Exploring the Impact of Engineering Choices on Fairness in Machine Learning Classifiers Across Intersectional Subpopulations

Leona Marie Schneider

Master's Thesis Psychology,
Methodology and Statistics Unit, Institute of Psychology
Faculty of Social and Behavioral Sciences, Leiden University
Date: 29.11.2024
Student number: [redacted]
Supervisor: Dr. Rüya Koçer

Abstract

Using machine learning (ML) classifiers in high-stakes situations, such as financial and health decisions, is becoming increasingly popular. Those classifiers can impact the outcome of essential matters for individuals and carry a social responsibility. There has been a demonstrated potential for classifiers showing disparate treatment for more vulnerable subpopulations. However, the classifier pipeline's design process often focuses on optimising prediction accuracy, disregarding the potential influence of engineering choices on the fairness outcome.

This research scrutinizes the influence of ML engineering choices on the fairness outcome of different subpopulations. For this, engineering choices and their impact on fairness were investigated in five decision points (missing data imputation, scarce outcome intervention, algorithm choice, hyperparameter tuning, and threshold setting). Furthermore, a simulated society modelling social benefit fraud detection was utilized and tested on various pipeline combinations. The pipeline fairness performances were tested under different data challenges characterized by scarcely distributed outcomes and non-linear decision boundaries.

The analysis demonstrated that the selected choices in the decision points influences the fairness outcome, often showing the potential to benefit advantaged subpopulations and harm disadvantaged subpopulations. The decision point of chosen scarcity intervention portrayed the overall highest impact on the fairness outcome.

Conclusively, it highlights how each step along the ML classifier pipeline should be evaluated for its fairness impact, taking a fairness-aware approach that does not solely focus on raising a model's accuracy but also incorporates ethical considerations.

Keywords: High-Stakes Algorithmic Decision-Making**,** Machine Learning Fairness**,** Bias in Machine Learning**,** Fairness Metrics**,** Intersectional Fairness**,** Disparate Impact and Treatment**,** Fairness-Accuracy Trade-off**,** Pipeline Design Choices**,** Hyperparameter Tuning and Fairness, Scarcity Intervention

Table of Content

Chapter 1 – Introduction

Algorithmic decision-making using machine learning (ML) is becoming increasingly prevalent in administrative and commerce applications, such as job applications, parole evaluations and loan requests (Pessach & Shmueli, 2022). These applications make the ML algorithms agents of high-stakes decisions in citizens' lives (Oneto & Chiappa, 2020). With these algorithms increasingly influencing important decisions, it is critical to evaluate that they are non-discriminatory and objective (Liang et al., 2022).

Initially, it is often assumed that algorithmic decision-making is inherently fairer and less biased than human decision-making as more data is included in evaluating cases and biases associated with human cognition are absent (Pessach & Shmueli, 2022). However, past research demonstrated that ML algorithms can be prone to biased decisions, which can lead to discrimination based on sensitive attributes of an individual, such as gender, ethnicity, and disability (Kearns et al., 2019). Amongst many examples, the societal relevance to unfair decisions facilitated by ML has been shown by the cooperation of Apple with Goldman Sachs for the Apple credit card, assigning lower creditworthiness to women compared to men with similar or even lower financial situations (Vigdor, 2019).

The societal consequences of implementing biased ML classifiers can be expressed in various domains. There are possible allocative harms, as the classifications decide who is granted opportunities and resources, with biases leading to the structural exclusion of protected groups (Barocas et al., 2017). Furthermore, representational harms can be perpetrated, as the discriminatory classifications could minimise protected identities over time (Barocas et al., 2017). ML classifiers can intensify their expressed biased tendencies through a feedback loop, as their decisions will likely affect future data collections and iterations of future training processes (Chouldechova & Roth, 2018).

Hence, when applying ML techniques for high-stakes decisions, it is essential to ensure non-biased and non-discriminatory classification (Liang et al., 2022). Researchers have contributed to developing fair ML theory and techniques to address those concerns.

Main debates in algorithmic fairness

The focus in fair ML research revolves around three themes. Firstly, it revolves around how fairness can be defined and measured in the context of ML, since fairness generally is more viewed as a social concept, rather than a mathematical one (Wan et al, 2023). Secondly, how

data can be seen as the main source of bias with human error and societal prejudice being introduced to it (Van Giffen et al., 2022). Lastly, how fairness interventions can be introduced into the pipeline to reduce bias (Caton & Haas, 2024). Those three topics are highly discussed in the fair ML research and are hence important to be understood when investigating algorithmic bias.

However, a commonality of those broad topics is that they are not concerned about the usual steps in an ML classifier's design process. The themes exclude the possibility of pipeline design choices being a source of bias, often treating the pipeline engineering process as value-free and mostly concerned with increasing the overall model accuracy.

Definitions of discrimination and fairness in machine learning

Legally, discrimination is differentiated between disparate impact and disparate treatment (Wan et al., 2023). Disparate treatment is defined as an explicit discrimination, meaning that the information of an individual belonging to an underrepresented group is deliberately used to treat them differently (Khodadadian et. al, 2021). Disparate impact, however, refers to implicit discrimination, where the information of an individual belonging to an underrepresented group is not used for decision-making, but they are still treated unfairly (Khodadadian et. al, 2021). This is mainly caused by proxy attributes, which are other features that hint at an individual's group membership but might not be easily removed since they contain information vital for accurate classification (Khodadadian et. al, 2021).

In ML, the population prone to discrimination is defined by sensitive attributes such as gender, ethnicity, and disability (Kearns et al., 2019). By this, models can be evaluated on their treatment towards binary subpopulations such as men/women, ethnic majority/minority, and non-disabled/disabled. However, only evaluating binary subgroups might not do justice to the complexity of fine-grained protected groups within society. Multiple sensitive attributes could be used to generate intersections of protected groups to, for example, investigate whether the classifier might not be biased towards white women but exhibits bias towards Black and disabled women (Kearns et al., 2019). However, a vast body of research on fair ML only considers binary populations, disregarding the societal structure of various subpopulations that might be advantaged or disadvantaged to different extents (Hutiri et al., 2023; Tubella et al., 2022; Cruz et al., 2020). Hence, considering the intersectional impact on fairness could give novel insight into the more differentiated impact of ML challenges on fairness.

In the context of fair ML algorithms, different fairness notions are broadly categorized into individual and group fairness. *Individual fairness* is the idea of individuals with similar feature values receive similar classifications regardless of their group membership defined by sensitive attributes (Morina et al., 2019). This primarily reflects the idea of disparate treatment by not using sensitive attributes and training an algorithm that does not differentiate based on these attributes, resulting in non-discriminatory results (Wan et al., 2023). A usual concern with individual fairness definitions is the disregard of proxy variables that the algorithm could use to identify patterns of group belongings and hence utilize to derive unfair classifications for individuals of protected groups (Khodadadian et al., 2021). For example, using education as a feature might be of disadvantage to a minority group in a society where the ethnic majority receives better education than the ethnic minority.

Group fairness suggests that protected and unprotected groups should have comparable outcomes, which are defined on different fairness metrics, with one of them being demographic parity, which implies having an equal probability of receiving a favourable outcome (Verma & Rubin, 2018). For example, when training a classifier to decide whether to grant a credit, the probability of being granted the credit should be similar between males and females. However, there are also many other suggested alternative measures to assess the fairness of a classifier and compare the treatment of different subpopulations. (Wan et al., 2023).

It is important to note that different definitions of fairness and measures represent different ideas of fairness. The choice of which measure to use is context-dependent and still open to the engineer's choice. Much research is being done to examine the development of fairness metrics, with more than 20 well-developed and discussed measures (Verma & Rubin, 2018). However, the fairness measures are not mutually exclusive; while a classifier might satisfy individual fairness, it might not satisfy group fairness (Wan et al., 2023). How fairness is supposed to be evaluated is still a topic of debate (Srivastava et al., 2019). In conclusion, fairness definitions in the context of ML are still debated and which definition to utilize in which context has yet to be unified.

Algorithmic bias caused by bias in the data

The second main focus of fair ML research concerns biased data, as ML models can be sensitive to bias in the data which can then potentially influence its performance and fairness. Hence, when training the model, it is vital to consider how representative the data is of the inspected population (Van Giffen et al., 2022).

One reason is the historical bias inherent to the training data, which stems from the data generation process and reflects the human bias towards protected groups. The algorithm will correctly mirror the pattern of the training data; however, the classification also carries the social bias towards the protected groups automatically and systematically (Chouldechova & Roth, 2018). In data collection and processing, systematic bias could be introduced, affecting the fairness of the predicted outcomes (Chadhari et al., 2022). Often data is approached as the sole source for biased outcomes and once the data is debiased this would inevitably lead to a fair algorithm, with the actual ML pipeline not asserting much influence on the resulting fairness (Rohani, 2021; Hutiri et al., 2023).

Particularly for unbalanced datasets, the bias can be exuberant, especially when the data contains attributes that hold discriminative information due to underlying factors (Davariaschtiyani et al., 2024; Chaudhari et al., 2022). Hence, this idea is based on the claim that social or historical biases are already present in the data before the utilization of an ML classifier and reflects the existing societal bias towards subgroups of the population (Van Giffen et al., 2022).

Fairness interventions in machine learning

Lastly, researchers on fairness interventions are developing pre-, in-, and post-processing methods, depending on where they are placed within the machine learning pipeline (Caton & Haas, 2024).

Pre-processing methods adjust the data before training the classifier to remove existing (historical) bias and extract information relevant for accurate classifications (Morina et al., 2019). Those methods build on the idea of biased data and aim to create a fairer dataset by addressing the distribution in relation to sensitive attributes (Badran et al., 2022).

In-processing methods modify the classifier's training process, training it for accuracy and fairness simultaneously, which incorporates the fairness assumption directly into the classifier's training (Morina et al., 2019). One method is adversarial debiasing, which aims to maximize the classifier's accuracy while limiting its ability to correctly guess protected attributes (Van Giffen et al., 2022). These techniques aim to train intrinsically fair models by making it an objective of the training process (Wan et al., 2023).

Post-processing methods utilize the outcomes generated from a biased classifier and adjust them towards the fairness ideal by assigning adverse outcomes to the unprotected groups and positive

outcomes to the protected groups when cases are located close to the decision boundary (Chakraborty et al., 2019). Hence, generated outcomes are transformed to decrease the classifier's discriminatory nature (Morina et al., 2019). These methods are popular as they improve fairness without interfering with the usual training process, which is especially useful when applying black-box algorithms (Nguyen et al., 2021).

Many researchers in fair ML have put a strong effort into developing various fairness interventions to address biases in ML predictions. However, while fairness intentions are valuable, they are introduced after other technical choices have been made, leaving open the possibility that other choices in the ML pipeline could have already influenced the fairness of the outcome.

Unseen influences of engineering choices

While fairness definitions and interventions have recently advanced, little attention has been paid to how ML engineering choices could influence a model's fairness. The introduced main aspects of fair ML research cover a broad area of ways to identify and mitigate bias. However, all three categories leave out the potential influence of typical engineering choices on the fairness outcome. However, this gives blindness towards the idea that an ML pipeline's actual design process can introduce bias or offer opportunities to mitigate bias. These decisions are usually treated as value-free and are not further regarded in their impact on the fairness of the classifier. This raises the need to identify further and investigate a classifier's standard components and evaluate those towards fairness outcomes.

For example, the chosen data imputation is often overlooked in terms of fairness. However, it could have an impact on algorithmic fairness. Patterns in missing data often indicate specific subpopulations, and certain imputation methods could potentially construct errors across those affected groups (Barret et al., 2022). Even the choice of fairness intervention itself could have ethical implications, as Tubella et al. (2022) argues that already the decision between in- and post-processing methods shows different impact on individual treatment. Furthermore, model hyperparameter tuning shows an impact on the fairness-accuracy trade-off in the model predictions (Cruz et al., 2020) showing that the connection between fairness and predictive performance is not parallel (Klaassen et al, 2024). This suggest that seemingly technical choices need to be considered on their fairness impact rather than just assuming that they are value-free.

In conclusion, there are many different decision points within an ML pipeline that could potentially influence the fairness of the resulting classifier. Hence, looking beyond the effect of

applied fairness interventions and metrices might be necessary to pin down detailed sources of biases within the pipeline. It is vital to explore whether, for example, different missing data imputation methods, could affect the outcome of individuals of different subpopulations.

Investigating the impact of engineering choices on fairness

This research explores the often-overlooked impact of engineering choices in ML pipelines on fairness across intersectional subpopulations, distinguishing itself from existing studies that primarily focus on fairness interventions and metrics. Traditionally, when optimizing a model purely based on accuracy, it might fail to do justice to fair treatment, especially for critical choices such as missing data imputation and under-sampling rare events, which could induce bias that is amplified for specific subpopulations. This raises the research question of this study: Are the engineering choices involved in a typical ML pipeline (missing data imputation, scarce outcome intervention, algorithm choice, hyperparameter tuning, threshold setting) truly value-free? This will be investigated with an informed exploration to show which choices might have the most considerable impact. It is hypothesized that engineering choices typically involved in an ML pipeline make a difference in the diagnosis of fairness problems across various subpopulations.

The focus is placed on high-stakes binary decisions, where biases in ML could lead to systematic disadvantage for protected groups. Specifically, the data will be simulated in the setting of social benefit fraud detection to exemplify the direct application of the methods.

In more detail, by incorporating intersectional analysis, investigating the effect of scarce outcomes, and testing both linear and non-linear decision boundaries, the research aims to provide a comprehensive understanding of how various decision points in the ML pipeline can affect fairness. The research is aiming to contribute towards the expanding conversation on algorithmic fairness, ultimately offering new insight into how commonly used engineering practices can exacerbate or mitigate bias. Broadly, it challenges the assumption that standard ML engineering choices such as imputation and model choice are value-free, aiming to investigate further those often-overlooked components towards their impact on model fairness. Overall, this study aims to investigate how different engineering choices could be a potential source of bias but could also provide an opportunity to mitigate these biases when appropriately handled. Conclusively, findings in this domain could inform data engineering practice and fairness guidelines, contributing to future ML systems that are less biased across different societal groups.

Chapter 2 – Theory

ML classification and prediction models are increasingly applied across various fields, offering an efficient tool for automated decision-making (Meda & Bhogapathi, 2022; Shukla, 2020). ML allows for processing complex data and identifying patterns to make predictions for unseen cases. However, ML pipelines often increase in complexity, which often comes at the cost of transparency (Heidemann et al., 2024). However, overall lack of transparency in the pipeline choices and their effect on the outcome could lead to concerns about fairness and biases. This often limits the possibility to explain a decision to an affected individual. Hence, this can be especially problematic in high-stakes scenarios, such as decision-making in the public sector, where accountability is crucial (Veale, 2017). When designing an ML classifier, there is a large inventory of possible choices that could express different changes in the model's fairness outcome. The ML pipeline intends to give a sequential framework to those choices to help organize decision points in the development process. This can help streamline the decision-making process and enhance the efficiency of the ML project (Katam et al., 2024).

The pipeline process organizes challenges an engineer faces in the development process, sequentially integrating components such as data pre-processing, model selection and hyperparameter tuning, allowing a holistic overview of the development process (Gomaa et al., 2022). The pipeline is mostly used as a tool to increase the performance of the developed model (Forescu et al., 2020). However, the decision points in this research are not primarily evaluated on enhancing the model performance, but on how they impact the resulting fairness. The overall aim is to consider each decision point and evaluate how much potential it holds to influence the resulting fairness of the model.

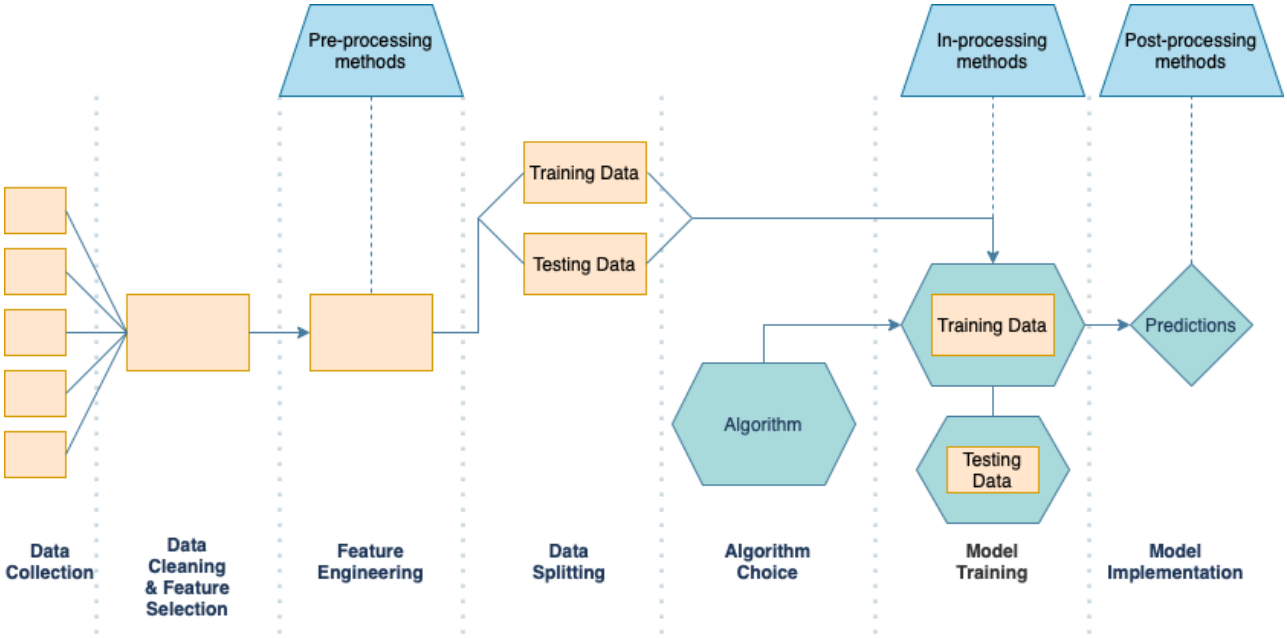
A Brief theory of machine learning classifier pipeline

Machine learning models build upon a pipeline that follows a general structure for any classification problem. Each pipeline follows a series of steps from data collection to model deployment, including various operational choices to ensure optimal fit to the data and are often optimized in isolation (Kunft et al., 2019).

The numerous options available could impact the predictive performance and fairness of the final model. Figure 1 shows the general flow of a classifier pipeline following steps from data collection to model implementation, with optional modules for fairness intervention methods. It demonstrated inevitable choices that arise when developing a classifier pipeline, with each decision point offering various options, of which the most suitable needs to be identified.

Figure 1.

General structure of a ML classifier pipeline



Firstly, data pre-processing and feature engineering, indicated in the first three columns of Figure 1, have the potential to enhance but also to diminish the quality of the data, which will affect the classifiers' ability to identify the underlying patterns correctly (Bilal et al., 2022). Hence, it is vital to carefully pre-process the data to enable the model to perform well (Bilal et al., 2022). However, for this current research, the simulated data that will enter the pipeline is assumed to be perfectly representative of the population. This representativeness is vital to avoid common discussions around data debiasing in ML and place the focus solely on engineering choices.

Additionally, well-considered feature selection can help to decrease the model's complexity to the most relevant variables while maintaining high accuracy and reducing computational efforts (Chemmakha et al., 2022). Furthermore, it is advised to consider fairness-aware feature selection to avoid using proxy variables that might cause bias towards subpopulations (Khodadadian et al., 2021).

Another vital step for pre-processing is the imputation of missing values. The engineer must be aware that implementing missing data does not contribute to any further bias within the data (Pessach & Shmueli et al., 2022; Jeong et al., 2022). The choice of imputation method can also affect the overall model's accuracy (Jadhav et al., 2019). Furthermore, in this step of the pipeline, there is the opportunity to incorporate pre-processing fairness interventions, which aim at debiasing the existing data from potential social bias (Morine et al., 2019).

Secondly, the data is partitioned into training and validation sets, a common practice to identify the best-working model options later, as indicated by the fourth column in Figure 1. However, this decision can affect the overall model accuracy with different proportions between the splits potentially influencing the model's performance and ability to estimate the model's performance (Birba, 2020). Furthermore, the split should consider the data's underlying distribution, and both splits should be equally representative of the underlying subpopulation, which is especially important when they are not evenly distributed (Salazar et al., 2022). For example, if a disadvantaged subpopulation is underrepresented in the test set, this could potentially inflate fairness metrics, and disparate treatment might go undetected (Salazar et al., 2022).

The next step is algorithm choice: an ML algorithm must be chosen for the data problem, as well as its complexity, to fit the data well enough but still generalize the outcome to unseen test cases and not be too computationally expensive, as indicated by the fifth column in Figure 1. With different ML algorithms, there is a chance to introduce algorithmic bias, which can lead to systematic discrimination of protected groups (Van Giffen et al., 2022). In general, algorithms replicate patterns detected within most data points in the training data and derive classifications based on this majority outcome. However, this can be a source of discrimination for minority subpopulations within the data (Barsotti & Koçer, 2022; Pessach & Shmueli, 2022). Since the main defined algorithmic goal is to minimize overall prediction errors, the majority groups will benefit over minority groups within their classifications as the algorithm cannot optimally fit multiple groups simultaneously (Pessach & Shmueli, 2022).

Furthermore, choices such as the tuning of hyperparameters could affect the fairness of the classifications derived from the algorithm (Tizpaz-Niari et al., 2022). This decision point is indicated in the sixth column in Figure 1. Overall, hyperparameter tuning can show the potential to amplify biases in disadvantaged subpopulations. However, tuning can also lead to fairer treatment while maintaining high accuracy (Kumar et al., 2022). Regardless, fairness-aware hyperparameter tuning remains an often-overlooked approach when developing an ML classifier (Tizpaz-Niari et al., 2022).

Lastly, when using the trained model to derive predictions for unseen cases, the engineer needs to set a decision boundary of what cases are accepted and which ones are rejected. The threshold setting can play a crucial role in the model's performance and fairness outcome (Yaseliani et al., 2024) and is indicated by the last column in Figure 1. By this, the threshold setting can show a trade-off between false-negative, false-positive predictions and accuracy (Birchha & Nigam,

2023). While fairness can be improved by adjusted threshold setting, this could be a trade-off towards the overall model performance (Radovanovic et al., 2021). In this process, post-processing methods can be applied to the cases close to the set decision boundary (Chakraborty et al., 2019). The threshold setting is an important decision, potentially negotiating between model performance and fairness.

In conclusion, Figure 1 gives a brief outline on the common decisions to make when designing an ML classifier, with each step offering many different methods to employ that potentially can influence the overall model's performance and fairness. Which choices are the most favorable depends on the data characteristics and the overall aim of the classifier; however, there are also generally more frequently applied choices in the industry that are used more often than others.

Common data characteristics

With the usual focus on historical bias in data as a topic of fair ML research, other challenges that a dataset can pose to the fairness implications might often remain disregarded. However, even with a perfectly representative dataset, there are still possible challenges for the ML pipeline in dealing with specific characteristics that might be present. These are reoccurring data characteristics that pose challenges to the development of a reliable ML classifier. Missing data, non-linearity and unbalanced outcomes could impact a classifier's performance and fairness and impose unique challenges on a pipeline. In realistic and applied settings, prevalent issues arise from the data and need to be addressed.

Firstly, missing data could be challenging depending on underlying patterns of missingness, as different subpopulations may be differently affected by missing values. Not carefully considered imputation methods could amplify biases towards subpopulation and impact the model's overall performance (Calmon et al., 2023). Missing data imputation gives another potential source of bias as missing data is often not missing completely at random, especially when it relates to humans. For instance, lower-income individuals might avoid answering income questions; minorities could lack responses because of language barriers, and disabled individuals might find forms not accessible enough to provide answers (Jeong et al., 2022). Generally, there are three types of missing data, beginning with *missing entirely at random*, implying that the missing value is external to other potential factors (García-Laencina et al., 2010). Secondly, *missing at random* implies that the missingness is traceable to other observed values (García-Laencina et al., 2010). Lastly, data can be *missing not at randomly*, where the missingness relates to unobserved factors (García-Laencina et al., 2010). The pattern of

missingness could imply important information that could be missed by choosing an unsuitable imputation method and hence significantly affect the fairness and accuracy of the resulting classifier, especially when the data is missing not at random (Mansoor et al., 2022).

Furthermore, data can portray different complexity with linear or non-linear separability of the features to the target variable. A linear decision boundary separates the binary classification by a hyperplane, while non-linear decision boundaries take the shape of a curved and irregular surface (Vaidyanathan et al., 2008). Different models and model choices might handle the increasing data complexity to different extents. Linear data can be effectively addressed by choices that assume linearity in the data and capture relationships sufficiently and more simply. However, non-linear approaches are more suitable for data with complex relationships. Hence, model decisions must be made considering the underlying data characteristics, as this can help increase both performance and fairness (Haghighat et al., 2024; Sun et al., 2024). Additionally, some algorithms might be more prone to give biased predictions than others, impacting the fairness of the resulting classifier differently. As Luengo et al. (2023) identified, models trained on the same data produce varyingly biased predictions for different subpopulations.

Lastly, the scarce distribution of the outcome variable can impact the model's fairness and overall performance (Badar et al., 2024). In the data circumstances examined in this study, the positive classification (1), which indicated committed social benefit fraud, overweighs the negative classification (0), which represents no committed fraud to different extends. This is a realistic condition in many settings, as some outcomes are rarer than others. For example, it can be assumed that fraud is only committed by a small proportion of the population. However, such a setting poses a challenge to the pipeline, as ML algorithms usually tend to minimize the mean loss on a single metric, which often neglects the scarce outcome, potentially impacting predictive performance and fairness (Yan et al., 2022). Various methods have been proposed to deal with data characteristics, including imbalanced outcomes.

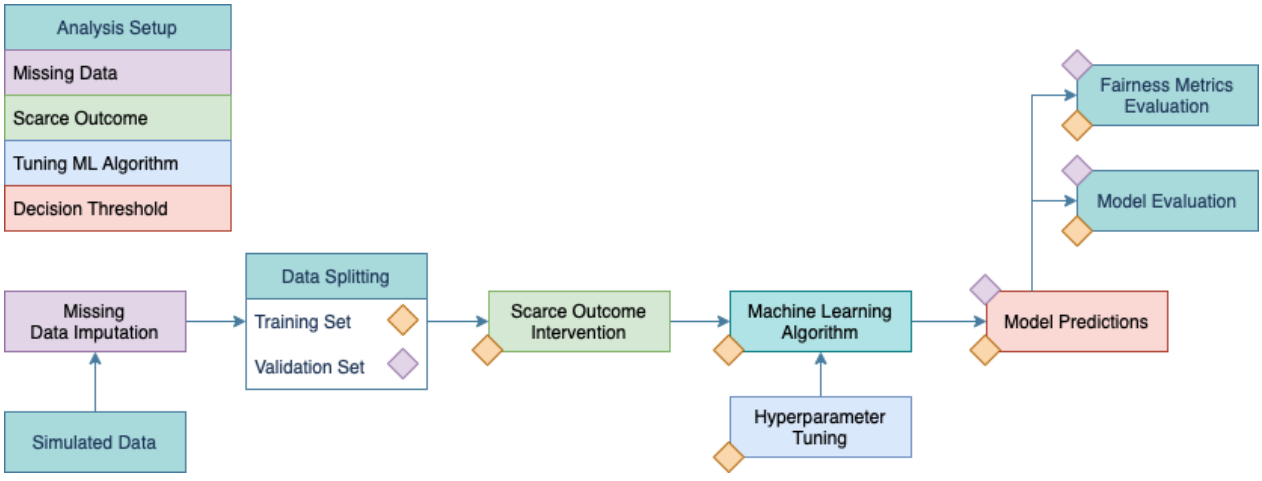
In conclusion, when faced with the decision points and their options, the practitioner needs to consider the data characteristics, including missing data, non-linearity, and distribution of the outcome variable. The chosen methods should adequately address challenges that arise from those characteristics and suit the context of the data.

Choices within the machine learning pipeline and their potential effect on fairness

The previously introduced decision points and their various options within give an explosively large map of possible combinations that could influence model performance and fairness in each point. This research aims to point out which decision points may show the largest potential in fairness outcomes and hence focuses on only a selected number of decision points and options within. With this strategy, the explored space is narrowed down for easier evaluation. The decision on what contingencies to explore was based on how commonly those methods find application in the industry for binary classification problems, to show how already frequently utilized approaches imprint the fairness of the resulting classifier (Chattopadhyay et al., 2021; Liao, 2023; Hong & Lynn, 2020, Viloria et al., 2020).

Figure 2 shows the assessed classifier pipeline with the explored five decision points, the colours indicated in the analysis setup block correspond to the colour of the decision point where this setup changes the decision point within the pipeline. Mainly, the study focuses on the effect of missing data imputation, unbalanced outcome intervention, choice of algorithm, hyperparameter tuning and the threshold setting for fairness of the model, which have been identified to be the main choices within a pipeline (Buczac & Guven, 2015; Van Giffen et al. 2022). It seeks to identify which of those decision points offer the potential to affect the resulting fairness outcomes. For this, in each decision point, two or three often applied methods are chosen to be further explored in terms of their different effects on how biased their predictions might be towards pre-defined subpopulations.

Figure 2.
ML classifier pipeline with explored decision points



Missing data imputation

The first investigated decision point is data imputation, as indicated on the left of Figure 2. When missing values are present in the data, a method to address these needs to be chosen. This choice of imputation could impact both model performance and fairness. It has been shown that different imputation methods could influence performance and fairness to different extents (Nezami et al., 2024). Nezami et al. (2024) found that some imputation methods introduce more bias, especially when the dataset already exhibits societal biases. Furthermore, imputation methods can improve a model's accuracy when chosen correctly (Poulus & Valle, 2018).

Two commonly used approaches are Random Forest-based and simple mean imputation (Hong & Lynn, 2020). Firstly, the Random Forest-based imputation, particularly MissForest, has already demonstrated to be better performing compared to simple mean imputation (Buczak et al., 2023). Its strength is mostly in how it can effectively handle complex data with interactions and non-linear relationships, making it a good choice for diverse data characteristics (Bühlmann & Stekhoven, 2011). MissForest imputation tends to outperform other imputation methods when applied to complex data (Stekhoven & Bühlmann, 2011). However, simple mean imputation is still a valid method, particularly for datasets with many features containing missing data (Buczak et al., 2023). Overall, this method is computationally more inexpensive than other methods, such as MissForest, which makes it less time-consuming, especially for larger datasets (Hong et al., 2020).

Whether these two imputation methods affect the fairness outcome differently should be explored in more detail. However, they are the first two evaluated choices on the contingency map of the ML pipeline.

Scarce outcome intervention

There is the common challenge of scarce datasets in machine learning, and various interventions have been developed to address this issue, with random over- and under-sampling being two frequently used approaches (Viloria et al., 2020). This makes this issue the second decision point to be further evaluated, indicated as a "scarce outcome intervention" in Figure 2. Scarce outcomes are characterized by a small proportion of positive classifications in the target variable (here, 1, "social benefit fraud committed") and a large proportion of negative classifications (here, 0, "no fraud committed").

Random over- and under-sampling methods have positively impacted model fairness and accuracy and are frequently applied methods (Viloria et al., 2020). Briefly, random over-

sampling resamples the minority outcome to the point where it is balanced with the majority outcome. In contrast, random under-sampling reduces the majority outcome to have the same frequency as the minority outcome. In application, random under-sampling has been demonstrated to improve the treatment of minority subpopulations without reducing the overall accuracy of the model (Naboureh et al., 2020). However, for more complex data random under-sampling faces the issue of potentially removing too much information from the majority class, especially around the decision boundary (Arslan et al., 2022). Hence, random under-sampling might struggle with more complex scenarios (Chennuru & Timmappareddy, 2017), especially when the data is noisy or includes a class overlap (Grina et al., 2021). The more beneficial method is context-dependent; however, in some areas, any resampling intervention shows improvement compared to no applied adjustment method; however, no method outperformed the others (Jacobucci & Li, 2022).

Some methods introduce synthetically generated instances of the minority classification, such as SMOTE (Synthetic Minority Oversampling Technique). SMOTE is shown to be very effective in increasing the overall predictive performance of the model (Bal & Kayaalp, 2023). However, in the context of high-stakes decision-making on human individuals it is essential to note that resampling the data brings great responsibility, and grounding classification on artificially generated data might be challenging to justify towards affected individuals because of rising concerns about data integrity and authenticity (Neves et al., 2022). This led to the decision to focus on the effect of using random over- and under-sampling compared to not applying any intervention to identify the impact on fairness.

Choice of machine learning algorithm

The third investigated decision point is indicated as “Machine Learning Algorithm” in Figure 2. For a classification problem, many different ML algorithms are available; however, for simplification, this study focuses on logistic regression, Random Forest and XGBoost (Extreme Gradient Boosting) and their impact on fairness. The algorithms were chosen based on their frequent utilization for classification problems (Chattopadhyay et al., 2021; Liao, 2023).

Firstly, logistic regression is a commonly used model. Grounded as a traditional statistical method, it offers a simple and easily interpretable approach to a classification problem (Li & Chen, 2020). It can be a robust model for linearly separable data; however, its performance could decrease for more complex and non-linear relationships (Li & Chen, 2020). Additionally,

there is some evidence of logistic regression potentially biasing minority groups within the data, giving less accurate predictions for those groups (Do et al., 2022).

Random Forest is an ensemble method combining multiple weak learners in the form of decision trees, increasing the overall predictive performance and susceptibility to more complex relationships within the data. More specifically, it is a bagging method, training multiple weak learners independently from each other. Depending on the data, Random Forest is one of the most robust methods to apply, with the capability of outperforming boosting methods (such as XGBoost, see below) in accuracy (Nagaraj & Ghosh, 2024). However, it is essential to note that the algorithm choice needs to be tailored to the currently assessed data, with Random Forest and XGBoost often outperforming other ML methods (Zhang et al., 2020).

XGBoost is a more advanced ensemble method that combines multiple weak learners, similar to Random Forest. However, this is a boosting method, meaning that the weak learners are built sequentially, with the subsequent one learning from the errors of the previous one. XGBoost often outperforms other ML algorithms in complex data situations but is also more prone to overfitting (Mahesh et al., 2023). However, both Random Forest and XGBoost show the potential to underfit minority classes within the data, potentially leading to systematic bias (Ugirumurere et al., 2024).

Each of the three algorithms has its justification within the industry, with XGBoost and Random Forest outperforming each other in accuracy for different contexts (Wu et al., 2024). Furthermore, due to its simplicity and explainability, logistic regression is still shown to be one of the best-performing baseline classifiers (Li & Chen, 2020). Hence, those three options are chosen for the evaluation decision point to further identify potential changes in fairness outcomes.

Tuning of the machine learning algorithm

Hyperparameter tuning is the third decision point within the evaluated pipeline, indicated in Figure 2. Hyperparameter tuning can impact a model's accuracy and fairness, with changes on single hyperparameters already showing potential to alter predictive performance and fairness (McCarthy & Narayanan, 2023).

Often the focus of hyperparameter tuning is to increase the model's accuracy which disregards its potential to affect the fairness of predicted classifications. By this, Cruz et al. (2020) established that with some small decrease in predictive accuracy, the fairness of the model can

be significantly increased, showing the importance of fairness-aware hyperparameter optimization and the consideration of the fairness-accuracy trade-off. Nevertheless, it is important to note that the hyperparameters are model-specific with different tuning aspects for different algorithms. Hence, the fairness implications of hyperparameter tuning might be different depending on the utilized ML algorithm.

This shows that it is recommendable to not only consider the traditional performance metrics but also incorporate fairness assessments into the process. This approach could help to increase the model performance while also avoiding possible negative effects for disadvantaged subpopulations (McCarthy & Narayanan, 2023).

Threshold setting for predictions

The threshold setting is the last engineering choice considered within the assessed pipeline, indicated in the "model prediction" in Figure 2. Threshold setting plays a crucial role in balancing accuracy and fairness. A threshold of 0.5 is commonly used by default; however, adjustments can be made context-dependently, impacting performance and fairness of the outcome. In some applications, the risks connected to false positive and false negative outcomes may require strong adjustments to the default value, by which threshold above 0.9 (avoiding false negative) or below 0.1 (avoiding false positive) could be a plausible choice.

The threshold setting could potentially impact different subpopulations leading to disparate impacts. It can help to balance towards a fairness definition, with lower thresholds decreasing the number of false positive classifications but also increasing false negative predictions and vice versa for higher thresholds. A survey has revealed that practitioners tend to be more concerned about avoiding false negatives than false positives (Cappelen et al., 2018). However, the threshold can be approached adaptively to the situation, depending on whether higher false negative rate (FNR) or false positive rate (FPR) could have an adverse impact.

Generally, in terms of fairness, an overall lower threshold has been shown to decrease the impact of discrimination for data with present historic bias (Aggarwal et al., 2019). However, in other domains, such as medical risk assessment, a lower threshold might be advisable as the damage for false negatives outweighs the damage of false positives (Kodama et al., 2022). In conclusion, the threshold should be based on the requirements of the model, the costs associated with different types of errors and the stakeholders' preferences (Bondugula et al., 2021). Regarding fairness consideration, the threshold might affect different subpopulations differently, which needs further investigation.

Conclusion

Considering all those decision points in an ML classifier pipeline with all those different options, this research argues that fairness in an ML model should not be considered as an afterthought but rather be approached critically in each step of designing the pipeline. By understanding how each decision point shows the potential to affect subpopulations differently and the ability to upsurge but also mitigate bias, the practitioner could better balance between predictive performance and justifiable outcomes.

Chapter 3 – Methodology

The overall purpose of the study is to scrutinize how engineering choices can show potential to enhance and mitigate systematic bias in a classifier and encourage fairness considerations during the pipeline design process. This is important since automated decision-making using ML has become ever-more prevalent for high-stake decisions on individuals (Pessach & Shmueli, 2022). This study aims to replace the focus on different engineering choices and avoid the usual attention of fairness research concerning definitions of fairness and data biases. Hence, this research starts by simulating data to inspect the impact of engineering choices. The simulated datasets are assumed to be fully representative of the population, not showing any indications of social bias or social prejudice elicited by the data collection process. Thus, it is assumed that the simulated data is fully representative of the current social reality, aligning with present inequalities and privileges connected to sensitive attributes of gender, ethnicity and disability (Betts & Null, 2011; Rana, 2024; Gee, 2018; Kim, 2022; Manzoor, 2023).

By this, the engineering choices are assessed in response to data that resamples the social reality and hence bias inherent to the data rather than data caused by poor data sampling. The data is simulated under different underlying assumptions about linearity and scarcity of the outcome. Since those data characteristics are frequent challenges an ML practitioner faces when designing a pipeline, those are chosen for the investigation. *Non-linear decision boundaries* are a frequent challenge that needs to be addressed by engineering choices that do not assume underlying linearity (Li & Chen, 2020). Furthermore, *scarce outcomes* are frequent for practitioners, where the target label is naturally less frequently present (Viloria et al., 2020).

The setting of the data simulation is concerned with social benefit fraud detection, considering the social reality of gender, ethnicity, and disability, further impacting the predictive variables and reflecting on an individual committing fraud. The binary-sensitive attributes are considered in their intersectionality, showing that different combinations of attributes lead to fine-grained subpopulations that might be more prone to lower social status. Different algorithm choices are tested with the generated and fully representative population to determine how they affect pre-defined fairness measures. Since the chosen fairness measure is still a choice that has yet to be unified by research, the choice of an indicator should not be taken quickly. There are over 20 defined fairness metrics, with no unified context-related suggestions on which one should be set (Verma & Rubin, 2018). Since this issue is already a well-discussed topic in fair ML, this project aims to deliberately avoid the metric choice debate (Verma & Rubin, 2018). Hence, for

this research, two fairness measures are fixed (false positive rate and false negative rate) to evaluate the impact of engineering choices, and the argument of which fairness metric might be the best for the researched setting is simplified. The two fairness measures are often used as critical predictors to indicate the prediction of undesired outcomes, both for the contracting authority as well as the individual that is evaluated by the ML pipeline (Wang et al., 2023)

Overall, five engineering decision points will be investigated: missing data imputation, scarce outcome intervention, choice of algorithm, hyperparameter tuning and threshold setting. Each of those decision points is evaluated based on their impact on the resulting fairness of the model to show which point offers the most potential to exuberate or mitigate the systematic bias of an ML classifier. This will offer insight into how considerations during the pipeline engineering process can already contribute to fair treatment of the evaluated individuals. In the following part of the chapter, the data simulation process will be explained, and then the data analysis process will be detailed in a fictional and simplified data situation.

Simulating the population

The simulation is designed to generate three protected attributes and seven features, of which three are grounded in real-world resemblance (income, education, house ownership), and four are a linear or non-linear combination of the prior. The data simulation follows a repetitive flow detailed in Figure 3, where specific steps are adjusted to accommodate scarce outcomes and non-linear characteristics while maintaining other assumptions about the population for each data set.

For the analysis, four different populations are simulated in R (code detailed in Appendix A), corresponding to the high-stakes environment of social security benefit fraud. This will showcase how decision points vary in their effect on fairness depending on the underlying data characteristics typically found within the industry. Table 1 shows the characteristics of each of the four datasets, assembled of either a linear or non-linear decision boundary for the outcome and a balanced or unbalanced distribution. This will later allow to extract the effect of non-linearity and unbalanced outcomes on the fairness of the resulting model. Overall, non-linearity and scarcity of outcome are two main challenges that engineers often face and, hence, should be further considered in terms of fairness implications when evaluating engineering choices (Li & Chen, 2020; Viloria et al., 2020).

Table 1.

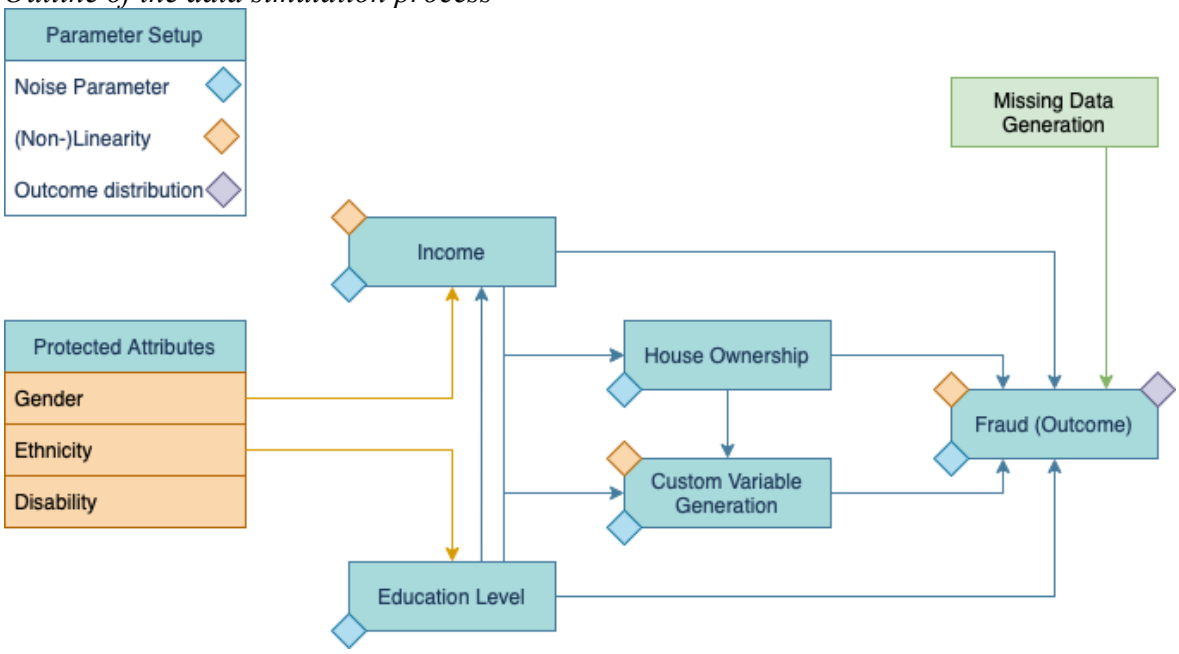
Overview of the characteristics of the four tested datasets

	Decision Boundary	
Distribution of outcome	Linear, Non-scarce outcome	Non-linear, Non-scarce outcome
	Linear, Scarce outcome	Non-linear, Scarce outcome

The four datasets are simulated in a data factory, which is a common simulations structure across all data situations and allows for adjustments in the shape of the decision boundary and the distribution of the target variable. The other factory settings remain constant throughout the simulation to ensure that the datasets only differentiate on the targeted data characteristics. The fixed factory settings are adjusted to ensure that the data models a realistic society, which anchors the tested data to reality. Hence, the chosen variables resemble real-world circumstances and relate to each other by realistically chosen probabilities, constants and coefficients (Betts & Null, 2011; Rana, 2024; Gee, 2018; Kim, 2022; Manzoor, 2023). Figure 3 outlines the connections between the sensitive attributes, features and outcome variables. The parameter setup in Figure 3 shows the adjusted parameters in the data simulations and are shown correspondently in the variables that they are influencing. The yellow arrows show which variables are directly influenced by the protected attributes, indicating that the influence of the protected attributes is indirect towards the fraud outcome.

Figure 3.

Outline of the data simulation process



Protected Attributes

The simulation starts with generating the three protected binary attributes that mimic the distribution found in a real society. The variables are distributed with fixed probabilities as follows:

- *Gender*: Let $G \in \{0,1\}$, where $G = 0$ represents male and $G = 1$ represents female. The probability of being female is $P(G = 1) = 0.5$.
- *Ethnicity*: Let $E \in \{0,1\}$, where $E = 0$ represents the majority and E = 1 represents the minority. $P(E = 1) = 0.35$.
- *Disability*: Let $D \in \{0,1\}$, where $D = 0$ represents no disability and $D = 1$ represents disability. $P(D = 1) = 0.07$.

The combination of protected attributes leads to intersectionality that creates a total of eight different subpopulations that are differently disadvantaged depending on how many underprivileged attributes they are made of. Generally, a subpopulation can take on different levels of disadvantage, with an additive effect of sensitive attributes, leading to the most advantaged group (male X majority X non-disabled) and the most disadvantaged group (female X minority X disabled). Table 2 details the cross-sectional nature of the sensitive attributes and their different levels of vulnerability, showing the most advantaged group in green and the most disadvantaged group in red.

Table 2.

Cross-sectional subpopulations and their vulnerability levels

Gender	Ethnicity	Disability	Level of Vulnerability
Male	Majority	Non-disabled	0
Male	Majority	Disabled	1
Male	Minority	Non-disabled	1
Male	Minority	Disabled	2
Female	Majority	Non-disabled	1
Female	Majority	Disabled	2
Female	Minority	Non-disabled	2
Female	Minority	Disabled	3

Education Variable

To express an individual's level of education, a categorical variable with six levels is introduced, which is denoted as shown in Table 3. The level of education is expressed numerically, assuming an ordering of education levels and also simplifying but still realistically reflecting the magnitude between education levels.

Table 3.
Numeric expression of the six education levels

Level of Education	Numeric Expression
Lower than High School	0
High School	1
Some College	2
Bachelor's	3
Master's	4
Doctorate	5

The probability for an individual to reach a particular education level k (where $k \in \{0,1,2,3,4,5\}$) is influenced by their gender G, ethnicity E, and disability D, expressed as:

$$P(Edu = k|G, E, D) \quad (1)$$

The educational level generation follows some societal assumptions to introduce variance that could match a real society. Those assumptions mainly follow:

- Women tend to achieve higher educational levels than men when controlling for ethnicity and disability (Betts & Null, 2011), expressed as:

  $$P(Edu = (3,4,5)|G = 1, E = 0, D = 0) > P(Edu = (3,4,5)|G = 0, E = 0, D = 0) \quad (2)$$

- Furthermore, the majority group has higher chances of obtaining higher levels of education compared to the minority group (Rana, 2024), denoted as:

  $$P(Edu = (3,4,5)|G = 0, E = 0, D = 0) > P(Edu = (3,4,5)|G = 0, E = 1, D = 0) \quad (3)$$

- Hence, women of minority ethnicity tend to receive lower education than men of majority ethnicity, showing that minority females face higher barriers to education (Rana, 2024), denoted as:

  $$P(Edu = (3,4,5)|G = 0, E = 0, D = 0) > P(Edu = (3,4,5)|G = 1, E = 1, D = 0) \quad (4)$$

- Individuals with disability face the highest barrier to education, with a higher probability of not finishing high school (Gee, 2018), with:

  $$P(Edu = 0|D = 1) = 0.4 \quad (5)$$

- Generally, the effect of sensitive attributes is additive; the more disadvantaged groups an individual belongs to, the lower their education tends to be (Kim, 2022; Manzoor, 2023).

In summary, the simulated society assumes that females tend to achieve higher education. However, other protected attributes lead to less access to education, resulting in lower education levels for individuals with minority ethnicity and disabilities.

Income Variable

To generate the income variable *I*, a combination of the base salary S_0 and the effects of gender *G,* ethnicity *E,* disability *D,* and education *Edu* are applied while also including possible non-linear interactions. Overall, the income variable can be expressed as:

$$I = S_0 + f_G(G) + f_E(E, Edu) + f_D(D) + f_{Edu}(Edu) + \epsilon \tag{6}$$

Where:

- S_0 is the base salary
- $f_G(G)$ represents the gender effect
- $f_E(E,Edu)$ represents the ethnicity effect
- $f_D(D)$ represents the disability effect
- $f_{Edu}(Edu)$ represents the education effect
- ε is a random noise term to introduce variability

The simulation considers a gender pay gap, with females showing the tendency to earn less than males which resembles the income structure of a realistic society (Betts & Null, 2011). The gender effect $f_G(G)$ shows that males (G = 0) receive a bonus of 7000 on their income, while females (G = 1) get a reduction of 3000, expressed as:

$$f_G(G) = \begin{cases} 7000, & G = 0 \\ -3000, & G = 1 \end{cases} \tag{7}$$

Individuals belonging to the minority group (E = 1) receive a deduction of 5000 unless they have reached higher education levels. This implies that the negative impact of belonging to the minority group can be avoided by higher education (Letterman et al., 2018), denoted as:

$$f_E(E, Edu) = \begin{cases} -5000, & Edu \leq 3 \\ 0, & Edu > 3 \end{cases} \tag{8}$$

The disability effect shows that individuals with a disability (D = 1) have a generally lower income (Cervini-Plá et al., 2016), expressed as:

$$f_D(D) = \begin{cases} -7000, & D = 1 \\ 0, & D = 0 \end{cases} \tag{9}$$

The income variable is generated from a combination of a base salary and the effects due to the protected variables and education, with income being affected by education in the sense that higher education leads to higher income (Wang et al., 2022), as followingly defined:

$$f(x) = \begin{cases} 0, if\ Edu = 0\ (No\ High\ School) \\ 5000, if\ Edu = 1\ (High\ School) \\ 10000, if\ Edu = 2\ (Some\ College) \\ 20000, if\ Edu = 3\ (Bachelor's) \\ 35000, if\ Edu = 4\ (Master's) \\ 50000, if\ Edu = 5\ (Doctorate) \end{cases} \tag{10}$$

When introducing non-linearity to the income, interaction terms between protected attributes and education are introduced as:

$$f_{non-linear}(G, E, Edu) = \beta_1 G \cdot Edu + \beta_2 E \cdot Edu^2 \tag{11}$$

Hence, the income *I* is composed by the income function with optional non-linear terms:

$$I = S_0 + f_G(G) + f_E(E, Edu) + f_D(D) + f_{Edu}(Edu) + f_{non-linear}(G, E, Edu) + \varepsilon \tag{12}$$

By this, the income variable captures linear and non-linear contributions to gender, ethnicity, disability and education, showcasing potential economic disadvantages within a society.

House Ownership Variable

The house ownership variable is H ∈ {0,1}, where H = 1 represents individuals owning a house and H = 0 means they do not. Whether an individual owns a house is determined by the base probability, which is dependent on education and income. Base probabilities increase with higher levels of education (Wang et al., 2022; Yao, 2017), following this mapping:

$$P(H = 1|Edu) = \begin{cases} 0.1, if\ Edu = 0\ (No\ High\ School \\ 0.2, if\ Edu = 1\ (High\ School) \\ 0.3, if\ Edu = 2\ (Some\ College) \\ 0.5, if\ Edu = 3\ (Bachelor's) \\ 0.6, if\ Edu = 4\ (Master's) \\ 0.7, if\ Edu = 5\ (Doctorate) \end{cases} \tag{13}$$

With complementary probabilities not given but implied, such as:

$$P(H = 0|Edu = 0) = 0.9 \tag{14}$$

The income effect adjusts the baseline probability to the individual's income by using the income values *I* as standardized z-scores and scaling the values by 0.1, as denoted by:

$$f_I(I) = \frac{I - \mu_I}{\sigma_I} \cdot 0.1 \tag{15}$$

Conclusively, the adjusted house ownership probability is:

$$P_{adjusted}(H = 1|Edu, I) = P_{base}(Edu) + f_I(I) + \varepsilon_H \tag{16}$$

This shows that with higher income, the likelihood of owning a house increase, and vice versa. To add realistic uncertainty, random error is introduced to the adjusted probability.

To determine house ownership H, a threshold is determined for each individual by drawing from a uniform random variable $U \sim \mathcal{U}(0,1)$ which reflects other unconsidered factors such as increased or decreased difficulty of obtaining a house depending on the area where an individual decides to reside. If the threshold is less than the individual's house ownership probability, the individual will be classified as a house owner (H = 1), as in:

$$H = \begin{cases} 1, & \text{if } U < P(H = 1|Edu, I, \varepsilon_H) \\ 0, & \text{otherwise} \end{cases} \quad (17)$$

Additional variables

Furthermore, two binary and interval variables are generated based on income, education and house ownership, with each variable influencing the values to different extents in linear or non-linear ways:

- $V_1 = Binary(0.5 \cdot I + 1.5 \cdot Edu + \varepsilon, \text{cut-off at mean})$ (18)
- $V_2 = \log(|3.6 \cdot I + 0.7 \cdot Edu + 2.3 \cdot V_1| + 1) + \varepsilon$ (19)
- $V_3 = Binary(\frac{1}{1+\exp(-(0.12 \cdot H + 0.7 \cdot V_1 + 0.03 \cdot V_2 + \varepsilon))}, \text{cut-off at median})$ (20)
- $V_4 = -6.2 \cdot H - 0.0004 \cdot I - 13 \cdot V_3 + \varepsilon$ (21)

Fraud Classification

The seven generated features are utilised to classify whether an individual committed social benefit fraud *F*. A linear combination of the variables obtains the baseline; if the generated dataset is supposed to follow a non-linear decision boundary, this effect is added later. Firstly, the linear combination *LC* is calculated as follows:

$$LC = -0.2 \cdot Edu - 0.00003 \cdot I - 1.7 \cdot H - 0.7 \cdot V_1 + 0.0007 \cdot V_2 + 0.065 \cdot V_3 + 0.0009 \cdot V_4 \quad (22)$$

If the setup requires a non-linear decision boundary, then non-linear components *NL* are added to the linear combination:

- *Income squared*: $NL_{I^2} = -0.05 \cdot I^2$ (23)
- *Interaction terms*: $NL_{V_1 \times V_2} = 0.02 \cdot (V_1 \cdot V_2)$ and $NL_{Edu \times I} = -0.03 \cdot (I \cdot Edu)$ (24)
- *Upper-middle-class criminality assumption:* If the income is between the 60th and 75th percentiles, there is an increase probability of fraud:
  $C_{increase} = 1$ if $Q_{0.6} < I < Q_{0.75}, 0$ otherwise (25)
- *Lower-middle-class criminality assumption:* If the income is between the 35th and 50th percentiles, there is a decrease in the probability of fraud:
  $C_{decrease} = -1$ if $Q_{0.35} < Q_{0.5}, 0$ otherwise (26)

All non-linear components get combined and added to the linear combination, giving a combined effect:

$$NL = NL_{I^2} + NL_{V_1 \times V_2} + NL_{Edu \times I} + C_{increase} + C_{decrease} \tag{27}$$

$$Combined\ Effect = LC + NL \tag{28}$$

To convert the combined effect into probabilities to commit fraud, a sigmoid function is applied to the values and a random noise parameter is added to model uncertainty:

$$P(F = 1|Edu, I, H, V_1, \dots, V_4) = \frac{1}{1 + \exp\big(-(Combined\ Effect)\big)} + \varepsilon \tag{29}$$

Finally, the binary fraud classification is derived by using a threshold T_F, which adjusts based on the desired proportion of fraud in the dataset, with the non-scarce condition taking on 25% of positive classifications and the scarce condition taking an overall 4%:

$$F = \begin{cases} 1, & \text{if } P(F = 1|\cdot) > T_F \\ 0, & \text{otherwise} \end{cases} \tag{30}$$

For the non-linear setup, this threshold selects 70% of the fraud cases from above the threshold and 30% from the 20% below the threshold to give a more flexible fraud classification, strengthening the concept of individual variability.

To introduce further noise, a proportion of cases is randomly selected, and their labels are flipped. This aims to simulate real-world uncertainty by incorrectly assigning some observations' outcomes.

Missing Value Generation

Lastly, missing data is induced to the income, education and house ownership variables, with different underlying relationships:

- 2% of income and education variables are missing completely at random.
- 6% of income missing, with a higher likelihood to be missing for minority and disabled group members, missing not at random.
- 75 missing cases on one of the generated variables, with a higher probability of missing for low-income and low education, missing at random.
- 85 missing cases on home ownership, with a higher probability of missing for females and minority ethnicity, missing not at random.

This means that there are missing completely at random and missing not at random values in the data, representative of real-world mechanisms involved in missing data. 2.7% of the data is missing, with approximately 16% of the cases having at least one missing value.

Data analysis

For the data analysis, the used fairness metrics are first fixed to later evaluate the fairness implication at each decision point. Following, the 30 different combinations of engineering decisions are tested on the four different data situations in terms of accuracy and fairness of the outcome. Those models are then used to isolate the effect of each decision point for each data situation, which should provide an accessible summary of where in the pipeline most of the changes in fairness outcomes are initiated.

Evaluation of fairness implications

All the selected choices at the five different decision points within the ML pipeline will be evaluated for their impact on fairness toward multiple subpopulations to identify potential sources of bias. To do this, a definition of fairness must be chosen. However, as previously mentioned, there are many fairness metrics to choose from, and they are not mutually exclusive (Verma & Rubin, 2018). Nevertheless, a choice must be made to assess the fairness impact of decision points. Therefore, this context chooses false positive rate (FPR) and false negative rate (FNR) as fixed fairness measures.

FNR and FPR are two basic fairness definitions that often underpin more complex measures, making them relevant indicators for biased predictions (Verma & Rubin, 2018). In the context of social benefit fraud detection, which is the focus of this research, FPR and FNR allow for the assessment of binary predictions related to undesired outcomes. Essentially, they represent both stakeholders of these classifiers. A low FNR ensures that the party seeking to identify fraudulent activity does not miss too many cases. In contrast, a low FPR reduces the risk of citizens falsely being accused of fraudulent activities by this classifier. In conclusion, FNR and FPR provide a good measure of fairness, considering the perspectives of both affected stakeholders.

In detail, FPR and FNR are derived by assuming that:

- $Y \in \{0,1\}$, true label ($0 =$ negative outcome, $1 =$ positive outcome)
- $\hat{Y} \in \{1,0\}$, predicted label ($0 =$ negative outcome, $1 =$ positive outcome)
- $A \in \{a_1, a_2, \dots, a_k\}$, sensitive attribute combinations (gender, ethnicity, disability)

The FPR indicates the probability of a model predicting a positive outcome ($\hat{Y} = 1$) when its true label is negative ($Y = 0$). Mathematically, fairness according to FPR is fulfilled when:

$$P(\hat{Y} = 1 | Y = 0, A = a_1) = P(\hat{Y} = 1 | Y = 0, A = a_2) = \cdots = P(\hat{Y} = 1 | Y = 0, A = a_k) \quad (31)$$

The FNR is the probability that the model assigns a negative outcome ($\hat{Y} = 0$), with the true label being positive ($Y = 1$). Fairness fulfilled through the FNR is denoted in the following way:

$$P(\hat{Y} = 0|Y = 1, A = a_1) = P(\hat{Y} = 0|Y = 1, A = a_2) = \cdots = P(\hat{Y} = 0|Y = 1, A = a_k) \quad (32)$$

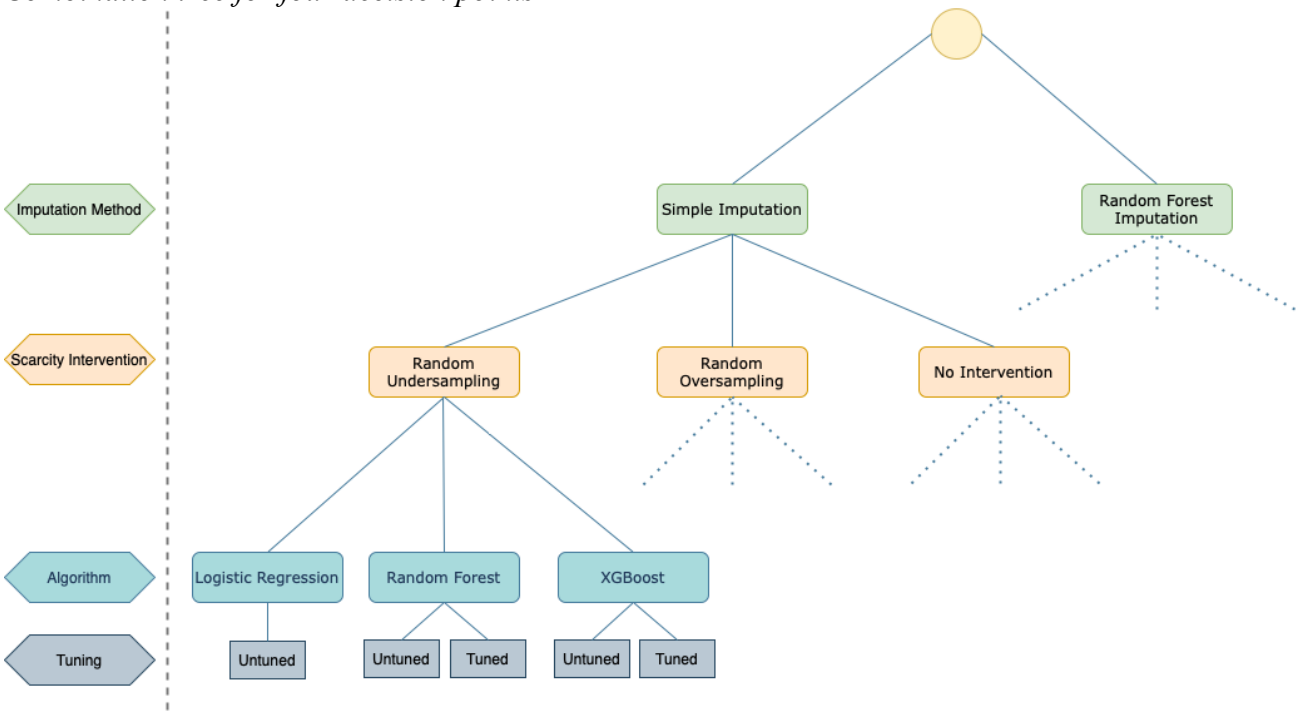
Fairness across subpopulations is often assumed using the 80% rule, stating that no subpopulation should have less than 80% of the usually achieved metric (Raghavan & Kim, 2024). Which follows as:

$$\frac{P(\hat{Y} = 1|A = a_i)}{P(\hat{Y} = 1|A = a_j)} \geq 0.8, \quad \text{for all } i, j \quad (33)$$

Combination of engineering choices

Overall, each decision point can contribute to a different semblance of the overall ML pipeline. The decision points of missing data imputation, scarcity intervention, algorithm choice, and tuning are all considered in their interplay with each other while only using the default threshold. The threshold will be evaluated separately in terms of its effect on the two fairness outcomes. Figure 4 shows how the different decision points generate different combinations of engineering choices, leading to 30 different ML pipelines for each data situation. Overall, Figure 4 demonstrates how each decision leads to a tree of subsequent choices, indicating the large landscape of possible ML pipeline configurations that result from the engineering choices.

Figure 4.
Combination tree for four decision points



For each of the 30 models for one of the data situations, general model performance in terms of accuracy and AUC are obtained alongside the FPR and FNR for each of the eight intersectional subpopulations. Those values are then used to evaluate the potential impact of each decision point on the fairness of the predicted outcome.

Demonstration of the data analysis

The overall effect of each decision point is isolated by obtaining the mean change of the fairness measures when changing the applied method in the decision point while keeping the other decisions constant. The higher the mean change $\overline{\Delta_{dp}}$, the more potential this decision point shows to affect the fairness outcome, where *dp* represents the decision point (e.g., algorithm choice, imputation method, scarcity intervention). Furthermore, the standard deviation of the mean change $\overline{\Delta_{dp}}$ indicates change potential within that decision point.

For each possible combination of engineering choices, the AUC, FNR, and FPR of the resulting model are recorded for the overall model and separately for each of the eight subpopulations. The resulting data structure is shown in Table 4 with some exemplary values; however, this example is strongly simplified by only showing a fraction of the resulting rows.

Table 4.

Exemplary data structure for the analysis

Algorithm	Tuning	Imputation	Scarcity	Metric	Value	Gender	Ethnicity	Disability
log	No	simple	No	FPR	0.287	*NA*	*NA*	*NA*
log	No	simple	Oversa.	FPR	0.314	*NA*	*NA*	*NA*
log	No	MissForest	No	FPR	0.253	*NA*	*NA*	*NA*
XGBoost	No	simple	No	FPR	0.531	*NA*	*NA*	*NA*
RF	No	simple	No	FPR	0.137	*NA*	*NA*	*NA*
log	No	simple	No	FPR	0.341	female	minority	disabled
log	No	MissForest	No	FPR	0.374	female	minority	disabled
XGBoost	No	simple	No	FPR	0.582	female	minority	disabled
XGBoost	No	MissForest	No	FPR	0.613	female	minority	disabled
RF	No	simple	No	FPR	0.196	female	minority	disabled
RF	No	MissForest	No	FPR	0.237	female	minority	disabled

Note. Introduced abbreviations are log – logistic regression, RF – Random Forest, Oversa - Oversampling

To explain the process in more detail, the decision point of algorithmic choice is used as an example using the values from Table 4 for the calculations of the FPR. Firstly, the overall mean metrics are assessed for each choice within the decision point. Let *alg*∈{logistic regression,

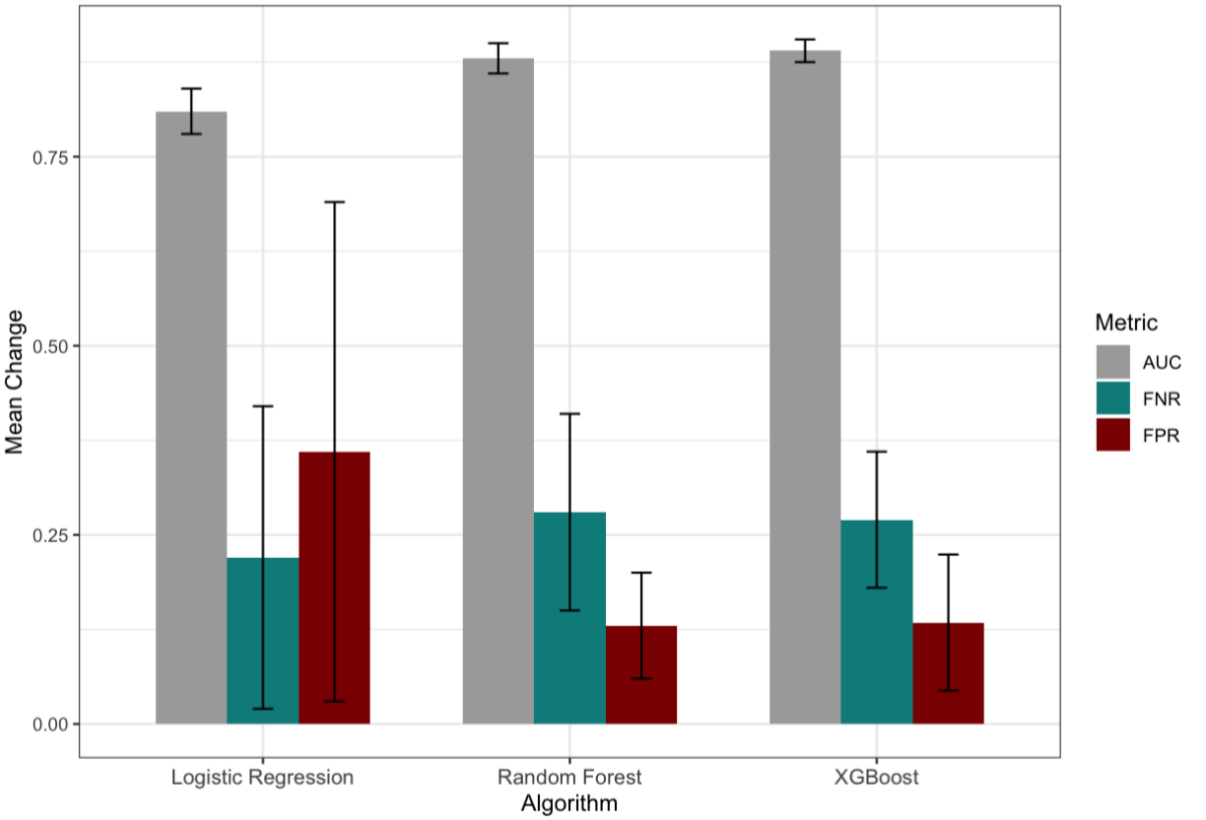
random forest, XGBoost} represent the available algorithms. Figure 5 shows the overall mean AUC_{alg}, FNR_{alg} and FPR_{alg}, with their standard deviation σ_{Δ} as error bars for each of the three investigated algorithms. This gives a broad insight into the fairness change depending on the choice of algorithm. For the example shown in Table 4 using only 3 of the 30 available FPR for the logistic regression, this value would be derived as follows:

$$\bar{X}_{alg} = \frac{1}{3}(0.287 + 0.314 + 0.253) = 0.285$$

$$\sigma_{alg} = \frac{(0.287 - 0.285)^2 + (0.314 - 0.285)^2 + (0.253 - 0.285)^2}{3} = 0.031$$

This means that Figure 5 shows the derived mean values as bars and standard deviations as error bars, considering the overall outcome of the 30 model combinations for AUC, FNR, and FPR.

Figure 5.
Mean values for AUC, FNR and FPR for the three algorithm choices



From those performances, differentiated by choice, an overall measure of mean change $\bar{\Delta}_{alg}$ is obtained to isolate the fairness potential within this decision point. By this, the mean change from the baseline setting (here logistic regression, $M_{logistic\ regression}$) to the other possible

options $M_{other\ alg}$ is calculated for configurations that keep the other choices constant. This mean change is given by:

$$\bar{\Delta}_{alg} = \frac{1}{n}\sum_{i=1}^{n}\left|metric^{i}_{other\ alg} - metric^{i}_{logistic\ regression}\right| \quad (34)$$

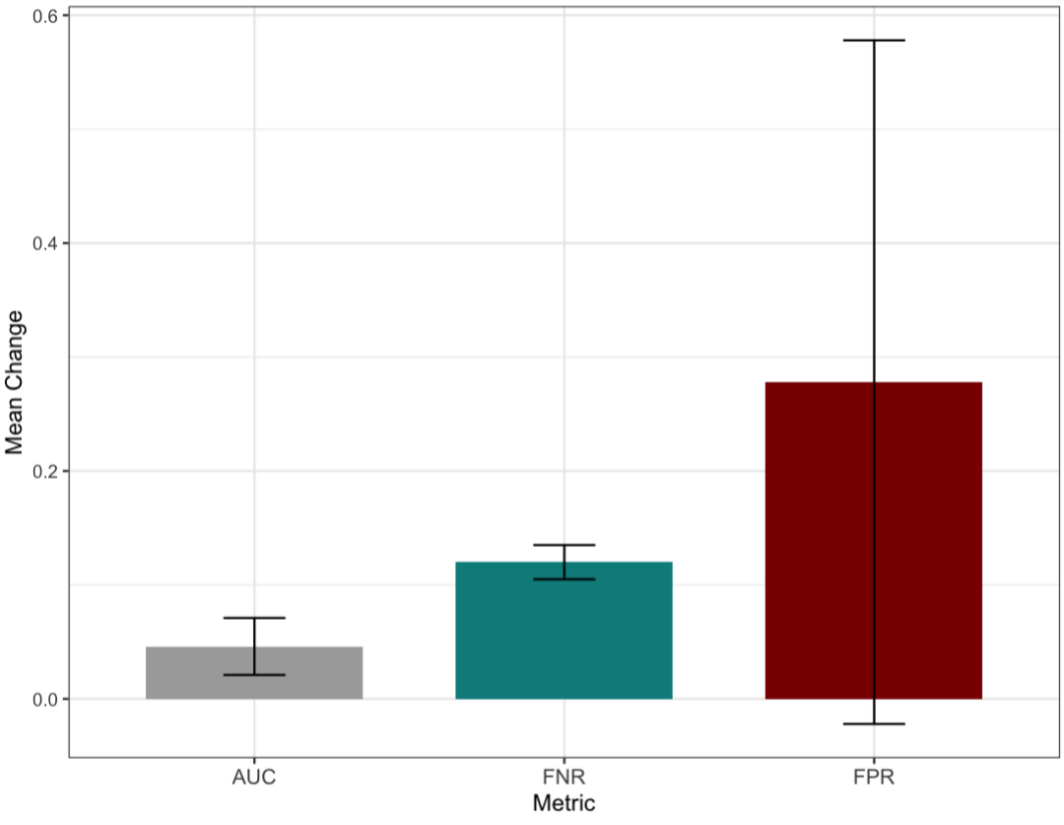
Where n is the number of applicable model constellations, and $metric \in \{AUC, FPR, FNR\}$, for the example data in Table 4, this would give the following values:

$$\bar{\Delta}_{alg} = \frac{1}{2}(|0.287 - 0.531| + |0.287 - 0.135|) = 0.274$$

$$\sigma_{alg} = \frac{(0.152 - 0.274)^2 + (0.244 - 0.274)^2}{2} = 0.065$$

When using all values available, this shows the overall changes in AUC, FPR, and FNR achieved through the changes in the algorithm alone, as detailed in Figure 6. This example would show that the algorithm choice seems to affect the resulting FPR more than the FNR.

Figure 6.
Mean change in AUC, FPR, FNR values for algorithm choice



Lastly, the change indicated in Figure 6 can be broken down into the different subpopulations, showing which subpopulations the algorithm changes leads to the biggest change in fairness outcome. This breakdown helps assess whether specific subpopulations experience larger impacts. For subpopulations defined by the intersectionality of gender G ∈ {male, female}, ethnicity E ∈ {majority, minority} and disability D∈ {disabled, non-disabled}. The calculations follow the same approach as in equation 34, repeated for each subpopulation. In this case, the exemplary numbers from Table 4 would provide insight into the subpopulation of disabled females from the minority population with the following calculations for FPR when only considering the four model constellations:

$$\bar{X}_{fem\,\min dis} = \frac{1}{4}(|0.341 - 0.582| + |0.341 - 0.196| + |0.374 - 0.613| + |0.374 - 0.237|)$$

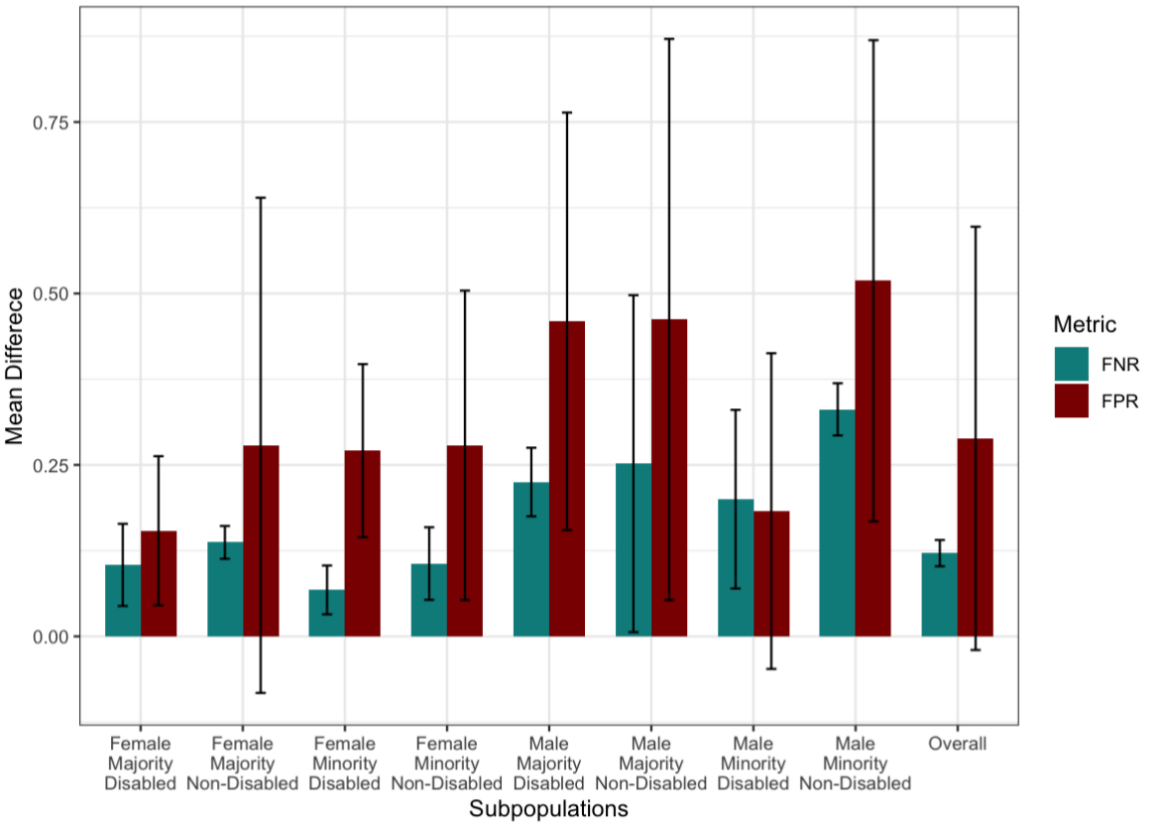
$$\bar{X}_{fem\ min\ dis} = 0.191$$

$$\sigma_{fem\,\min dis} = \frac{(0.241 - 0.191)^2 + (0.145 - 0.191)^2 + (0.239 - 0.191)^2 + (0.137 - 0.191)^2}{4}$$

$$\sigma_{fem\ min\ dis} = 0.058$$

Figure 7 shows this mean change $\bar{\Delta}_{alg}$ by subpopulation, detailing possible differentiating effects between the eight cross-sectional subpopulations.

Figure 7.
Mean change in fairness caused by algorithm choice for different subpopulations



The analysis of the mean of fairness and performance metrics and the mean changes within the decision point will be repeated for each decision point on the default threshold of 0.5. First, a broad overview of the average outcome for each option in the decision point is given, and then the potential for changing the fairness outcome that this decision point holds is shown. Lastly, this change potential is segregated into the eight subpopulations to showcase how different groups might benefit through changes within that decision point.

Chapter 4 – Analysis

The data analysis aims to reveal which decision points can influence the overall model's fairness outcome in terms of mean change between similar models that only differentiate within the investigated decision points. By this, mean change values and their standard deviations are generated from the 30 pre-defined pipeline set-ups for each of the four data situations (with the combinations indicated in Figure 4). This leads to a repetitive investigation of how the decision points might vary in their influence depending on the specific challenges that arise from a given data set in terms of non-linearity of the decision boundary and scarce outcome. To derive the performance for each unique engineering choice combination for each of the datasets, the 30 unique pipelines were run on each dataset. The computational time took around two hours for one data situation and eight hours in total (calculations were conducted in R version 4.3.2 on a MacBook Air (Apple M1, 2020) with macOS Monterey 12.4 and RAM 8GB, the R code is detailed in Appendix A). The derived performance values were then used to follow the analysis as outlined in the example analysis in Chapter 3 for each data situation separably.

The analysis procedure follows an overarching outline: first, it introduces the characteristics of the focused data situation, and the possible challenges associated with it. Then, it structurally interprets the mean change of fairness metrics for each of the five decision points in the order of occurrence within the typical ML pipeline. Lastly, it provides a summary showing the key insights from this data situation before moving to the next data situation.

First data situation - Linear decision boundary with non-scarce outcome

The following data situation is characterized by linear separability, showing that the decision boundary linearly depends on the seven features shown in Figure 12. It shows how income and education have a negative linear relationship to the fraud probability and disability has a positive relationship, with the other generated variables also exhibiting linear connections to the outcome. The first and second generated variables negatively affect the crime outcome. The fourth variable shows a positive relationship to the outcome, while the third variable shows no relevant impact on the probability of fraud. Overall, the fraud classification is scarcely distributed but more balanced, with 25% of the cases within the data being classified as fraudulent. Conclusively, this data situation is the easiest of the four combinations detailed in Table 5. With the relationship between fraud commitment and the features being linearly separable, this could lead to logistic regression to be a reliable classifier. It offers an environment where the classifiers can assume a linear separability and a non-scarce outcome,

which helps avoid a majority class bias. Even though the distribution of the binary outcome is not fully balanced, with only a quarter belonging to the fraudulent classification, the challenge is less pressing. There is still a risk of overfitting for the majority class. However, the algorithm will not disregard an outcome that makes up a quarter of the cases as much as with the scarcity setting.

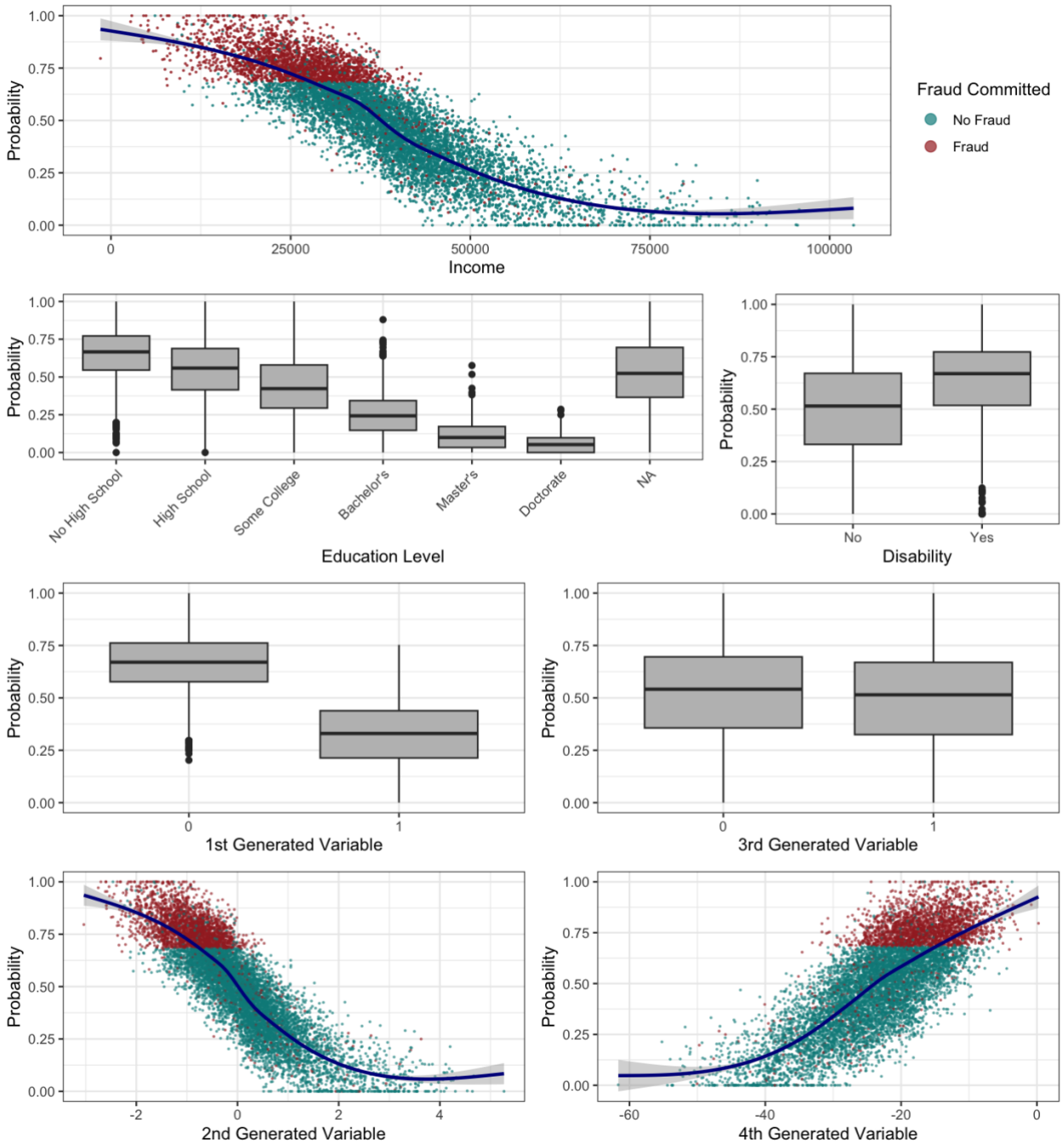
Table 5.

Overview of the currently tested data characteristics

	Decision Boundary	
Distribution of outcome	Linear, Non-scarce outcome	Non-linear, Non-scarce outcome
	Linear, Scarce outcome	Non-linear, Scarce outcome

The connections between fraud probability and the predictive features are detailed in Figure 8. The data was simulated according to the data factory outline of Chapter 3. The sensitive attributes influenced the outcome of the predictive features. While they do not directly relate to the fraud outcome, they impress on it indirectly through their relations to the provided model features. This creates a realistic and representative society that models the underlying influences of sensitive attributes on an individual's outcome on factors such as income and education.

Figure 8.
Relationship of fraud probability and classification by features



Overview of decision point impact

Before introducing the mean change potential in each decision point, the overall mean performance for each engineering choice is assessed in terms of AUC, FNR and FPR. The calculations were done according to the exemplary analysis in Chapter 3. Figure 9 demonstrates those mean performances (bars) and their standard deviations (error bars). They start with the imputation method, which shows similar outcomes on overall and fairness performance for both choices in terms of mean and standard deviation, which indicates only a slight difference in

fairness performance between the two imputation methods. Secondly, the scarce outcome intervention shows similar mean performances as the missing data imputation, with more visible variability between the methods. When no scarcity intervention is applied, this results in a higher FNR. However, random oversampling shows a balance between FNR and FPR, while random undersampling has a higher FPR than the FNR. Overall, the variability is similar between the scarcity interventions.

Following the choice of algorithm, the logistic regression shows some increase in AUC compared to the tree-based methods and additionally shows a balanced outcome for FNR and FPR. The tree-based methods exhibit similar performance and fairness. Lastly, the tuning procedure elicits an increase in the FNR compared to untuned models. Both for the choice of algorithm and tuning, the standard deviations are similar between the methods.

Figure 9.
Mean performance for each investigate engineering choice within the pipeline

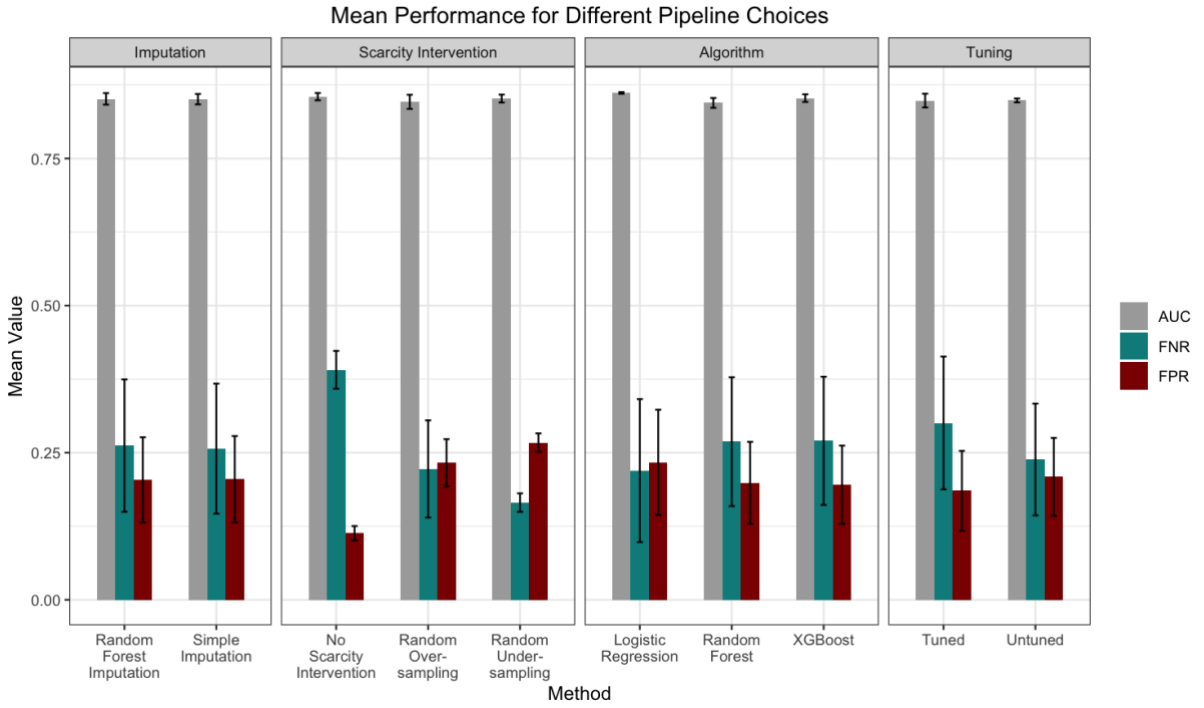
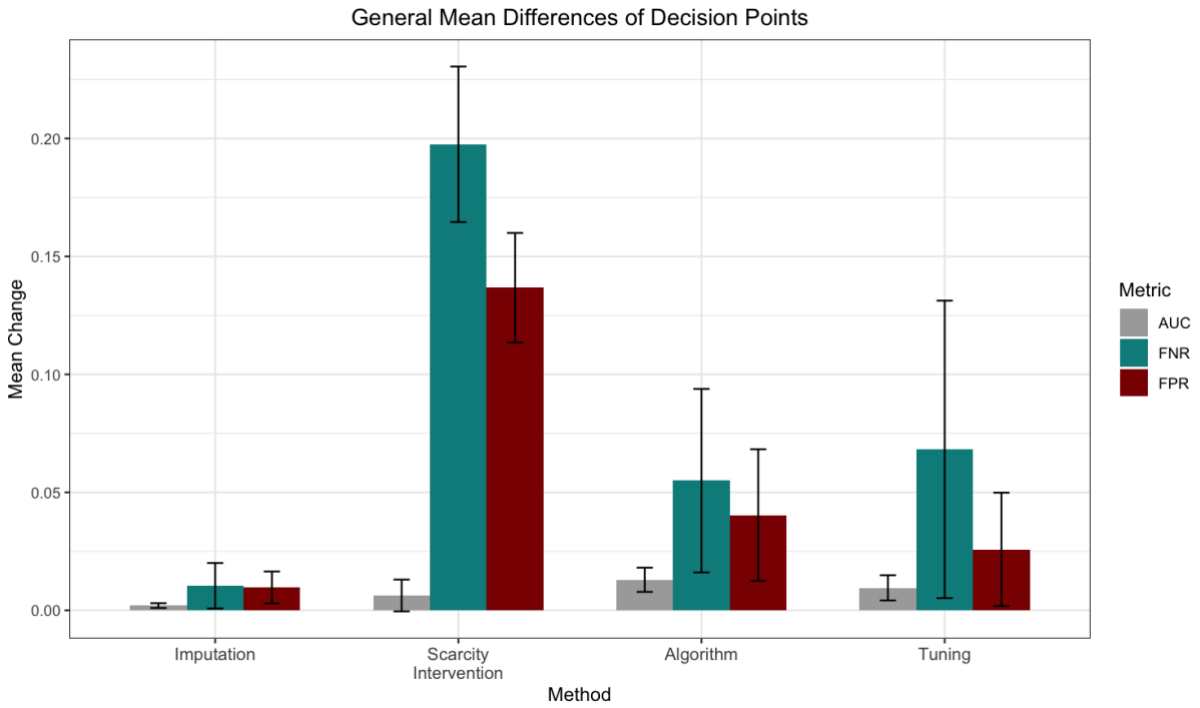


Figure 10 demonstrates the mean changes (bars) and their standard deviations (error bars) that occur within each of the decision points while keeping the other engineering choices constant. The Figure is calculated using the analysis outline of Chapter 3. It shows that the greatest change potential is in the scarcity intervention method. Overall, the decision points assert a greater influence on the FNR outcome than the FPR. The greatest variability is observed in the FNR of the hyperparameter tuning decision point.

Figure 10

Mean differences in performance and fairness for different decision points



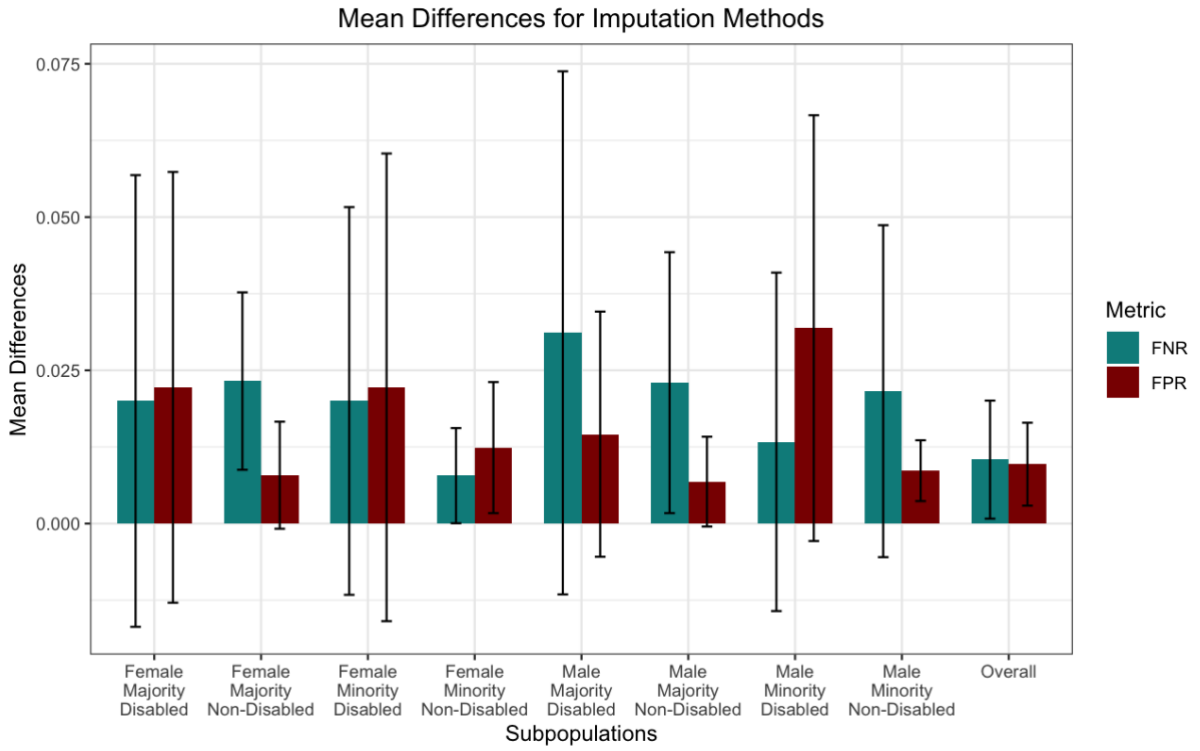
Those mean differences are further investigated for the subpopulations on each decision point to see whether the engineering choices can treat some subpopulations differently while others might not experience change. The exploration of the singular decision points will follow the order of the usual ML classifier pipeline.

First decision point: missing data imputation

The missing data imputation was the decision point with the most minor global influence on performance and fairness measures. Overall, judging by the standard deviations of the values shown in Figure 11, the disabled subpopulations show greater variety in the fairness outcome compared to the non-disabled subpopulations. This highlights how the imputation affects the individuals with their data not missing randomly, as disability was a defined factor for omitted data points. Generally, it also shows that the non-disabled subpopulations tend to exhibit greater change in the FNR, while the disabled subpopulations tend to have greater change in the FPR.

Figure 11.

Mean difference in FNR and FPR for subpopulations created by missing data imputation



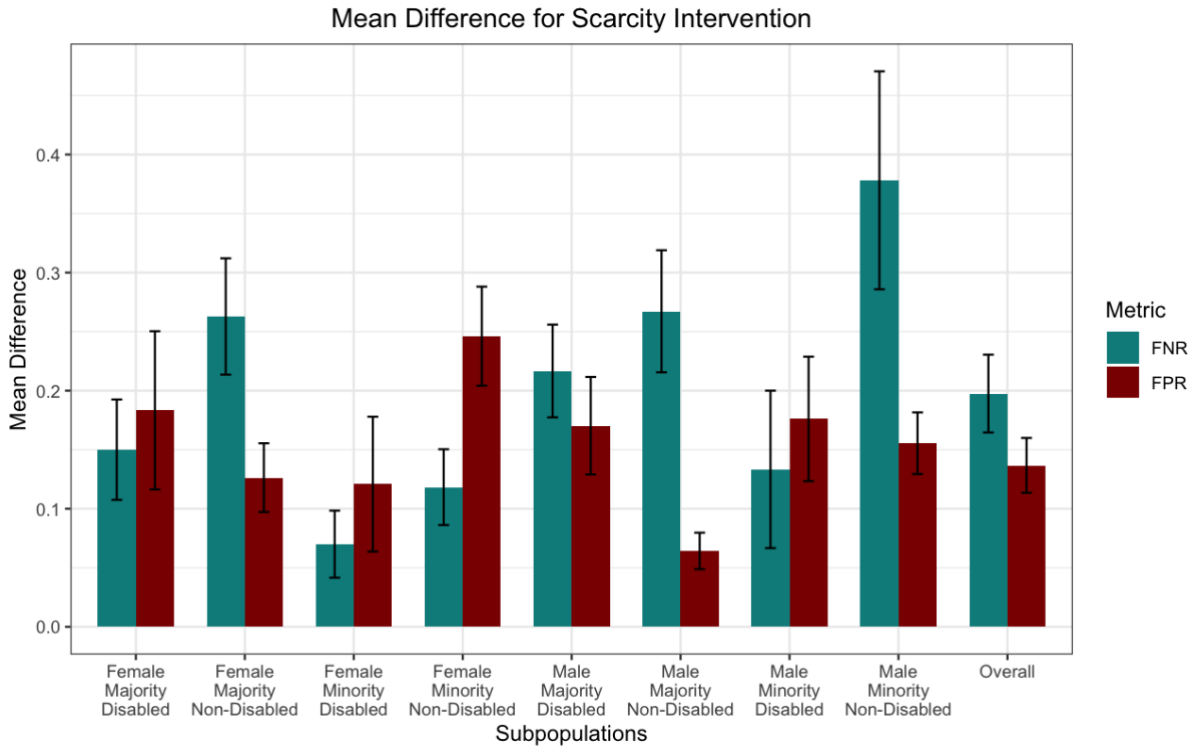
Overall, this shows that the influence of the data imputation can be different regarding fairness depending on whether an individual is categorized as disabled or not.

Second decision point: scarcity intervention

The scarce outcome intervention demonstrated the greatest change potential among all the decision points. Figure 12 breaks down this mean change of FNR and FPR by the eight subpopulations. It shows the greatest change and variability in FNR for the male X minority X non-disabled group. Conversely, lower change in FNR is experienced by the female X minority X non-disabled group, showing a gender effect of the scarcity intervention. This subpopulation is also experiencing the greatest change in FPR, an additional indicator of a possible gender effect. The most advantaged group (male X majority X non-disabled) shows the lowest change

in FPR, showing that the classifier threatens them to be more likely to wrongfully categorize them as fraudulent despite the choice of scarcity intervention. However, it asserts a greater influence on their FNR, showing that they have greater potential to benefit from the changes as being wrongfully classified as non-fraudulent benefits to the individual. Overall, the variability, expressed through the standard deviations of the mean change, remains similar across groups.

Figure 12.
Mean difference in FNR and FPR for subpopulations created by scarcity intervention



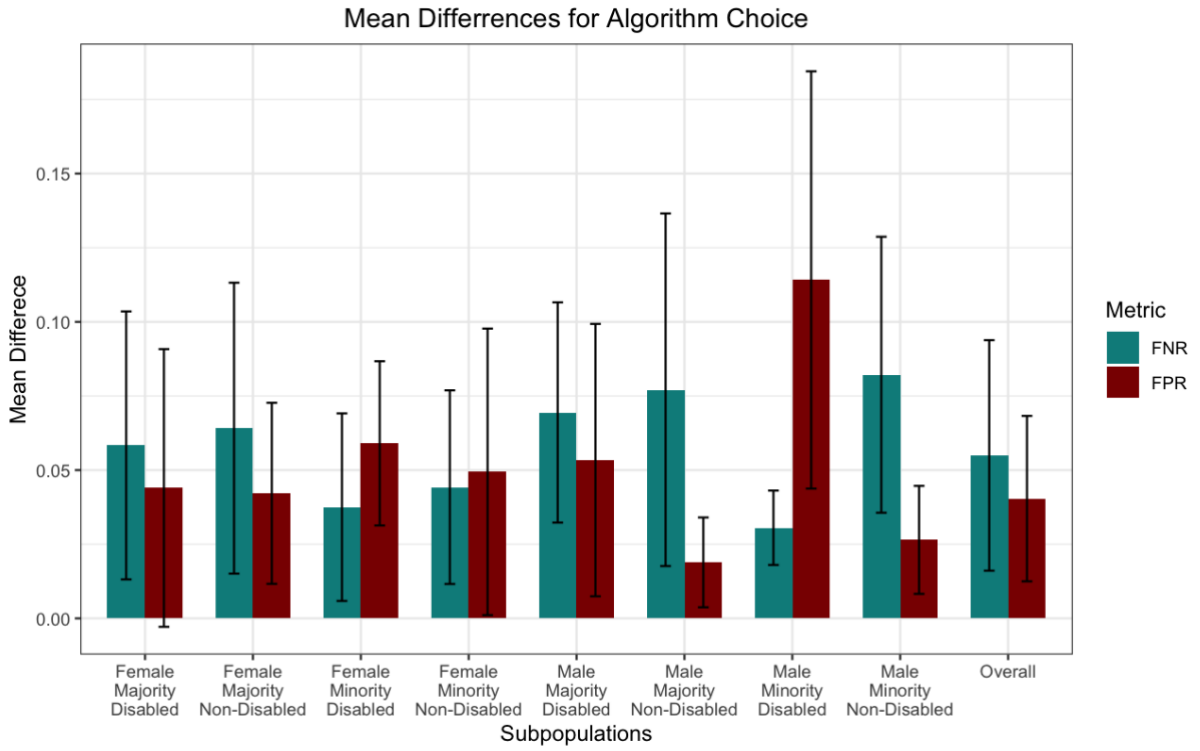
Conclusively, the scarcity intervention shows some effect on the fairness outcomes of the pipelines despite the non-scarce distribution of the target variable. Additionally, some vulnerable groups experience greater changes depending on the chosen scarcity intervention. Lastly, the most advantaged group shows less risk of fluctuations in the FPR but greater potential for changes in FNR, which makes individuals of the group more likely to benefit.

Third decision point: algorithm choice

Furthermore, the choice of algorithm can assert changes in the fairness metrics, as shown by the average changes in Figure 10. The mean change by subpopulation is demonstrated in Figure 13. Overall, the patterns of FNR and FPR seem similar for all the female subpopulations, showing a balanced change for both measures. However, comparing the most advantaged male population (male X majority X non-disabled) with the least advantaged male population (male X minority X disabled), a contrast in how the algorithm affects those subpopulations is revealed.

The advantaged group demonstrates low FPR and higher FNR changes, showing a higher possibility for the positive impact of the choice of algorithm rather than a negative impact. However, this effect is reversed for the disadvantaged group with a higher mean increase in the FPR, showing the threat of more wrongfully fraudulent classification for individuals of this population and resulting adverse treatment. Furthermore, the standard deviations (error bars) indicate more substantial variability in the outcome of the female subpopulations compared to the males.

Figure 13.
Mean difference in FNR and FPR for subpopulations created by algorithm choice



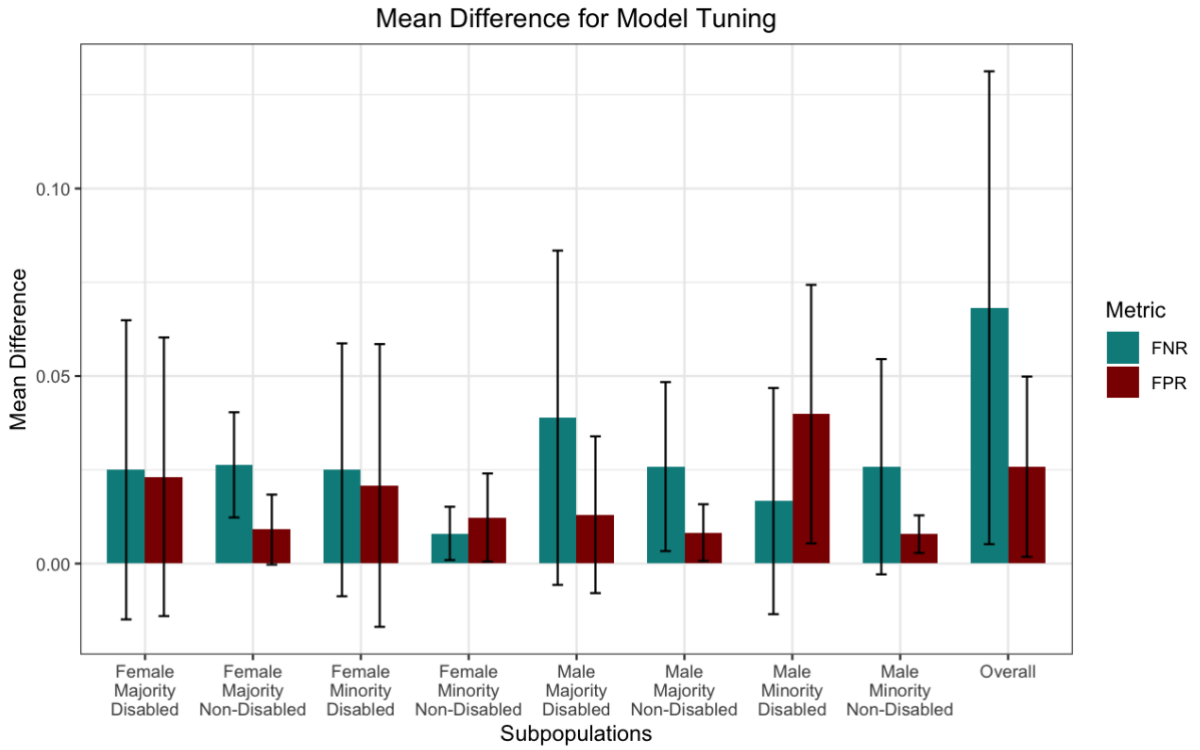
The female subpopulations experience similar change patterns depending on the algorithmic choice. However, the male subpopulations vary in impact depending on their vulnerability, with more vulnerable individuals being more prone to adverse impact. In contrast, more advantaged individuals have more potential for beneficial impact of the algorithm choice.

Fourth decision point: hyperparameter tuning

The effect of hyperparameter tuning for the tree-based methods is demonstrated by the subpopulation in Figure 14. Overall, the tuning does not greatly impact the fairness measures compared to the other data situations that will be introduced; overall, it contributes more towards the FNR than the FPR. Generally, judging by the standard deviation of the values, the variability in change is higher for the disabled subpopulations, showing that the tuning process affects this group more than the non-disabled individuals. Additionally, the FNR change is higher than the FPR change for all the subpopulations apart from the disabled males of minority ethnicity, showing that they have greater potential for adverse treatment depending on the tuning process.

Figure 14.

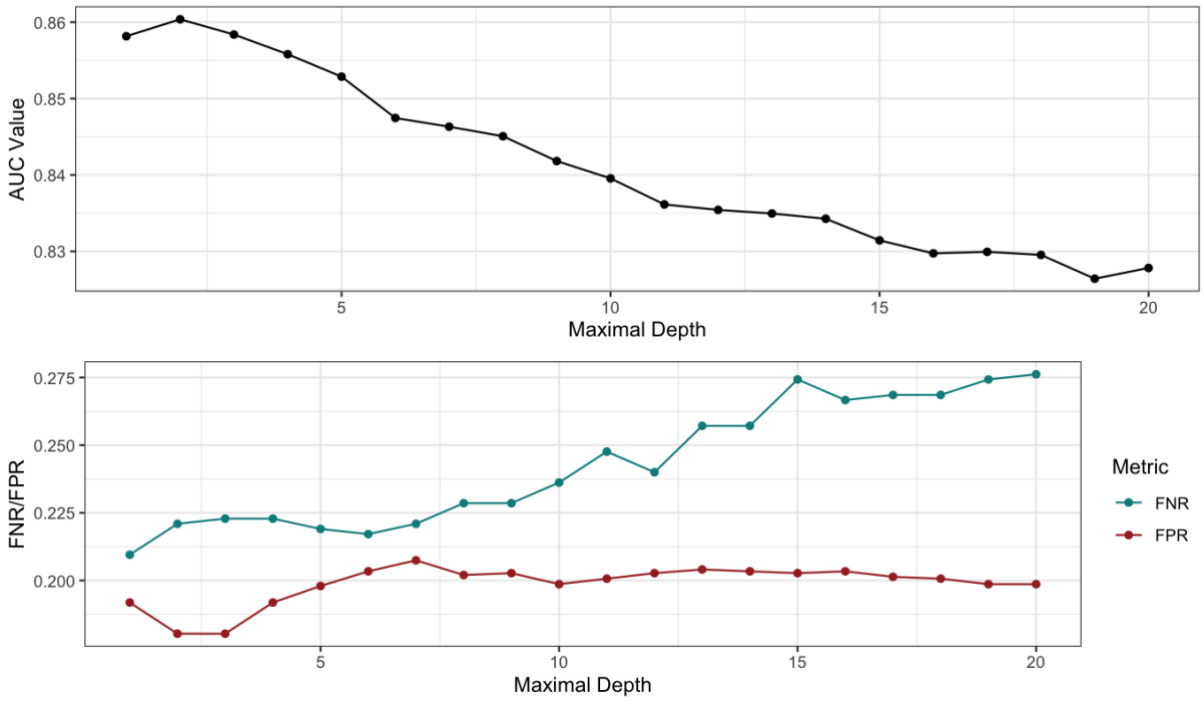
Mean difference in FNR and FPR for subpopulations created by hyperparameter tuning



The tuning process of XGBoost and Random Forest is different since both methods have different hyperparameters. To probe into the process, a hyperparameter is tuned for each algorithm while keeping the other hyperparameters at their default values. Starting with XGBoost, Figure 15 shows the tuning of the maximal depth trees are allowed to grow into and their performance on AUC and the fairness metrics. It shows that the AUC peaks early at the maximum depth of two, while the fairness measures develop differently through the maximum depth range. For the maximum depth of one, the fairness measures approach each other more closely than for the maximum depth of two, while the AUC value remains similar. This shows

the potential to adjust for fairness measures with hyperparameter tuning at a low cost for the AUC performance. However, the change in the fairness measures is relatively small, which decides how vital the minimal change in fairness is balanced to the slight loss in performance of a domain-specific one. Regardless, this indicates the potential benefit of evaluating the hyperparameter tuning process from the perspective of performance and fairness separately.

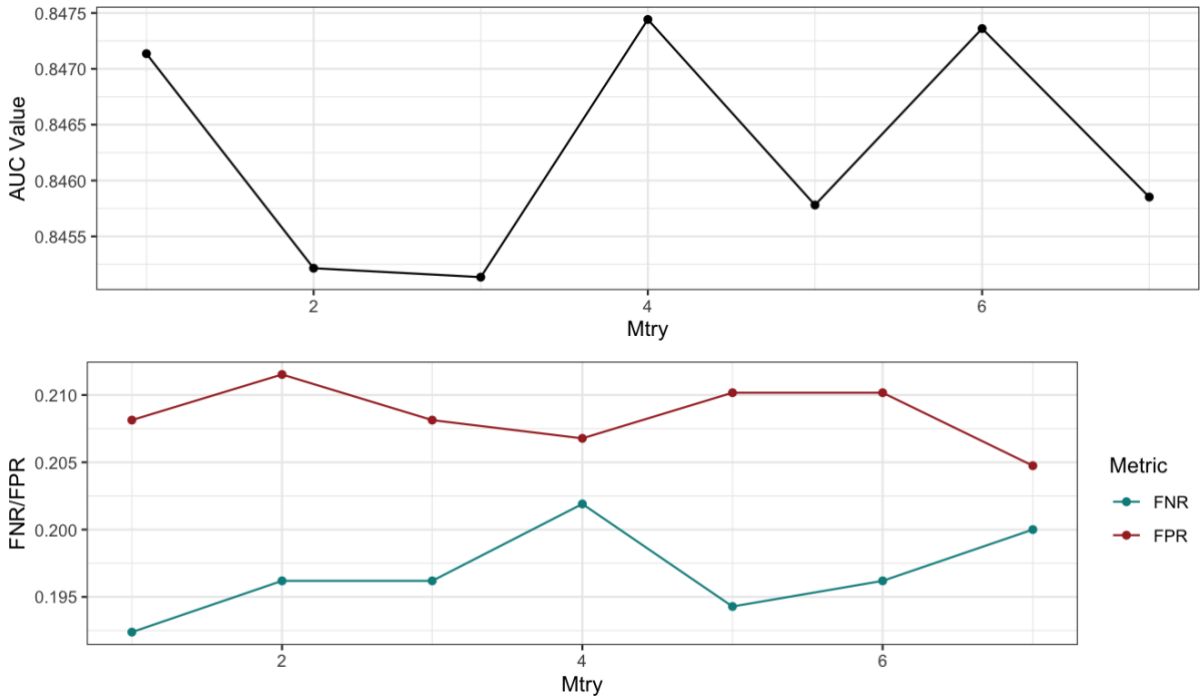
Figure 15.
Isolated tuning of XGBoost's maximal depth and its impact on AUC, FNR and FPR



Furthermore, the change for the Random Forest in performance and fairness measures is assessed depending on the hyperparameter that indicates the maximum number of features selected at each split (mtry). Figure 16 shows that the overall change in the AUC value is minimal depending on the hyperparameter value, as the scare of the y-axis only records small value changes. Overall, it gives two peaks at a maximum number of features set at four or six. Looking at the fairness outcome, there is some difference in the performance, with the lower maximum depth value resulting in a comparably higher FNR of approximately 5% and a lower FPR of approximately 2.5%. This allows the engineer to set the hyperparameter to either achieve a higher FPR or a higher FNR while keeping the overall performance constant. This potential for changing the fairness outcome can be easily overlooked when only tuning for increased performance while disregarding the fairness impact.

Figure 16.

Isolated tuning of Random Forest's mtry and its impact on AUC, FNR and FPR



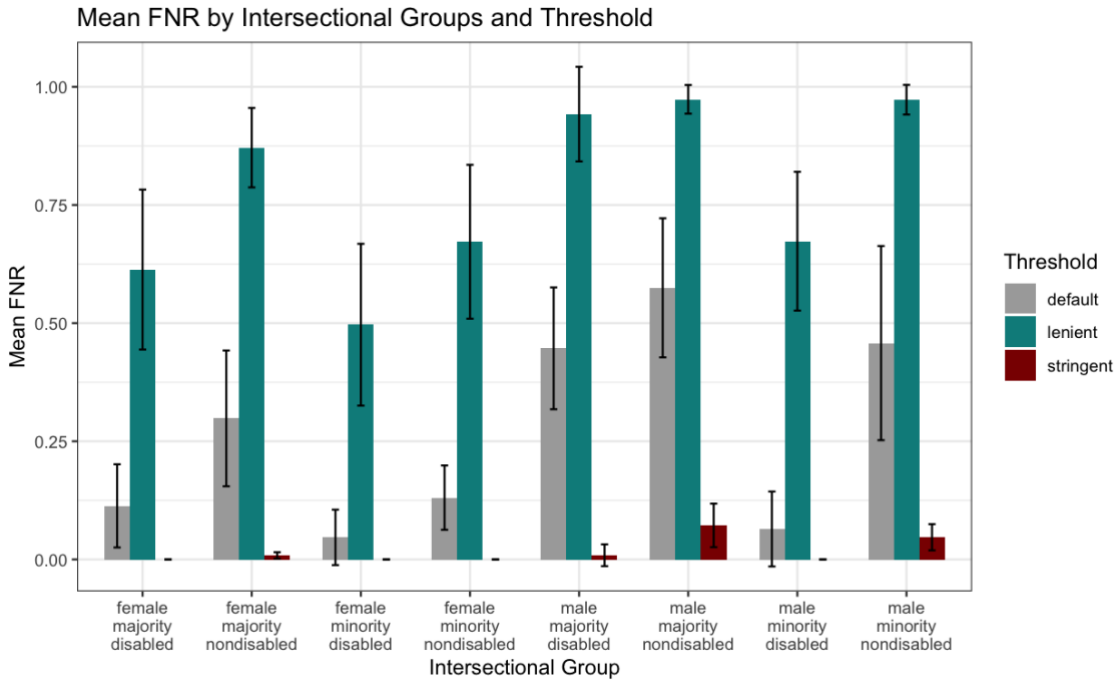
Overall, the hyperparameter tuning shows some potential to balance FNR and FPR differently while the overall model performance remains consistent. The choice of which hyperparameter value to set can be domain-specific; however, the fairness impact can be considered without experiencing a loss in model performance.

Fifth decision point: threshold setting

Lastly, the threshold setting is assessed, and all the other decision points have been evaluated based on the default threshold (cut-off at 0.5). Two different threshold settings are introduced and compared to the default threshold, giving a lenient option (cut-off above the 0.9 quantile of fraud probability) and a stringent option (cut-off above the 0.1 quantile). Firstly, Figure 17 shows the impact of the lenient and stringent threshold setting compared to the default setting on the FNR. The lenient threshold increases the FNR for more advantaged subpopulations with a one or lower vulnerability level. More vulnerable subpopulations still experience an increase in the FNR; however, the overall difference between the groups is lower. The stringent threshold minimizes the FNR across all subpopulations, with the most advantaged subpopulations still having the highest FNR.

Figure 17.

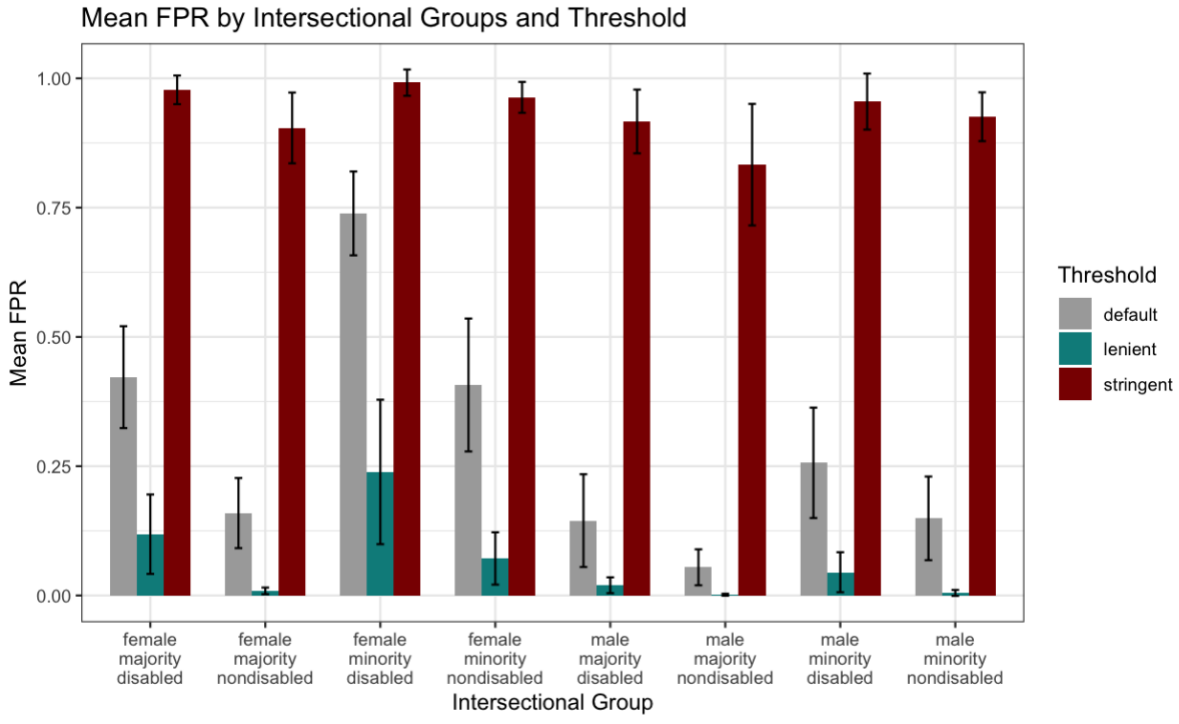
Effect of threshold setting on FNR for different subpopulations



Regarding FPR, the stringent threshold increases the value compared to the default value and shows an overall high FPR for all subpopulations in Figure 18. The lenient threshold mainly lowers the FPR for the male subpopulations. However, the most disadvantaged group will still have a higher FPR with the lenient threshold, showing that the more advantaged a group is, the more it will benefit from the lenient threshold setting.

Figure 18.

Effect of threshold setting on FPR for different subpopulations



Summary

Overall, this data situation was optimal for an ML pipeline, as it did not face the challenge of non-linearity and scarce outcome distribution. The missing data imputation asserted greater influence on the disabled subpopulations compared to the non-disabled, showing how belonging to a minority group that is more prone to missing data might show a greater impact on the outcome depending on the chosen imputation method.

Additionally, the scarcity intervention had a minor effect than the scarce linear data situation. However, it still asserted the biggest overall effect within the ML pipeline. It showed that the most advantaged subpopulations are more likely to benefit from the scarcity intervention than the disadvantaged subpopulations.

Furthermore, the choice of algorithm had the greatest impact on the male subpopulations, with the more vulnerable male subpopulations more prone to adverse algorithm choice effects.

The hyperparameter tuning revealed greater change potential for the disabled subpopulations. More detailed probing into single hyperparameters for each tree-based algorithm also demonstrated the potential to consider the FNR and FPR during tuning and the possibility of influencing the overall fairness outcome to a certain extent while keeping the model performance consistent.

Lastly, the threshold setting demonstrated that lenient and stringent thresholds can modulate fairness outcomes. However, the more disadvantaged groups will be more affected by the stringent threshold, and the more advantaged groups will benefit more from the lenient thresholds in terms of fairness outcomes.

Table 6.

Summary of insights from data with non-scarce outcome and linear decision boundary

Decision Point	Impact on Fairness (FNR, FPR)	Subpopulation Variability and Key Observations
Missing Data Imputation	Similar overall performance between methods (minimal differences)	Disabled subpopulations display more variability and higher FPR, while non-disabled groups have higher FNR
Scare Outcome Intervention	Lack of intervention results in higher FNR, while oversampling balances FNR and FPR	Vulnerable groups show greater fairness variability, and advantaged groups show less FPR fluctuations
Algorithm Choice	Logistic regression shows more balanced fairness measures, compared to tree-based methods	Primarily affects male subpopulations, with less vulnerable males benefitting and more vulnerable males facing adverse effects
Hyperparameter Tuning	Increased FNR compared to untuned models	Disabled subpopulations show greater variability due to tuning. FNR and FPR adjustments are possible with minimal AUC changes
Threshold Setting	Lenient threshold increases FNR in advantaged groups more, disadvantaged groups face higher FPR under stringent threshold	More advantaged groups benefit more from both threshold compared to disadvantaged groups

Second data situation – Linear decision boundary with scarce outcome

The second data situation is characterized by a linear decision boundary and scarce outcome, as depicted in Table 7, generated through the data factory detailed in Chapter 3. Figure 19 demonstrates the simulated society generated from this contingency. One can see here how the decision boundary of committing fraud is negatively linearly related to higher income and education. However, disability raises the probability of an individual to commit fraud. Furthermore, the generated variables further contribute to a linear relationship between them and the individual's fraud classification. The first and second generated variables reduce the fraud probability, while the fourth variable linearly increases the fraud probability. The third generated variable shows no major contribution to the decision boundary. All variables have an underlying direct or indirect relation to the protected attributes (as indicated in Figure 3 in Chapter 3). This means that while the protected attributes are not present in the decision boundary setting, they are still indirectly linked to it through the proxy variables depicted in Figure 19. Hence, the Figure shows realistic connections between the features and fraud outcome, considering underlying and non-visible societal factors.

Table 7.

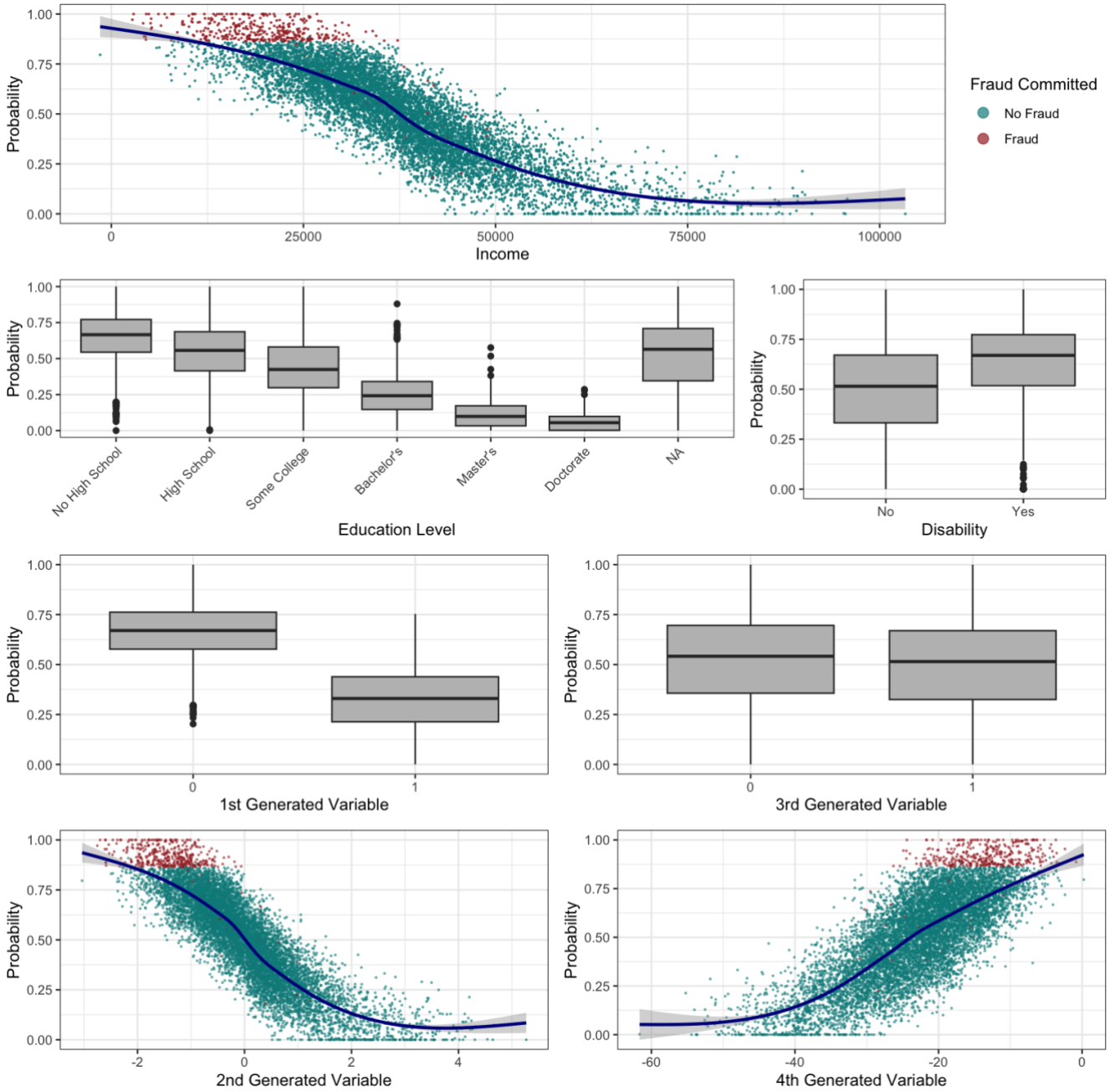
Overview of the currently tested data characteristics

	Decision Boundary	
Distribution of outcome	Linear, Non-scarce outcome	Non-linear, Non-scarce outcome
	Linear, Scarce outcome	Non-linear, Scarce outcome

In Figure 19, the positive cases (red, committed fraud) are linearly separable from the negative cases (blue, no fraud committed). However, this data set's challenge arises from its scarce distribution, with only 4% of the cases being classified as fraudulent. This mimics the often-present challenge of the target variable being rarely distributed. Conclusively, this might allow algorithms that assume linearity within the data to work as well as more flexible algorithms. However, the challenge of picking up on patterns associated with positive classification sufficiently affects all the applied algorithms. The class imbalance poses the problem that the classifier might be biased towards the majority outcome, achieving high accuracy while failing to detect fraudulent instances reliably.

Figure 19.

Relationship of fraud probability and classification by features



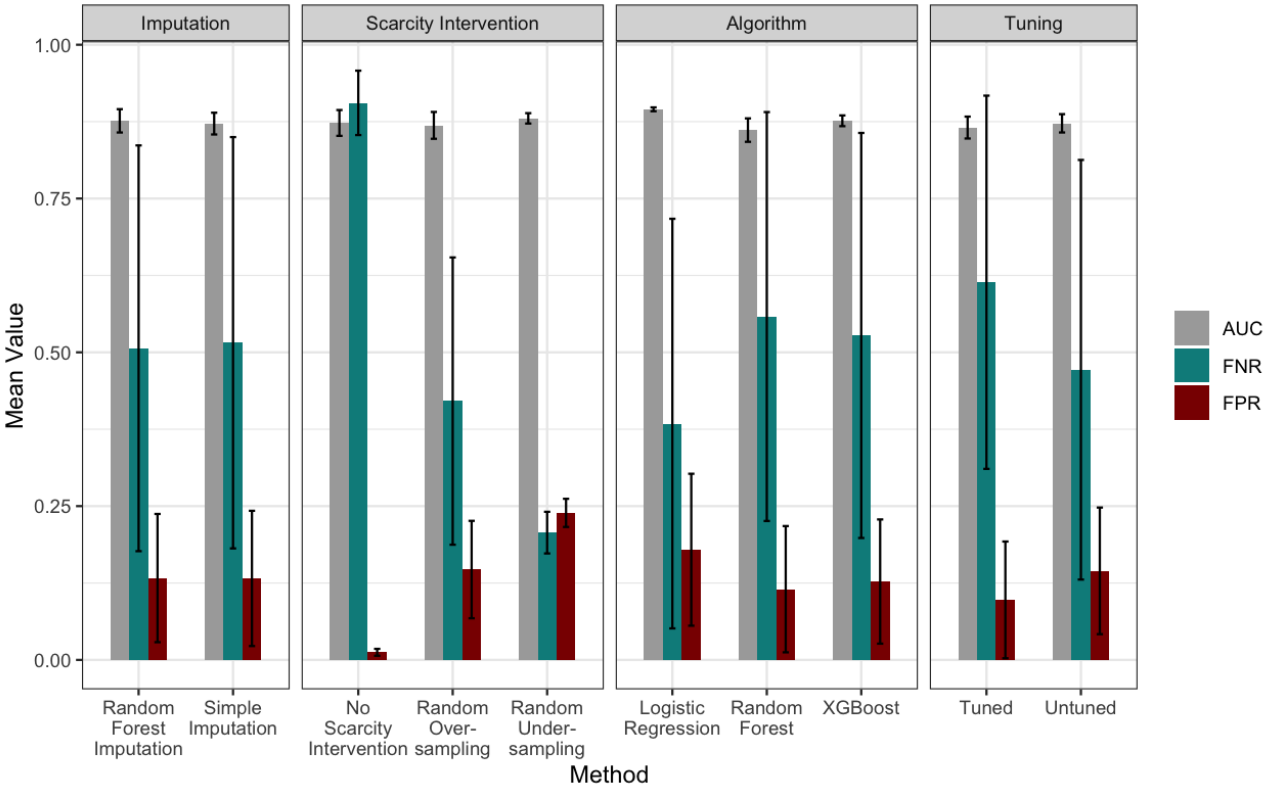
Overview of decision point impact

Before introducing the mean change within each decision point, the overall performance for each method in each decision point is explored. Figure 20 shows the four decision points performing under the default threshold of 0.5 for AUC, FNR, and FPR. It shows the mean values (bars) and their standard variations (error bars) calculated from model combinations that contain the engineering choice, as detailed in the exemplary analysis in chapter 3.

Beginning with the choice of imputation method, there seems to be no remarkable difference between any performance metric and the chosen imputation method, with mean values and standard deviation being similar across both methods. However, the chosen scarcity intervention shows the highest effect on the fairness metrics, while the AUC remains substantially consistent. Not introducing any scarcity intervention has a high impact on the FNR, which reflects the challenge of a classifier not being biased towards the majority class and failing to identify fraudulent cases. Furthermore, for scarcity intervention, the error bars show no high variation in the differentiating outcomes, apart from the FNR performance in pipelines using random oversampling.

The algorithm choice reflects that the logistic regression works well under the linearly separable environment, indicated by the high AUC values. Furthermore, the other algorithm choices show a higher impact on the FNR. However, the FPR behaves similarly between the algorithm choices. The high impact of algorithm choice also manifests in greater variability of mean values shown by the standard deviations. Lastly, the hyperparameter tuning shows the potential to increase the FNR further. This is also reflected in the standard deviations for the FNR, showing great potential for changing FNR.

Figure 20.
Mean performance for each investigate engineering choice within the pipeline



The mean changes that occur when changing one decision while keeping the remaining decision points constant can reveal the change potential within the inspected decision point. This mean change for each decision point is detailed in Figure 21 and derives according to the detailed analysis in Chapter 3.

Figure 21.
Mean differences in performance and fairness for different decision points

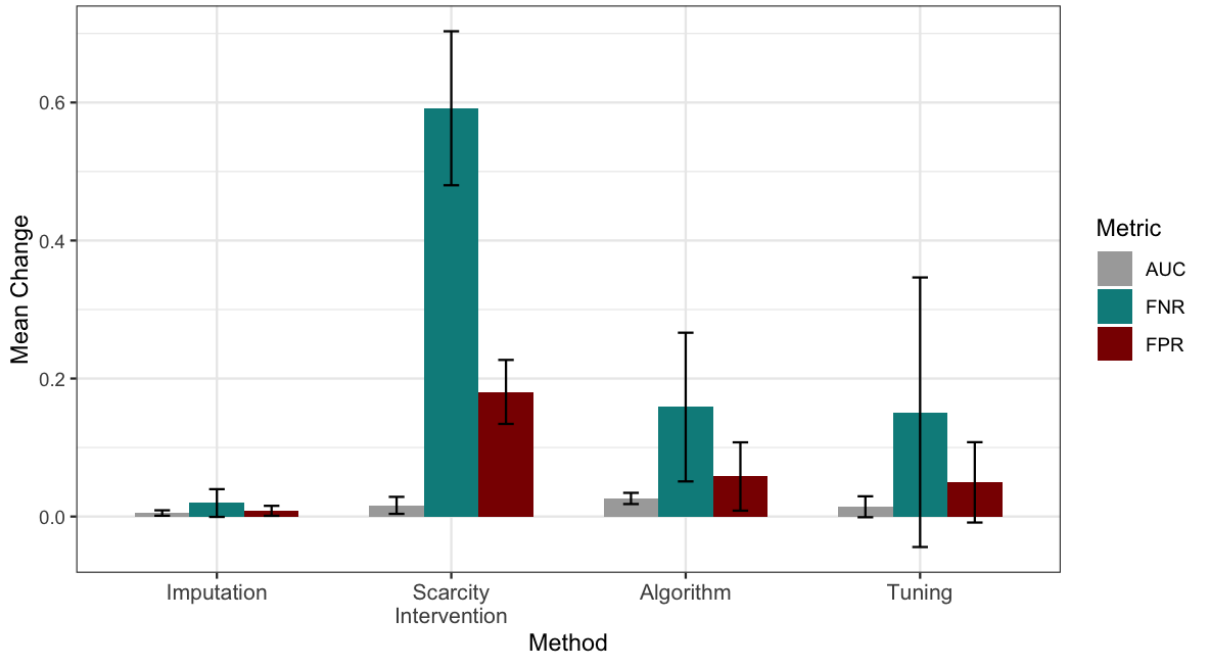


Figure 21 shows the main potential to change the FNR and FPR in the model's predictions, which lies within the scarcity intervention. This is consistent with the common challenge that arises from unbalanced datasets, as not addressing the imbalance can introduce a high FNR. Algorithm choice and tuning similarly impact the model's fairness, with hyperparameter tuning showing the highest variability as indicated by the standard deviations. However, imputation does not enforce change potential on the model's fairness as much as the other decision points do. The following sections will detail the change potential within each decision point, following the order of the standard ML classifier pipeline.

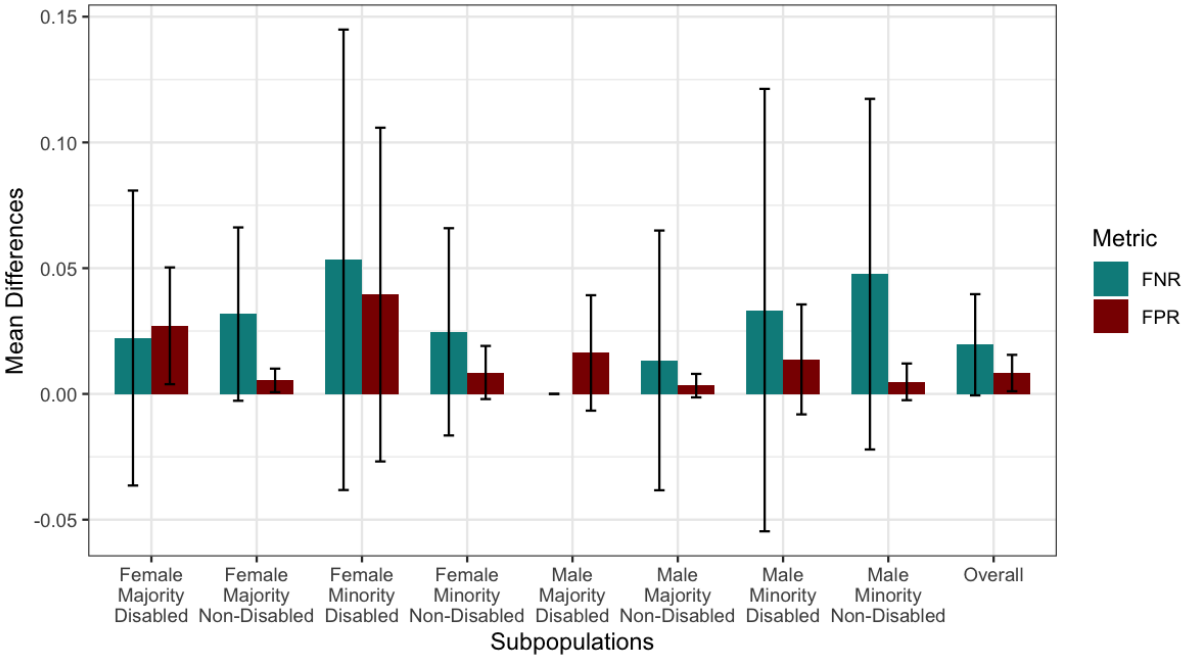
First decision point: data imputation

Data imputation is the first decision point within the pipeline that is assessed for its effect on the classifier's fairness outcome. The formerly introduced mean difference in fairness on the decision point and its standard deviation are separated between the eight subpopulations that are cross-sectional between the binary sensitive attributes of gender, ethnicity, and disability.

Figure 22 shows how different groups are affected by the imputation choice. Overall, the choice of imputation does not have a large effect on the fairness outcome of the overall model compared to the other decision points. However, in terms of change in FPR and FNR, the female

X minority X disabled subpopulation seems to experience the largest change. This shows that imputation affects the most disadvantaged group within the population in terms of mean values and standard deviation. Generally, the more vulnerable a subpopulation is, the higher the variability, as the error bars indicate. This might indicate that with data not missing at random, especially those subpopulations more prone to omitting their data are affected by the chosen imputation method. Similarly, cases of male minority groups show higher change in FNR depending on the data imputation, with minority classes having a higher probability of missing data.

Figure 22.
Mean difference in FNR and FPR for subpopulations created by data imputation



Conclusively, there are differences in how subpopulations are treated depending on the missing data imputation. With vulnerable groups more often omitting data, they seem to have more of an effect on changing fairness outcomes by the chosen imputation method. However, the mean differences are not significant compared to the differences created within the other decision points.

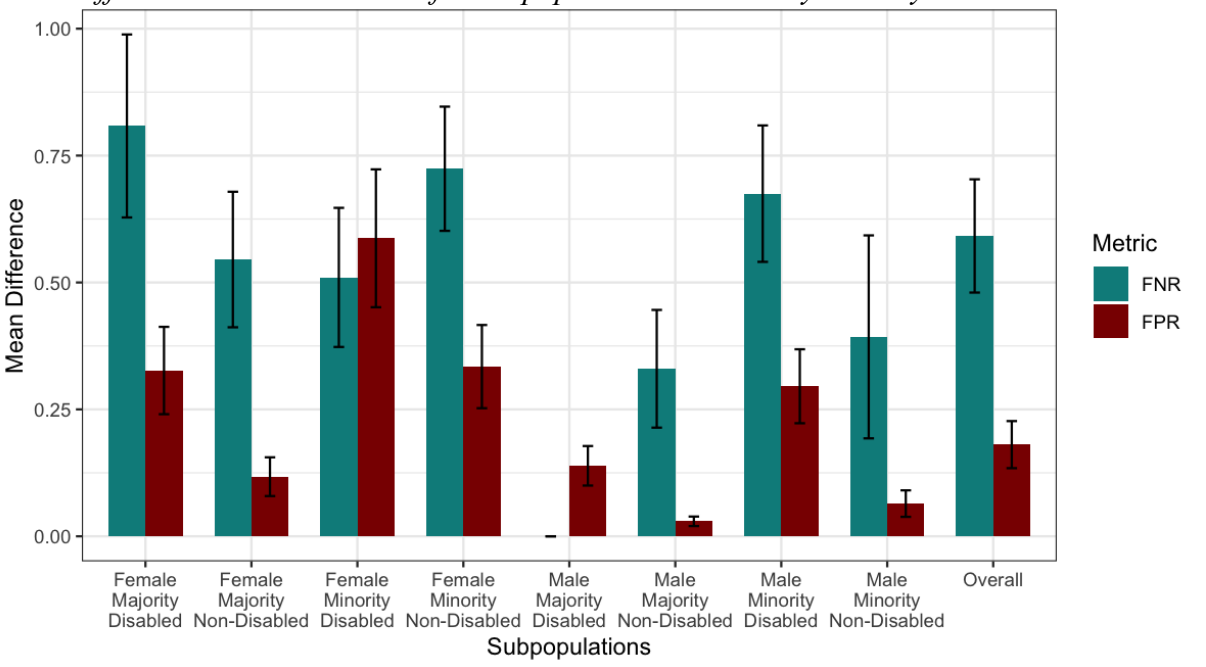
Second decision point: scarcity intervention

The second explored decision point is the scare outcome intervention, which already demonstrated the greatest fairness change potential on the initial overview, especially for FNR (see Figure 21).

Figure 23 shows this mean change separated by the eight subpopulations to show how different subpopulations might be affected differently by the scarcity intervention. It shows for the most disadvantaged group (female X minority X disabled) that the change in FPR and FNR is both relatively high, showing that this subpopulation can experience very different treatment depending on the chosen intervention with the vulnerability of having increased FPR more than any other subpopulation. Generally, the FNR changes the most, showing that scarcity intervention is vital to make the classifier more sensitive towards positive classifications. Overall, the impact of the FNR appears higher for the female subpopulations than for the males, potentially, as in the simulated society, the females commit more fraudulent cases compared to the males. Hence, it is also vital to consider the intervention's effect on the FPR, as it is higher for female individuals with an additional vulnerable attribute. In conclusion, the scarcity intervention affects individuals with a vulnerability level of two or higher (as indicated in Chapter 3, Table 2). In terms of variability, expressed in standard deviation, the performance remains similar across the subpopulations.

Figure 23.

Mean difference in FNR and FPR for subpopulations created by scarcity intervention



In conclusion, while the scarcity interventions might counter the increased FNR, they also show the potential to affect the FPR for the more vulnerable groups with two or more sensitive attributes.

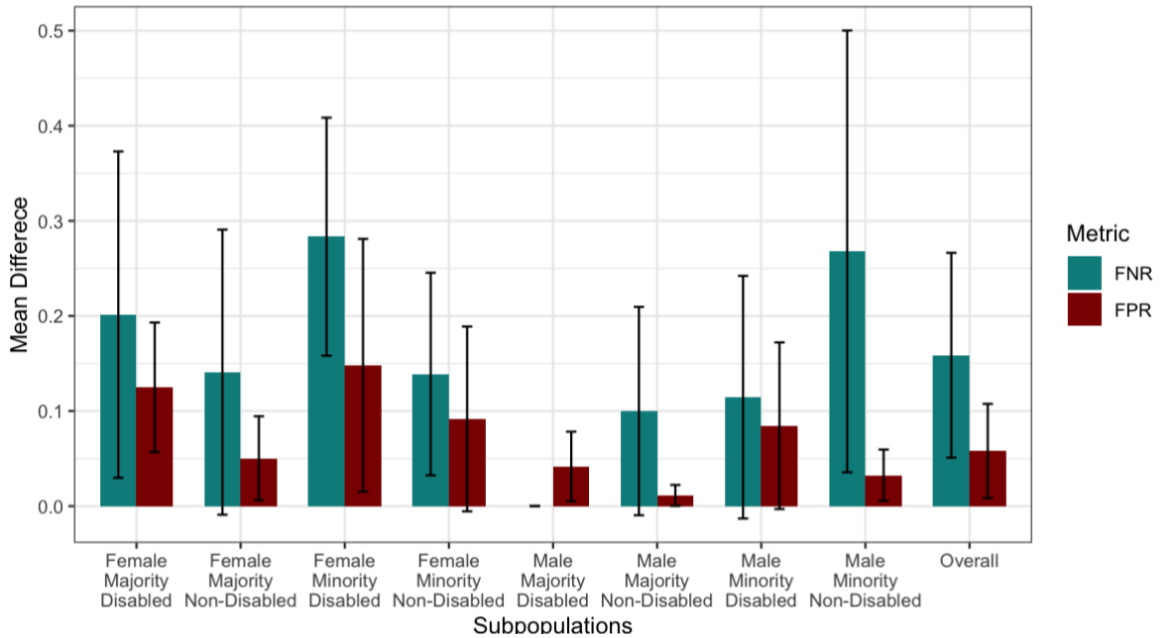
Third decision point: algorithm choice

The next decision point elaborated in greater detail is the algorithm choice, which is investigated on their fairness impact on different subpopulations. Figure 24 again demonstrates the overall change potential of the algorithm decision point separated into the eight populations. Overall, it shows that the most significant potential change in fairness outcome is within the FNR, which aligns with the common challenge of the scarcely distributed outcome. Subpopulations with a vulnerability level of two or higher show higher FNR and FPR, presenting that they will be the most impacted individuals by the decision for an algorithm. The most advantaged group (male X majority X non-disabled) showed the lowest FPR, showing that the choice of the algorithm showed the lowest possible risk of adverse treatment for them compared to the other groups.

Similarly, as to the previous decision point, the most vulnerable group (female X minority X disabled) shows the highest potential impact of algorithm choice in terms of risk of adverse treatment. The variability in mean outcome remains similar between most of the subpopulations. However, the group of male X minority X non-disability shows a higher standard deviation, indicating that their FNR results vary the most depending on the algorithm choice.

Figure 24.

Mean difference in FNR and FPR for subpopulations created by algorithm choice



In conclusion, the choice of algorithm can impact the FPR more for more vulnerable individuals than for less vulnerable individuals. Furthermore, the greater change in FNR reflects how different algorithms might be more successful in dealing with the scarce outcome distribution, especially affecting the female subpopulations.

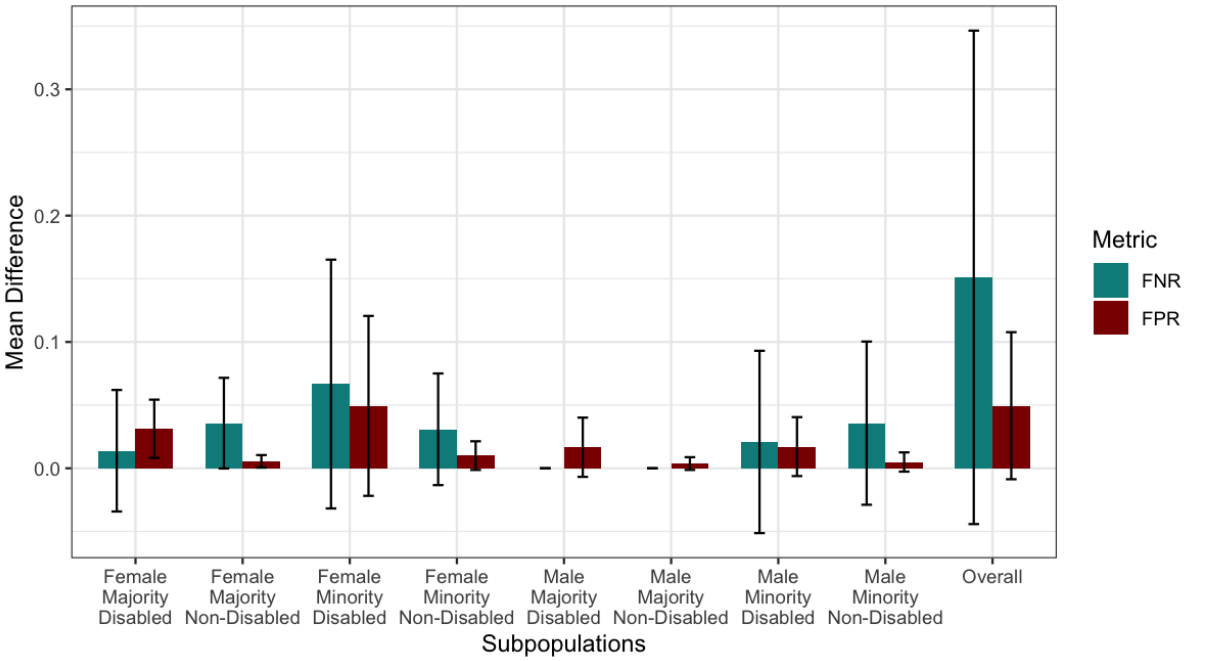
Fourth decision point: hyperparameter tuning

For the investigation of hyperparameter tuning, the models with hyperparameters to be tuned are considered, excluding logistic regression, as logistic regression has no typically tunable hyperparameters. Overall, the tuning shows some impact on the fairness measures, which exceeds the impact of missing data imputation. However, it does remain low compared to the impact of scarcity intervention. Figure 25 shows that the impact mainly resides in the FNR, which reflects that the tuning is affected by the challenge of the scarce outcome distribution.

Overall, the female subpopulations show a greater change in FNR by tuning than the male subpopulations. Again, the greatest change is experienced by the most disadvantaged subpopulation (female X minority X disabled). Generally, Figure 25 shows the additive effect of vulnerable attributes; the more vulnerable attributes an individual carries, the more affected it is by the tuning of the model. In conclusion, subgroups with a vulnerability level of at least two show greater change potential through the model-tuning process.

Figure 25.

Mean difference in FNR and FPR for subpopulations created by hyperparameter tuning



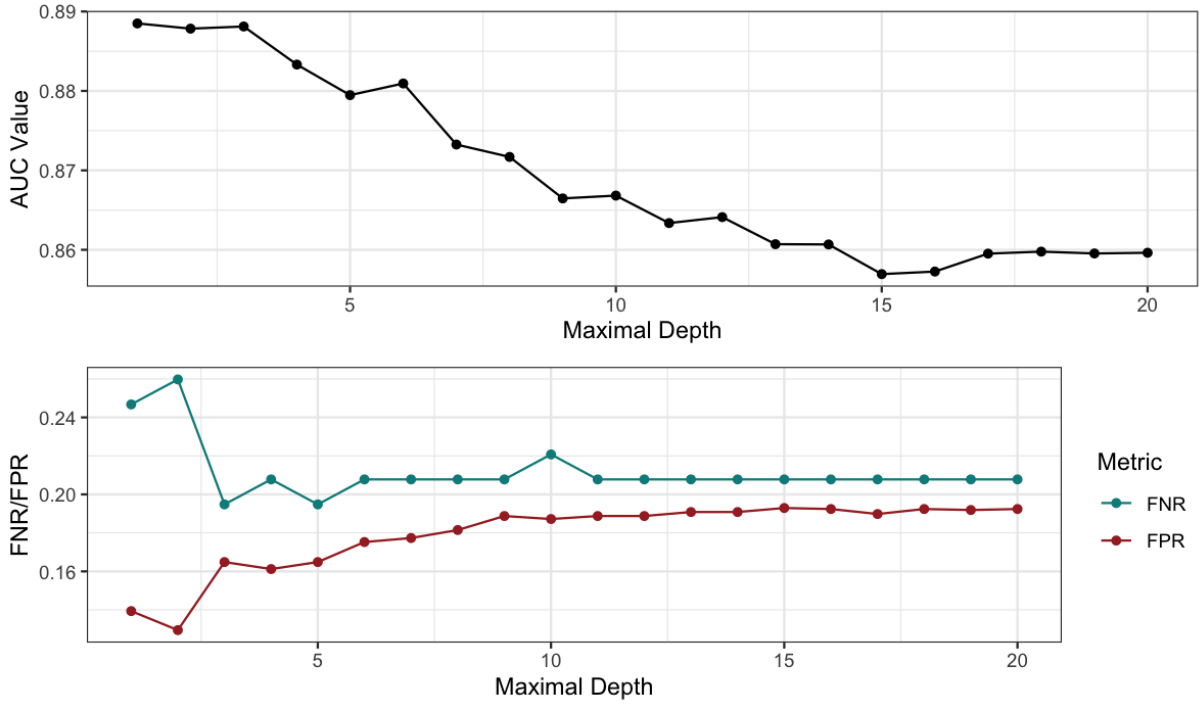
Since two different models are applied that have different hyperparameters to be optimized, the optimization impact is probed for one hyperparameter for each tunable algorithm, XGBoost

and Random Forest. This aims to give some insight into the potential fairness differences elicited by hyperparameter tuning, acknowledging a diverse range of hyperparameters for different ML algorithms.

Beginning with XGBoost, the maximal depth the individual trees are set to grow to is optimized while keeping the other hyperparameter values (such as gamma, eta, and subsample) in the default setting. Figure 26 shows that the optimal point for the AUC value is set at a low maximal depth. However, the FNR and FPR experience greater changes within the range of 1 to 6 in maximal depth and stabilize after. It shows that the set of the maximal depth parameter can influence the overall FNR and FPR balance. For instance, while the AUC value is comparable between the maximum depth of one and three, the outcome in terms of fairness is different. At the maximum depth of one, the fairness outcome would show a higher FNR and lower FPR, while for the maximum depth of three, it shows a higher FPR and a lower FNR, with both metrics approximating each other. Depending on the context, a higher FNR and lower FPR might be fairer for the model outcome; in other contexts, an approached balance between FNR and FPR might be desirable. In this case, the engineer would have two choices for the hyperparameter value that barely affects the performance in terms of AUC; however, it can show different overall fairness outcomes for the model.

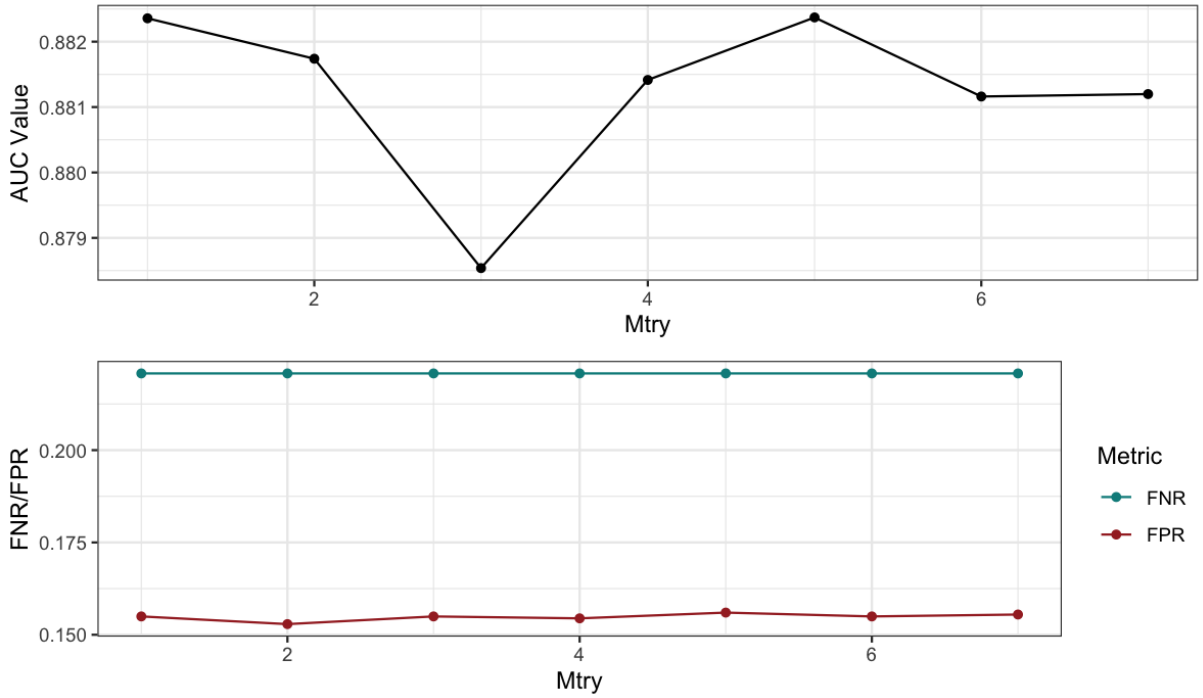
Figure 26

Isolated tuning of XGBoost's maximal depth and its impact on AUC, FNR and FPR



Secondly, the Random Forest is tuned, and to probe into its tuning effect, the hyperparameter of the maximum number of features chosen for each split (mtry) is tested, showing how the number of selected features in each tree can affect the performance outcome. Figure 27 shows this tuning process in isolation while setting the other hyperparameters (such as number of trees, criterion and minimum number of samples) at their default values. It shows two optimal maximum number of features in terms of AUC performance. Interestingly, in contrast to XGBoost, the overall change in FNR and FPR remains absent regardless of the chosen hyperparameter value. Hence, the decision for the hyperparameter value does not influence the fairness outcome of the model.

Figure 27.
Isolated tuning of Random Forest's mtry and its impact on AUC, FNR and FPR



Overall, this shows that the hyperparameter tuning of mtry for Random Forest is value-free regarding fairness outcome for the current data situation. However, the tuning of XGBoost shows that the choice of hyperparameter value can change the overall fairness outcome while not affecting the overall AUC performance of the model. This could indicate a diverse effect of hyperparameter tuning, with the tuning process being different for different ML models; the effect might differ depending on the data situation and the chosen model.

Fifth decision point: threshold setting

Lastly, the threshold setting is assessed. While the previous decision points were all evaluated on the default threshold (cut-off at 0.5), the threshold might be adjusted depending on the overall goal of the classifier and domain-specific considerations. For this, a more stringent (cut-off at 0.9 quantile of fraud probability) and more lenient (cut-off at 0.1 quantile of fraud probability) threshold compared to the default threshold is introduced to see how different groups might benefit or be harmed differently by them.

First, the impact on the FNR, depending on the threshold, is considered in Figure 28. It shows a high FNR for the more advantaged male subpopulations using the lenient threshold (above 0.9 quantiles). The more vulnerable attributes the defined subgroup has (level two or higher), the lower their FNR under the lenient threshold compared to the default threshold. More advantaged groups derive the more considerable benefit of a more lenient threshold, showing that rising the threshold does not eliminate potential bias towards disadvantaged subpopulations. Overall, the stringed threshold (above 0.1 quantile) lowers the FNR to comparable levels across subpopulations. However, the most advantaged subpopulation shows an overall higher FNR even under the stringent threshold, indicating that they benefit from this threshold setting. The variability in the outcome is similar across the subpopulations, apart from the male X majority X disabled group, where the FNR remains consistently at 1.0, and hence the standard deviation remains low.

Figure 28

Effect of threshold setting on FNR for different subpopulations

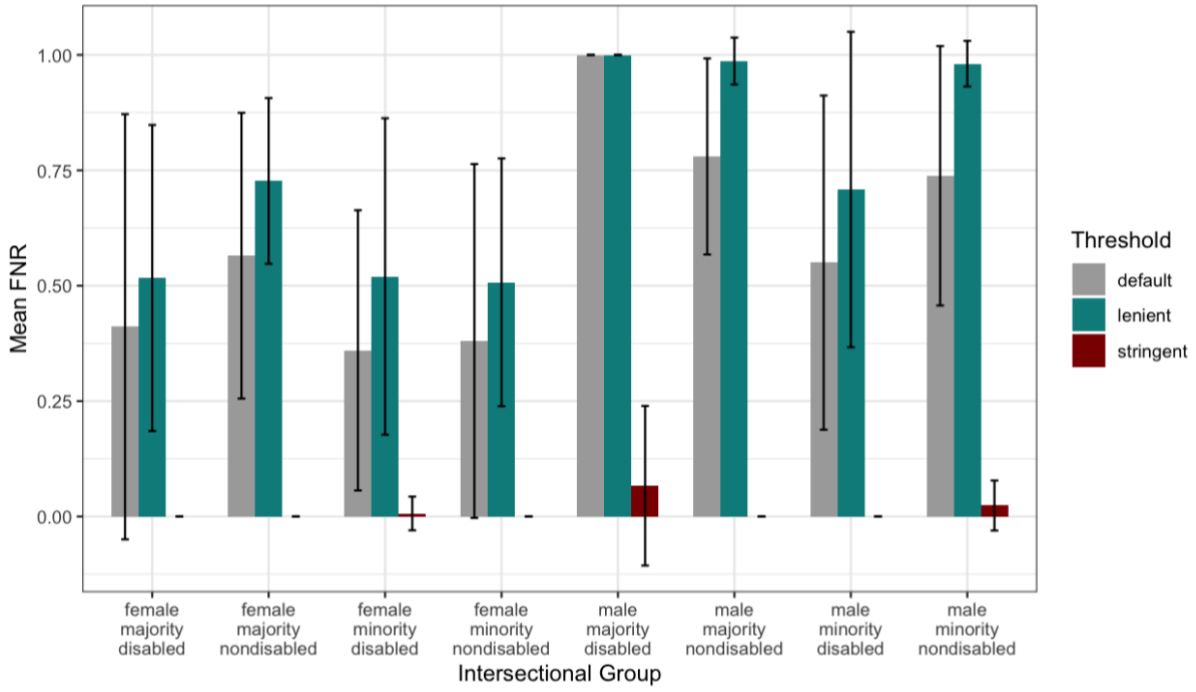
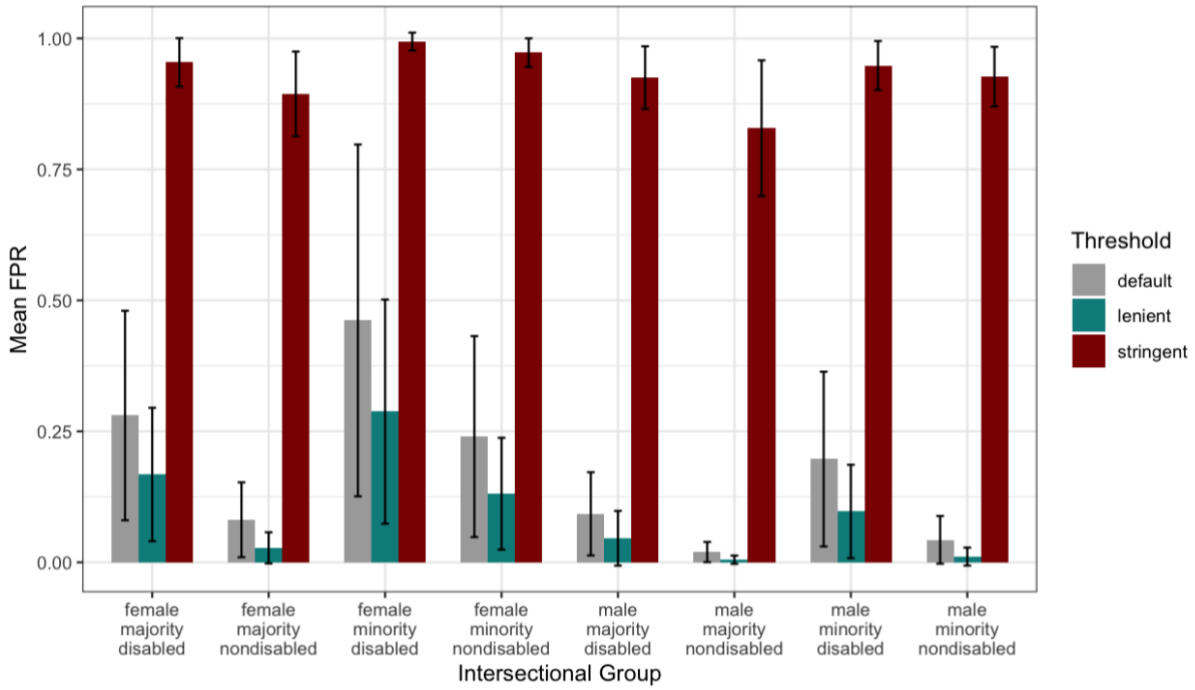


Figure 29 shows the impact of the threshold setting on the FPR. Overall, the lenient threshold lowers the FPR. However, this decrease is more substantial for more advantaged groups (with a vulnerability level of one or lower). Additionally, the variability in outcome is higher for more vulnerable groups, as the error bars indicate. The stringent threshold raises the FPR between all subpopulations, and there seems to be no pattern related to how different subpopulations might be affected differently in terms of mean outcome and standard deviation.

Figure 29.
Effect of threshold setting on FPR for different subpopulations



In conclusion, the overall change of FPR and FNR introduced by the stringent threshold does not discriminate between subpopulations. However, the lenient threshold shows a higher impact for advantaged groups, giving them a higher FNR and a lower FPR than the more vulnerable groups. It shows an additive effect of several vulnerable attributes defining the subpopulations. Hence, raising the threshold mainly benefits the advantaged subpopulations while introducing less benefit to the more vulnerable groups.

Summary

Overall, the decision point of scarcity intervention shows the highest impact on the model's fairness outcome, especially in terms of FNR. Scarcity interventions counteract the increased FNR that the scarce outcome distribution illicit. However, they also have the potential to change the FPR for vulnerable groups, introducing bias towards vulnerable individuals while benefiting advantaged individuals.

The missing data imputation demonstrated the lowest impact on fairness change among all decision points. However, it also demonstrated that vulnerable groups can be differently impacted by the choice of imputation method, while advantaged groups show more stable performance across the choice of method.

The choice of algorithm significantly impacts the FNR, showing how different algorithms better respond towards the linearity of the data and the scarce distribution of the outcome. Furthermore, the choice of algorithm again shows more change potential in fairness measures for the vulnerable subpopulations compared to the advantaged subpopulations.

Hyperparameter tuning exhibits a more noteworthy change in vulnerable groups. Isolated observations of single hyperparameters showed that XGBoost's maximal depth hyperparameter could give varying results in terms of FPR and FNR. In contrast, for Random Forest's maximal number of features hyperparameter, the fairness outcome remains stable regardless of the chosen value.

Lastly, the threshold setting shows that the stringent threshold shows similar fairness outcomes across all subpopulations, while the lenient threshold mostly benefits more advantaged individuals and approaches similar results as the default threshold for the vulnerable subpopulations.

Table 8.

Summary of insights from data with scarce outcome and linear decision boundary

Decision Point	Impact on Fairness (FNR, FPR)	Subpopulation Variability and Key Observations
Missing Data Imputation	Minimal overall fairness impact; generally low on FNR and FPR	Most disadvantaged groups experience highest change, likely due to higher likelihood of data omission
Scare Outcome Intervention	Highest overall impact, without intervention, classifier shows high FNR (majority class bias)	Vulnerable groups face increased FPR
Algorithm Choice	High impact on FNR, logistic regression performs well under linear separability	Vulnerable groups show greater FNR and FPR and variability
Hyperparameter Tuning	Primarily affects FNR	Most vulnerable groups experience greatest change, XGBoost tuning shows effect, while Random Forest tuning is value-free
Threshold Setting	Lenient threshold benefits advantaged groups with higher FNR, stringent threshold gives more consistency across groups	Vulnerable groups benefit less from lenient threshold. Stringent threshold reduces FNR disparities

Third data situation – Non-linear decision boundary with non-scarce outcome

This data situation is defined by a non-linear decision boundary and a non-scarcely distributed outcome variable. Table 9 shows the characteristics of the currently inspected data among the four different datasets. Overall, the non-linear relationships between the fraud probability and the features show a non-linear decision boundary between the features and the fraud classification. Those non-linear relationships are visible in Figure 30. The outcome of the fraud classification is non-scarce, making up 25% of the overall cases, which poses less of a challenge to the classifier pipeline to fail to identify fraud classifications. This reduces the challenge to the non-linearity of the decision boundary with more complex relationships between fraud outcomes and the features that might not be easily identified by engineering choices that assume linearity.

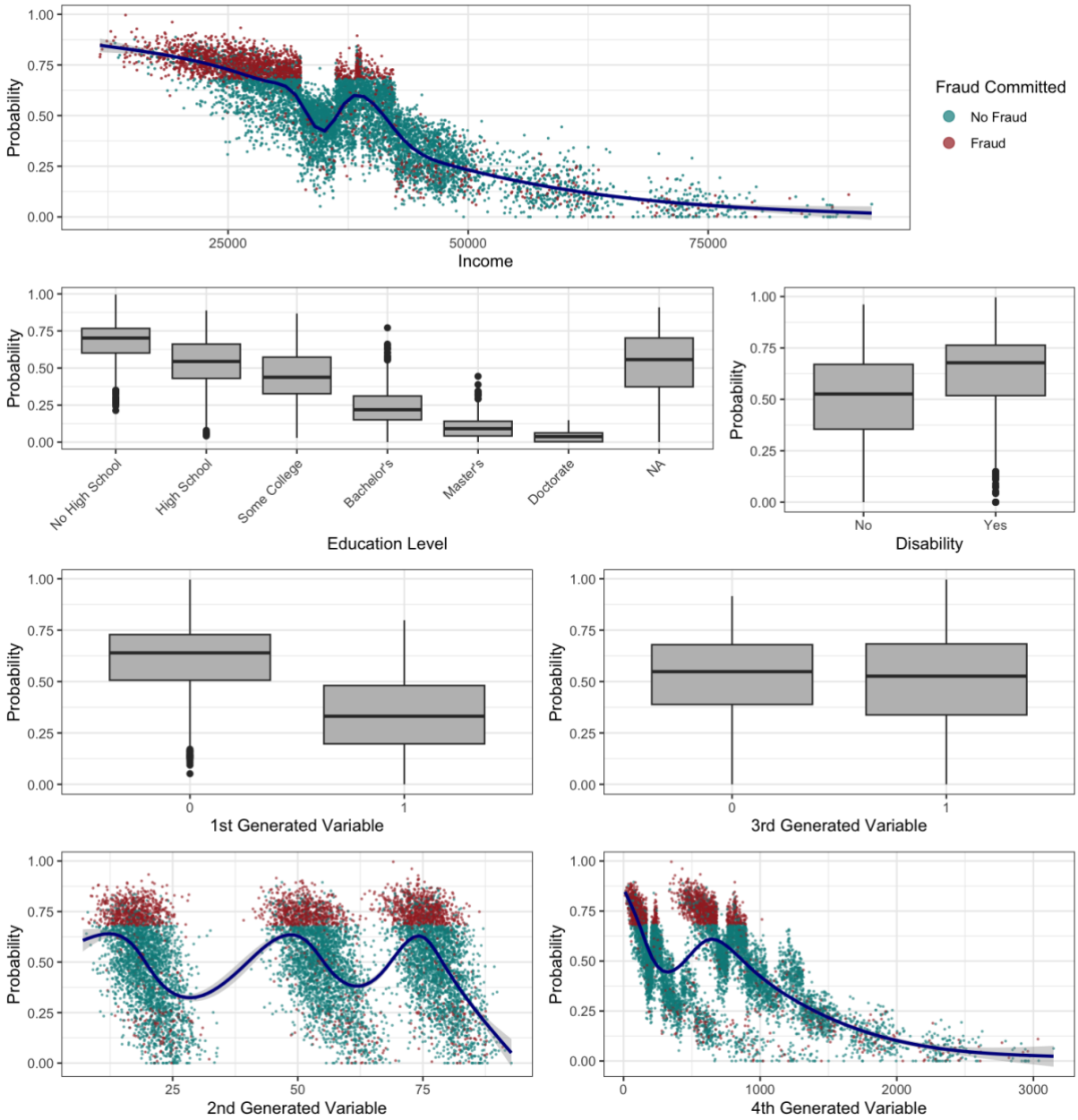
Table 9.
Overview of the currently tested data characteristics

	Decision Boundary	
Distribution of outcome	Linear, Non-scarce outcome	Non-linear, Non-scarce outcome
	Linear, Scarce outcome	Non-linear, Scarce outcome

A closer look at Figure 30 shows the non-linear relationship between the fraud outcome and the features. The income variable shows a higher risk for fraud commitment in lower-class households, with a higher middle-class increase in fraud rate and a dip for the lower middle class. Higher levels of education generally decrease the probability of fraud; however, the differences between the levels of education are non-linear. Disability positively contributes to fraud probability. Furthermore, the first and fourth generated variables negatively affect the outcome of fraud. However, the second and third variables show no straightforward relationship to the classification outcome. Figure 30 shows how a linear decision boundary cannot be assumed for this dataset, which could lead to poorer performance on models that assume this underlying linearity.

Figure 30

Relationship of fraud probability and classification by features

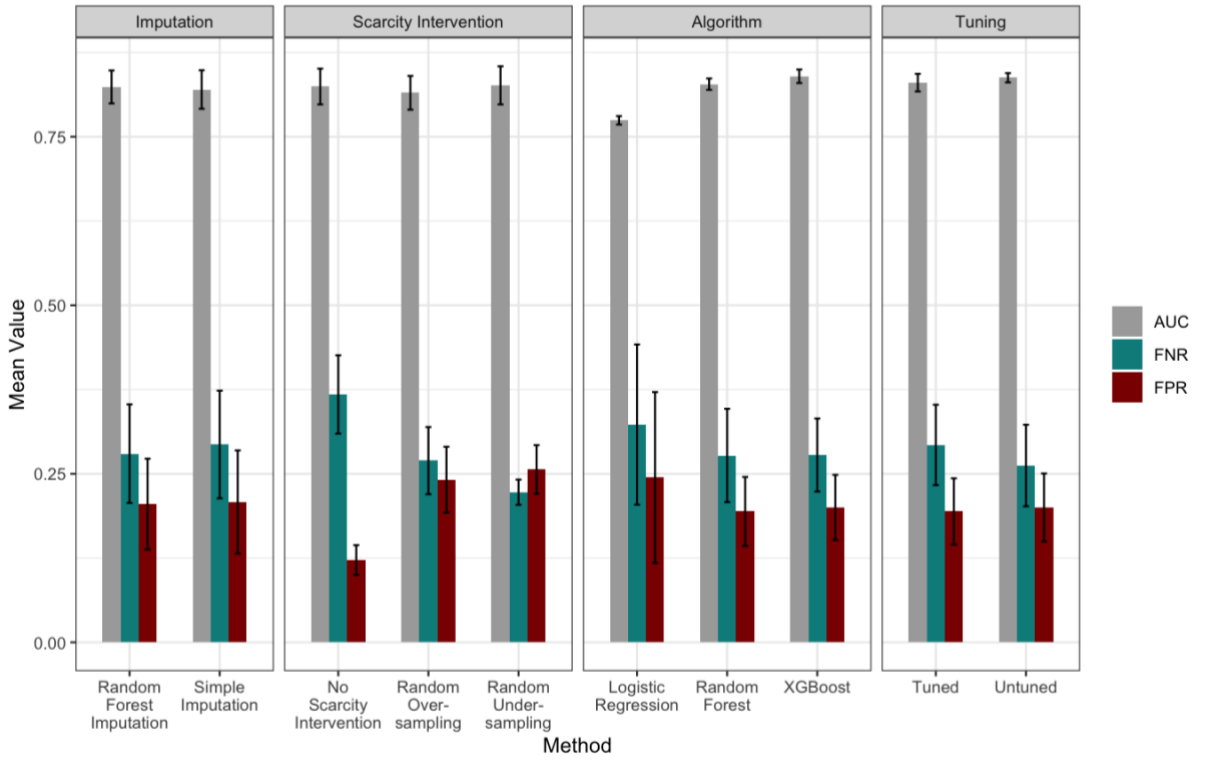


Overview of decision point impact

Before investigating the mean change in each decision point, Figure 31 presents the mean (bars) performance and fairness outcome and their standard deviations (error bars) for each tested engineering choice. It shows that both missing data imputation methods show similar outcomes in terms of performance and fairness outcomes. The scarcity intervention shows less effect than the third data situation with scarce outcomes. However, not applying any scarcity intervention method still increases the FNR compared to utilized interventions. The algorithm choice shows that the logistic regression underperforms compared to the tree-based methods in terms of AUC,

which shows that this algorithm could be more optimal for coping with the non-linearity of the data. However, between algorithm choices, the fairness outcome remains relatively consistent. The fairness outcome remains consistent for tuned and untuned model configurations. Overall, the standard deviations of the performances remain similar across all contingencies.

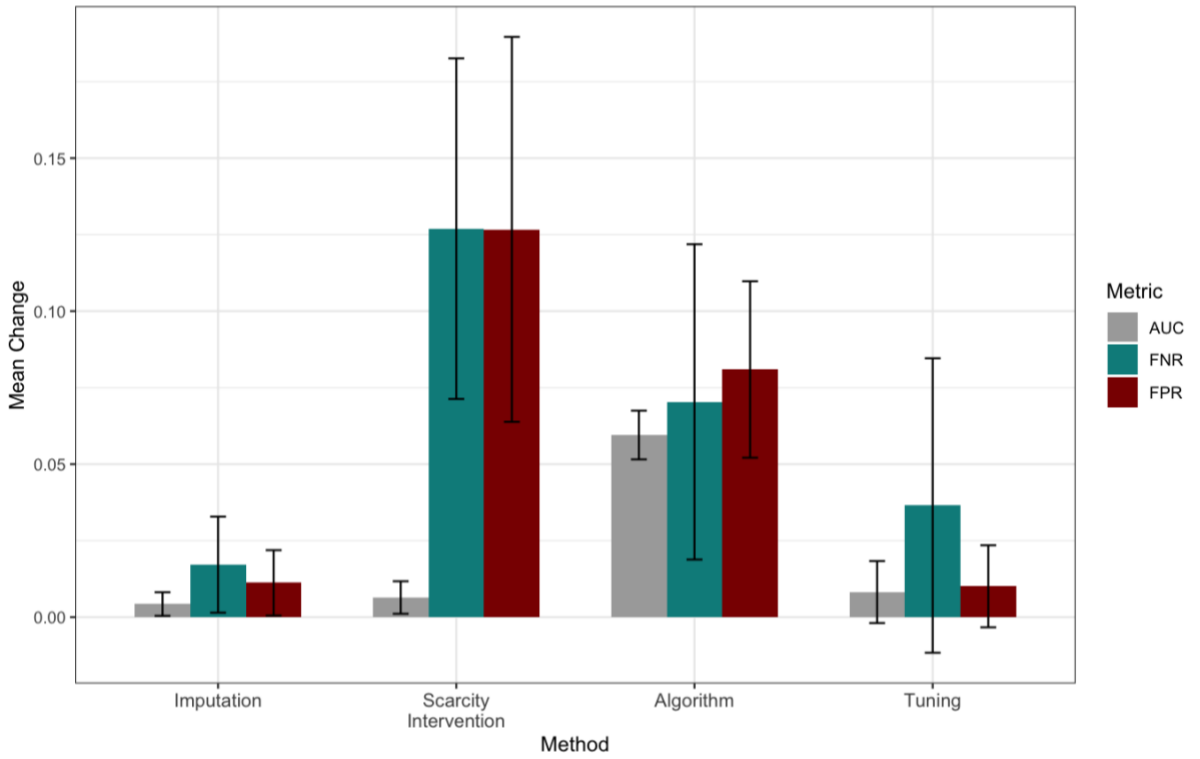
Figure 31.
Mean performance for each investigate engineering choice within the pipeline



Moving between different pipeline setups considering the previous choices, there is a mean difference in the fairness outcome between them while keeping the other engineering choices constant. Those mean differences (bars) and their standard deviations (error bars) are illustrated in Figure 32, showing that scarcity intervention and algorithm choice show the most considerable change in fairness outcome regarding mean change and their standard deviation. However, compared to the previous data situations, the overall potential to change fairness outcomes within the decision points is lower. Additionally, the algorithm decision point shows the most remarkable change in AUC value, reflecting that the tree-based methods outperform the logistic regression by identifying more complex relationships between the features and the fraud outcome.

Figure 32.

Mean differences in performance and fairness for different decision points



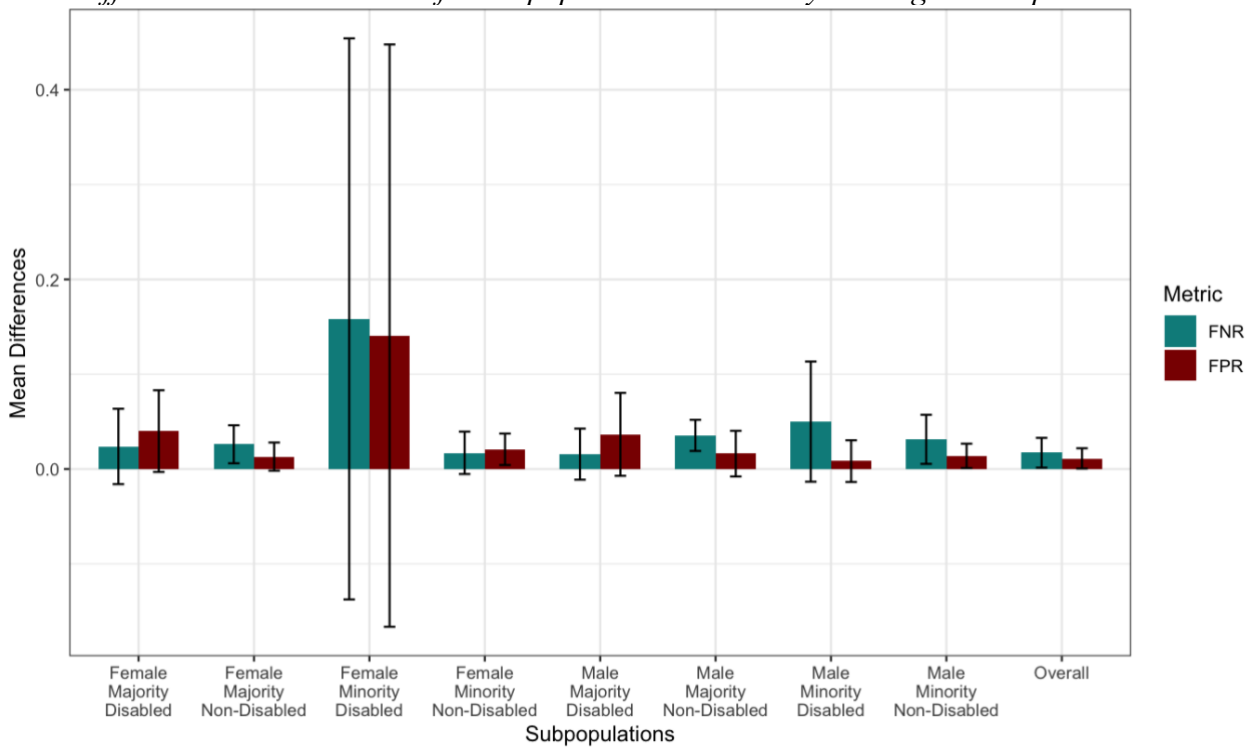
Those mean changes will be further explored by investigating each decision point individually in the order of that standard ML classifier pipeline.

First decision point: missing data imputation

The first decision point is the missing data imputation, which has the lowest impact on the fairness outcome out of all the decision points. Figure 33 shows that most of the subpopulations' mean differences in FNR and FPR are relatively consistent. However, the most disadvantaged group (female X minority X disabled) demonstrate a higher mean change in both fairness outcomes, showing that the imputation method similarly impacts the subgroups apart from the most disadvantaged one. Furthermore, their standard deviation is also the highest amongst the group, confirming that their change potential elicited by the imputation method is the highest.

Figure 33.

Mean difference in FNR and FPR for subpopulations created by missing data imputation

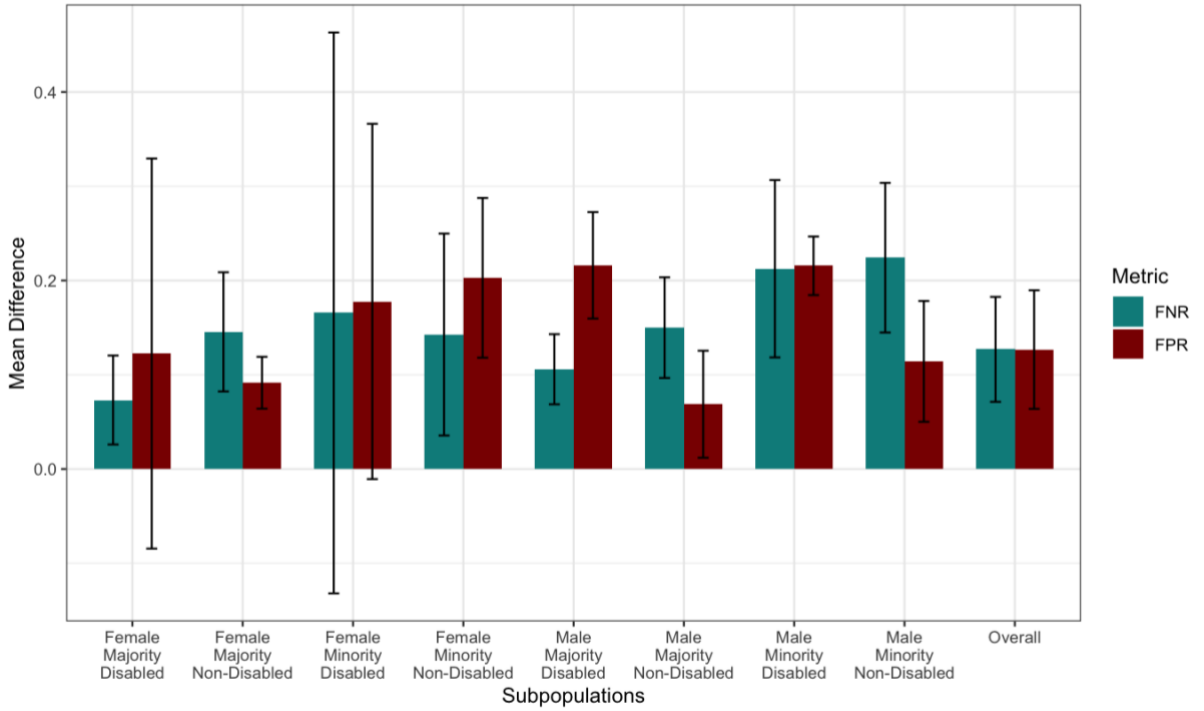


Second decision point: scarcity intervention

The scarcity intervention demonstrated the highest change potential among the decision points. Figure 34 shows that the difference between scarcity interventions is higher for individuals belonging to subpopulations with at least one vulnerable attribute. The most disadvantaged group especially experiences the greatest variability in change, while the most advantaged group shows lower levels of mean change. This shows that the scarcity intervention mainly affects the more vulnerable subgroups.

Figure 34

Mean difference in FNR and FPR for subpopulations created by scarcity intervention

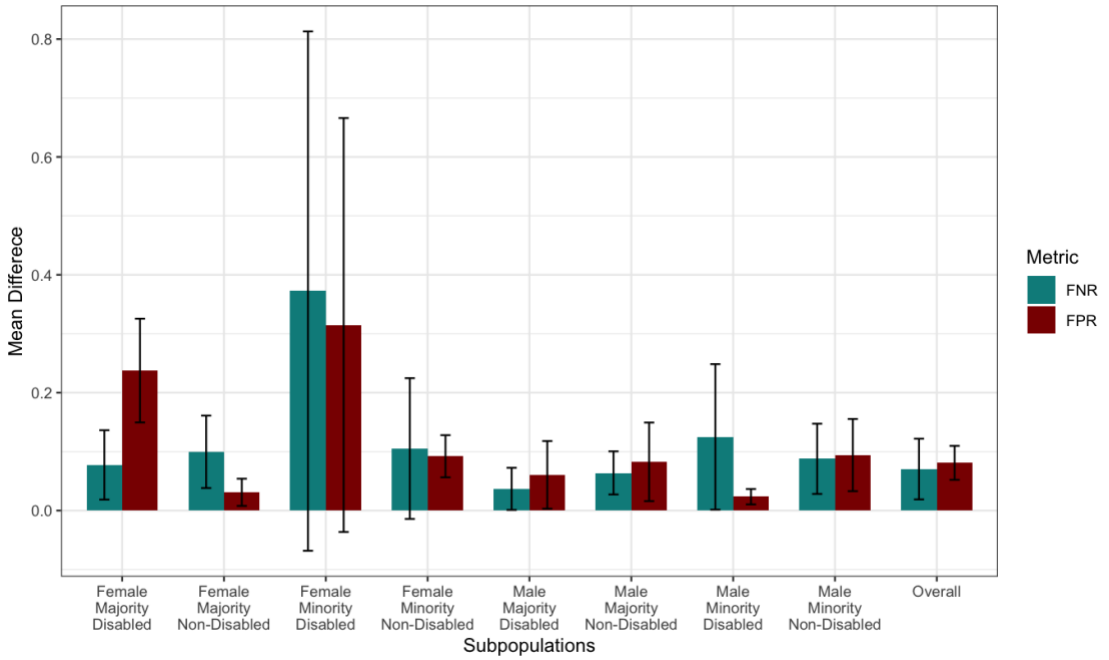


Third decision point: algorithm choice

The mean difference in fairness outcome by choice of algorithm is shown in Figure 35, showing that the choice of algorithm impacts the resulting fairness similarly between subgroups. However, the most disadvantaged subpopulation (female X minority X disabled) experiences the greatest change, differentiating itself compared to the impact on the other groups. Depending on the chosen algorithm, individuals from this subpopulation are more likely to experience varying fairness outcomes for both FNR and FPR. Additionally, the subpopulation composed of female X majority X disability shows higher changes in FPR for algorithm choice, showing that this group can potentially have an adverse fairness treatment depending on the chosen algorithm. In conclusion, the algorithm choice mainly affects female subpopulations with disability.

Figure 35

Mean difference in FNR and FPR for subpopulations created by algorithm choice

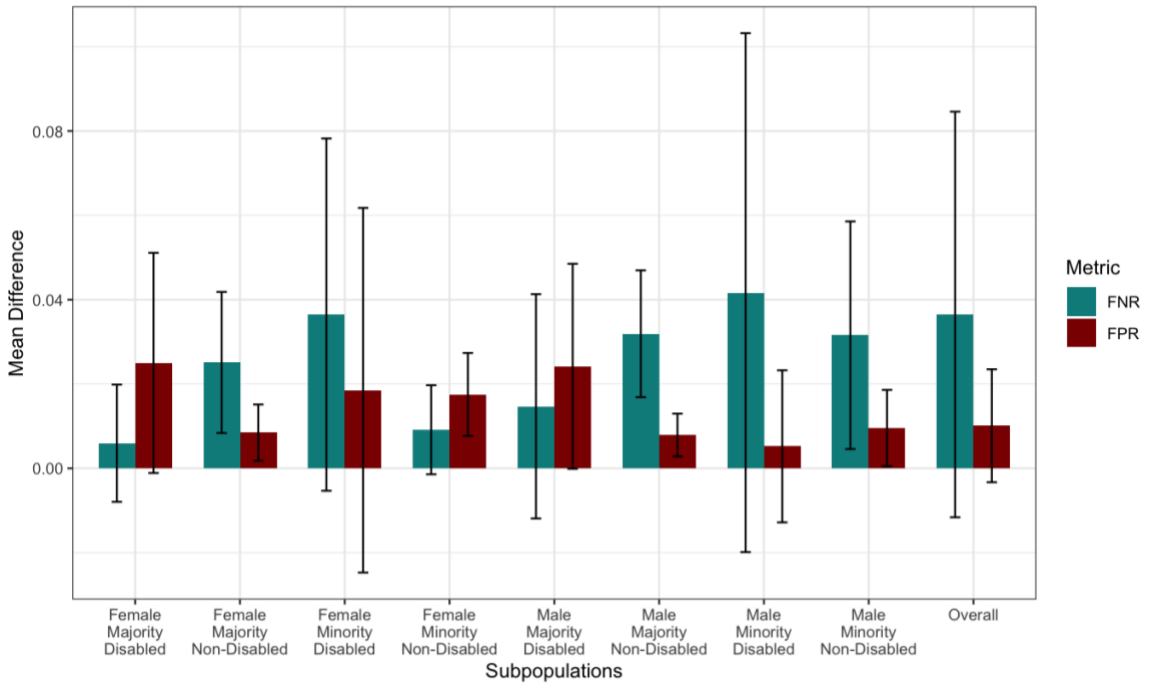


Fourth decision point: hyperparameter tuning

Hyperparameter tuning has a greater impact on the FNR than on the FPR. Figure 36 shows that the impact is comparably similar between the subpopulations. The subpopulations defined by disability and minority ethnicity show higher values and variation in the FNR; however, apart from this instance, the fairness outcome depending on tuning does not largely differentiate between the subpopulations.

Figure 36.

Mean difference in FNR and FPR for subpopulations created by hyperparameter tuning

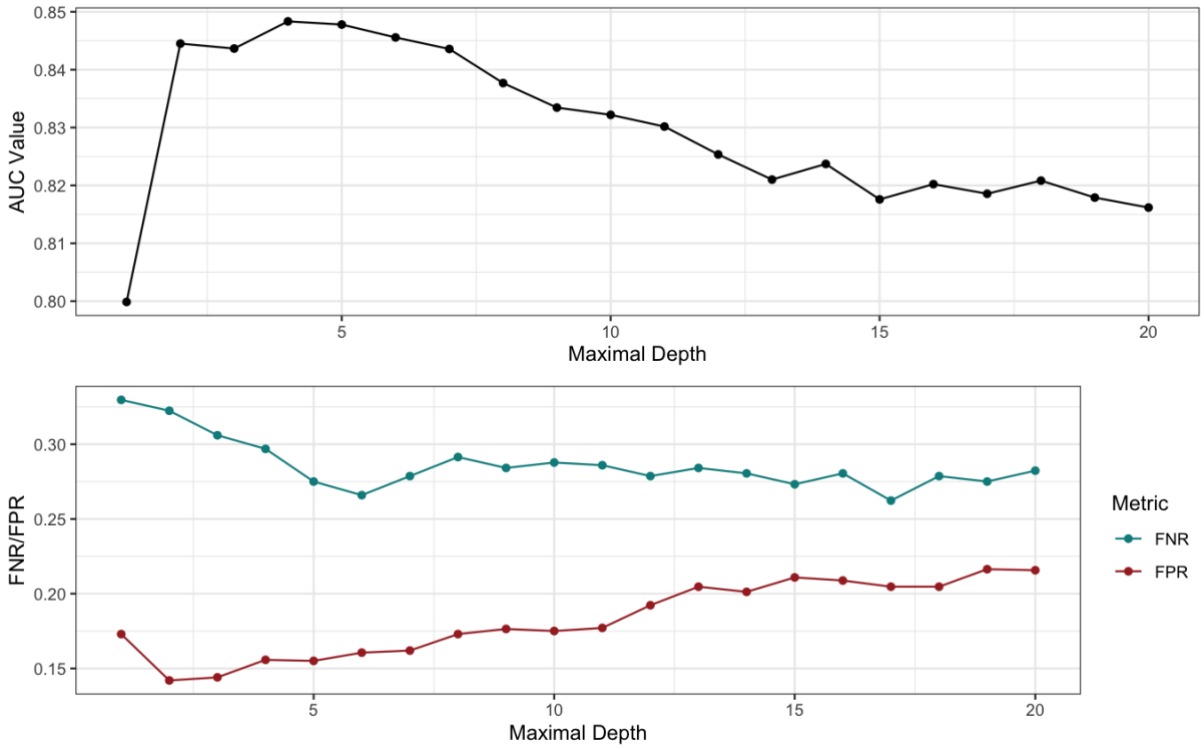


Probing into the tuning process more in detail, the XGBoost hyperparameter maximal depth and the Random Forest hyperparameter maximal number of features in each split are investigated more in detail while keeping the other hyperparameter values at their default.

Figure 37 shows XGBoost's maximal depth tuning, which peaks at four in terms of AUC value. However, if the engineer is interested in lowering the FPR, some difference could be achieved by lowering this hyperparameter setting from four to two at the cost of some decrease in AUC. Similarly, the FNR could be lowered by setting the maximal depth to six. Generally, the AUC performance between maximal depth set to four or five is comparable. However, it could decrease the overall FNR of the classifier. The engineer and the domain-specific responsibility need to consider these choices to achieve a lower FNR in the classifier.

Figure 37

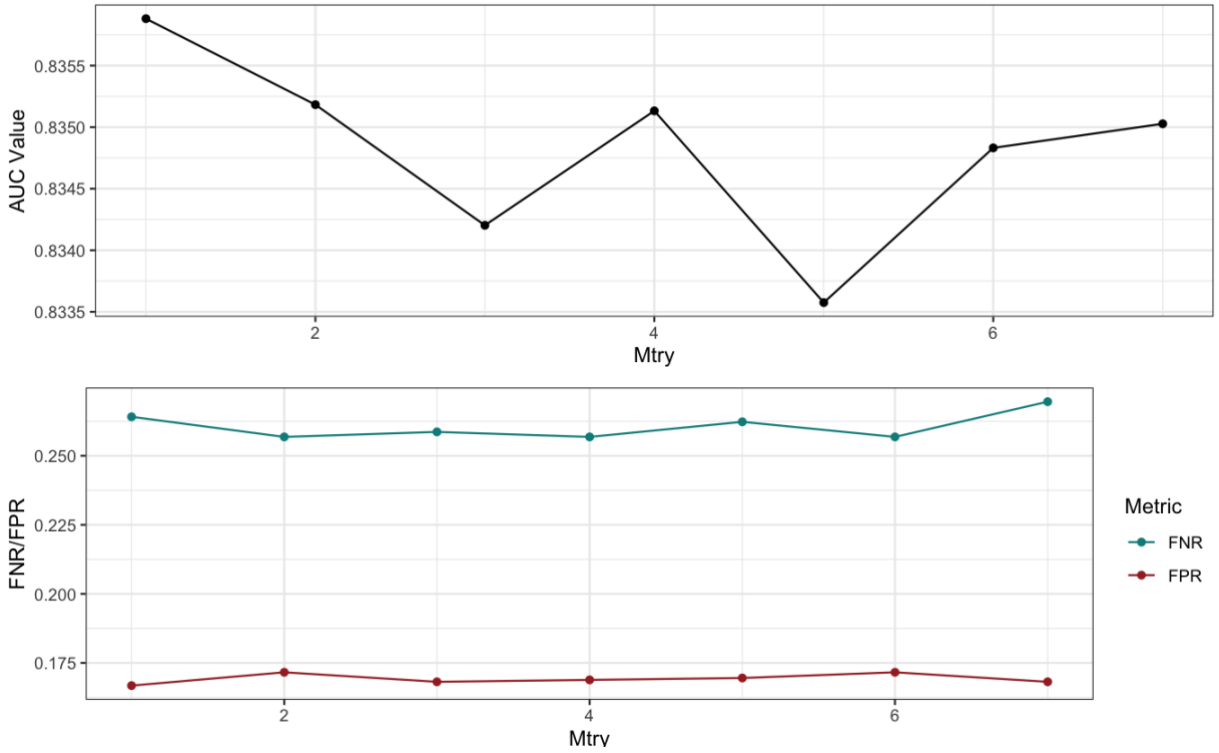
Isolated tuning of XGBoost's maximal depth and its impact on AUC, FNR and FPR



Similarly, the number of features selected in each node (mtry) is tuned for the random forest with results presented in Figure 38. Overall, the hyperparameter value only shows minimal change in the AUC value, and the fairness measures behave in a constant manner regardless of the hyperparameter setting. Hence, in this instance, the hyperparameter tuning of the number of selected features does not hugely impact the model performance, nor does the overall fairness.

Figure 38

Isolated tuning of Random Forest's mtry and its impact on AUC, FNR and FPR



Fifth decision point: threshold setting

Lastly, the threshold setting is considered, deviating from the previously only considered default threshold (cut-off at 0.5). Figure 39 shows the different impacts of threshold values on the FNR. Overall, the lenient threshold (cut-off at 0.9 quantiles) increases the FNR; however, the increase is higher for those groups that have no more than one sensitive attribute. The stringent threshold (cut-off at 0.1 quantiles) lowers the FNR and shows only a comparably increased FNR for the most advantaged group.

Figure 39.

Mean FNR by subpopulation and threshold setting

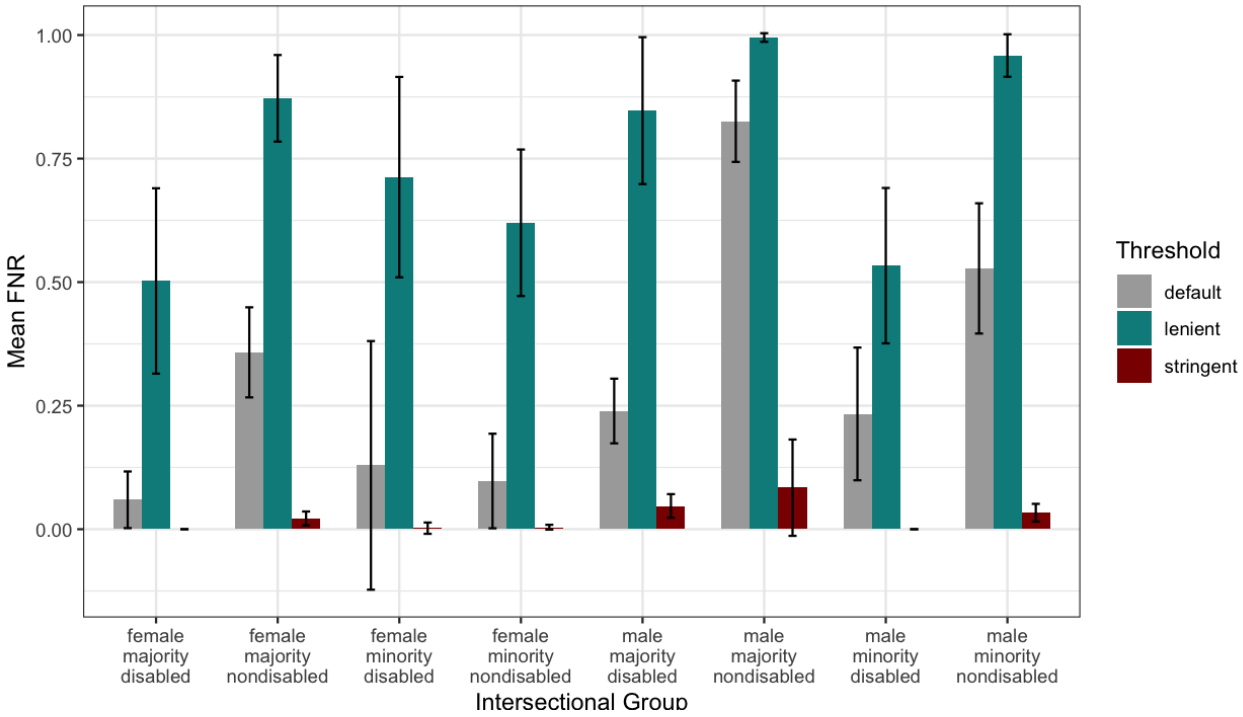
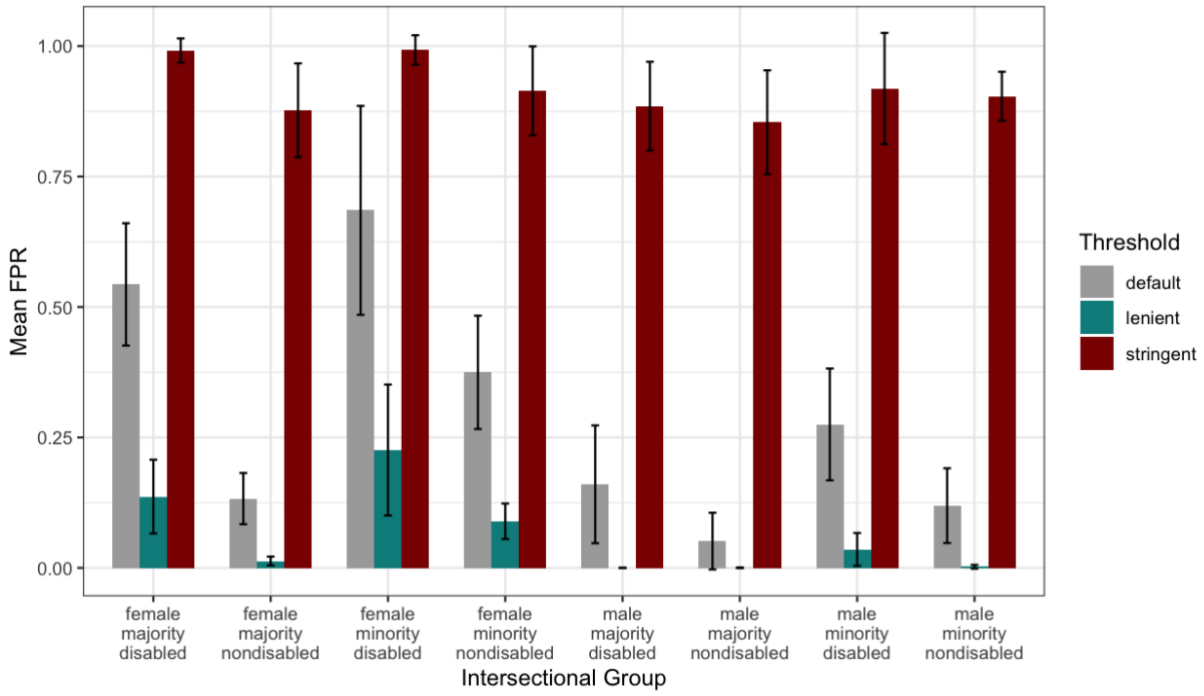


Figure 40 shows how the threshold setting influences the FPR across groups. The stringent threshold consistently increases the FPR for the subpopulations. The lenient threshold almost diminishes the FPR for the male subpopulations. However, the female subpopulations still have a higher FPR, even though it is reduced compared to the default threshold.

Figure 40.
Mean FPR by subpopulation and threshold setting



Summary

Overall, the non-linear decision boundary creates the greatest challenge in this data situation. However, this data set's characteristics are less challenging than those of the last data situation will be, as the outcome could be more accurately distributed.

Beginning with the missing data imputation, the overall change potential in this decision point was the lowest. Its impact was constantly low, except for the most disadvantaged subpopulation, which experienced increased change potential in both fairness measures by the imputation method.

The scarcity intervention had the largest impact of all decision points; however, it was lower compared to the scarce data situations. Subgroups defined by two or more sensitive attributes had greater changes and variability in the changes of the fairness measures, showing that scarcity intervention has the largest impact on disadvantaged groups.

The algorithm choice showed consistent changes in fairness across the subpopulations. However, the most disadvantaged subpopulation was disproportionally impacted on FNR and FPR by the choice of algorithm compared to the other subpopulations, showing that the combination of multiple vulnerable attributes might increase the risk of different treatments depending on the choice of algorithm.

The tuning process only minimally impacted the overall fairness outcome, with a somewhat higher impact on subpopulations defined by disability and minority ethnicity. The tuning of the XGBoost hyperparameter showed some options to achieve some adjustments to the fairness outcome without decreasing the AUC value. However, the Random Forest maximum number of features did not reveal any potential to impact the fairness outcomes.

Lastly, the threshold setting demonstrated that the stringent threshold successfully lowers the FNR and increases the FPR across subpopulations. However, it seems to impose harder judgement on more disadvantaged groups. The lenient threshold benefits all subpopulations; however, the benefit is higher if the subgroup is defined by fewer sensitive attributes.

Table 10.

Summary of insights from data with non-scarce outcome and non-linear decision boundary

Decision Point	Impact on Fairness (FNR, FPR)	Subpopulation Variability and Key Observations
Missing Data Imputation	Overall lowest impact	Most disadvantaged group shows highest mean and variability
Scare Outcome Intervention	Highest impact, mitigates increase FNR, lower impact than in scarce outcome situation	Higher impact on groups with multiple sensitive attributes
Algorithm Choice	Outcomes consistent between different algorithms	Most disadvantaged group experiences greater change in FNR and FPR, data is handled better by tree-based methods
Hyperparameter Tuning	Minimal effect, slightly higher impact on FNR than on FPR	Higher impact on groups of minority and disability
Threshold Setting	Lenient threshold reduces FPR especially for advantaged groups	The more advantaged the group, the more positively it is affected by the lenient threshold and negatively affected by the stringent threshold

Fourth data situation – Non-linear decision boundary with scarce outcome

The last data situation poses the greatest challenge to the ML pipeline among all four data configurations. It is characterized by a non-linear decision boundary and a scarce outcome distribution, with non-fraudulent cases (96%) outweighing fraudulent cases (4%). Table 11 shows the last investigated data situation in the context of the four different data configurations.

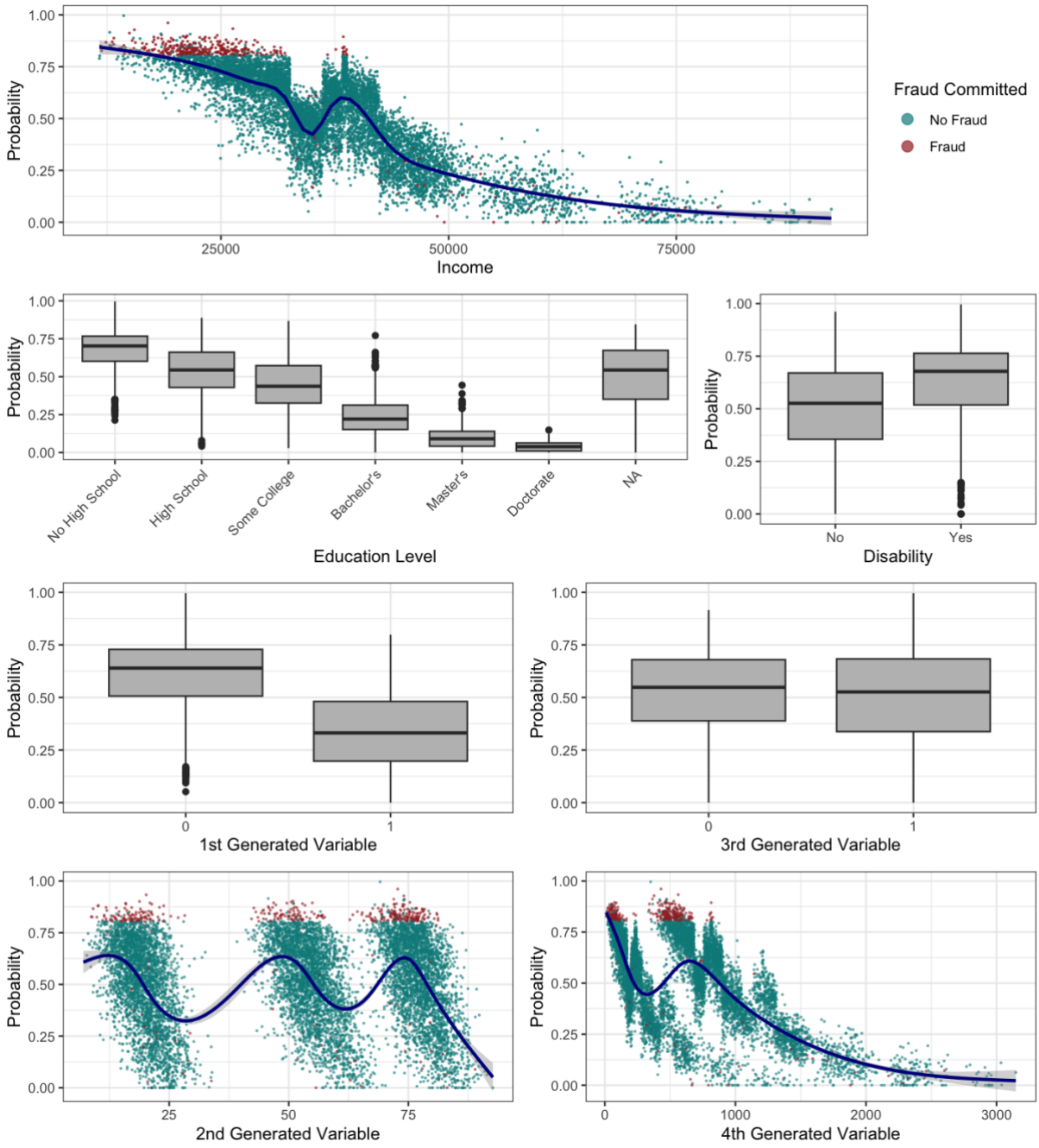
Table 11.

Overview of the currently tested data characteristics

	Decision Boundary	
Distribution of outcome	Linear, Non-scarce outcome	Non-linear, Non-scarce outcome
	Linear, Scarce outcome	Non-linear, Scarce outcome

Figure 41 displays the relationship between income and fraud commitment, demonstrating that lower income leads to higher fraud probability. However, it also assumes the non-linear trend of the lower middle class being less prone to fraud and the upper middle class being more prone to committing fraud. Additionally, the continuously generated variables exhibit non-linear relationships to the fraud outcome; compared to the first two data situations with a linear decision boundary (see Figure 19 and Figure 30), the patterns of the decision boundary are strongly distinguished. Furthermore, higher levels of education assume lower levels of fraud. However, the decrease between levels of education is non-linear. Disability increases the fraud probability. Additionally, the first generated variable decreases the fraud probability, while the third generated variable shows no meaningful connection to the fraud outcome. The fourth generated variable shows an overall decrease in the probability of fraud, with a non-linear connection.

Figure 41.
Relationship of fraud probability and classification by features

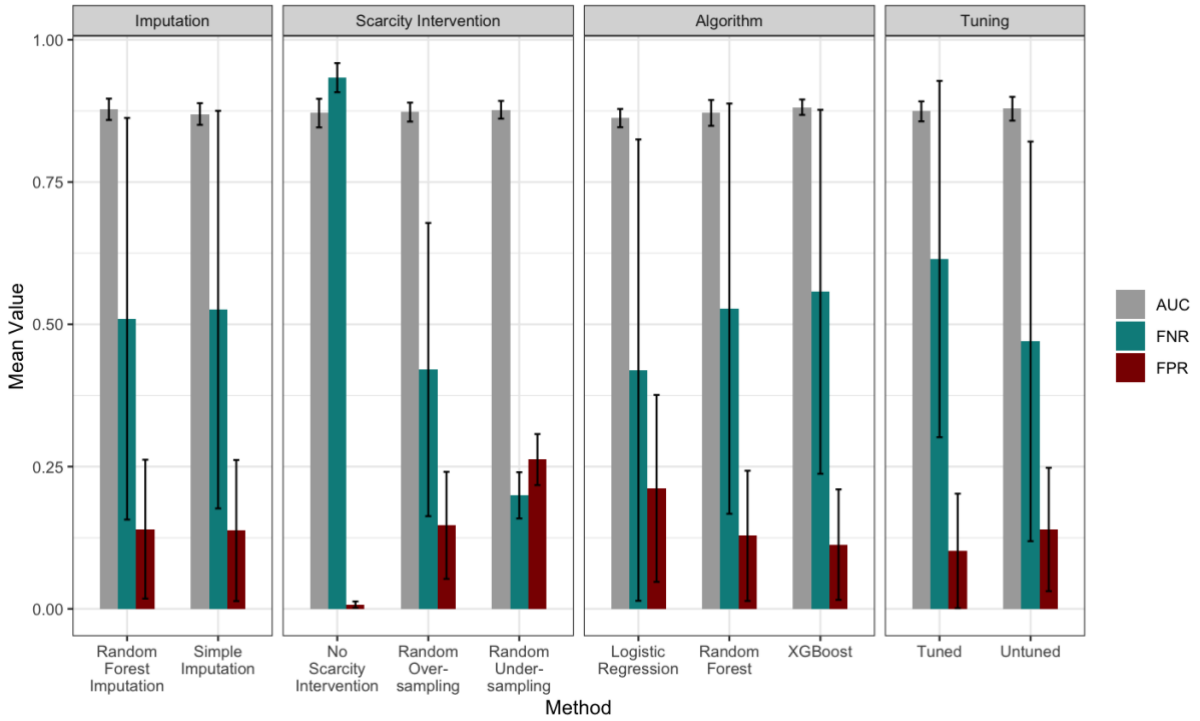


Overview of decision point impact

Before investigating the potential change each decision point asserts towards the fairness measures, the overall performance and fairness outcome for each decision point are detailed in Figure 42. Unlike the previous data situations, the mean value for FNR on the missing data imputation methods is higher. However, the differences between the methods remain little. The scarcity intervention shows that when not applying an intervention, FNR increases, which shows how the models do not sufficiently pick up on the fraudulent cases, resulting in a low

FPR. In this case, the models show a majority class bias, missing the patterns around the 4% of fraudulent classifications. The chosen scarcity intervention can lead to an overall higher FNR with random oversampling or a balance between FNR and FPR for random undersampling. The algorithm choice affects the AUC value, with the tree-based methods achieving higher AUC values than the logistic regression. Overall, the tree-based methods have an increase in FNR and a decrease in FPR compared to the logistic regression. Lastly, tuning the hyperparameters shows a tendency to increase FNR and decrease FPR compared to the untuned models. The variability in fairness measures is higher for algorithm choice and hyperparameter tuning than for the imputation method and scarcity intervention.

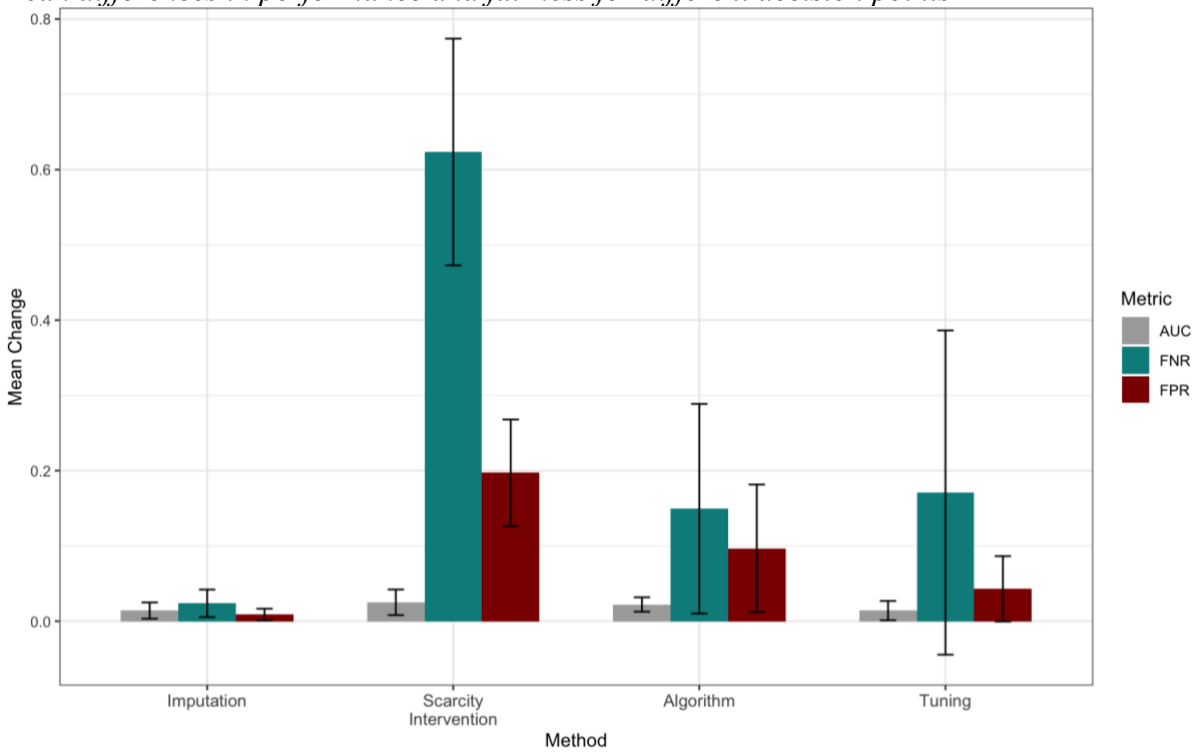
Figure 42.
Mean performance for each investigate engineering choice within the pipeline



The differences in the performance of options within the same decision points are visualized in Figure 43 as mean change (bars) and their standard deviations (error bars) while keeping the other engineering choices constant. The scarcity intervention exhibits the most remarkable change potential, especially for the FNR. Algorithm choice and tuning further impact the fairness measures; however, missing data imputation only has a limited impact on the overall fairness outcome of the pipeline. These tendencies match the outcome of the scarce data situation with a linear decision boundary, which has a higher impact on algorithm choice and tuning.

Figure 43.

Mean differences in performance and fairness for different decision points



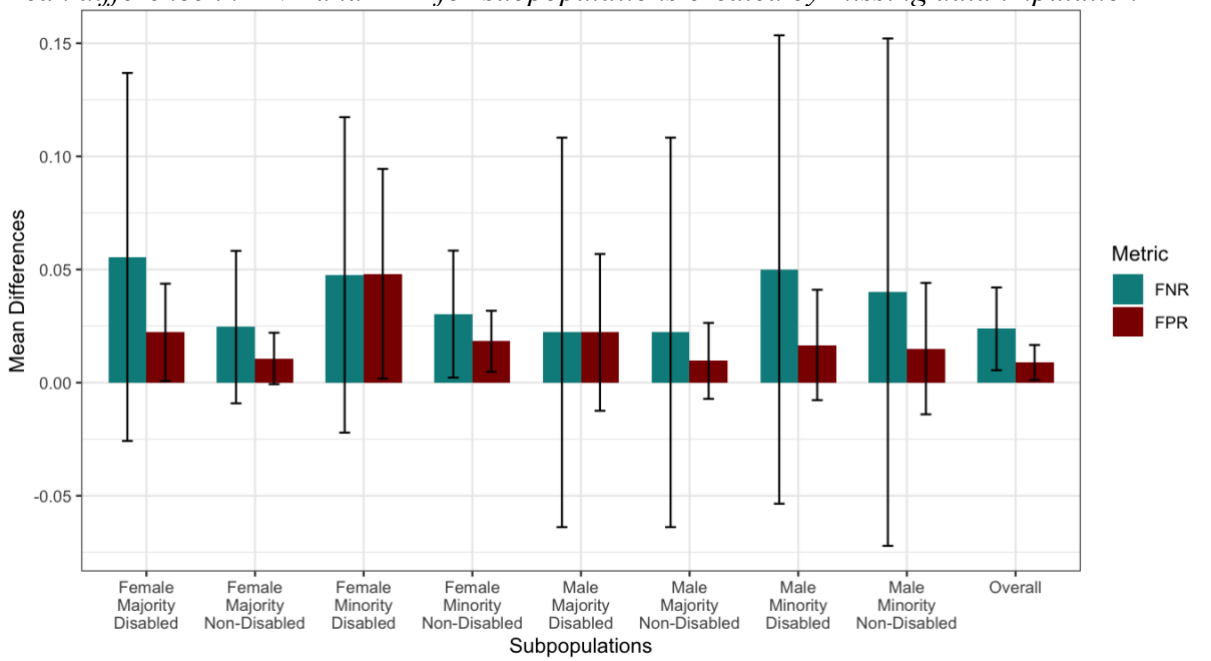
The mean differences in each decision point are assessed in more detail by separating the changes according to the eight subpopulations. This will give insight into whether the change potential is more prevalent in specific subpopulations than others. The decision points will be explored in the standard ML classifier pipeline order.

First decision point: missing data imputation

The missing data imputation shows the smallest effect on fairness measures among all evaluated decision points. Figure 44 shows how the chosen data imputation method expressed differently in the eight subpopulations regarding fairness outcome. Generally, it shows relatively similar patterns across the subpopulations since it exhibits little impact on the fairness outcome. However, there is a tendency for greater variability for the FNR compared to FPR, and more considerable changes in this fairness measure are expected depending on the chosen imputation method. Generally, subpopulations with two or more vulnerable attributes show greater differences in the impact of the imputation method, which could be explained by their values not missing at random but being related to their group membership.

Figure 44.

Mean difference in FNR and FPR for subpopulations created by missing data imputation

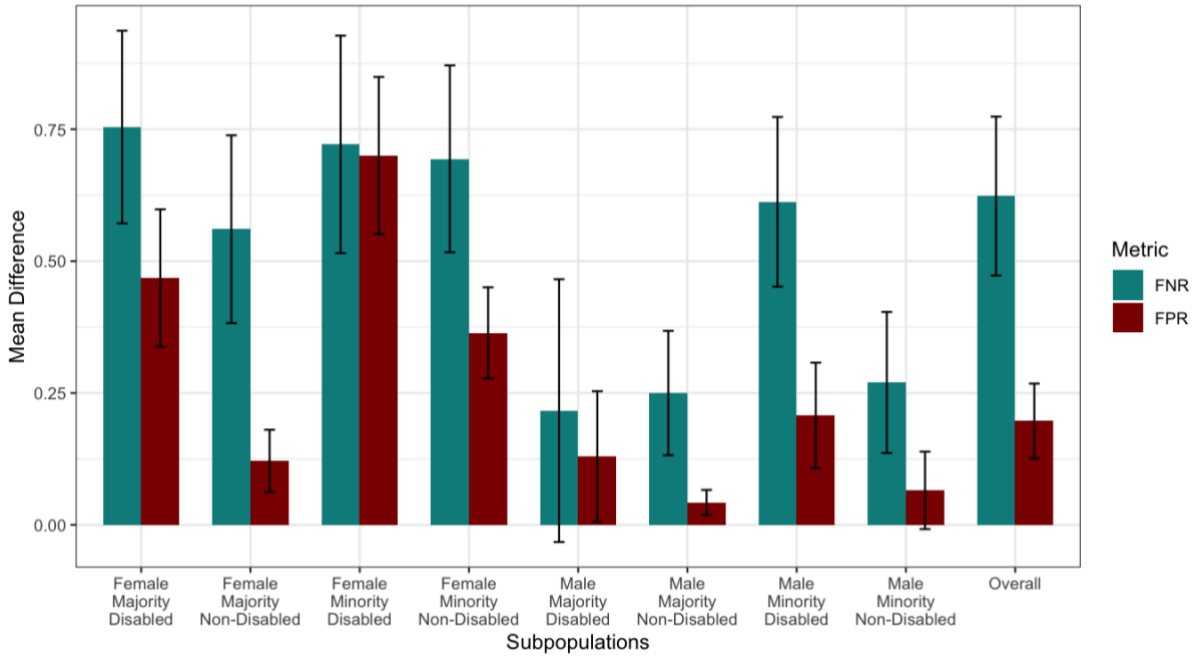


In conclusion, the impact of the missing data imputation on the fairness outcome are limited. However, the overall mean change and variability is higher in the FNR. Additionally, groups with a vulnerability level of two or higher experience higher change potential due to the imputation methods than less vulnerable groups.

Second decision point: scarcity intervention

The scarce outcome intervention exhibits the greatest change potential in fairness outcome amongst all inspected decision points. Figure 45 demonstrates how this change is differentiated between the subpopulations. There is a variable impact on the chosen intervention for the different populations, with populations of vulnerable level one or lower being less impacted by the choice. Hence, the male population with one or less vulnerable attributes are less affected by the scarcity intervention. However, the most vulnerable male subpopulation (male & minority X disabled) shows an increased impact of the scarcity intervention, with increased change mainly for the FNR and the FPR. The female subpopulations are affected the most, exhibiting higher FPR and FNR than the male subpopulations. The more vulnerable attributes are collected in a subpopulation, the higher the change in fairness outcome will be, depending on the chosen scarcity intervention. The variability in terms of standard deviation remains similar across subpopulations.

Figure 45.
Mean difference in FNR and FPR for subpopulations created by scarcity intervention



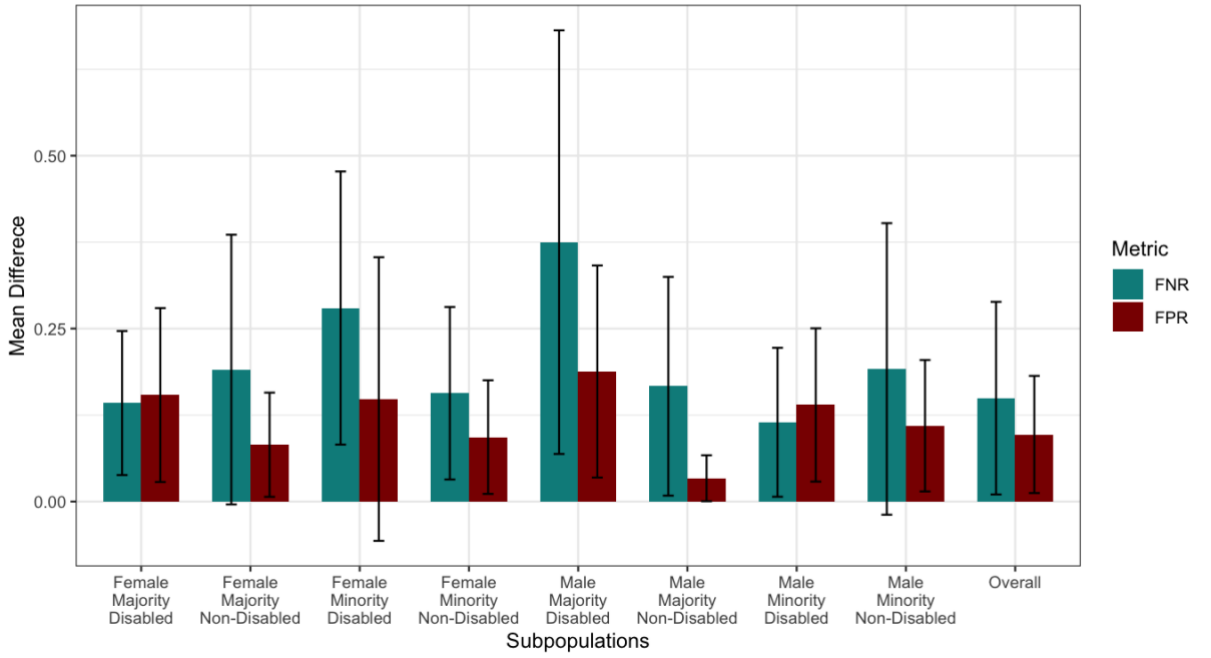
In conclusion, scarcity intervention affects the most vulnerable subpopulations, especially amongst the female populations. However, the more advantaged populations experience less impact by the scarcity intervention, showing that they are less vulnerable to experiencing adverse effects depending on the chosen intervention.

Third decision point: algorithm choice

The overall impact of the algorithm's choice on the fairness measure's mean change is separated by subpopulation in Figure 46. It shows a relatively consistent impact between most subpopulations, with balanced changes in FPR and FNR. The greatest change is experienced by the male X majority X disabled group, with increased change in FNR and FPR. A similar impact can be seen on the most disadvantaged group (female X minority X disabled). The most advantaged group has the lowest impact (male X majority X non-disabled). Lastly, the variability (expressed as standard deviation in error bars) is higher for higher levels of vulnerability.

Figure 46.

Mean difference in FNR and FPR for subpopulations created by algorithm choice



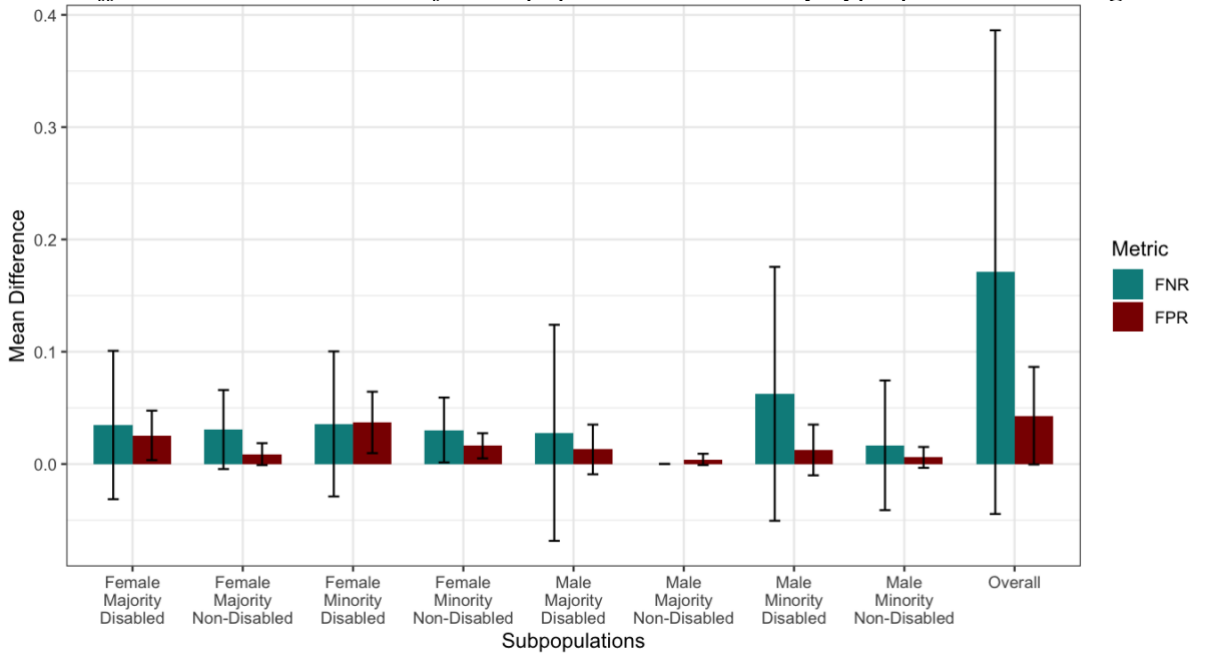
Overall, the effect of algorithm choice is approximately consistent across most subpopulations. However, the most advantaged subpopulation experiences less risk of changes in the FPR. Meanwhile, two disadvantaged groups show greater changes in both fairness measures.

Fourth decision point: hyperparameter tuning

Generally, the hyperparameter tuning process shows greater FNR changes than the FPR. Looking at the changes by subpopulation, the mean change in the fairness measure is approximately similar between the groups. However, the error bars indicate that the standard deviation is higher in the FNR for the disabled subpopulations, showing that their fairness results vary more than those of the non-disabled subpopulations.

Figure 47.

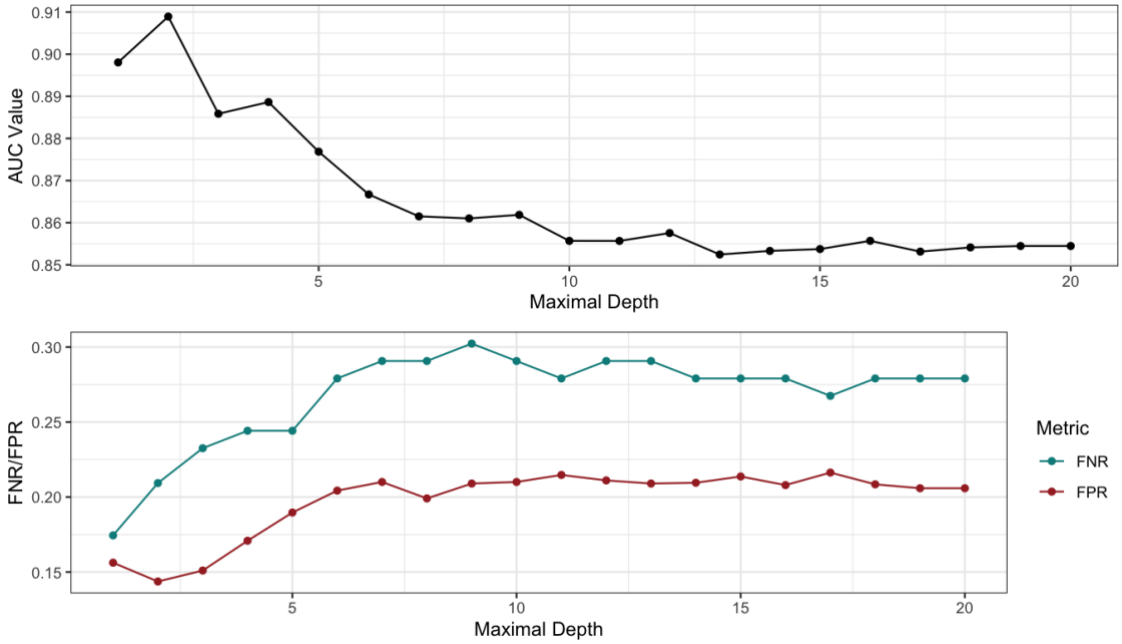
Mean difference in FNR and FPR for subpopulations created by hyperparameter tuning



The hyperparameter tuning process is further probed by isolating the tuning of one hyperparameter for XGBoost and Random Forest while keeping the other hyperparameters at their default value. Figure 48 shows the tuning process of XGBoost, which only considers the maximal depth parameter. It shows that the AUC value peaks at the setting of two, while the fairness measures show varying outcomes depending on the chosen value. However, the AUC peak is relatively isolated, with other choices leading to a more considerable decrease in the AUC value, meaning that choosing the hyperparameter according to the fairness outcome will introduce a loss in the model performance. At the optimal maximal depth value, the FNR is increased compared to one value lower, which would lead to a decreased FNR. However, if an engineer made this decision, the AUC value would decrease more than in previous data situations. On the other side, making this decision on the sole basis of engineering perspective without paying attention to fairness concerns, the overall AUC performance would be higher.

Figure 48.

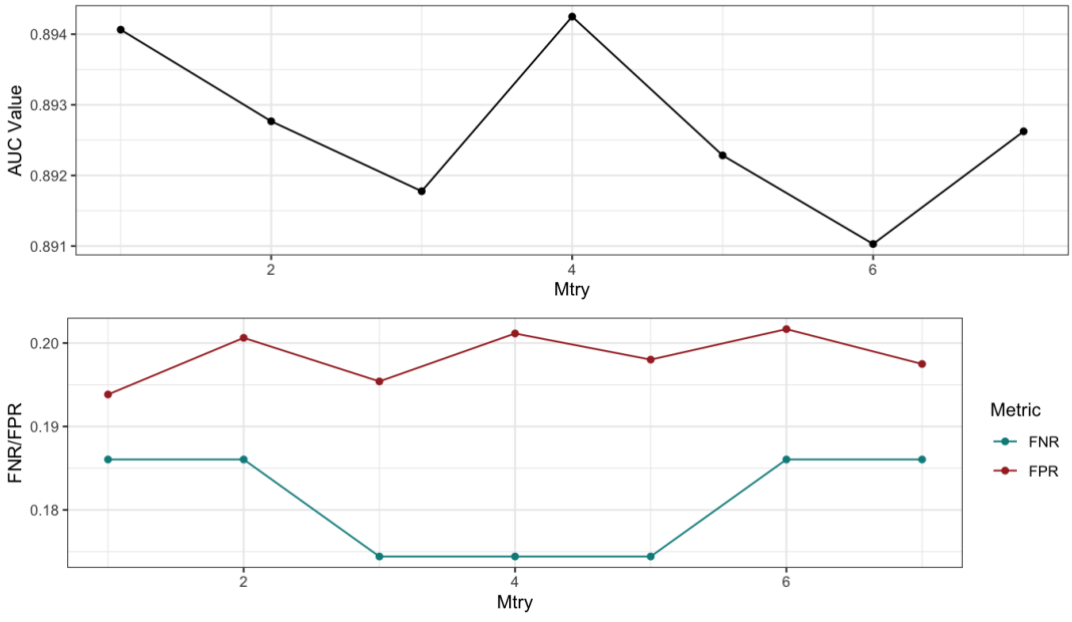
Isolated tuning of XGBoost's maximal depth and its impact on AUC, FNR and FPR



For the Random Forest, the hyperparameter that sets the number of features selected in each split (mtry) is probed with results shown in Figure 49. The plot shows only minimal changes in the AUC value for this hyperparameter, with two peaking values. With the first value setting the maximum number of features to one, the FNR and FPR would be reasonably balanced; however, setting the hyperparameter to four, the FPR would be higher, and the FNR would be lower. Hence, the engineer could decide on the approximation of the fairness measures. However, it is essential to note that the difference in both AUC and the fairness measures is minimal, depending on the value of the hyperparameter.

Figure 49.

Isolated tuning of Random Forest's mtry and its impact on AUC, FNR and FPR

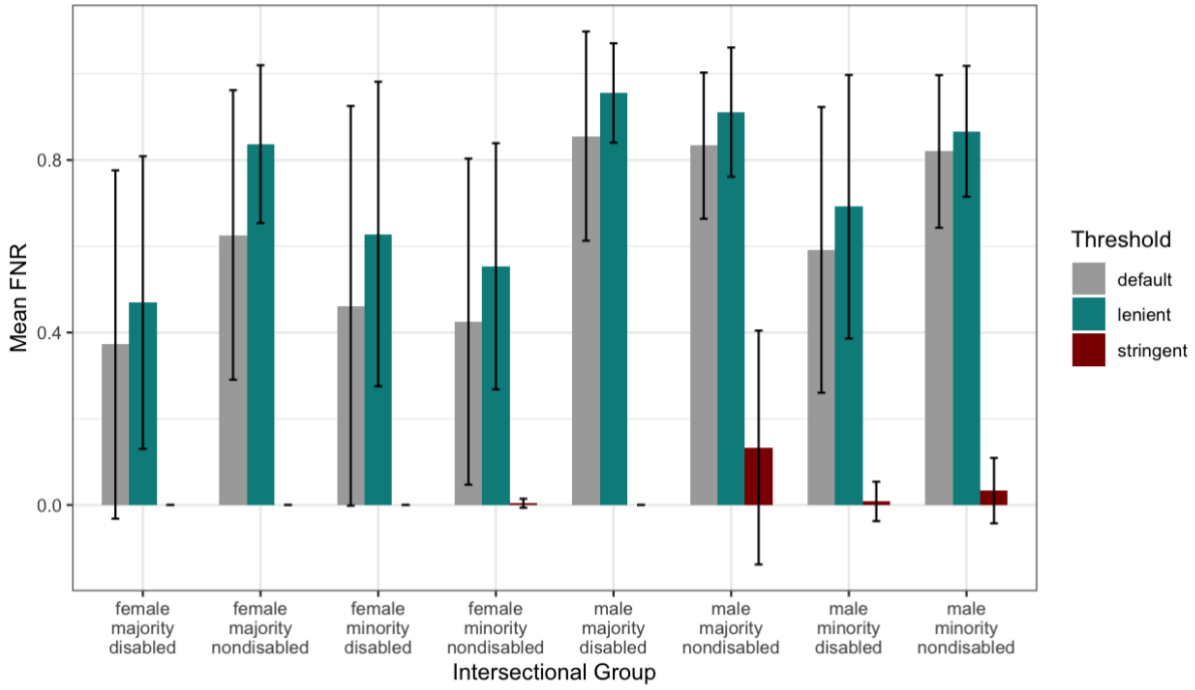


Fifth decision point: threshold setting

Lastly, the impact of threshold setting is assessed separately from the previous decision points that only considered the default threshold (cut-off at 0.5 fraud probability). The FNR is increased by the lenient threshold (cut-off over 0.9 quantiles). However, the increase stays relatively close to the default outcome. However, with the stringent threshold (cut-off over 0.1 quantile), the FNR is minimized, with mostly only the most advantaged subpopulation having varying FNR. This shows that the stringent threshold can still benefit the advantaged subpopulation while correcting for wrongfully negative classifications in all other subgroups. The standard deviations of the mean fairness performances remain relatively similar between the subpopulations.

Figure 50.

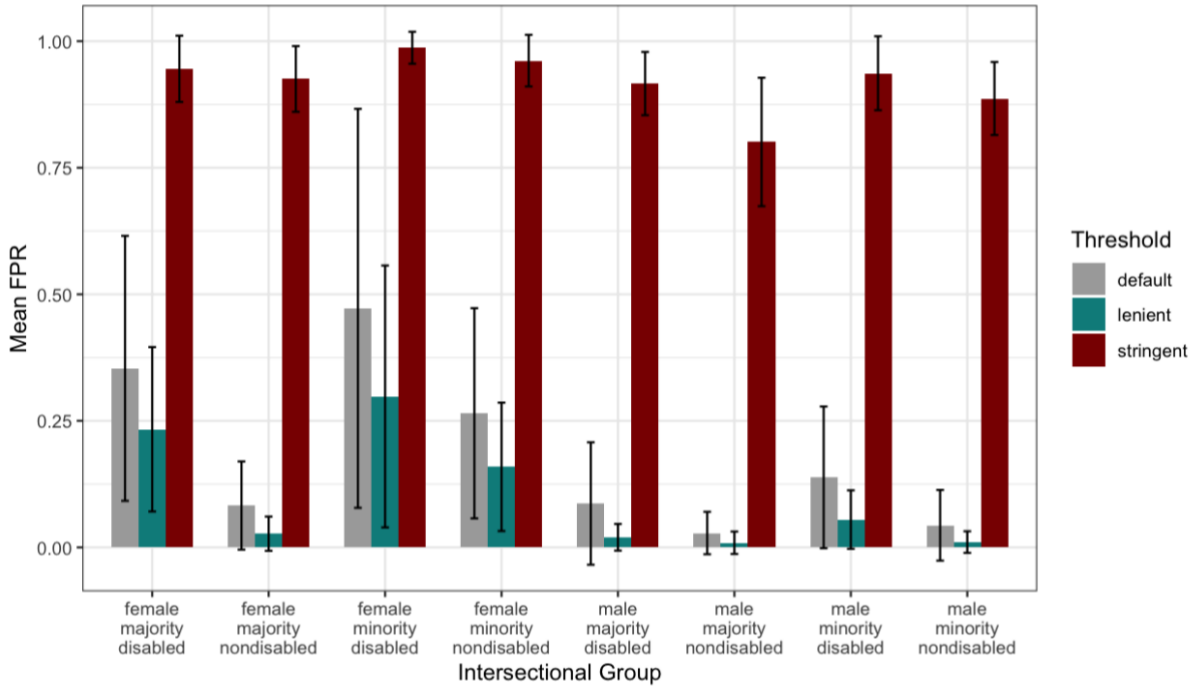
Effect of threshold setting on FNR for different subpopulations



The stringent threshold increases the FPR for all groups, with a lower impact on the most advantaged subpopulation. The lenient threshold almost diminishes the FPR for the male groups; however, the more sensitive attributes a group is defined by, the more the FPR of the lenient threshold approaches the values of the default threshold.

Figure 51

Effect of threshold setting on FPR for different subpopulations



In conclusion, the more privileged a subpopulation is, the more it benefits from different threshold settings, making the threshold setting ineffective as a moderator for between-group discrepancies in fairness outcomes.

Summary

Overall, this data situation was the most challenging one for the ML pipeline to tackle, with a non-linear decision boundary and scarce outcome distribution. The different decision points exerted different influences on the fairness outcomes for the subpopulations.

Generally, the missing data imputation had the smallest impact; however, it showed a tendency to have a greater impact on groups defined with two or more vulnerable attributes.

The scarcity intervention had the most considerable impact on fairness outcome change. Generally, the chosen scarcity intervention showed the most prominent difference for vulnerable female subpopulations, with less vulnerable groups showing fewer mean changes in fairness outcome depending on the scarcity intervention.

The choice of algorithm had an overall consistent impact on the fairness outcome for most subpopulations; however, two vulnerable groups experienced greater mean changes, while the most advantageous subpopulation was less prone to changes in FPR.

The hyperparameter tuning did show that disabled subpopulations experience greater variety in fairness outcomes compared to the non-disabled groups. However, probing singular

hyperparameters did not reveal feasible options to impact the fairness outcome without reducing the overall model performance.

Lastly, the threshold setting can rebalance the FNR and FPR; however, the least advantaged groups are more negatively affected by the stringent threshold than the advantaged groups. Additionally, the lenient threshold is of larger benefit for the more advantaged subpopulations.

Table 12.

Summary of insights from data with non-scarce outcome and linear decision boundary

Decision Point	Impact on Fairness (FNR, FPR)	Subpopulation Variability and Key Observations
Missing Data Imputation	Minimal overall fairness impact with slight tendency for higher FNR	Higher impact on vulnerability level two or higher for FNR
Scare Outcome Intervention	Highest overall impact Not applying method raises the FNR and lowers FPR (majority class bias)	Greater impact on vulnerable subgroups, especially females Undersampling balances FNR and FPR Oversampling raises FNR
Algorithm Choice	Moderate impact Tree-based methods increase FNR and decrease FPR compared to logistic regression	Higher impact on disabled and minority groups, advantaged groups experience least change
Hyperparameter Tuning	Increases FNR compared to untuned, trade-off between performance and fairness	Disabled groups show higher variability, limited potential to increase fairness without reducing performance
Threshold Setting	Stringent and lenient threshold moderate FNR and FPR	Least advantaged groups see mire negative impact, still favoring advantaged groups with higher FNR and lower FPR

Chapter 5 – Discussion

The following chapter aims to summarize the principal highlights from the analysis section and draw overarching conclusions on what those results could imply for future fair ML considerations. Each decision point will be discussed separately, considering the different data situations and showing how engineering choices can impact fairness codependently with the underlying data structure. It will highlight similarities and differences between the change potential in decision points depending on the data situations.

The decision points were analyzed using a simulated population that is assumed to be fully representative. The simulation depicts how certain factors, such as income and education, could impact an individual's likelihood of committing social benefit fraud. The population contains underlying connections to sensitive attributes of gender, ethnicity, and disability, modelling realistic societal influences of sensitive attributes towards the outcome of variables such as income and education level. Missing data was introduced, with individuals from protected groups having a higher probability of omitting data, indicating that data is missing not at random. As social benefit fraud is assumed to be a rare instance, scarce outcome distribution only contains 4% of fraudulent cases, while the non-scarce data situations contain 25%. In the modelled society, individuals with protected attributes (e.g., women, minority ethnicity, disabled) tend to more often commit social benefit fraud, with the accumulation of protected attributes increasing this inclination, as the proxy variables (e.g. income and education) are affected differently depending on the subpopulation status. Using ML classifiers to predict whether an individual would commit fraud, different decision points with different options are evaluated towards their impact on the model performance and fairness.

The model fairness is assessed through the FNR and FPR, which are central fairness measures used in fair ML. In this case, they represent both the stakeholders of the ML classifier, which would be the contracting authority and the evaluated individuals. The FNR represents the proportion of false negative classifications, meaning that an individual would not be classified as fraudulent even though they committed fraud. A higher FNR is in the interest of the evaluated individuals, giving them a higher chance to remain undetected. Hence, a group that shows substantial changes in FNR within a decision point shows the potential to get favoured through it. Reversely, the contracting authority would be interested in maintaining a low FNR as they are interested in not missing potentially fraudulent individuals. The second measure, FPR, shows the proportion of wrongful positive classifications, which, in this case, means that an

innocent individual would be accused of fraudulent activity. The interest in a low FPR is higher for the evaluated individuals, as a wrongful accusation could harm them. Hence, if a group experiences higher FPR changes in a decision point, this shows that this decision point can elicit adverse treatment for them.

Change potential of decision points

The average change in fairness measures was obtained for four different data situations for each decision point, showing how individuals might be impacted depending on the options chosen within that decision point. Figure 52 shows the previously presented overall change potential of the decision points on the datasets using the default threshold (0.5), giving a broad insight into how the influence of decision points can vary depending on the underlying data situation. The impact of threshold setting is investigated in a separate section below.

Figure 52.

Mean change in each decision point contrasted by data situation

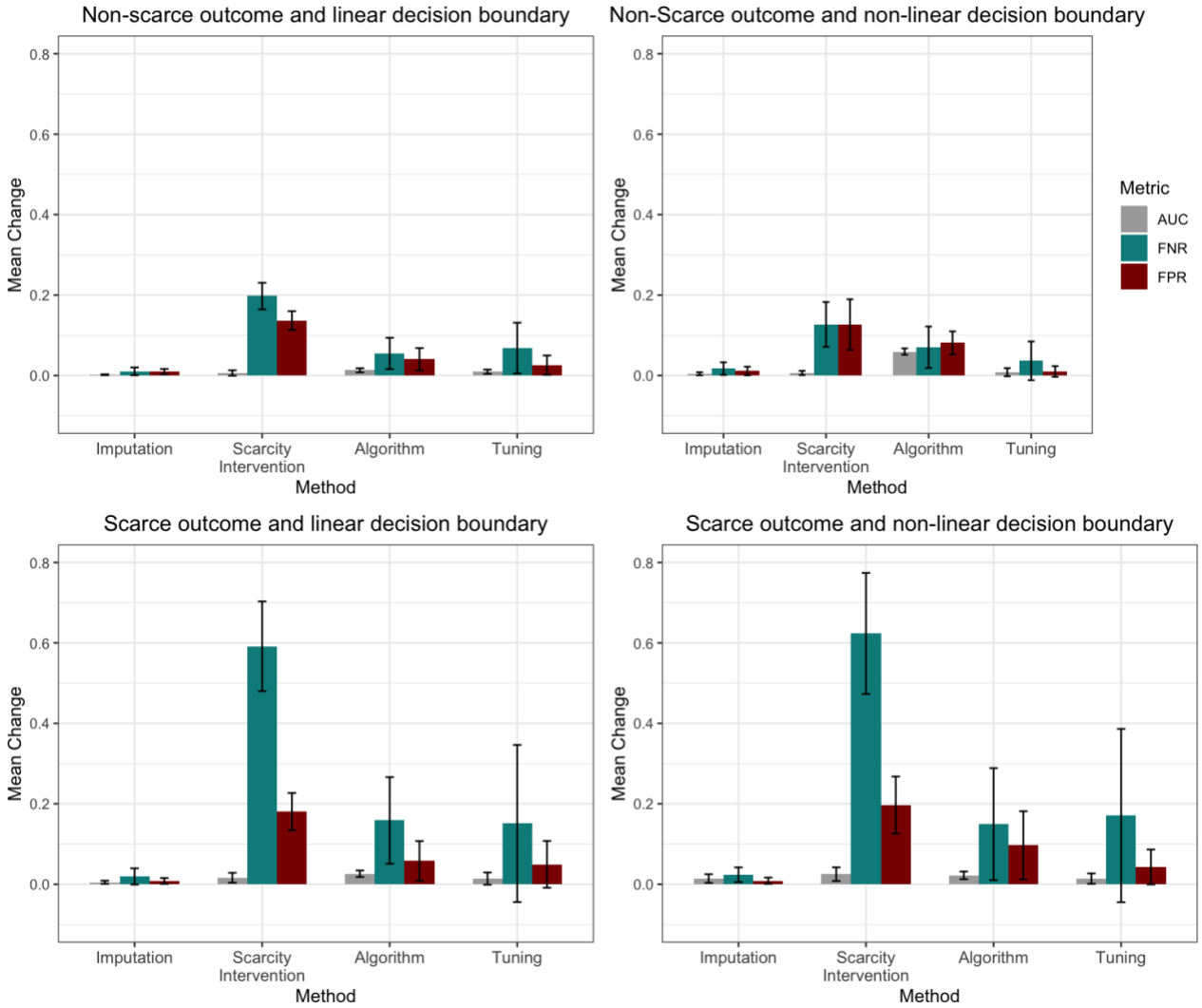


Figure 52 shows that the highest change potential in fairness measures is contained within the dataset with scarcely distributed outcomes and a non-linear decision boundary, showing that the engineering choices might show the highest impact under more complex data characteristics. Overall, the scarcity intervention demonstrated the highest effect in all datasets, although its impact is more substantial within the scarce outcome data. The missing data imputation had the most negligible impact among all the decision points. However, this influence may increase considering a scenario with a higher proportion of missingness. The algorithm choice showed a higher impact with more complex data situations, again having the highest impact on fairness measures for the data situation with scarce outcomes and non-linear decision boundaries.

Lastly, tuning showed the most impact on the scarce datasets. Generally, the decision points influenced the FNR more than the FPR. This gives an overview of the general change potential of each decision point, showcasing that the choices elicit more considerable change for more complex data characteristics and making scarcity intervention the most influential decision point on fairness measures while missing data imputation barely carries a change potential. Those decision points will be discussed in further detail, considering their differentiated influences on the underlying subpopulations defined by their combination of protected attributes.

Missing data imputation

Across all data situations, missing data imputation has a negligible effect on the fairness outcome. However, there are still some visible different impacts depending on the subpopulation. Depending on the imputation method, more disadvantaged groups show higher variability in their fairness outcome. This pattern demonstrates that imputation methods can impact fairness outcomes, particularly for disadvantaged groups whose data is more likely to be missing not at random. The overall impact was higher for the data situations with scarce outcomes, demonstrated by a larger variety in fairness outcomes, especially for the disabled subpopulations. Overall, non-linear and scarce data contexts amplify fairness disparities more for the most disadvantaged subpopulations (e.g., female, minority, disabled). This indicates that tailored imputation strategies considering group-specific risks could mitigate adverse fairness impacts in more complex decision scenarios.

A potential weakness in considering this decision point is that only one proportion of missing data was considered. For more insight into the fairness impact of missing data, higher

proportions of missingness could have been considered. In this case, the number of missing values might have been too low to unveil substantial revelations on the effect of imputation.

Nevertheless, while missing data imputation has a lower overall fairness impact, it disproportionately affects certain vulnerable groups. Engineers should consider data sensitivity to missingness when selecting imputation methods, as this could mitigate or exacerbate fairness disparities depending on the subpopulation.

Scarcity intervention

Scarcity interventions consistently influenced fairness outcomes across all situations, especially for the vulnerable subpopulations. These interventions influenced fairness outcomes in scarce and non-scarce settings, though their impact is generally more intense in the scarce context. The effect on FNR was consistently high, with some interventions effectively mitigating increased FNR under scarce outcome distributions. However, this often came with a trade-off which introduced bias in FPR against vulnerable groups. In the linear data situations, the more advantaged subpopulations experienced less risk for FPR fluctuations but higher changes in FNR, offering them potential benefits compared to the vulnerable groups. In the non-linear data situations, scarcity interventions showed more significant variability, especially for disadvantaged groups with multiple sensitive attributes. These groups saw the highest mean change in fairness outcomes, with the scarcity intervention's effects consistently more pronounced than in the linear scenarios.

Generally, the influence of scarcity interventions on fairness is more visible for non-linear and scarce data characteristics, where subpopulations with multiple sensitive attributes undergo the most substantial changes. In contrast, the linear/non-scarce setup shows significant effects but less variability among disadvantaged groups, which suggests that linear models may handle scarce data distributions with less adverse fairness impact. In conclusion, scarcity interventions are vital for moderating fairness for subpopulations, especially for the disadvantaged ones. However, scarcity interventions effectively reduce FNR, but they can also introduce FPR biases that disproportionately affect vulnerable groups. Adjusting for scarcity should be carefully considered to avoid unintended bias towards vulnerable groups.

Choice of algorithm

Algorithm choice consistently affects fairness across data situations with variations in FPR and FNR notably different between male and female subpopulations. Overall, vulnerable groups are more likely to experience changes in fairness metrics based on the algorithm choice, showing

the potential impact of this decision point on the fairness outcome. In both scarce and non-scarce linear data, female subpopulations saw similar fairness outcomes regardless of algorithm choice. In contrast, male subpopulations experienced variability depending on their level of vulnerability, with minority ethnicity and disability negatively impacting their fairness outcome. The vulnerable groups saw higher FNR, reflecting differences in how algorithms handle linear, scarce outcomes. For the non-linear data situations, algorithm choice significantly impacted the FNR and FPR of the most vulnerable groups particularly females with disabilities. Advantaged groups showed less variability, while the most disadvantaged groups saw increased mean changes in FNR.

Overall, in non-linear contexts, algorithm effects become more prominent for individuals with double vulnerability (specifically for female subpopulations with disabilities), indicating that non-linear decision boundaries may introduce unique fairness challenges for specific demographic traits. Linear models show relatively more stability, still the fairness impact varies by vulnerability level. In conclusion, algorithm selection can substantially impact fairness especially for the more vulnerable subpopulations. Fairness considerations should inform the choice of algorithm, as specific algorithms may offer better performance in managing fairness outcomes for groups with multiple vulnerabilities.

Hyperparameter tuning

Hyperparameter tuning shows a relatively consistent impact across situations, which gives options for fairness adjustments with minimal effect on AUC. This consistency highlights tuning as a feasible strategy for improving fairness without compromising model performance. For linear and non-scarce data, the lower maximum depth for XGBoost led to more minor differences in fairness metrics. At the same time, scarce data scenarios showed that tuning could balance FPR and FNR differently. Vulnerable subpopulations tended to benefit less from tuning adjustments, highlighting that tuning might not fully mitigate fairness issues for these groups. In non-linear situations, particularly with scarce data, maximum depth tuning in XGBoost showed isolated AUC peaks, making fairness tuning feasible with limited performance loss. However, disadvantaged groups saw more remarkable FNR changes, especially under non-linear decision boundaries, where fairness adjustments had a more considerable impact. In contrast, tuning Random Forest's number of features used in each split shows that there is not always an impact on the fairness outcome. Hence, considering hyperparameter tuning, in some data situations, some hyperparameters can behave value-free and should be tuned towards maximum accuracy. In other situations, they offer several good choices with different outcomes

regarding fairness for a small accuracy trade-off. However, there are also scenarios where the choice dependent on the fairness outcome would result in more substantial accuracy losses. Overall, depending on the scenario, hyperparameter tuning may offer limited potential for fairness adjustments, while in others, it might show opportunities for fairness modifications.

Overall, the investigation of the effect of hyperparameter tuning on fairness was heuristic and by far not extensive. For the pipelines' tuning process, only a few possible values were assessed due to computational constraints, leading to a simplified approach to the tuning process. Hence, the tuning set-up could have been more exhaustive for the investigated models. Furthermore, only two hyperparameters were investigated in more detail, giving a glimpse into the potential fairness impact but disregarding potential interplay with the tuning process of other hyperparameters kept at default.

Nonetheless, the results indicate that hyperparameter tuning is valuable for balancing performance and fairness, especially in non-linear data situations. While tuning can improve fairness with minimal impact on AUC, it may still fail to address disparities for the most vulnerable groups.

Threshold setting

Across all data situations, the lenient threshold elicited some change compared to the default threshold and demonstrated disparate treatment between subpopulations, while the stringent threshold varied in its impact. Generally, the lenient thresholds across data situations tend to benefit more advantaged subpopulations by lowering their FPR, while stringent thresholds impacted groups similarly, showing less disparity in fairness metrics. In linear, non-scarce scenarios, lenient thresholds benefited male subpopulations more, while disadvantaged groups still faced higher FPR. In scarce data, lenient thresholds showed a similar pattern, with more advantaged subpopulations gaining more. In non-linear settings, lenient thresholds increased FNR more for advantaged groups but reduced FPR substantially for male subpopulations. Stringent thresholds, however, increased FPR across subpopulations, showing minimal discrimination across groups.

Overall, for all data characteristics the lenient threshold disproportionately favors the advantaged groups. This means, the lenient threshold exacerbates FNR disparities for advantaged subpopulations. This indicates that threshold adjustments alone may not be sufficient to achieve equitable outcomes across all demographics. In conclusion, the threshold setting has mixed effects on fairness, often benefiting advantaged groups when set to lenient

values. While stringent thresholds provide consistency across groups in some data situations, they do not fully address biases, especially under non-linear decision boundaries.

Overall conclusions

Three overarching conclusions can be drawn from the observation of each singular decision point and its influence on the fairness outcome. Firstly, data situations and engineering choices are interdependent. The effect of each decision point varies across linear vs. non-linear and scarce vs. non-scarce data. Non-linear and scarce data scenarios heighten fairness disparities, making fairness-focused adjustments in ML pipelines more necessary. Hence, the engineering choices were not value-free in terms of the fairness outcome, confirming the hypothesis. However, the extent to which those choices affect the fairness outcomes depends on the complexity of the data characteristics. However, the analysis also gives insight into the consideration that there is no single solution to the fairness problem. Depending on the data characteristics, a decision point shows different effects on the fairness outcome, showing that the considerations must be weighted individually depending on the applied situation.

Secondly, there is a need to tailor engineering choices while considering multiple subpopulations, as the overarching fairness outcome does not consistently apply to all the groups represented in the data. This is showcased in how vulnerable subpopulations face the highest disparities across all decision points. Adjusting choices such as algorithms, hyperparameters and thresholds alone is essential but insufficient. Engineers must assess each pipeline component's impact on fairness to achieve equitable outcomes across all subpopulations. Hence, it is essential to remember the impact of fairness throughout the configuration of the entire pipeline.

Lastly, with some decisions the engineer should consider balancing fairness and performance, as in some scenarios, fairness can be adjusted with only minimal losses in performance. However, while some adjustments (e.g., tuning, algorithm choice) allow for improved fairness without significant AUC loss, scarcity interventions and threshold settings reveal trade-offs that prioritize fairness for vulnerable groups at a performance cost for advantaged groups. Decision-makers must weigh these trade-offs based on application-specific priorities. Overall, the analysis showed that engineering decisions in different parts of the ML pipeline could influence the fairness outcome and introduce disparate treatment between subpopulations. Hence, the hypothesis that engineering choices are not value-free in their

impact on the fairness outcome has been supported. This implies that an engineer should consider the fairness outcome at each step of the pipeline to ensure a fair ML classifier.

Implications on the fair machine learning discussion

In conclusion, this study provided evidence that engineering choices influence the overall fairness outcome of a model and that fine-grained subpopulations are differently affected by those choices. The change potential of pipeline decision points depends on the underlying data characteristics, with complex data showing more leverage in the fairness impact of engineering choices. The subpopulations were affected differently, with vulnerable groups prone to adverse impact, while advantaged groups usually benefitted more from different engineering choices. Furthermore, there is some potential to change the fairness outcome without losing the overall accuracy of the classifier in some scenarios. Overall, the analysis showcased that engineers should not treat choices along the pipeline as value-free towards fairness and should have an ethical approach towards engineering choices, not solely focusing on raising the model's accuracy.

The model's FNR and FPR being affected by the engineering choices cannot be treated lightly, especially not in the context where the classifications directly relate to society and give individuals adverse treatment depending on their vulnerable attributes (Das et al., 2021). Depending on the context, an ML classifier might be used for decisions that do not affect humans substantially (e.g. consumption recommendation algorithms); hence, those pipelines do not carry the same social responsibility. In those cases, the engineer might treat the engineering choices as value-free. However, in high-stakes scenarios such as fraud detection, this simplified view of training the model solely for high accuracy does not represent the social responsibility of non-discriminatory treatment. Hence, context-dependent, ML engineers have a social responsibility to assess whether the developed classifier is fair towards different subpopulations represented in the data (Kenfack et al., 2021). By this, fairness considerations are not a "nice-to-have" step to include in the model-building process but should be a critical consideration for any high-stakes application that can shape individual's lives depending on the outcome (Das et al., 2021; Kenfack et al., 2021).

Rather than focusing only on debiasing the entered data or applying fairness intervention methods, this analysis showcases that the ML engineer can consider the fairness outcome at each pipeline step. Similarly, Dat et al. (2021) propose a fairness-aware ML pipeline that considers various fairness measures at different stages of the pipeline setup. Rather than seeing

the fairness enhancing process as an optional module to the pipeline, it should be approached as a second optimization goal next to accuracy. With this approach, the engineer would be aware of the consequences of fairness at each step of the pipeline, and the engineering choices would test the effect on the overall model performance. Related to this process, tools such as FairLay-ML (Yu et al., 2023) and ArgusEyes (Schelter et al., 2023) have been developed to help data scientists screen an ML pipeline for fairness violations and give insight into the source of bias. Conclusively, the results suggest that the seemingly technical decisions in a pipeline are embedded in a model's overall fairness outcome and can have a real-world impact on different groups.

Furthermore, this model building with fairness outcome in mind should not be simplified. One could simply inspect the overall FNR and FPR for the model, assuming that they are acceptable in a given context and move on. However, this might not do justice to the underlying subpopulations that should be well-defined at the beginning of the pipeline construction (Davoudi et al., 2024; Li et al., 2022). The analysis showcased how the fairness outcome can strongly vary between subpopulations, and this should be controlled for when considering different options for the classifier pipeline. Overall, a proactive approach should be taken to define the context-relevant subpopulations, which would help ensure that fairness is consistently evaluated rather than relying on aggregate metrics that might mask disparate treatment.

The current research approach has certain limitations, as it aims to give a broad overview of the landscape of technical choices along the ML pipeline. Because of this, the choices within the decision points were limited, and expanding this space could reveal more generalizable influences on each decision point. Furthermore, the data characteristics were simplified and could be expanded by adding different proportions of missing data. Furthermore, the tuning process of the models was simplified because of computational considerations. This gives room for future research, which can also investigate more detail of one specific decision point rather than considering several decision points. Furthermore, the analysis could be considered for metrics beyond FNR, FPR, and real-world datasets for better generalizability. Lastly, the investigation of resampling methods could be expanded, as this decision point showed the highest impact. Future research could consider to not only resample towards the outcome variable, but also resample to balance the protected attributes.

In conclusion, this research argues for a fairness by-design approach, where fairness considerations are embedded throughout the ML pipeline rather than as an afterthought. The findings support this approach as they demonstrate that fairness adjustments can be made at different decision points and show effects like post-hoc fairness interventions. In conclusion, the seemingly technical choices offer the possibility to mitigate unfair treatment but also show the threat to exuberate it. ML engineers should be encouraged to view fairness as an optimization goal alongside accuracy, promoting a more holistic approach to model evaluation.

References

Aggarwal, A., Lohia, P., Nagar, S., Dey, K., & Saha, D. (2019, August). Black box fairness testing of machine learning models. In *Proceedings of the 2019 27th ACM joint meeting on european software engineering conference and symposium on the foundations of software engineering* (pp. 625-635).

Arslan, Y., Klein, J., Allix, K., Lefebvre, C., Boytsov, A., & Bissyandé, T. F. (2022, December 1). *Exploiting Prototypical Explanations for Undersampling Imbalanced Datasets*. https://doi.org/10.1109/icmla55696.2022.00228

Badar, M., Fisichella, M., Nejdl, W., & Sikdar, S. (2024). FairTrade: Achieving Pareto-Optimal Trade-Offs between Balanced Accuracy and Fairness in Federated Learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, *38*(10), 10962–10970. https://doi.org/10.1609/aaai.v38i10.28971

Badran, K., Côté, P.-O., Collante, A., Costa, D., Kolopanis, A., Bouchoucha, R., Shihab, E., & Khomh, F. (2022). *Can Ensembling Pre-processing Algorithms Lead to Better Machine Learning Fairness?* cornell university. https://doi.org/10.48550/arxiv.2212.02614

Bal, F., & Kayaalp, F. (2023). *Comparing SMOTE-based-ML Methods on the unbalanced dataset: A case study on a date and pistachio fruits*. authorea. https://doi.org/10.22541/au.169583589.98531349/v1

Barocas, S., Crawford, K., Shapiro, A., & Wallach, H. (2017, October). The problem with bias: Allocative versus representational harms in machine learning. In *9th Annual conference of the special interest group for computing, information and society* (p. 1).

Barrett, J., Jeanselme, V., Zhang, Z., De-Arteaga, M., & Tom, B. (2022). *Imputation Strategies Under Clinical Presence: Impact on Algorithmic Fairness*. cornell university. https://doi.org/10.48550/arxiv.2208.06648

Barsotti, F., & Koçer, R. G. (2022). MinMax fairness: from Rawlsian Theory of Justice to solution for algorithmic bias. AI & SOCIETY, 1-14.

Betts, K., & Null, N. (2011). Career Advancement: Ten Negotiation Strategies for Women in Higher Education. *Academic Leadership: The Online Journal*. https://doi.org/10.58809/ellb8284

Bilal, M., Anwar, M., Ali, G., Kadir, R. A., Iqbal, M. W., & Malik, M. S. A. (2022). Auto-Prep: Efficient and Automated Data Preprocessing Pipeline. *IEEE Access*, *10*, 107764–107784. https://doi.org/10.1109/access.2022.3198662

Birba, D. E. (2020). A Comparative study of data splitting algorithms for machine learning model selection.

Birchha, V., & Nigam, B. (2023). Performance Analysis of Averaged Perceptron Machine Learning Classifier for Breast Cancer Detection. *Procedia Computer Science*, *218*, 2181–2190. https://doi.org/10.1016/j.procs.2023.01.194

Bondugula, R. K., Udgata, S. K., & Bommi, N. S. (2021). A Novel Weighted Consensus Machine Learning Model for COVID-19 Infection Classification Using CT Scan Images. *Arabian Journal for Science and Engineering*, *41*(2), 11039–11050. https://doi.org/10.1007/s13369-021-05879-y

Buczak, A. L., & Guven, E. (2015). A survey of data mining and machine learning methods for cyber security intrusion detection. *IEEE Communications surveys & tutorials*, *18*(2), 1153-1176.

Buczak, P., Pauly, M., & Chen, J.-J. (2023). Analyzing the Effect of Imputation on Classification Performance under MCAR and MAR Missing Mechanisms. *Entropy*, *25*(3), 521. https://doi.org/10.3390/e25030521

Calmon, F., Wang, H., & Feng, R. (2023). *Adapting Fairness Interventions to Missing Values*. cornell university. https://doi.org/10.48550/arxiv.2305.19429

Cappelen, A. W., Tungodden, B., & Cappelen, C. (2018). Second-Best Fairness Under Limited Information: The Trade-Off between False Positives and False Negatives. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3243332

Caton, S., & Haas, C. (2024). Fairness in Machine Learning: A Survey. *ACM Computing Surveys*, *56*(7), 1–38. https://doi.org/10.1145/3616865

Cervini-Plá, M., Silva, J. I., & Castelló, J. V. (2016). Estimating the income loss of disabled individuals: The case of Spain. *Empirical Economics*, *51*, 809-829.

Chakraborty, J., Xia, T., Fahid, F. M., & Menzies, T. (2019). Software engineering for fairness: A case study with hyperparameter optimization. arXiv preprint arXiv:1905.05786.

Chattopadhyay, S., & Kishore, S. (2021). Classification of Mobile Price Range with Different Machine Learning Algorithms and Optimized Hyperparameters. *American Journal of Electronics & Communication*, *2*(2), 17–18. https://doi.org/10.15864/ajec.2204

Chaudhari, B., Agarwal, A., & Bhowmik, T. (2022). *Simultaneous Improvement of ML Model Fairness and Performance by Identifying Bias in Data*. cornell university. https://doi.org/10.48550/arxiv.2210.13182

Chemmakha, M., Habibi, O., & Lazaar, M. (2022). Improving Machine Learning Models for Malware Detection Using Embedded Feature Selection Method. *IFAC-PapersOnLine*, *55*(12), 771–776. https://doi.org/10.1016/j.ifacol.2022.07.406

Chennuru, V. K., & Timmappareddy, S. R. (2017). *MahalCUSFilter: A Hybrid Undersampling Method to Improve the Minority Classification Rate of Imbalanced Datasets* (pp. 43–53). springer. https://doi.org/10.1007/978-3-319-71928-3_5

Chouldechova, A., & Roth, A. (2018). The frontiers of fairness in machine learning. *arXiv preprint arXiv:1810.08810*.

Cruz, A., Soares, C., Bizarro, P., Saleiro, P., & Belém, C. (2020). *A Bandit-Based Algorithm for Fairness-Aware Hyperparameter Optimization*. https://doi.org/10.48550/arxiv.2010.03665

Das, S., Yilmaz, P., Donini, M., Gelman, J., Larroy, P., Haas, K., Katzman, J., Zafar, M. B., Kenthapadi, K., & Hardt, M. (2021). Fairness Measures for Machine Learning in Finance. *The Journal of Financial Data Science*, *3*(4), 33–64. https://doi.org/10.3905/jfds.2021.1.075

Davariashtiyani, A., Wang, B., Hajinazar, S., Zurek, E., & Kadkhodaei, S. (2024). *Impact of Data Bias on Machine Learning for Crystal Compound Synthesizability Predictions*.

Davoudi, A., Chae, S., Evans, L., Sridharan, S., Song, J., Bowles, K. H., Mcdonald, M. V., & Topaz, M. (2024). Fairness gaps in Machine learning models for hospitalization and emergency department visit risk prediction in home healthcare patients with heart failure. *International Journal of Medical Informatics*, *191*, 105534. https://doi.org/10.1016/j.ijmedinf.2024.105534https://doi.org/10.48550/arxiv.2406.17956

Do, H., Zhong, J., Putzel, P., Smyth, P., & Nandi, S. (2022). A joint fairness model with applications to risk predictions for underrepresented populations. *Biometrics*, *79*(2), 826–840. https://doi.org/10.1111/biom.13632

Florescu, D., & England, M. (2020). *A machine learning based software pipeline to pick the variable ordering for algorithms with polynomial inputs*. https://doi.org/10.48550/arxiv.2005.11251

García-Laencina, P. J., Sancho-Gómez, J. L., & Figueiras-Vidal, A. R. (2010). Pattern classification with missing data: a review. *Neural Computing and Applications*, *19*, 263-282.

Gee, K. A. (2018). Minding the Gaps in Absenteeism: Disparities in Absenteeism by Race/Ethnicity, Poverty and Disability. *Journal of Education for Students Placed at Risk (JESPAR)*, *23*(1–2), 204–208. https://doi.org/10.1080/10824669.2018.1428610

Gomaa, I., El-Tazi, N., Mokhtar, H. M. O., & Zidane, A. (2022). *SML-AutoML: A Smart Meta-Learning Automated Machine Learning Framework*. research square platform llc. https://doi.org/10.21203/rs.3.rs-2085778/v1

Grina, F., Lefevre, E., & Elouedi, Z. (2021). *Evidential Undersampling Approach for Imbalanced Datasets with Class-Overlapping and Noise* (pp. 181–192). https://doi.org/10.1007/978-3-030-85529-1_15

Haghighat, P., Anahideh, H., Kang, L., & Gándara, D. (2024). Fair Multivariate Adaptive Regression Splines for Ensuring Equity and Transparency. *Proceedings of the AAAI Conference on Artificial Intelligence*, *38*(20), 22076–22086. https://doi.org/10.1609/aaai.v38i20.30211

Heidemann, A., Hülter, S. M., & Tekieli, M. (2024). Machine learning with real-world HR data: mitigating the trade-off between predictive performance and transparency. *The International Journal of Human Resource Management*, *35*(14), 2343–2366. https://doi.org/10.1080/09585192.2024.2335515

Hong, S., & Lynn, H. S. (2020). Accuracy of random-forest-based imputation of missing data in the presence of non-normality, non-linearity, and interaction. *BMC Medical Research Methodology*, *20*(1). https://doi.org/10.1186/s12874-020-01080-1

Hong, S., Sun, Y., Li, H., & Lynn, H. (2020). *Influence of parallel computing strategies of iterative imputation of missing data: a case study on missForest*. https://doi.org/10.48550/arxiv.2004.11195

Hutiri, W. (Toussaint), Kawsar, F., Mathur, A., & Ding, A. Y. (2023). Tiny, Always-on, and Fragile: Bias Propagation through Design Choices in On-device Machine Learning Workflows. *ACM Transactions on Software Engineering and Methodology*, *32*(6), 1–37. https://doi.org/10.1145/3591867

Jacobucci, R., & Li, X. (2022). Does Minority Case Sampling Improve Performance with Imbalanced Outcomes in Psychological Research? *Journal of Behavioral Data Science*, *2*(1), 59–74. https://doi.org/10.35566/jbds/v2n1/p3

Jacobusse, G., & Veenman, C. (2016). On selection bias with imbalanced classes. In *Discovery Science: 19th International Conference, DS 2016, Bari, Italy, October 19–21, 2016, Proceedings 19* (pp. 325-340). Springer International Publishing.

Jadhav, A., Pramod, D., & Ramanathan, K. (2019). Comparison of performance of data imputation methods for numeric dataset. *Applied Artificial Intelligence*, *33*(10), 913-933.

Jeong, H., Wang, H., & Calmon, F. P. (2022, June). Fairness without imputation: A decision tree approach for fair prediction with missing values. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 36, No. 9, pp. 9558-9566).

Katam, B. R. (2024). Optimizing Data Pipeline Efficiency with Machine Learning Techniques. *INTERANTIONAL JOURNAL OF SCIENTIFIC RESEARCH IN ENGINEERING AND MANAGEMENT*, *08*(07), 1–15. https://doi.org/10.55041/ijsrem36850

Kearns, M., Neel, S., Roth, A., & Wu, Z. S. (2019, January). An empirical study of rich subgroup fairness for machine learning. In Proceedings of the conference on fairness, accountability, and transparency (pp. 100-109).

Kenfack, P. J., Khan, A. M., Kazmi, S. M. A., Hussain, R., Khattak, A. M., & Oracevic, A. (2021). *Impact of Model Ensemble On the Fairness of Classifiers in Machine Learning*. *12*, 1–6. https://doi.org/10.1109/icapai49758.2021.9462068

Khan, T. A., Maqbool, T., Hamid, W., & Jahangir, M. S. (2020). Disabled Students Seeking Higher Education in Kashmir: A Study of their Experiences. *Higher Education for the Future*, *7*(2), 132–146. https://doi.org/10.1177/2347631120932241

Khodadadian, S., Nafea, M., Ghassami, A., & Kiyavash, N. (2021). Information Theoretic Measures for Fairness-aware Feature Selection. *ArXiv*, abs/2106.00772.

Kim, M. (2002). Has the Race Penalty for Black Women Disappeared in the United States? *Feminist Economics*, *8*(2), 115–124. https://doi.org/10.1080/13545700210160997

Klaassen, S., Bach, P., Schacht, O., Chernozhukov, V., & Spindler, M. (2024). *Hyperparameter Tuning for Causal Inference with Double Machine Learning: A Simulation Study*. https://doi.org/10.48550/arxiv.2402.04674

Kodama, S., Fujihara, K., Kato, K., Matsuzaka, T., Sone, H., Watanabe, K., Kitazawa, M., Shimano, H., Horikawa, C., Iwanaga, M., & Nakagawa, Y. (2022). Predictive ability of current machine learning algorithms for type 2 diabetes mellitus: A meta-analysis. *Journal of Diabetes Investigation*, *13*(5), 900–908. https://doi.org/10.1111/jdi.13736

Kumar, A., Tizpaz-Niari, S., Trivedi, A., & Tan, G. (2022). *Fairness-aware Configuration of Machine Learning Libraries*. https://doi.org/10.48550/arxiv.2202.06196

Kunft, A., Rabl, T., Schelter, S., Markl, V., Breß, S., & Katsifodimos, A. (2019). An intermediate representation for optimizing machine learning pipelines. *Proceedings of the VLDB Endowment*, *12*(11), 1553–1567. https://doi.org/10.14778/3342263.3342633

Letterman, M. R., Clifford, M. T., & Brown, J. L. (2018). Major Choice and the Wage Differential between Black and White Women. *Journal of Applied Social Science*, *12*(2), 145–163. https://doi.org/10.1177/1936724418785411

Li, Y., & Chen, W. (2020). A Comparative Performance Assessment of Ensemble Learning for Credit Scoring. *Mathematics*, *8*(10), 1756. https://doi.org/10.3390/math8101756

Li, Y., Wang, H., & Luo, Y. (2022). Improving Fairness in the Prediction of Heart Failure Length of Stay and Mortality by Integrating Social Determinants of Health. *Circulation: Heart Failure*, *15*(11). https://doi.org/10.1161/circheartfailure.122.009473

Liang, Y., Shu, K., Tian, T., & Chen, C. (2023). Fair classification via domain adaptation: A dual adversarial learning approach. *Frontiers in Big Data*, *5*. https://doi.org/10.3389/fdata.2022.1049565

Liao, C. (2023, February). Employee turnover prediction using machine learning models. In *International Conference on Mechatronics Engineering and Artificial Intelligence (MEAI 2022)*(Vol. 12596, pp. 227-231). SPIE.

Luengo, V., Lallé, S., Verger, M., & Bouchet, F. (2023). *Is Your Model "MADD"? A Novel Metric to Evaluate Algorithmic Fairness for Predictive Student Models*. cornell university. https://doi.org/10.48550/arxiv.2305.15342

Mahesh, T. R., Geman, O., Margala, M., & Guduri, M. (2023). The stratified K-folds cross-validation and class-balancing methods with high-performance ensemble classifiers for breast cancer classification. *Healthcare Analytics*, *4*, 100247.

Mansoor, H., Ali, S., Alam, S., Hassan, U., Khan, M., & Khan, I. (2022). *Impact Of Missing Data Imputation On The Fairness And Accuracy Of Graph Node Classifiers*. cornell university. https://doi.org/10.48550/arxiv.2211.00783

Manzoor, M. B. (2023). EDUCATION AND DISABILITY: A STUDY ON ACCESS TO HIGHER EDUCATION FOR STUDENTS WITH DISABILITY IN JAMMU AND KASHMIR. *Towards Excellence*, 507–529. https://doi.org/10.37867/te150344

Mccarthy, M. B., & Narayanan, S. (2023). Fairness–accuracy tradeoff: activation function choice in a neural network. *AI and Ethics*, *3*(4), 1423–1432. https://doi.org/10.1007/s43681-022-00250-9

Meda, S., & Bhogapathi, R. (2022). An Integrated Machine Learning Model for Heart Disease Classification and Categorization. *Journal of Computer Science*, *18*(4), 257–265. https://doi.org/10.3844/jcssp.2022.257.265

Morina, G., Oliinyk, V., Waton, J., Marusic, I., & Georgatzis, K. (2019). Auditing and achieving intersectional fairness in classification problems. *arXiv preprint arXiv:1911.01468*.

Naboureh, A., Bian, J., Amani, M., Lei, G., & Li, A. (2020). A Hybrid Data Balancing Method for Classification of Imbalanced Training Data within Google Earth Engine: Case Studies from Mountainous Regions. *Remote Sensing*, *12*(20), 3301. https://doi.org/10.3390/rs12203301

Nagaraj, N., & Ghosh, T. (2024). *Evaluating the Determinants of Mode Choice Using Statistical and Machine Learning Techniques in the Indian Megacity of Bengaluru*. https://doi.org/10.48550/arxiv.2401.13977

Neves, D. T., Alves, J., Naik, M. G., Proença, A. J., & Prasser, F. (2022). From Missing Data Imputation to Data Generation. *Journal of Computational Science*, *61*, 101640. https://doi.org/10.1016/j.jocs.2022.101640

Nezami, N., Haghighat, P., Anahideh, H., & Gándara, D. (2024). Assessing Disparities in Predictive Modeling Outcomes for College Student Success: The Impact of Imputation Techniques on Model Performance and Fairness. *Education Sciences*, *14*(2), 136. https://doi.org/10.3390/educsci14020136

Nguyen, D., Gupta, S., Rana, S., Shilton, A., & Venkatesh, S. (2021). Fairness improvement for black-box classifiers with Gaussian process. *Information Sciences*, *576*, 542–556. https://doi.org/10.1016/j.ins.2021.06.095

Oneto, L., & Chiappa, S. (2020). Fairness in machine learning. In Recent trends in learning from data: Tutorials from the inns big data and deep learning conference (innsbddl2019) (pp. 155-196). Springer International Publishing.

Pessach, D., & Shmueli, E. (2022). A review on fairness in machine learning. *ACM Computing Surveys (CSUR)*, *55*(3), 1-44.

Poulos, J., & Valle, R. (2018). Missing Data Imputation for Supervised Learning. *Applied Artificial Intelligence*, *32*(2), 186–196. https://doi.org/10.1080/08839514.2018.1448143

Radovanović, S., Suknović, M., Petrović, A., & Delibašić, B. (2021). A fair classifier chain for multi-label bank marketing strategy classification. *International Transactions in Operational Research*, *30*(3), 1320–1339. https://doi.org/10.1111/itor.13059

Raghavan, M., & Kim, P. T. (2024). Limitations of the" Four-Fifths Rule" and Statistical Parity Tests for Measuring Fairness. *Geo. L. Tech. Rev.*, *8*, 93.

Rana, D. K. (2024). Quality Education for Underrepresented Groups: Bridging the Gap. *International Journal of English Literature and Social Sciences*, *9*(1), 212–219. https://doi.org/10.22161/ijels.91.28

Rohani, A. (2021). *Bias measurement in small datasets* [northeastern university library]. https://doi.org/10.17760/d20403633

Salazar, J. J., Garland, L., Ochoa, J., & Pyrcz, M. J. (2022). Fair train-test split in machine learning: Mitigating spatial autocorrelation for improved prediction accuracy. *Journal of Petroleum Science and Engineering*, *209*, 109885.

Schelter, S., Karlas, B., Zhang, C., Guha, S., & Grafberger, S. (2023). *Proactively Screening Machine Learning Pipelines with ARGUSEYES*. *41*, 91–94. https://doi.org/10.1145/3555041.3589682

Shukla, A. K. (2020). *Patient Diabetes Forecasting Based on Machine Learning Approach* (pp. 1017–1027). springer singapore. https://doi.org/10.1007/978-981-15-4032-5_91

Srivastava, M., Heidari, H., & Krause, A. (2019, July). Mathematical notions vs. human perception of fairness: A descriptive approach to fairness for machine learning. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining* (pp. 2459-2468).

Stekhoven, D. J., & Bühlmann, P. (2011). MissForest—non-parametric missing value imputation for mixed-type data. *Bioinformatics*, *28*(1), 112–118. https://doi.org/10.1093/bioinformatics/btr597

Sun, Y., Fung, B. C. M., & Haghighat, F. (2022). In-Processing fairness improvement methods for regression Data-Driven building Models: Achieving uniform energy prediction. *Energy and Buildings*, *277*, 112565. https://doi.org/10.1016/j.enbuild.2022.112565

Tizpaz-Niari, S., Kumar, A., Tan, G., & Trivedi, A. (2022, May). Fairness-aware configuration of machine learning libraries. In Proceedings of the 44th International Conference on Software Engineering (pp. 909-920).

Tubella, A. A., Barsotti, F., Koçer, R. G., & Mendez, J. A. (2022). Ethical implications of fairness interventions: what might be hidden behind engineering choices?. *Ethics Inf. Technol.*, *24*(1), 12.

Ugirumurera, J., Bensen, E. A., Severino, J., & Sanyal, J. (2024). Addressing bias in bagging and boosting regression models. *Scientific Reports*, *14*(1). https://doi.org/10.1038/s41598-024-68907-5

Vaidyanathan, S. G., Kar, B., & Kumaravel, N. (2008). A regression based approach to a maximum margin classifier for separation of linearly inseparable pattern classes. *Journal of Interdisciplinary Mathematics*, *11*(2), 237–245. https://doi.org/10.1080/09720502.2008.10700555

Van Giffen, B., Herhausen, D., & Fahse, T. (2022). Overcoming the pitfalls and perils of algorithms: A classification of machine learning biases and mitigation methods. *Journal of Business Research*, *144*, 93-106.

Veale, M. (2017). *Logics and practices of transparency and opacity in real-world applications of public sector machine learning*. center for open science. https://doi.org/10.31235/osf.io/6cdhe

Verma, S., & Rubin, J. (2018, May). Fairness definitions explained. In *Proceedings of the international workshop on software fairness* (pp. 1-7).

Vigdor, N. (2019). *Apple Card Investigated after Gender Discrimination Complaints*. Retrieved from https://www.nytimes.com/2019/11/10/business/Apple-credit-cardinvestigation.html. Accessed January 6, 2022.

Viloria, A., Lezama, O. B. P., & Mercado-Caruzo, N. (2020). Unbalanced data processing using oversampling: machine learning. *Procedia Computer Science*, *175*, 108-113.

Wan, M., Zha, D., Liu, N., & Zou, N. (2023). In-processing modeling techniques for machine learning fairness: A survey. *ACM Transactions on Knowledge Discovery from Data*, *17*(3), 1-27.

Wang, H. E., Saria, S., Kharrazi, H., & Weiner, J. P. (2023). Evaluating Algorithmic Bias in 30-Day Hospital Readmission Models: Retrospective Analysis. *Journal of Medical Internet Research*, *26*, e47125. https://doi.org/10.2196/47125

Wang, H., Cheng, Z., Smyth, R., Sun, G., Li, J., & Wang, W. (2022). University education, homeownership and housing wealth. *China Economic Review*, *71*, 101742. https://doi.org/10.1016/j.chieco.2021.101742

Wu, Y., Hu, W., Qi, X., Si, S., & Zhang, Z. (2024). Prediction of flood sensitivity based on Logistic Regression, eXtreme Gradient Boosting, and Random Forest modeling methods. *Water Science and Technology : A Journal of the International Association on Water Pollution Research*, *89*(10). https://doi.org/10.2166/wst.2024.146

Yan, B., Seto, S., & Apostoloff, N. (2022). *FORML: Learning to Reweight Data for Fairness*. cornell university. https://doi.org/10.48550/arxiv.2202.01719

Yao, Y. (2017). Accounting for the Decline in Homeownership Among the Young. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3050714

Yaseliani, M., Noor-E-Alam, M., & Hasan, M. M. (2024). Mitigating Sociodemographic Bias in Opioid Use Disorder Prediction: Fairness-Aware Machine Learning Framework. *JMIR AI*, *3*, e55820. https://doi.org/10.2196/55820

Yu, N., Tan, G., & Tizpaz-Niari, S. (2023). *FairLay-ML: Intuitive Remedies for Unfairness in Data-Driven Social-Critical Algorithms*. https://doi.org/10.48550/arxiv.2307.05029

Zhang, W., Wu, C., Zhong, H., Li, Y., & Wang, L. (2020). Prediction of undrained shear strength using extreme gradient boosting and random forest based on Bayesian optimization. *Geoscience Frontiers*, *12*(1), 469–477. https://doi.org/10.1016/j.gsf.2020.03.007

Appendix

Appendix A – R code for simulation and analysis

The R code for the data simulation and analysis is available through GitHub under the following link:

https://github.com/Hectine/Thesis-Fair-ML.git

The repository contains all the code to run the detailed simulation and analysis. Furthermore, the four simulated societies and their output for their 30 individual pipelines are saves as CSVs.