

Rachel K. E. Bellamy, Kuntal Dey, Michael Hind, Samuel C. Hoffman, Stephanie Houde, Kalapriya Kannan, Pranay Lohia, Jacquelyn Martino, Sameep Mehta, Aleksandra Mojsilovic, Seema Nagar, Karthikeyan Natesan Ramamurthy, John Richards, Diptikalyan Saha, Prasanna Sattigeri, Moninder Singh, Kush R. Varshney, and Yunfeng Zhang. AI Fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias, October 2018. URL `https://arxiv.org/abs/1810.01943`.

Rachel KE Bellamy, Kuntal Dey, Michael Hind, Samuel C Hoffman, Stephanie Houde, Kalapriya Kannan, Pranay Lohia, Jacquelyn Martino, Sameep Mehta, Aleksandra Mojsilović, et al. Ai fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias. *IBM Journal of Research and Development*, 63(4/5):4–1, 2019.

Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

Alessandro Castelnovo, Riccardo Crupi, Greta Greco, Daniele Regoli, Ilaria Giuseppina Penco, and Andrea Claudio Cosentini. A clarification of the nuances in the fairness metrics landscape. *Scientific Reports*, 12(1):1–21, 2022.

Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. page 214–226, 2012.

Michael Feldman, Sorelle A Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. Certifying and removing disparate impact. pages 259–268, 2015.

Gabriel Goh, Andrew Cotter, Maya Gupta, and Michael P Friedlander. Satisfying real-world goals with dataset constraints. *Advances in Neural Information Processing Systems*, 29, 2016.

Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29, 2016.

James E Johndrow and Kristian Lum. An algorithm for removing sensitive information: application to race-independent recidivism prediction. *The Annals of Applied Statistics*, 13(1): 189–220, 2019.

Toshihiro Kamishima, Shotaro Akaho, Hideki Asoh, and Jun Sakuma. Fairness-aware classifier with prejudice remover regularizer. pages 35–50, 2012.

Nikita Kozodoi and Tibor V. Varga. *fairness: Algorithmic Fairness Metrics*, 2021. URL `https://CRAN.R-project.org/package=fairness`. R package version 1.2.1.

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. pages 2847–2851, 2019.

Sérgio Moro, Paulo Cortez, and Paulo Rita. A data-driven approach to predict the success of bank telemarketing. *Decision Support Systems*, 62:22–31, 2014.

Devin G. Pope and Justin R. Sydnor. Implementing anti-discrimination policies in statistical profiling models. *American Economic Journal: Economic Policy*, 3(3):206–31, August 2011.

Stefan Wager and Susan Athey. Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523):1228–1242, 2018.

# Appendices

## A Causal Forest

CF are a version of General Random Forest (GRF), which are designed to efficiently estimate heterogeneous outcomes, and thus an extension of the Random Forest (RF) concept explained by Breiman (2001). The key differences between GRF and RF are the way splits are determined and how a final estimate is obtained.

In a RF, splits are created by minimising the in-sample prediction error of the node. However, in a CF this split is based on the divergence in the outcome of interest. As specified in Ascarza and Israeli (2022), the GRF split criterion as proposed by Athey et al. (2019), can be formulated as

$$\Delta(C_1, C_2) = \frac{n_{C_1} n_{C_2}}{n_P} \left( \hat{\theta}_{C_1}(\mathcal{J}) - \hat{\theta}_{C_2}(\mathcal{J}) \right)^2, \tag{7}$$

where $n_{C_1}, n_{C_2}$ and $n_P$ denote the number of observations in the children nodes $C_1, C_2$ and parent node $P$. $\hat{\theta}_{C_i}(\mathcal{J})$ denotes the solution to the estimating equation computed in the children node $i$, as specified in equation (4) by Athey et al. (2019).

The final estimate $\hat{\theta}(x)$ of a CF is calculated by combining similarity weights $\widetilde{\alpha}_i(x)$ and the local CATE estimate $\widehat{\tau}_i$, given by

$$\hat{\theta}(x) = \sum_{i=1}^{N} \widetilde{\alpha}_i(x) \widehat{\tau}_i.$$

Here $\widetilde{\alpha}_i(x)$ is computed as $\widetilde{\alpha}_i(x) = \frac{1}{B} \sum_{b=1}^{B} \widetilde{\alpha}_{bi}(x)$ with $B$ the number of trees. Moreover, $\widetilde{\alpha}_{bi}(x)$ is the frequency with which observation $i$ falls into the same leaf as $x$ and is computed as $\widetilde{\alpha}_{bi}(x) = \frac{\mathbf{1}(\{X_i \in L_b(x)\})}{|L_b(x)|}$. Here, $L_b(x)$ denotes the leaf in tree $b$ containing $x$.

## B Programming code

For the implementation of the explained methods and evaluation methods, a hybrid of programming in R and Python is used. The initiation of the estimation is done in R, where both data sets are prepared and formatted according to the input for the GRF methods. Next, the estimation of targetability scores with the Causal Forests, Regression Forests, Debiasing and BEAT methods is done in R. The corresponding optimal targeting policies, based on explained decision rule, are derived in R as well. To implement the Equalised Odds Post Processing method, the created train and test subsets and derived targeting policies (in each iteration) from GRF without protected variables and BEAT are exported as CSV files to Python. In Python, the original targeting policies are adjusted to achieve Equalised Odds and these are exported back

to R, where the final computation of different evaluation metrics is performed. A brief overview of the code resulting in Table 4 and 5 is given below.

1. First, the data is imported and formatted according to the GRF inputs
2. Protected variable Age is split into two categories based on quantiles, needed for Equalised Odds
3. Parameters are initiated (target rate, train/test split, BEAT penalties, random seed etc)
4. The following process is repeated 100 iterations to obtain the average values shown in Table 4 and 5
   (a) Random train/test split is made
   (b) The six methods (GRF-FD/GRF-NP/GRF-DB/BEAT3/BEAT5/BEAT8) are trained on the training sample
   (c) For each method, targetability scores (mu) are predicted for each individual
   (d) Targeting policies are derived from predicted scores using a decision rule (target rate)
   (e) For individual fairness, the Age attribute is adjusted for each individual
   (f) Evaluation metrics are computed: Efficiency, Imbalance and Delta-Policy
   (g) CSV with the random train/test split and targeting policies of GRF-NP and BEAT8 exported for Python implementation of Equalised Odds Post Processing method
   (h) For each method, Equalised Odds metric is computed with respect to each protected variable
5. Finally, from the obtained matrices including all results from the 100 iterations, means can be computed resulting in the final values for the six 'replication' methods given in Table 4 and 5
6. The code in "Thesis_Extension_EOPP_method_Otto_Haanappel.ipynb" implements the Equalised Odds Post Processing method from Hardt et al. in Python. This codes executes the following process 100 times
   (a) Import CSV files containing the train/test splits from R and get correct formats
   (b) Create dataset classes, which will be the input for the EOPP method
   (c) Define the derived targeting policies from R and add to the previously obtained classes
   (d) Define the privileged and unprivileged groups in a protected variable

(e) Implement the Equalised Odds Post Processing algorithm and obtain adjusted targeting

(f) Compute the evaluation metrics, containing the Equalised Odds metric

(g) Repeat the computation of Equalised Odds with respect to the other protected variables

7. Finally, export the obtained adjusted targeting policies and computed EO metrics in each iteration in CSV files to R

8. In R, import the created CSV files from Python to evaluate the adjusted policies with the evaluation metrics. This results in the final values for the EOPP methods in Table 4 and 5