Master thesis on Intelligent Interactive Systems

Universitat Pompeu Fabra

An Interactive Framework to Evaluate Algorithmic Discrimination

Aideen Farrell

Supervisor: Carlos Castillo

September 2020

upf. Universitat Pompeu Fabra *Barcelona*

Master thesis on Intelligent Interactive Systems

Universitat Pompeu Fabra

An Interactive Framework to Evaluate Algorithmic Discrimination

Aideen Farrell

Supervisor: Carlos Castillo

September 2020

upf. Universitat
Pompeu Fabra
Barcelona

Contents

Abstract

Machine-learning decision-making model and ranking algorithms are used daily for a variety of tasks. They make predictions and decisions about our potential abilities and trustworthiness. The decision could be a simple yes/no, or a score intended to rank potential to succeed, in a job, an educational setting, as a law-abiding citizen, in the repayment of a loan and so on. Such decisions often have lifelong consequences for individuals, groups and society as a whole. The quality of any ML model is reliant upon the accuracy of the training data and the previous outcomes or decisions reflected therein. Suppose we accept that society is prone to bias, and this bias is mirrored in outcomes and decisions. In that case, we must also accept that any ML model trained to assimilate historically 'optimal' human decisions and outcomes will also assimilate the outright, subtle or pervasive prejudices entrenched within.

The first step towards mitigating prejudice is by identifying if, where, and why disparities exist across subpopulation in the data or any resulting model or ranking output. This work focuses on providing a framework to reflect upon fairness as it pertains to a particular ML implementation and subpopulation, through the frame of a context-driven worldview. It will be argued that the various existing mathematical and statistical measurements of fairness, although useful and necessary, do not translate well as a basis for clear and meaningful conversations between all stakeholders in the ML delivery pipeline. Neither are they conducive to an easily interpretable statement of accountability. To this end, an analysis was performed on the law school dataset through the proposed framework, and the results presented.

Keywords: Ranking; Bias; Fairness; Data Science; Ranking; Positional bias; Equal opportunity; Discrimination; Toolkit, Accountability; Worldview

Chapter 1

Introduction

1.1. Motivation

As the tools necessary to develop, train and deploy ML models become increasingly accessible through pre-trained models, libraries and off the shelf tools for training and deploying models, such as *Azure ML*[1], *IBM Watson Studio*[2], *Amazon SageMaker*[3] the possibility of encountering an ML model created with no oversight from a domain expert or data scientists becomes increasingly likely, as does the potential for the increased oversight of unfairness against various subpopulations. Even with the collaboration of a data scientist, there are many fairness definitions to consider. Companies need to balance their responsibilities to employees, customers, the community, and shareholders, especially when they conflict. The decisions around fairness, therefore, require collaboration. At the same time, a predictive model can have severe positive and negative consequences for individuals as well as for groups and entire societies. Therefore, the concept of fairness needs to be taken seriously.

As fairness is a social construct with no single agreed-upon definition, an active discussion is necessary between all stakeholders in the project delivery pipeline. The incentive has been to bridge the gap between the mathematical and academic definitions of bias, fairness and unfairness which data scientists and academics may be familiar with, and the real-world sentiments of fairness that a bias mitigation strategy (or lack thereof) signifies in terms of company values.

[1] https://azure.microsoft.com/en-us/services/machine-learning/

[2] https://www.ibm.com/uk-en/cloud/machine-learning

[3] https://aws.amazon.com/sagemaker/

1.2. Objectives

Many mathematical and statistical definitions of fairness exist, as do several techniques to diminish the impact of unfairness through the lens of each definition. As discussed in *Improving fairness in machine learning systems: What do industry practitioners need?* [1], confusion still exists in terms of the specific measurement of fairness to consider. Fairness is a complex topic with many interpretations and legal considerations. The very same model can also serve a variety of needs, makes decisions around fairness unique to each implementation. The objective is the development of a framework utilising statistical and mathematical interpretations of fairness while framing the discussion in the context of a ‘worldview’. As an algorithm does not have the luxury of critical thinking, and cannot independently develop its own beliefs, it is the responsibility of all those involved in delivering such solutions to be intentional in their effort. Furthermore, there should be a conscious effort to ensure the model reflects the values that the organisation collectively believes in, that the customer of the application expects and accepts, and that these values are transparent to society and the humans who will be ranked by the model.

1.3. Structure

Chapter 1: Introduction, provides context around the topic and a brief description of motivations and objectives

Chapter 2: Related work, reviews areas of related work and existing frameworks.

Chapter 3: Framework, provide an overview of the proposed framework and the concepts that have motivated it.

Chapter 4: Implementation, outlines the technical details of the implementation and distribution of the framework and the location of the code.

Chapter 5: Use Case, outlines the analysis of a dataset using the framework followed by the analysis of a logistic regression model for fairness.

Chapter 6: Conclusion, includes discussion and future work.

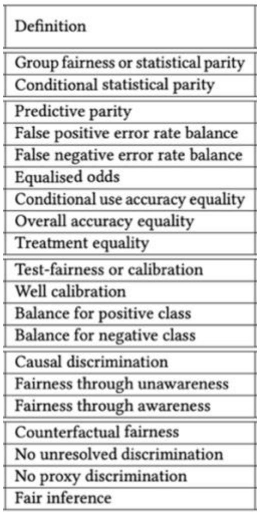
Chapter 2

Related Work

The purpose of this chapter is to describe the related research areas and place the proposed framework in context. Reviewing problems raised in the field and work which proposes different or similar methods to solve the problem. It is divided into sections 2.1, describing research in the field and section 2.2, describing existing conceptual and physical frameworks.

2.1. Research

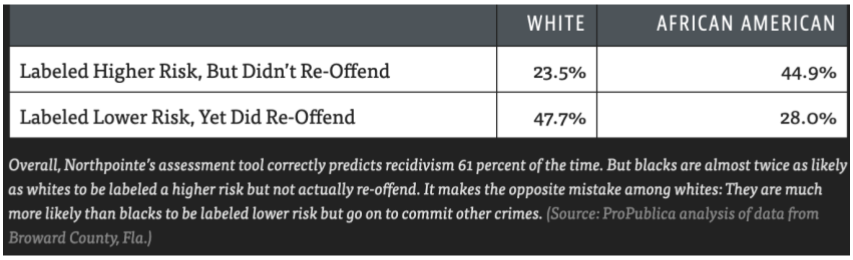
Much work has been carried out related to the detection and mitigation of bias at the various stage of the ML pipeline, with a particular focus on the learning algorithms and ranking algorithms [39, 40, 41, 42]. *Holstein et al.* [1] focused their research on what industry practitioners need. It was identified that many challenges are faced not just by ML practitioners but also by the commercial product teams who are aware that their products have real-world outcomes. Through investigations carried out by the authors, it was noted that both alignment and the necessity for collaborative discussions across these organisations is often an issue. It was also noted that, as the focus from the research community is primarily fixed on de-biasing methods, the data itself is often considered to be fixed. At the same time, practitioners believe the data warrants more consideration. Another area identified to be of concern was the bias blind spots of the practitioners working on the various stages of the delivery pipeline. *Mehrabi et al.* [43] outline twenty-three definitions of bias, while concluding that lack of synthesis in the definitions creates an impossibility of understanding and is as such an open research problem, as too is the detection of bias in the dataset. While in *Fairness definitions explained* [3], as illustrated in Figure 1, there were also over twenty mathematical definitions of fairness identified. Fulfilling one often has the consequence of ensuring that the other cannot be satisfied. This is a relatively intuitive observation when considering group vs individual fairness. However, this apparent conflict is repeated across several definitions such that any particular ML implementation can be considered unfair according to one definition and fair according to the next.



Definition
Group fairness or statistical parity
Conditional statistical parity
Predictive parity
False positive error rate balance
False negative error rate balance
Equalised odds
Conditional use accuracy equality
Overall accuracy equality
Treatment equality
Test-fairness or calibration
Well calibration
Balance for positive class
Balance for negative class
Causal discrimination
Fairness through unawareness
Fairness through awareness
Counterfactual fairness
No unresolved discrimination
No proxy discrimination
Fair inference

Figure 1: Definitions of fairness (from Verma and Rubin 'Fairness definitions explained' [3])

This conflict has been quite apparent in the 2016 report [27][28] published by ProPublica concerning COMPAS, a recidivism algorithm used by the U.S. courts to assess the likelihood of a defendant re-offending. One observation in the report focused on fairness through the frame of *false-positive rates*. According to the ProPublica report, and as can be seen in *Figure 2*, the algorithm produced a much higher false-positive rate for Black people than it did for White.



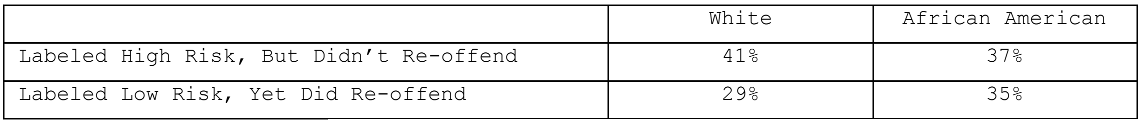
	WHITE	AFRICAN AMERICAN
Labeled Higher Risk, But Didn't Re-Offend	23.5%	44.9%
Labeled Lower Risk, Yet Did Re-Offend	47.7%	28.0%

Overall, Northpointe's assessment tool correctly predicts recidivism 61 percent of the time. But blacks are almost twice as likely as whites to be labeled a higher risk but not actually re-offend. It makes the opposite mistake among whites: They are much more likely than blacks to be labeled lower risk but go on to commit other crimes. (Source: ProPublica analysis of data from Broward County, Fla.)

Figure 2: ProPublica report 2016 [27]

During testing (when the model's performance was assessed using known outcomes) the model incorrectly predicted that Black people would re-offend, when in fact they had not, at a rate of 44.9%. This incorrect prediction occurred at a rate of 23.5% for White people. An incorrect prediction of *will re-offend* could have serious real-world consequences for those being judged. Pretrial incarceration imposes high costs on individuals, as evident from the case of *Kalief Browder* [28] who was sixteen when he was accused of stealing a backpack. Although he was never convicted of the crime, he spent almost three years on Rikers Island awaiting trial because he could not afford to pay the 3,000USD bail.

Northpointe (now Equivant), the manufacturer of the algorithm, released a response [29] addressing ProPublica's analysis. Northpointe sited the values, as seen in Figure 3.



	White	African American
Labeled High Risk, But Didn't Re-offend	41%	37%
Labeled Low Risk, Yet Did Re-offend	29%	35%

Figure 3: What Northpointe refer to as the 'corrected table' in response [29] to ProPublica report.

While the tables in both Figure 2 and Figure 3 use the terminology '*Labeled High Risk, but did not re-offend*' and '*Labeled Low Risk, but did not re-offend'* the underlying mathematical equations and concepts used to obtain the results are entirely different. Northpointe calculated the true positive rate, i.e. the measure of how well the model correctly predicted that a person had re-offended when they had re-offend. For white people, there is a 100 - 29 = 71%, chance that the model will predict this correctly, and for Black people, there is a 100 - 35 = 75% chance. Conversely, the false positive rate (negative predictive value) is a measure of how correctly the model predicted that a person would not re-offend when they had not re-offended. In this case, there was a rate of 100 - 41 = 59% for White people and of 100 – 37 = 63% for Black people. If these four measurements are compared from a natural language perspective it can be determined what the model is 'saying' about its prediction making capabilities as it relates to a Black person and a White person.

Table 1: Natural language interpretation of Northpointe and ProPublica's statements.

	Black	**White**
Northpointe	If you are really going to re-offend there is a 75% chance that I will guess this correctly based upon the personal information I have about you.	If you are really going to re-offend there is a 71% chance that I will guess that correctly based upon the personal information I have about you.
Northpointe	If you are really not going to re-offend there is a 63% chance that I will guess that correctly based upon the personal information I have about you.	If you are really not going to re-offend there is a 59% chance that I will guess that correctly based upon the personal information I have about you.

ProPublica	There is a 44.9% chance I will incorrectly guess that you will re-offend, the consequence being that you will receive a harsher judgment you do not deserve.	There is a 23% chance I will incorrectly guess that you will re-offend, the consequence being that you will receive a harsher judgment you do not deserve.
ProPublica	There is a 28% chance I will incorrectly guess that you will not re-offend the consequence being that you will receive a leniency you apparently do not deserve.	There is a 47% chance I will incorrectly guess that you will not re-offend the consequence being that you will receive a leniency you apparently do not deserve.

Reviewing the natural language interpretation of the results, in conjunction with the already identified disparities in parole and sentencing decisions across racial groups [30] gives a clearer picture of why the difference in these two interpretations of fairness is so significant. It also becomes apparent that what differentiates Northpointe from ProPublica is the interpretations of fairness. Both interpretations can be mathematically justified, however both place importance on different aspects of real-world outcomes. There is a different *worldview* at play.

Concepts of fairness as it relates to machine learning have largely centred around the topics of individual vs group fairness. As a result, these concepts have been addressed extensively in the field since they were first discussed by *Dwork et al.* in *Fairness Through Awareness* [2]. However, in the paper *On the Apparent Conflict Between Individual and Group Fairness* [32], it is argued that individual and group fairness are not fundamentally in conflict when viewed through the lens of political and legal philosophy. It is proposed that any conflict between individual and group fairness may lurk in the failure to fully describe the assumptions behind them and the reasons for applying them in a particular context. Individual fairness, as per the definition provided in this thesis, relies on the assumption that we have *captured* and *correctly quantified* those characteristics of a person which *matter* within the context of decision when comparing individuals. However, as mentioned by the author in the paper, both group and individual fairness *‘fail to satisfy the principle of individual justice (despite that principle’s surface-level similarities with individual fairness)’* [32].

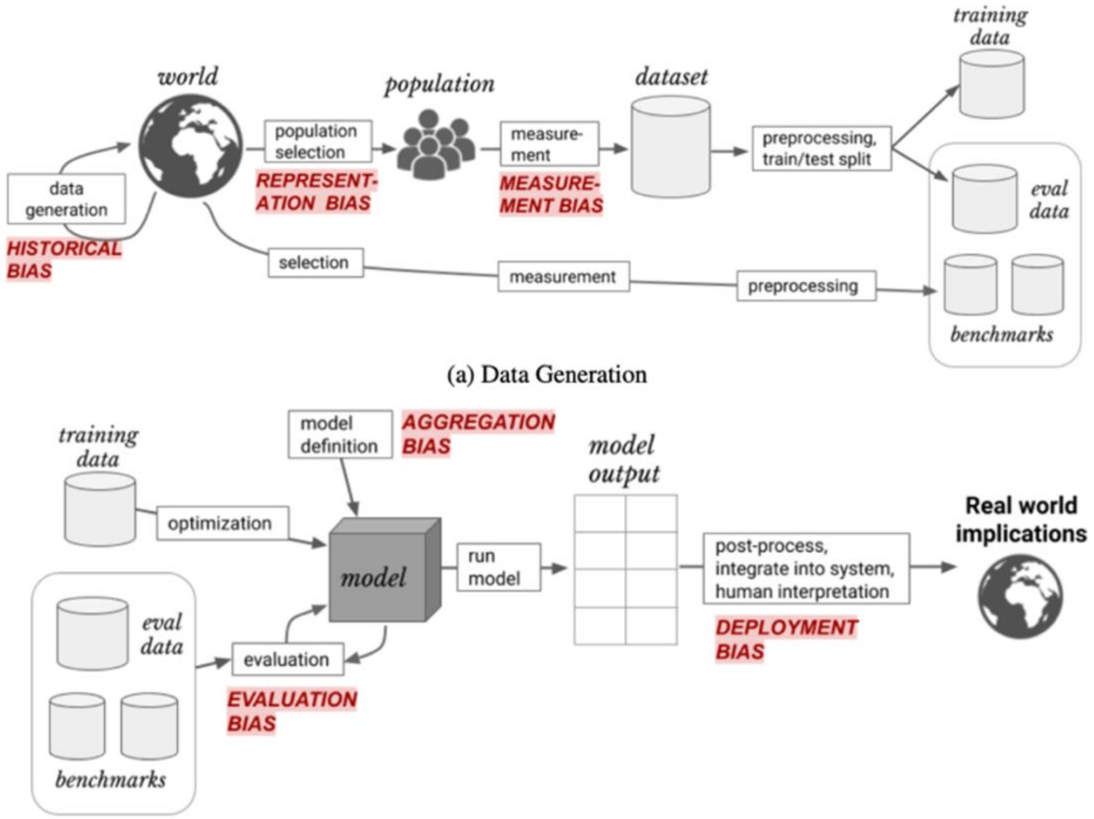
Friedler et al. noted in *on the (im)possibility of fairness* [33], that these quantifiable features of a person do not necessarily reflect the *true* features of a person, which are often not observable. Besides, there may not even be enough measurable and observable features to make accurate decisions or predictions between individuals. Notably, they also designed a mathematical expression for two worldviews. The first, '*What you see is what you get (WYSIWYG)*', posits that whatever the data reflects constitutes an objective picture of the world, even if inequalities appear in the results. The second '*We're all equal (WAE)',* despite its naming convention which may give the impression that it refers to individual fairness, posits that data is not inherently objective. A WAE worldview supposes that data does not account for the subjective, bias entrenched influence under which the data was created. WAE, therefore, entails that there be a mathematical compensation for any significant discrepancies in output. *Yeom and Tschantz*[52] further mathematically compare three definitions of group fairness (demographic parity, equalized odds, and predictive parity) through the lense of the theoretical framework[33] referencing the criterion of '*disparity amplification'* to argue that the difference between demographic parity and equalized odds can be reduced to the selected worldview.

In Counterfactual fairness [34], *Kushner et al.* speak on the topic of causal relationships. They describe counterfactual fairness as that which captures *'the intuition that a decision is fair towards an individual if it is the same in (a) the actual world and (b) a counterfactual world where the individual belonged to a different demographic group'.* They too are veering into the area of worldviews, with the assumption that in an unfair world there are other non-quantifiable, non-visible, causal features at play, such as social biases. Counterfactual fairness recognises that these unquantified variables have influenced the quantified features. Using a causal framework and knowledge of the population and domain in which any resulting ML model will be deployed a relationship between the unquantified causal feature(s) and the known and quantified input features is modelled. This, in effect, promotes reflection with the objective of forming a more inclusive worldview of the problem at hand. Any features which the newly introduced causal features have had any influence upon are then assessed as candidates for removal from the training data. This could be considered a type of fairness driven feature engineering. This approach, however, can easily result in the removal of too many, or even all of the available input features. To cater for this scenario an alternative approach is taken which involves a calculation of the potential effect of this unquantified causal feature, by examining the differences in the distribution of outcomes across groups, and an adjustment of the model to ‘cancel out’ this effect.

In *Troubling trends in machine-learning scholarship* [35], the authors identify four patterns which have been observed in the academics of machine learning. Two of which may compound the challenges faced in any attempt to bridge the gaps between the data scientists or engineer working on a particular ML implementation and the broader collection of stakeholders and policymakers at the organisation level. These patterns have been identified as ‘*Mathiness’* and ‘*Misuse of language’.* They hold particular relevance in the area of fair ML where, in addition to the standard measurements of accuracy, exist several mathematical definitions of various interpretations of fairness. This ‘mathiness’, as apparent in the COMPAS example, can allow organisations to hide behind the maths of fairness rather than facilitating full transparency with a natural language statement of the *worldview* that both the company and the ML model are entrenched in. Even with the best of intentions, the heavy mathematical based discussions around fairness often obfuscates the *real-world* point of view that comes hand in hand with the desire to satisfy a particular measurement of fairness. The author also points out that within the Fairness ML literature terminology that has originated from complex legal doctrine is often appropriated, resulting in *‘a literature where “fairness,” “opportunity,” and “discrimination” denote simple statistics of predictive models, confusing researchers who become oblivious to the difference and policymakers who become misinformed about the ease of incorporating ethical desiderata into ML.’* This inevitably leads to confusion when definitions are altered to fit within the mathematical context of machine learning.

2.2. Existing frameworks

In *A Framework for Understanding Unintended Consequences of Machine Learning* [36], the authors provide a conceptual framework that partitions potential areas of bias into six categories. The purpose of the framework is to facilitate the development of solutions that ‘*stem from an understanding of application-specific populations and data. generation processes, rather than relying on general statements about what may or may not be “fair.”*. However, there is no practical implementation of the concept.



(a) Data Generation

Figure 4: Model building and implementation, from A Framework for Understanding Unintended Consequences of Machine Learning [36]

Many machine learning fairness toolkits exist such as *AIF360*[45][1] which provides a library of fairness metrics for integration into the model development pipeline. *Aequitas*[2] provides an API for reviewing fairness metrics in much the same way as *AIF360*. It also provides a web-based UI which can be used to upload a trained model’s true (y) and predicted (ŷ) values for evaluation of quality over groups, with no necessity for coding and the minimum necessity for column preparation supporting binary and multi-class classifiers. *Fairlearn*[3] also provides a fairness library for integration into the development pipeline, similar to Aequitas, it provides an *IpyWidgets*[4] based UI for the evaluation of

[1] https://github.com/Trusted-AI/AIF360

[2] http://aequitas.dssg.io/, https://github.com/dssg/aequitas

[3] https://github.com/fairlearn/fairlearn

[4] https://ipywidgets.readthedocs.io/en/latest/developer_docs.html

the models true and predicted values over groups. *MithraRanking*[1] supports the evaluation of ranked lists; however, it is fully web-based and no longer appears to be operational. The *what-if*[2] tool can also be used to analyse a model after it has been created. From explainability point of view, some common tools are *SHAP*[3], *LIME*[4] and *Microsoft's InterpretML*[5]. The Canadian government also launched an algorithmic impact assessment[6] which, although not explicitly referencing the ideas of worldview does require reflection upon the training data and the context within which any resulting model will be used. Its creation was driven by the need for a centralised framework to allow technical and non-technical people to have open and transparent discussions about the social impacts of solutions. There is, however, no interaction with data but rather a manual audit involving a series of questions.

[1] http://mithra.eecs.umich.edu/demo/ranking/

[2] https://pair-code.github.io/what-if-tool/get-started/

[3] https://github.com/slundberg/shap

[4] https://github.com/marcotcr/lime

[5] https://interpret.ml/

[6] https://dataresponsibly.github.io/documents/mde783-guanA.pdf

Chapter 3

Framework

The purpose of this chapter is to provide an overview of the proposed framework and the concepts that have motivated it. It is divided into 3 sections. Section 3.1 outlines some definitions useful for the discussion. Section 3.2 provides a background of both the ranking of humans and prejudice in general. 3.3 outlines the proposed framework where the discussion is further divided into three sections. 3.3.1 discusses the high-level concept of the framework, 3.3.2 introduces the concept of the fairness worldview. 3.3.3 provides a discussion on counterfactual thinking and finally 3.4 briefly discusses strategies to increase adoption.

3.1. Definitions

Before continuing the discussion, it is practical to define some of the terms that will be used throughout. The focus of the thesis is on the identification of bias in machine learning models which rank, score or otherwise make decisions about humans. The framework delivered as part of this work supports supervised learning models using either binary classification or regression. It is within this context that the definitions are given.

Artificial Intelligence*:* is the human-like intelligence demonstrated by automated systems. It is a vast field. Machine learning is a subfield of *artificial intelligence* which is premised on learning patterns from *data*. Machine learning has close ties to data science and statistics.

Classification and Regression Algorithms*:* are types of supervised machine learning algorithms. Classification algorithms make predictions that are restricted to discrete numbers. Binary classifiers are further restricted to two discrete classes (*0 and 1*). Regression algorithms make predictions in the form of a continuous number within a range. A regression algorithm can naturally be used for ranking due to the continuous nature of the output. A probabilistic binary classifier, which predicts the probability distribution across classes (rather than merely outputting the most likely class) can also be used for the purpose of ranking, where ranking is based on the probability of belonging to Class 1.

Target**:** the data used to train supervised ML models is labelled; each data point comprises of an input and the known corresponding output, known as the Target. The proximity of each prediction to this target determines the models perceived accuracy.

Features: features are measurable and quantifiable properties of input dataused to train models via learning algorithms. When we speak about features in terms of models used to rank humans, these features will likely consist of easily quantifiable human attributes such as height, weight, age or perhaps self-identified preferences. Arbitrary features such as name, address, marital status, income or similar might also be present. Measurements of other complex human characteristics such as intelligence, resilience, potential, trustworthiness are often converted, by one means or another, to abstract quantities. Measurements may also exist which have been directly influenced by the judgment of others such as prison terms, the number of times stopped by the police, performance review results (meets expectations, above expectation) to name but a few. It is essential to keep in mind that the quantifiable features related to a human are frequently assumed to be objective when, as a matter of fact, they are often influenced by circumstance and subjectivity. There should be no immediate assumption that those characteristics of a person which really matter within the context of a decision-making system have been captured correctly or fairly quantified.

Protected Group*:* a group is a collection of individuals who share specific characteristics. Protected groups are groups of people who often qualify for additional protection by a law, policy, or other authority in order to prevent discrimination or harassment based on their membership of the group in question. Examples of protected groups include but is not limited to those of a specific race, sex, national origin, age, religion, disability, or colour. Most countries have discrimination laws. However, fairness may be extended beyond the confines of the law and applied for social and ethical purposes.

Protected Feature: the protected feature is the feature containing the values which indicate whether a samples data-point is referencing a member of a protected group or not. For example, race might be a protected feature with possible values {White, Black, Asian}. Depending on the context, black, might be the protected group. Gender might be a protected feature with possible values {male, female, other}, where female or other might be protected group(s). During the feature engineering phase of the ML pipeline, a decision must be made as to include the protected feature in model training or not. Even if the feature is not included in the training, it is necessary to keep the information available for group fairness analysis.

Privileged group: the privileged group is non-protected and typically refers to the non-protected group, which has historically held the majority of political and social power within the population. For example, White and Male are typical choices for the privileged group for race and gender features, respectively.

Reference Group: the reference group is the group against which bias and prejudice will be measured if the fairness measurement in question is a comparative measure. The reference group is usually, although not always, the privileged group. We might want to generate comparisons amongst protected groups (e.g. between Black and Hispanic as opposed to between Black and White) if the particular context warrants it.

Individual Fairness: individual fairness was initially proposed in the context of ML by *Dwork et al.* in *'Fairness through Awareness'* [2]. The notion of individual fairness centres on the concept that similar individuals should be treated similarly. Therefore, a model outcome should be equal for individuals whose input features are similar. If a small change in one or more of the input features resulted in the arbitrary placement of two very similar individuals on different sides of the decision boundary. E.g. if an income of 100,000 results in a *'yes'* and an income of 99,999 results in a *'no'*, then the model does not comply with individual fairness.

Similarly, if groups of individuals were treated differently by the model in order to ensure equalised outcomes across groups, then the axiom of individual fairness would also be removed. In the context of human-based decision making, individual fairness may be represented through a view that everything '*is what it is'* and that exceptions or *'judgement calls'* can be made when a borderline case exists. This view implies the quantifiable and measurable features of each individual are what they are with no consideration for non-quantifiable, non-observed, causal influences, there is an assumption that individuals are similar in all ways *we think* matter.

Group Fairness: an alternative axiom to individual fairness is that of group fairness, a fairness goal which centres on the grouping of individuals based on the protected feature and ensuring that the distribution and type of outcomes are equitably distributed across groups. Specifically, what defines 'equitable', depends on the type of group fairness to be employed.

Group Fairness - Demographic parity: is a representation related fairness which is satisfied if the absolute count of beneficial (or non-punitive) outcomes for each group is equal. If no biological or inherent reason accounts for the absence of demographic parity, then a desire to satisfy it indicated a recognition of the existence of prejudice. It also intimates a wish to create a 'decision-maker' willing to apply corrective measures to counter historical discrimination and ensure that all groups are represented equally in beneficial outcomes. The 'decision-maker' is aware that such an intervention may have a risk attached. The notion of affirmative action promotes demographic parity, for example, by employing males and females at the same rate. The implied risk is that as this situation has never occurred in the past, the employer has no reference to gauge what the outcome may be in terms of utility. Adjusting an ML model to comply with demographic parity will almost surely result in a loss of accuracy (which is usually perceived to correlate to utility).

Group Fairness - Proportional parity (Impact Parity): Is a representation related fairness which is satisfied if the proportion of beneficial (or non-punitive) outcomes for each group matches that of the geographical location, or the domain in question. If no biological or inherent reason accounts for the absence of demographic parity, then a desire to satisfy it indicated a recognition of the existence of prejudice. It also intimates a wish to create a 'decision-maker' willing to apply corrective measures to counter historical discrimination and ensure that all groups are represented proportionately in beneficial outcomes. The 'decision maker' is aware that such an intervention may have a risk attached. The notion of affirmative action promotes demographic parity, for example, by employing males at 49% and females at 51% if this is the proportion in the general population, or hiring females at 16% and males at 84% for tech positions if this is the proportion of STEM graduates in the population.

Group Fairness - Equality of opportunity: is an accuracy-related fairness goal which is satisfied if the model correctly predicts an individual is deserving of an outcome when they are labelled as deserving the outcome (assistive or punitive) at equal rates across groups. In other words, if the true positive rate is equal across all groups. If no biological or inherent reason accounts for the absence of demographic parity, then a desire to satisfy it indicated a recognition of the existence of prejudice. It also indicates a desire to prevent further prejudice. However, it does not indicate a desire to apply corrective measures to counter historical discrimination proactively. A human decision-maker might think about this as *'levelling the playing field'* such that when, for example, women are qualified for a job, they will be hired at the same rate as men who are qualified for the job.

Group Fairness - Equalised odds: is similar to equality of opportunity in that it is satisfied if the model correctly predicts an individual is deserving of an outcome when they are labelled deserving of the outcome (assistive or punitive) at equal rates across groups. However, it must also satisfy the condition that those who do not deserve the outcome are incorrectly labelled as deserving the outcome at equal rates across groups. A human decision-maker might think about this as 'levelling the playing field' such that when, for example, women are qualified for a job, they will be hired at the same rate as men who are qualified for the job. It further implies that women be promoted without merit at the same rate men are promoted without merit (equal amounts for favouritism). The implications of an effort to satisfy equalised odds is similar to that of equality of opportunity; however, when the outcome is punitive, equalised odds has additional importance.

Fairness through unawareness: to satisfy *'fairness through unawareness'* the protected features are removed from the training data. An assumption is made that that the features are entirely independent of the remaining input features. In terms of human decision making, this is the equivalent to taking a 'colour-blind' or 'gender blind' approach. An example might be the use of a 'blind CV' process, removing the protected feature (such as race or sex) from view, however membership of a protected group can often be implied through relationships with other features such as name, address, school attended or similar.

Disparate Treatment: a decision-making process suffers from disparate treatment if its decisions are based, or partly based on the individual's membership of a protected group. In a human-based decision making this would be the equivalent of an outcome being partly based upon someone's gender, skin colour or age, a practice which is outlawed in many situations today. However, this also means that an attempt to satisfy proportional or demographic parity (e.g. via affirmative action) may also be considered illegal as in the case of *Ricci v. DeStefano* [4]

Disparate impact: is a group fairness assertion. A decision-making model suffers from disparate impact if the outcome of the decision disproportionately benefits one group or disproportionately hurts another group. In ML models it generally results in unintentional discrimination, although in human decision making, it was historically very often intentionally introduced as a means to bypass anti-prejudice laws. The practice of redlining [5] in the US is a good example of this. It is something that could be introduced unintentionally to an ML model via the use of a zip or post that has a strong relationship with a particular demographic group.

Eighty per cent rule: this is a 'rule of thumb' or a guideline applied to the measurement of disparate impact. Suppose a selection rate for a protected group a is less than 80% of the selection rate for group b, where group b is the group with the highest selection rate. This will generally be considered evidence of adverse impact against group a. The eighty per cent rule does not incorporate probability distribution to determine if the disparity is as a result of chance and therefore is not a definitive test. It is often used as a measurement for disparate impact.

The probability ranking principle: states that the ideal ranking should order items in the decreasing order of their probability of relevance, as this maximises utility or usefulness.

3.2. Background

3.2.1. Ranking

The importance of ranking by machine learning models should not be overlooked. Ranking in information retrieval is the position at which a particular result appears in the results list of a search query. ML models are used extensively to predictions but also to assign scores and provide rankings. Lists appeal to a human tendency for categorisation by displaying information in short, easily consumable components. They indicate that the heavy mental lifting of analysis and categorisation is achieved and that consumption can begin. Michaela Wänke's *paradox of choice* [6] investigation concluded that humans feel better when the amount of work necessary to process large amounts of information is reduced, in other words, the happiness that results from the process of decision-making increases with the speed at which decisions are made. *Walter Kintsch* [7, 8] pointed out back in 1968 that we can process information more easily when it is in a list as opposed to when it is *'clustered and undifferentiated'*. A list is therefore perfectly designed for the human brain, a ranked list even more so.

There is little doubt that the order of ranking is correlated with the result which receive the majority of the user attention and interaction and that our cognitive inclinations tend to place a higher value on results positioned at the top of a list. The results of a 2008 study by *Keane et al*. [9] has shown definite evidence of bias in a users' interaction with a list returned by a search, in that the user tends to give preferences to items at the top of the list, although they do sometimes seek out relevant results with a lower ranking. The results also showed that the items Google presents as '*the best'* are considered by the user to be the best. There is also a tendency to place a certain level of unquestioned faith in technology, a 2015 study by psychological scientists at the University of Missouri came to the conclusion that people show an implicit association between technology and success, a phenomenon they call the *'technology effect'* which '*has conditioned decision-makers to be overly optimistic about the potential for technology to drive successful outcomes'*[10], this false sense of security could lead to an over-reliance on decisions made by imperfect algorithms in decision-making tools, whilst the decision-maker remains unaware of exactly how the algorithm works or what the ramifications of relying on this effortlessly acquired data might be.

Search engines and other automated or semi-automated decision-making systems use algorithms to determine the ranking of results. An algorithm, as defined by the Cambridge English dictionary is *'a set of mathematical instructions or rules that, especially if given to a computer, will help to calculate an answer to a problem'*. As it relates to ranking the algorithm refers to the rules which decide in what order the results in the ranked list are returned. Many website owners engage with expensive search engine optimisation campaigns, using knowledge about how search engines work with the objective of improving their search engine ranking, to move their website closer to the top of the results. Websites that are ranked higher typically get larger per cent of click-throughs and attract more visitors. Search engines, such as Google and Yahoo, process almost every page on the Internet, assigning a 'rank' to each page based on the perceived usefulness of the content. The pages with the highest rank are displayed higher in the search engine results and are therefore given higher regard and receive more interaction. Humans are obviously many degrees more complex than web-pages however we too are often categorised or ranked via decision-making algorithms, reduced to a collection of entities albeit without the possibility of engaging in an expensive SEO campaign to ensure a beneficial position in the response. The probability ranking principle states that ranking should be in the decreasing order of their probability of relevance so as to maximise utility; however, this does not always result in the definition of fairness that an organisation wishes to reflect. When humans are being processed by ML models, there are several 'gates' to pass before the chance for attention exists.

3.2.2. Prejudice

Algorithms based on machine learning reflect the human process for decision making. Humans innately generalise learning from patterns detected in the surroundings; this improves the efficiency of learning and decision making. The study of psychology, in particular as it relates to decision making, discusses the use of heuristics in the decision-making process with a consensus that there are two modes of cognitive function. As described by *Kahneman* [11], there is *'an intuitive mode in which judgments and decisions are made automatically and rapidly and a controlled mode, which is deliberate and slower'.* It is this intuitive or automated mode that is often referred to as heuristic and which it has been determined is mainly responsible for unconscious bias in decision making. Implicit bias makes its presence felt in a variety of social scenarios and situations, '*For example, people prefer the young to the old, and pair women with the home more often than they pair women with the laboratory. People show a negative implicit association with members of a racial group other than their own*'. [12] Humans also tend to extend preferential treatment, in the form of affinity bias, towards those whom they believe they have something in common with. [13, 14] This can result in unfair outcomes in many scenarios, including recruitment, retention, and advancement such that white male dominance in leadership can, for example, fuel more white, male leadership. This ability to generalise can be useful, but it can also cause intentional and unintentional prejudice in human decision-making.

Similarly, machine learning algorithms also make decisions based on patterns. In supervised learning, the algorithm will '*learn'* a mathematical model from a training data set which contains both the input and the desired output. Successful learning is based on the assumption that patterns will be repeated; therefor, there is an intrinsic reliance on the past. The reality is that in a world where underrepresented groups are discriminated against, the data generated and the technology created in this environment cannot avoid reflecting these inherent biases. This bias frequently prioritises outcomes that benefit those who have traditionally held power and privilege. Many examples of such bias have been uncovered. For example, lending discrimination was noted in the 2017 *FinTech study* [15] which found that both human-driven and ML-based mortgage decisions resulted in interest rates significantly higher for Latinx and African American, at the cost of over 750 million USD per annum in aggregated extra interest charges. Other cases include recruitment discrimination, such as the amazon recruiting algorithm which recognised through the existence of historical patterns that male profiles had a greater success rate within Amazons’ recruitment process and therefore penalised profiles that appeared too female [16]. Apple had similar issues with its credit card, which offered less credit to women [17]. Many more examples of algorithmic bias are discussed in Cathy O’Neil book *Weapons of Math Destruction* [18], including examples relating to civic life, insurance, and education. Safiya Nobel focuses on examples of search engine reinforced racism in *Algorithms of Oppression* [19] where she argues that bias in discoverability is incredibly impactful towards women of colour, providing examples in textual and media searches and online advertising.

Both implicit/unconscious bias in the form of prejudice, bigotry, or unfairness directed by someone from a privileged group, intentionally or unintentionally, towards individuals from an oppressed or marginalised group, and systemic bias directed by health, educational, government, judicial, legal, religious, political, financial, media, or cultural institutions play their part. Examples of implicit and systemic bias surround us daily. According to the U.K. governments statistics [20] between April 2018 and March 2019, there were four stop and searches for every 1,000 White people compared with 38 stop and searches for every 1,000 Black people. A police officer may stop and search on the subjective proviso that *‘the officer has a reasonable cause to suspect they will find something’*. Black people are almost ten times more likely to be stopped, albeit the likelihood of offence detection is similar regardless of ethnicity, even with a slightly higher chance of offence detection for white compared to black people [21, 22]. If this historic data from 2008 to 2019 were to be used to train a model to predict ‘*reasonable cause*’ it is possible that this disproportionality may result in a model which reflects the subjective grounds for reasonable cause applied by the police force. A recent report, *Race and Racism in English Secondary Schools* [23], raises concerns over racism and the role of the police in society. It focused in particular on the impact of school-based policing on black and minority kids given the disproportionate attention placed by police on non-white people and the impact this may have. It is worth remembering that like the world, the data gathered from the world has centuries of historical context and historical prejudice.

Lack of equality leads to underrepresentation, or overrepresentation in many areas of life. This further leads to disparities in data, which may then lead to further disparities in life. For example, a review of UCLA's annual diversity reports from 2020 shows that while minorities make up an ever-increasing portion of the U.S. population and the per cent of minorities in leading roles in the film and entertainment industry is making headways, when it comes to writing and directing, minorities and women have gained little ground, in 2019, women directed 15.1%, and minorities directed 14.4% per cent of the top box office movies, women earned 17.4% of writing credits and minorities 13.9%[49]. According to a 2016 study from the TIAA Institute, faculty positions at U.S. universities have reflected an improvement in the representation of under-represents groups. However, this increase was manifested in a more extensive growth in the uncertain and insecure non-tenure-track positions [25], indicating that when representation does occur it does not always occur at the levels within an organisation that can affect change. According to a 2018 LinkedIn global survey [26], seventy per cent of people have been hired at a company where they had a connection. This reiterates the importance of representation, as Cathy O'Neil points out in Weapons of Math Destruction, 'The privileged are processed more by people, the masses by machine' [18].

During the 1960s, almost fifty per cent of white respondents in a US study [52] indicated that they would move if a Black family were to move next door. During the 1970s, the *Marriage Bar* [44] in Ireland forced women to resign from their job upon marriage. Often, what has happened in the past is brushed aside with remarks such as 'it was a different time' or 'it was a different world'. However, yesterday's injustices become today's inequalities. The median wealth among white households in America is $171,000 while amongst black households, it's $17,600 []. The wage gap between men and women in Ireland was still 7.5 per cent in 2018[50], with boardroom representation at only 24% [50]. Machine learning enables the creation of powerful applications but with great power comes great responsibility. If algorithms are to make decisions which may have severe consequences for society then it is imperative that the societies in which they are being used have full transparency as to the 'worldview' of the creator and of the model.

3.3. Proposal

One way in which we humans work to prevent unfair in decisions is by taking deliberate and intentional action to detect bias in ourselves and others and remove the influence of such bias from the process. Those organisations responsible for the creation of automated decision-making tools must take the same steps.

3.3.1. Overview

Drawing on the observations raised through the related work, it appears there is a necessity for a practical framework to facilitate inter-organisation alignment. The proposal is the delivery of such a framework, to evoke reflection upon, and audit of, fairness at various points in the machine learning pipeline. The framework should facilitate conversations pertaining to the fairness philosophies that the creators of the model believe the model should reflect. The framework will help to identify disparities across protected groups using data analysis, statistical techniques, unfairness detection and explainability methodologies to prompt conversations. These conversations should steer decisions around unfairness mitigation strategies in line with the agreed-upon philosophy in such a way that organisational values, and not mathematics, are taking the lead in the conversation. Reflections prompted by the framework will demand discussion, collaboration and agreement amongst various stakeholders within the business, including relevant domain experts. Answers provided during data analysis will be stored in the form of a report which can serve as a reference point for discussions. As the framework also includes a demo dataset it can additionally be used as a teaching aid on the topic of fairness and of ML feature engineering in general.

Fairness should be considered at various stages in the ML pipeline, including:

- When preparing data to train a model,
- When analysing the resulting learned model for accuracy
- When reviewing any resulting ranked list for accuracy.

To this end the framework consists of three components:

- A user interface for the analysis and pre-processing of the training data.

- A user interface for the analysis of the resulting model via the Aequitas fairness package and the SHAP explainability package.
- A user interface for the analysis of any resulting ranked list.

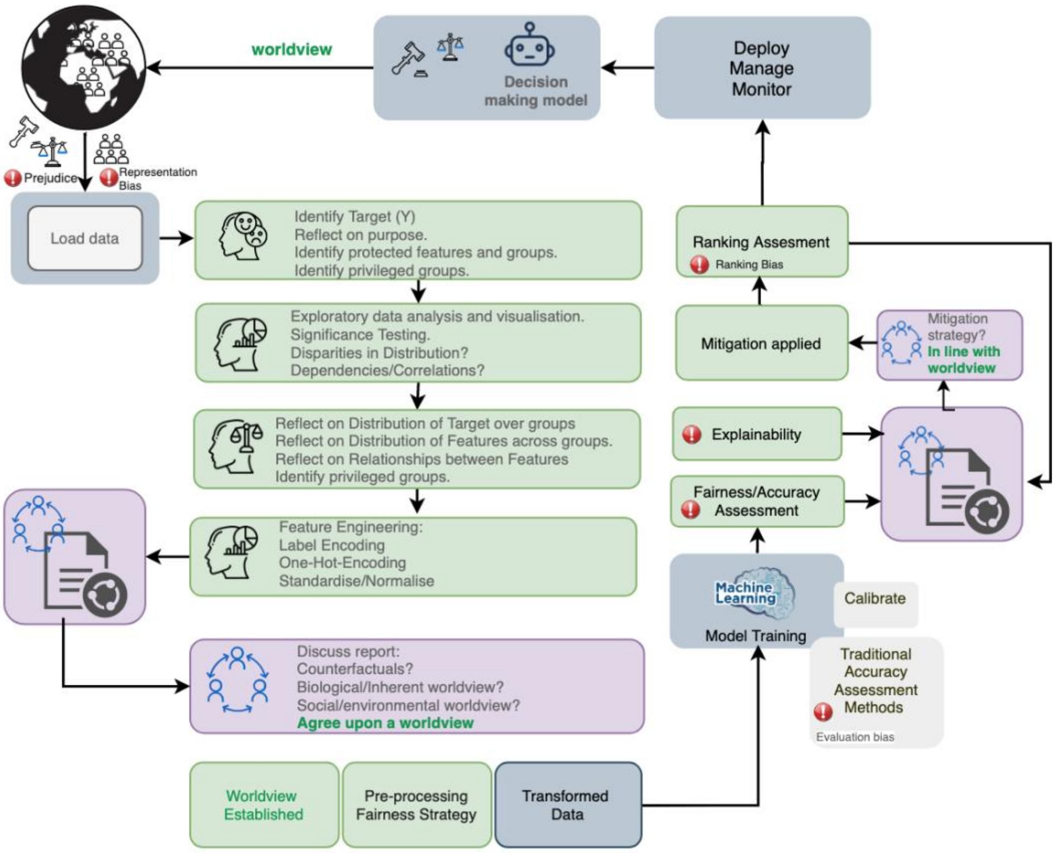

worldview
Decision making model
Deploy
Manage
Monitor
Prejudice
Representation Bias
Load data
Identify Target (Y)
Reflect on purpose.
Identify protected features and groups.
Identify privileged groups.
Exploratory data analysis and visualisation.
Significance Testing.
Disparities in Distribution?
Dependencies/Correlations?
Reflect on Distribution of Target over groups
Reflect on Distribution of Features across groups.
Reflect on Relationships between Features
Identify privileged groups.
Feature Engineering:
Label Encoding
One-Hot-Encoding
Standardise/Normalise
Discuss report:
Counterfactuals?
Biological/Inherent worldview?
Social/environmental worldview?
Agree upon a worldview
Worldview Established
Pre-processing Fairness Strategy
Transformed Data
Ranking Assesment
Ranking Bias
Mitigation applied
Mitigation strategy? In line with worldview
Explainability
Fairness/Accuracy Assessment
Machine Learning
Model Training
Calibrate
Traditional Accuracy Assessment Methods
Evaluation bias

Figure 5 Proposed workflow for bias reflection framework.

3.3.2. Fairness Worldview

Fairness as a philosophy has no objective definition, and as such there is no consensus on a mathematical formulation for fairness. When training a Machine learning model to predict an outcome and hence influence decisions that will have positive or negative consequence for a person or group it is necessary to reflect on the worldview or philosophy of fairness that the model will reflect. The framework centres around the premise of a worldview which will represent the desired fairness philosophy.

The philosophers Norman Geisler described worldview as *‘an interpretive framework through which or by which one makes sense out of the data of life and the world.’* His personal worldviews aside the definition seems quite fitting to the area of Machine learning and ‘the data of life and the world’. There have been extensive discussions around fairness, prejudice, discrimination, and justice throughout history. In the context of this framework a "Worldview" is a set of assumptions about a physical and social reality pertaining to a human feature or attribute, or to the measurement of same. As context must be taken into consideration there is no one fundamentally correct worldview but rather a reflection of a particular philosophy of life, or a conception of the world, as it relates to each of an individuals' apparently quantifiable features or attributes. In the case of this framework, the focus is, in particular, on the worldview held concerning any disparities in features or attributes that might be detected across groups within protected features such as race, gender or age. A disparity may, for example, refer to a non-proportionate representation or a significant difference in distribution. Two worldviews have been defined for this purpose, an inherent or biological worldview, and a social and environmental worldview.

3.3.2.1. Inherent or biological worldview

This worldview postulates that either chance or innate, inherent physiological, biochemical, neurological, cultural and/or genetic factors influence any disparities in features or attributes that might be detected across groups. This worldview could be quite easily applied to the measurements of height, or similar easily quantifiable features to be used as predictors for a specific outcome. The worldview, however, becomes more complex for those human attributes or features which are harder to quantify, such as grit, determination, intelligence, cognitive ability, self-control, growth mindset, reasoning, imagination, reliability or similar. This Inherent or biological worldview is closely aligned with the concept of individual fairness, where the fairness goal is to ensure that people who are ‘similar’ obtain similar results.

3.3.2.2. Social and environmental worldview

This worldview postulates that social and environmental factors, such as family income, parental educational backgrounds, school, peer group, workplace, community, environmental availability of nutrition, correct environment for sleep, stereotype threat (and other cognitive biases) often perpetuated by racism, sexism and other prejudices have influenced outcomes in terms of any detected disparities across groups. Differences in outcome may be a reflection of inequalities in a society which has led to these outcomes. Identifying this has important implications for the financial, professional, and social futures of particular protected groups within the population. Discrimination, privilege, institutional racism, sexism, ablism are examples of causal influences which may impact outcomes or representation. Disparities may have been caused by intentional, explicit discrimination against a protected group or by subtle, unconscious, automatic discrimination as the result of favouritism towards the reference group, or by other social and systemic factors.

3.3.3. Counterfactual thinking

In *Explaining intersectionality through description, counterfactual thinking, and mediation analysis'* [37] the concept of counterfactual thinking is heavily discussed. The author is very clear that 'simply observing that a disparity exists does not tell us how to reduce it or what certain interventions might achieve', conversely the starting point to considering what interventions may achieve must then be to observe if, where and why disparities exist. It is hoped that the framework will encourage counterfactual thinking on the discovery of any disparities within the data or the model produced. Machine learning relies on historical data, when we work with historic data, we can only observe the data without affecting it. The historic outcomes and decisions used to train machine learning models have already been made, there is no possibility of running A/B tests to determine what if, or what would have happened had historic conditions been altered, had historic outcomes or decisions been different. What if there were no historic bias? What if an individual had been a member of group one rather than of group two? What if systemic bias had not played a part in life and subsequently the formation of the data? Would additional members of a protected or minority group have a more favourable representation, or higher levels of visibility in this data now? Social equality often requires an active effort to detect and then change patterns in decision making, this involves a certain level of awareness of disparities, reflection and acknowledgment upon the cause of the disparities and action towards correcting that pattern and mitigating any unfairness that has resulted. This action comes with a potential inherent risk due to the unknown implications of a change. In the world of machine learning this risk is measurable based on the difference in accuracy or utility maximisation. It should, however, be noted that measuring the utility maximisation of a new decision-making model based on historic outcomes only provides a measure of perceived utility maximisation with no consideration for ‘counterfactual regret’, a form of causal reasoning which laments the difference between the actual outcome and the outcome that might have been, when a decision, which may have had a regretful outcome, was made. In the case of historic bias, the regretful outcome may have been a loss of utility, caused by the bias. The goal of utility-maximation, as measured by traditional accuracy assessment techniques, leaves no room for ‘counterfactual regret’ and therefore although a quantifiable goal, may not be the most desirable goal.

3.4. Feature engineering

Data scientists spend a significant amount of time on data preparation techniques include imputation of missing data, outlier detection, label encoding, one hot encoding, binning, grouping operations, scaling and transforming. Feature engineering techniques will be included in the tool to encourage the use and adoption of the framework.

Chapter 4

Implementation

The purpose of this chapter is to provide details regarding the technologies used for delivery of the application, the distribution methodology, location of the source code, dependencies and instructions pertaining to use of the framework.

4.1. Technology

This work has been developed using Python version 3.7 and comprises of approximately 8,000 lines of code. The package is based upon an interactive HTML interface which is powered by the open-source ipywidgets[1] packet for Jupyter notebooks[2], JupyterLab[3] and the IPython kernel[4]. Ipywidgets takes advantage of the Python architecture for interactive widgets that connects Python code running in the kernel and JavaScript/HTML/CSS running in the browser. These widgets allowed for the creation of a user interface which dynamically adapts to whatever *.csv* datafile is uploaded by the user. In addition to several bespoke bias detection features it also utilises functionality from the Aequitas Bias and Fairness Audit Toolkit[5], the SHAP (Shapley Additive explanations) package[6] for explainability and the FA*IR ranking package[7].

4.2. Distribution

The code has been packaged for distribution via PyPI and registered on the Python Package Index repository test site with an MIT license under the open-source initiative,

[1] https://ipywidgets.readthedocs.io/en/latest/developer_docs.html

[2] https://jupyter.org/

[3] https://jupyterlab.readthedocs.io/en/stable/

[4] https://ipython.org/ipython-doc/3/development/kernels.html

[5] http://www.datasciencepublicpolicy.org/projects/aequitas/

[6] https://christophm.github.io/interpretable-ml-book/shap.html

[7] https://github.com/MilkaLichtblau/FA-IR_Ranking

and with the package name ‘Machnamh’, a Gaelic (Irish) word which signifies wonder, reflection, or contemplation. The latest version can be found on PyPI[1]. The package can also be installed via pip, the de-facto standard package-management system that ships with most versions of Python, using the command *‘pip install machnamh’*.

4.3. Source code

The source code is available on GitHub[2].

4.4. Dependencies

The ‘*Machnamh’* package is dependent on the following libraries which are automatically installed by the pip distribution system at the time of package installation. 'kaleido', 'numpy', 'matplotlib', 'seaborn', 'pandas>=1.0.5', 'scikit-learn', ‘pandas-profiling>=2.9.0', 'phik>=0.10.0', 'ipywidgets', 'plotly', 'ipyfilechooser', 'dill', 'IPython', ‘shap’, ‘aequitas’, 'scipy', 'typing>=3.7.4.3', 'benfordslaw>=0.1.3', 'missingno', ‘fairsearchcore’

4.5. Running the Framework

To run the framework the user must start Jupyter Notebook, import the ‘machnamh’ library and invoke the User Interface.

```
import machnamh
from machnamh import pre_process as mpp
dpUI = mpp.data_pre_process_UI()
dpUI.render(use_demo_data = True)
```

Demo code is available in the GitHub repository in the ‘demo_jupyter_notebooks’ folder which contains three notebooks, one for each supported step of the ML pipeline.

```
machnamh_step_one_review_prepare_data.ipynb
machnamh_step_two_train_model_analyse_output.ipynb
machnamh_step_three__analyse_ranking.ipynb
```

[1] https://pypi.org/project/machnamh/

[2] https://github.com/aideenf/machnamh.

Chapter 5

Use Case

The purpose of this chapter is to demonstrate the framework by means of a use case. The chapter is divided into six sections 5.1 describes the dataset used, section 5.2 describes the task at hand. 5.3 describes a step by step review of the dataset analysed through the framework. 5.4 provides an overview of the fairness analysis carried out on the trained model using the frameworks integration with Aequitas API and Shap explainability as referenced in the technology section. 5.5 Shows the fairness in ranking analysis and finally 5.6 provides an overview of the discussions the analysis might raise.

5.1. Dataset

The Law School Admission Council surveyed 163 law schools in the United States. The dataset contains information on 21,790 law students over five years starting from 1991. For three years the study tracked those students who entered law school in Autumn 1991 in addition to tracking up to 5 rounds of the students' BAR examinations. Note that the original purpose of gathering this data was not to create a machine learning model, however, the data does serve as a good example of how existing data might be used for such a purpose and of the necessity to reflect upon the use of such data for the creation of decision-making models which have real-world consequences on humans. The LSAT exam has undergone extensive change since this data was collected in 1991 and therefore results are not necessarily applicable to the current exam. The data includes:

UGPA (Undergraduate grade point average): The undergraduate grade-point average collected before law school. GPA is a standard way of measuring academic achievement in the U.S. Each course is given a certain number of "credits". In secondary school it is usual for all courses/subjects to have the same number of credits, college or university courses however can have between 1 and 5 credits per course, the average is 3 which corresponds to 3 hours of lectures and approx. 6 hours of homework per week. GPA assumes a grading scale of A, B, C, D, F, each of which is assigned a number of grade points. A=4, B=3, C=2, D=1, F=0. The GPA is the (sum of the credits x grade)/sum of the credits. e.g. a 3-credit class with A grade and a 4-credit class with a C grade would result in (3*4+4*2)/ (3+4)

LSAT (Law school admissions test): The Law School Admission Test is an integral part of law school admission in the United States, Canada and several other countries, according to the admissions council the test is designed to assess critical reading, analytical reasoning, logical reasoning, and persuasive writing skills which are the skills deemed necessary to succeed in law school. Some Law schools do accept other tests, however, as of 2020, the LSAT remains the dominant test for law school admission. The sponsors of the LSAT (The Law School admissions council) suggest that it is a useful measure for predicting first-year average grades or the likelihood of ultimately passing the BAR exam.

Sander_index*:* Sander proposes combining the LSAT and GPA scores into a single weighted average, using weights that correspond roughly to those used in many law schools. This average is referred to as the "Sander Index." Because the units of this index are difficult to interpret it has been converted into a per centile score. This score ranges from 0 to 100 and represents the per centage of law school matriculants with lower index scores. That is, a student with a per centile score of 75 has better academic credentials than three-quarters of law school matriculants, but worse credentials than the remaining quarter.

ZFYA (Z-score of First-year average grade): This is a measure of first-year academic achievement in law school, presented as a z-score. Z-Scores are raw scores expressed in standard deviation units, relative to the mean score. Positive Z-scores indicate a raw score that is above the mean. Negative Z-scores indicate a raw score that is below the mean. Zero Z-score indicates a raw score that is equal to the mean. In a normally-distributed set of data, the general rule states that 68% of all scores will fall within ±1 SD of the mean; 95% of all scores will fall within ±2 SD, and 99.7% of all scores within ±3 SD. Z-scores between -2.00 and +2.00 are therefore considered relatively ordinary, while values greater than -2.00 and +2.00 are considered unusual.

First_pf: Whether or not the BAR exam was passed or failed on the first attempt, the BAR exam is facilitated by the American Bar Association and apparently designed to test knowledge and skills that every lawyer should have before becoming licensed to practice law. It is worth noting that in most states, a law school graduate cannot take the bar exam without having attended a law school accredited by the American Bar Association (ABA).

Region_first: The geographic region in which the first BAR exam was taken. The jurisdictions used in this study match those used by the Law School Admission Council (LSAC) in Regional Statistical Reports. Definitions of regional groups are:

- New England [NG] and Northeast [NE].
- Midsouth [MS], Southeast [SE], and South Central [SC] (South)
- Northwest [NW], Far West [FW], and Mountain West [Mt](West)
- Great Lakes [GL] and Midwest [MW].

A feature such as this might be used as an intentional proxy for the geographical location of the student, or an unintentional proxy for the race of a student.

Race: The race of the students has been categorised as 'White', 'Hispanic', 'Asian', 'Black', 'Other', 'Mexican', 'Puerto Rican', 'Amerindian'.

Sex: The original dataset contains a column titled “sex” which categorised the data into two genders those being either Female (1) or Male (2).

A practitioner of machine learning may initially consider this data a good starting point for the creation of a model used to predict an individual’s potential for success in law school. We can imagine such a model being used by an admissions department as part of the process for deciding who to accept into the program and who to reject. Inputs such as GPA, LSAT and Sanders-Index on first impressions appear to be objective and measurable ground truths. They appear to be reflecting real-world outcome and do not intuitively appear to be the result of an obviously subjective decision taken by a possibly prejudiced human. The data could be used either to create a model to predict the outcome of the first year of law school by using 'ZFYA' as the training target or to create a model to predict the likelihood of passing the BAR exam by using 'first-fp' as the training target.

5.2. Use case description

For the purposes of the use case the goal will be to create a model to predict the outcome of the first year of law school by using *'ZFYA'* as the target. Race and Gender will be defined as the protected features. The framework allows for the simultaneous review of both however for the purpose of brevity the focus will be on race only

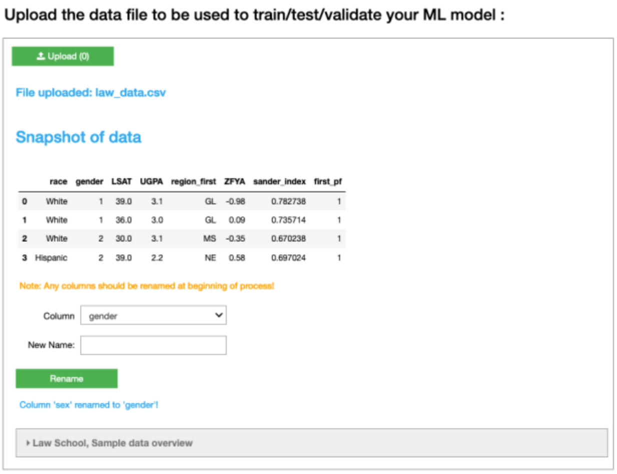
Following the chronologic order of education in the U.S. as pertaining to the exams scores reflected in the dataset, first the undergraduate grade point average (UGPA), followed by the standardised test necessary to enter into law school (LSAT), followed by the annual grade point average. In this case, we have access to the first-year grade point average which we have selected as the target (ZFYA). The final step is the BAR exam which is another standardised test necessary to practice law and which is represented by '*first_pf'*. As the goal is to predict ZFYA and ZFYA is prior to the BAR exam the use of 'first_pf' is not possible. The region in which the bar exam was taken is stored in the 'region_first' feature and therefor this will also be omitted.

First the dataset will be analysed and any necessary feature engineering will be applied. A Logistic regression model will then be trained and the output will be binarized by setting Y = 1 whenever ZFYA ≥ 0.09 (the median value.) Finally, the Logistic regression score will be used to create a ranked list so that we may examine most of the framework's functionality.

5.3. Dataset analysis via framework

To invoke the fairness analysis functionality of the framework package it is necessary to import the helper.

```
import machnamh
from machnamh import pre_process as mpp
dpUI = mpp.data_pre_process_UI()
dpUI.render(use_demo_data = True)
```



Upload the data file to be used to train/test/validate your ML model :

File uploaded: law_data.csv

Snapshot of data

	race	gender	LSAT	UGPA	region_first	ZFYA	sander_index	first_pf
0	White	1	39.0	3.1	GL	-0.98	0.782738	1
1	White	1	36.0	3.0	GL	0.09	0.735714	1
2	White	2	30.0	3.1	MS	-0.35	0.670238	1
3	Hispanic	2	39.0	2.2	NE	0.58	0.697024	1

Note: Any columns should be renamed at beginning of process!

Column: gender

New Name:

Column 'sex' renamed to 'gender'!

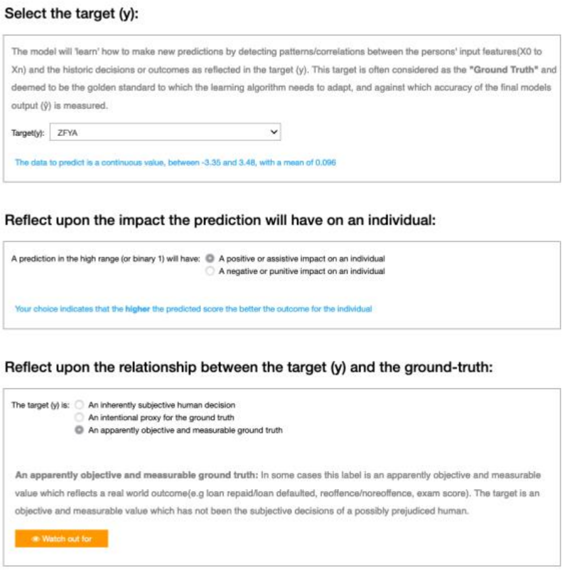
Step 1: Upload the data.
The first in the process is to upload the data to be analysed. Any .csv file may be uploaded.

Functionality at step 1 includes:

Display a sample of the data.

Facilitate Rename of columns.

If using demo mode, a summary of the demo data will also display here.



Select the target (y):

The model will 'learn' how to make new predictions by detecting patterns/correlations between the persons' input features(X0 to Xn) and the historic decisions or outcomes as reflected in the target (y). This target is often considered as the **"Ground Truth"** and deemed to be the golden standard to which the learning algorithm needs to adapt, and against which accuracy of the final models output (ŷ) is measured.

Target(y): ZFYA

The data to predict is a continuous value, between -3.35 and 3.48, with a mean of 0.096

Reflect upon the impact the prediction will have on an individual:

A prediction in the high range (or binary 1) will have: ● A positive or assistive impact on an individual
○ A negative or punitive impact on an individual

Your choice indicates that the **higher** the predicted score the better the outcome for the individual

Reflect upon the relationship between the target (y) and the ground-truth:

The target (y) is: ○ An inherently subjective human decision
○ An intentional proxy for the ground truth
● An apparently objective and measurable ground truth

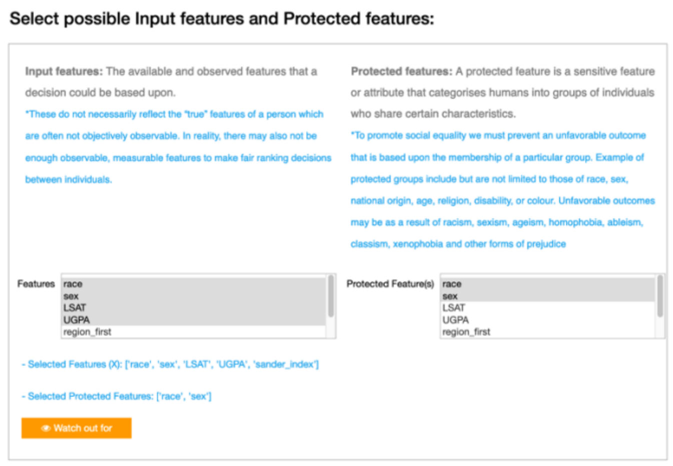
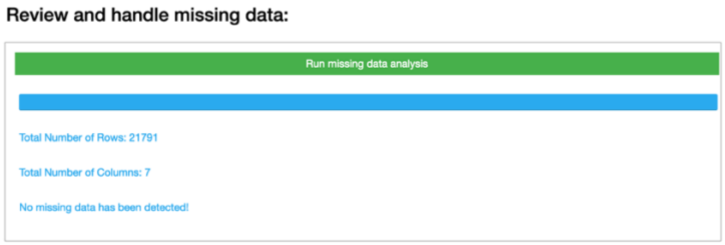
An apparently objective and measurable ground truth: In some cases this label is an apparently objective and measurable value which reflects a real world outcome(e.g loan repaid/loan defaulted, reoffence/noreoffence, exam score). The target is an objective and measurable value which has not been the subjective decisions of a possibly prejudiced human.

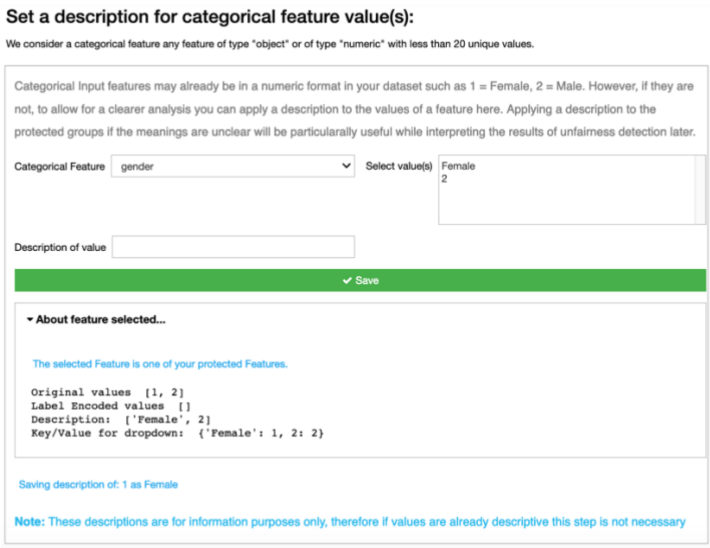
Step 2: Select target. Continuous and Binary output is supported.

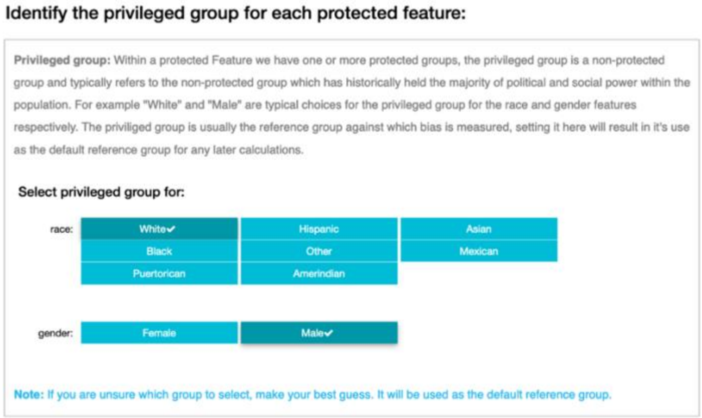
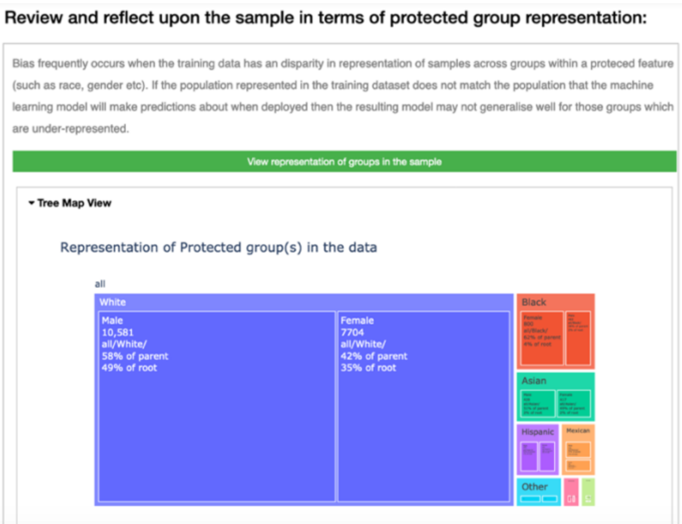
If the target consists of two non-binary values a ‘convert to binary’ option will appear

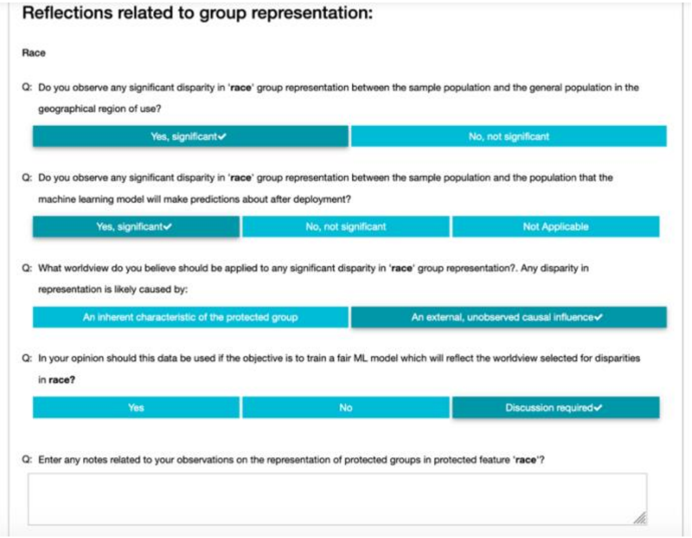
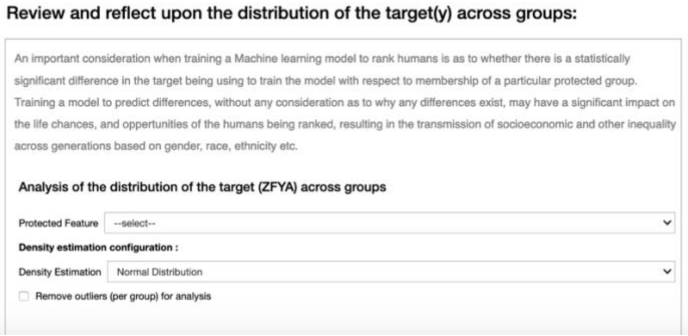
Step 3: Reflect on target.
Here we already begin to invoke reflection from the practitioner prompting reflection as to the real-world consequences of a high v’s low outcome on an individual (assistive or punitive)

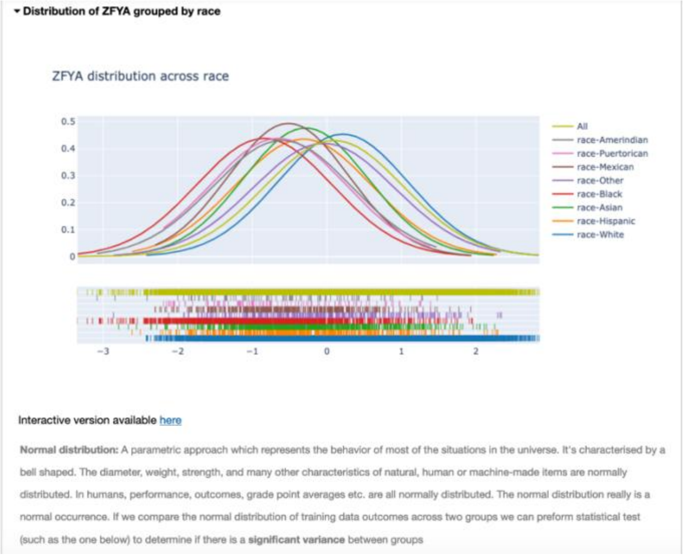
Step 4: Reflect on the ground truth.
Historical or Measurement bias considerations. Here we prompt the practitioner to consider the origin of the target, is it a proxy, is it objective, is there a chance that it has been influenced by human bias.

 Select possible Input features and Protected features: **Input features:** The available and observed features that a decision could be based upon. *These do not necessarily reflect the "true" features of a person which are often not objectively observable. In reality, there may also not be enough observable, measurable features to make fair ranking decisions between individuals. **Protected features:** A protected feature is a sensitive feature or attribute that categorises humans into groups of individuals who share certain characteristics. *To promote social equality we must prevent an unfavorable outcome that is based upon the membership of a particular group. Example of protected groups include but are not limited to those of race, sex, national origin, age, religion, disability, or colour. Unfavorable outcomes may be as a result of racism, sexism, ageism, homophobia, ableism, classism, xenophobia and other forms of prejudice Features: race, sex, LSAT, UGPA, region_first Protected Feature(s): race, sex, LSAT, UGPA, region_first - Selected Features (X): ['race', 'sex', 'LSAT', 'UGPA', 'sander_index'] - Selected Protected Features: ['race', 'sex'] 	**Step 5: Identify protected features.** Here again the practitioner is prompted to identify what groups may be at risk of prejudice, the tool also provides explanations as to what groups these may be. Once the protected feature or features are identified here the tool will automatically utilise this information during the rest of processing. For the analysis Race, Gender, LSAT, UGPA and Sanders_Index have been defined. Although later we can decide not to use Race and Gender while training the model.
 Review and handle missing data: Total Number of Rows: 21791 Total Number of Columns: 7 No missing data has been detected!	**Step 6: Handle missing data.** There is no missing data in the use case dataset, however the tool provides functionality to: View missing data per protected group. Impute missing data per group based on min/mean/max/most frequent/choose. Remove entire rows or columns if proportion of missing data per column or row does not meet a dynamically defined threshold.


Set a description for categorical feature value(s):
We consider a categorical feature any feature of type "object" or of type "numeric" with less than 20 unique values.
Categorical Input features may already be in a numeric format in your dataset such as 1 = Female, 2 = Male. However, if they are not, to allow for a clearer analysis you can apply a description to the values of a feature here. Applying a description to the protected groups if the meanings are unclear will be particularally useful while interpreting the results of unfairness detection later.
Categorical Feature: gender
Select value(s): Female 2
Description of value:
✓ Save
▾ About feature selected...
The selected Feature is one of your protected Features.
`Original values  [1, 2]`
`Label Encoded values  []`
`Description:  ['Female', 2]`
`Key/Value for dropdown:  {'Female': 1, 2: 2}`
Saving description of: 1 as Female
Note: These descriptions are for information purposes only, therefore if values are already descriptive this step is not necessary
	Step 7: Add description to feature values.
	
	Often categoric features are already converted to, or represented by numeric values in the dataset. However, this often causes a struggle when interpreting statistics around group fairness e.g. if it is necessary to remind oneself 1 = male, 2 = female, 0 = White, 1 = Non-White etc. This feature allows for the tagging of values with a description. These descriptions will then be used throughout the analysis for ease of interpretation.

 **Identify the privileged group for each protected feature:** **Privileged group:** Within a protected Feature we have one or more protected groups, the privileged group is a non-protected group and typically refers to the non-protected group which has historically held the majority of political and social power within the population. For example "White" and "Male" are typical choices for the privileged group for the race and gender features respectively. The priviliged group is usually the reference group against which bias is measured, setting it here will result in it's use as the default reference group for any later calculations. **Select privileged group for:** race: White✓ Hispanic Asian Black Other Mexican Puertorican Amerindian gender: Female Male✓ **Note:** If you are unsure which group to select, make your best guess. It will be used as the default reference group.	**Step 8: ID Privileged Groups.** This step prompts reflection in terms of who the privileged group might be. A description of 'privileged group' is provided. The group selected as the privileged group here is also used as the default reference group for any later comparative fairness measures (although it is also possible to select any other group)
 **Review and reflect upon the sample in terms of protected group representation:** Bias frequently occurs when the training data has an disparity in representation of samples across groups within a proteced feature (such as race, gender etc). If the population represented in the training dataset does not match the population that the machine learning model will make predictions about when deployed then the resulting model may not generalise well for those groups which are under-represented. View representation of groups in the sample ▾ Tree Map View Representation of Protected group(s) in the data all White Male 10,581 all/White/ 58% of parent 49% of root Female 7704 all/White/ 42% of parent 35% of root Black Asian Hispanic Mexican Other	**Step 9: Review the sample for representation.** This step provides a graphical and interactive view of the representation of groups in the sample. There is also an explanation as to why disparities may cause bias. Three examples of how representation bias may have real-world consequences are also available by selecting the orange buttons that appear throughout the screen.

 **Reflections related to group representation:** Race Q: Do you observe any significant disparity in '**race**' group representation between the sample population and the general population in the geographical region of use? Yes, significant✓ | No, not significant Q: Do you observe any significant disparity in '**race**' group representation between the sample population and the population that the machine learning model will make predictions about after deployment? Yes, significant✓ | No, not significant | Not Applicable Q: What worldview do you believe should be applied to any significant disparity in '**race**' group representation?. Any disparity in representation is likely caused by: An inherent characteristic of the protected group | An external, unobserved causal influence✓ Q: In your opinion should this data be used if the objective is to train a fair ML model which will reflect the worldview selected for disparities in **race?** Yes | No | Discussion required✓ Q: Enter any notes related to your observations on the representation of protected groups in protected feature '**race**'?	**Step 10: Reflect on representation in the sample.** This step encourages accountability for model fairness. The step also enables the practitioner to identify that collaboration with domain expert or stakeholder may be necessary. The answers selected here will form part of a report that the practitioner can use to engage in a discussion with stakeholders and domain experts. There is an intentional absence of '*don't know'*, or similar options, and a text area where observations may be noted.
 **Review and reflect upon the distribution of the target(y) across groups:** An important consideration when training a Machine learning model to rank humans is as to whether there is a statistically significant difference in the target being using to train the model with respect to membership of a particular protected group. Training a model to predict differences, without any consideration as to why any differences exist, may have a significant impact on the life chances, and oppertunities of the humans being ranked, resulting in the transmission of socioeconomic and other inequality across generations based on gender, race, ethnicity etc. **Analysis of the distribution of the target (ZFYA) across groups** Protected Feature --select-- **Density estimation configuration :** Density Estimation Normal Distribution □ Remove outliers (per group) for analysis	**Step 11: Analyse the Target across groups.** This step offers several methodologies for analysing the target as it pertains to group membership. For the purpose of the use case we will review protected feature 'race'. This step is useful for detecting historical bias. It is possible to analyse the target with or without the removal of outliers. It is also possible to select a distribution type (Normal or KDE)



▾ Distribution of ZFYA grouped by race

ZFYA distribution across race

Interactive version available here

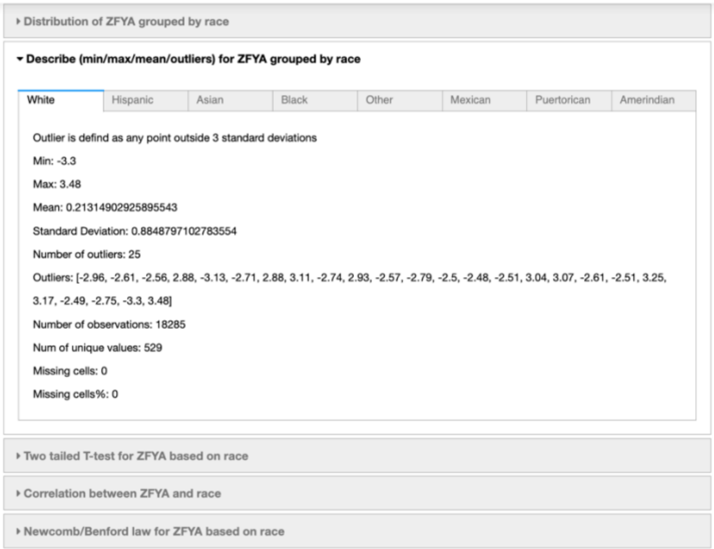
Normal distribution: A parametric approach which represents the behavior of most of the situations in the universe. It's characterised by a bell shaped. The diameter, weight, strength, and many other characteristics of natural, human or machine-made items are normally distributed. In humans, performance, outcomes, grade point averages etc. are all normally distributed. The normal distribution really is a normal occurrence. If we compare the normal distribution of training data outcomes across two groups we can preform statistical test (such as the one below) to determine if there is a **significant variance** between groups

Step 11(a): Analyse the distribution of the target.

The target is (ZFA) in this case.

For the purpose of the use case we have analysed ZFA as a normally distributed output with outliers removed.

Just looking at the distribution e.g. white applicants (the blue line v's black applicants (the red line) we can already visualise a significant difference in the distribution of outcomes across the various groups.



▸ Distribution of ZFYA grouped by race

▾ Describe (min/max/mean/outliers) for ZFYA grouped by race

Outlier is defind as any point outside 3 standard deviations

Min: -3.3

Max: 3.48

Mean: 0.21314902925895543

Standard Deviation: 0.8848797102783554

Number of outliers: 25

Outliers: [-2.96, -2.61, -2.56, 2.88, -3.13, -2.71, 2.88, 3.11, -2.74, 2.93, -2.57, -2.79, -2.5, -2.48, -2.51, 3.04, 3.07, -2.61, -2.51, 3.25, 3.17, -2.49, -2.75, -3.3, 3.48]

Number of observations: 18285

Num of unique values: 529

Missing cells: 0

Missing cells%: 0

▸ Two tailed T-test for ZFYA based on race

▸ Correlation between ZFYA and race

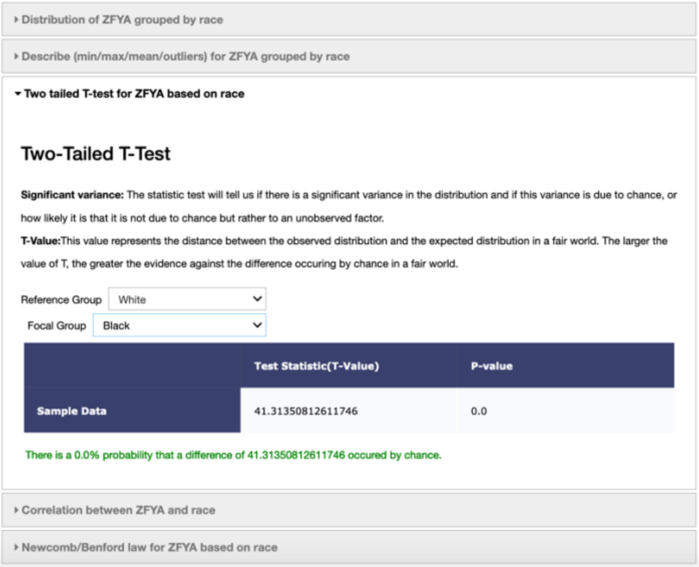
▸ Newcomb/Benford law for ZFYA based on race

Step 11(b): Descriptions.

Looking more closely under the descriptions section we can see that:

<u>White:</u>
Min: -3.3
Max: 3.48
Mean: 0.213
Standard Deviation: 0.88
Number of outliers: 25

<u>Black:</u>
Min: -3.35
Max: 2.6
Mean: -0.823
Standard Deviation: 0.92
Number of outliers: 4


Distribution of ZFYA grouped by race
Describe (min/max/mean/outliers) for ZFYA grouped by race
Two tailed T-test for ZFYA based on race
Correlation between ZFYA and race
Newcomb/Benford law for ZFYA based on race

Two-Tailed T-Test

Significant variance: The statistic test will tell us if there is a significant variance in the distribution and if this variance is due to chance, or how likely it is that it is not due to chance but rather to an unobserved factor.

T-Value: This value represents the distance between the observed distribution and the expected distribution in a fair world. The larger the value of T, the greater the evidence against the difference occuring by chance in a fair world.

Reference Group: White

Focal Group: Black

	Test Statistic(T-Value)	P-value
Sample Data	41.31350812611746	0.0

There is a 0.0% probability that a difference of 41.31350812611746 occured by chance.

Step 11(c): T-Test.

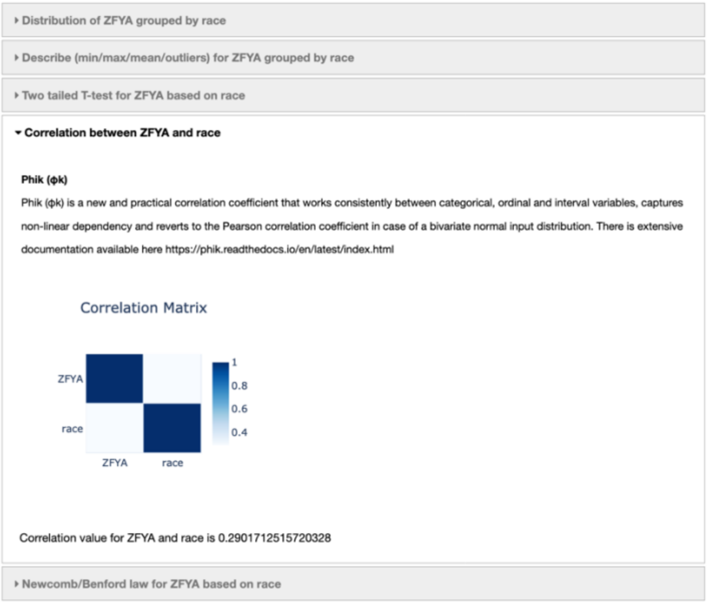
The tool provides a statistical significance test to determine if the difference in distribution is significantly different.

The T-Value and P-values are presented for those who are familiar with the test.

A natural language interpretation of the test is also provided for those less familiar with significance tests.

In this case when we analyse Black applicant's vs White applicants (as White is set as the privileged group). The T-value is seen to be 41.3.

The tool interprets this result as '*There is a 0% chance that this difference has occurred by chance*.'


Distribution of ZFYA grouped by race
Describe (min/max/mean/outliers) for ZFYA grouped by race
Two tailed T-test for ZFYA based on race
Correlation between ZFYA and race
Correlation Matrix
ZFYA
race
ZFYA
race
1
0.8
0.6
0.4
Newcomb/Benford law for ZFYA based on race

Phik (ϕk)

Phik (ϕk) is a new and practical correlation coefficient that works consistently between categorical, ordinal and interval variables, captures non-linear dependency and reverts to the Pearson correlation coefficient in case of a bivariate normal input distribution. There is extensive documentation available here https://phik.readthedocs.io/en/latest/index.html

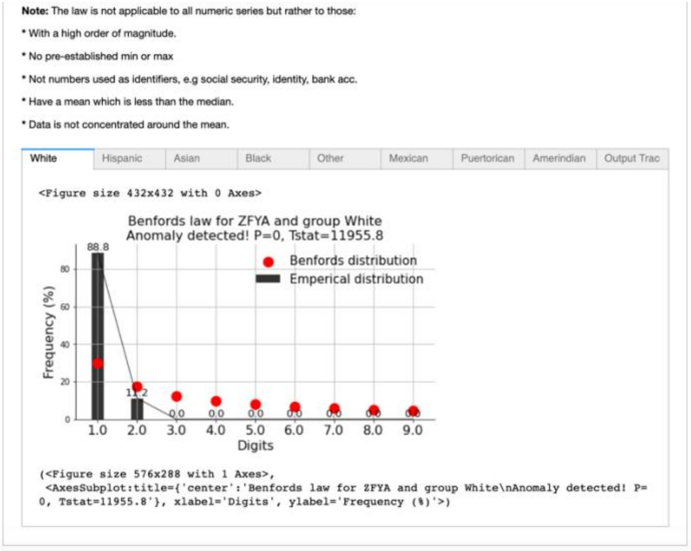
Correlation value for ZFYA and race is 0.2901712515720328

Step 11(d): Correlation.

The correlation section shows the correlation between the target and the protected feature being analysed.

In our use case we can see a **.29** correlation between ZFYA (the target) and race.

Phik (φk) correlation is used as it works consistently between categorical, ordinal and interval variables, it also captures non-linear dependency and reverts to the *Pearson correlation coefficient* in case of a bivariate normal input distribution.



Note: The law is not applicable to all numeric series but rather to those:

* With a high order of magnitude.

* No pre-established min or max

* Not numbers used as identifiers, e.g social security, identity, bank acc.

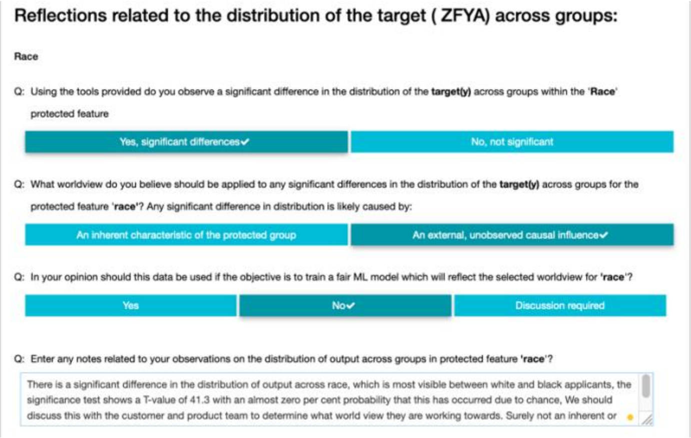
* Have a mean which is less than the median.

* Data is not concentrated around the mean.

```
<Figure size 432x432 with 0 Axes>
```

```
(<Figure size 576x288 with 1 Axes>,
 <AxesSubplot:title={'center':'Benfords law for ZFYA and group White\nAnomaly detected! P=
0, Tstat=11955.8'}, xlabel='Digits', ylabel='Frequency (%)'>)
```

Step 11(e): Bedford's' law.
Also known as *the law of first digits* or the phenomenon of significant digits. This law is the finding that the first numerals of the numbers found in series of records, of the most varied sources, do not display a uniform distribution, but rather a distribution as shown on the tool. This analysis is not useful for our use case, as Bedford's law is only applicable to numbers with a high order of magnitude. However, it is being shown here for completion. The law of first digits is often used in fraud detection, or to determine if numbers have been manually altered. Here we also provide results across groups.



Reflections related to the distribution of the target (ZFYA) across groups:

Race

Q: Using the tools provided do you observe a significant difference in the distribution of the **target(y)** across groups within the '**Race**' protected feature

[✓ Yes, significant differences] No, not significant

Q: What worldview do you believe should be applied to any significant differences in the distribution of the **target(y)** across groups for the protected feature '**race**'? Any significant difference in distribution is likely caused by:

An inherent characteristic of the protected group [✓ An external, unobserved causal influence]

Q: In your opinion should this data be used if the objective is to train a fair ML model which will reflect the selected worldview for '**race**'?

Yes [✓ No] Discussion required

Q: Enter any notes related to your observations on the distribution of output across groups in protected feature '**race**'?

There is a significant difference in the distribution of output across race, which is most visible between white and black applicants, the significance test shows a T-value of 41.3 with an almost zero per cent probability that this has occurred due to chance, We should discuss this with the customer and product team to determine what world view they are working towards. Surely not an inherent or

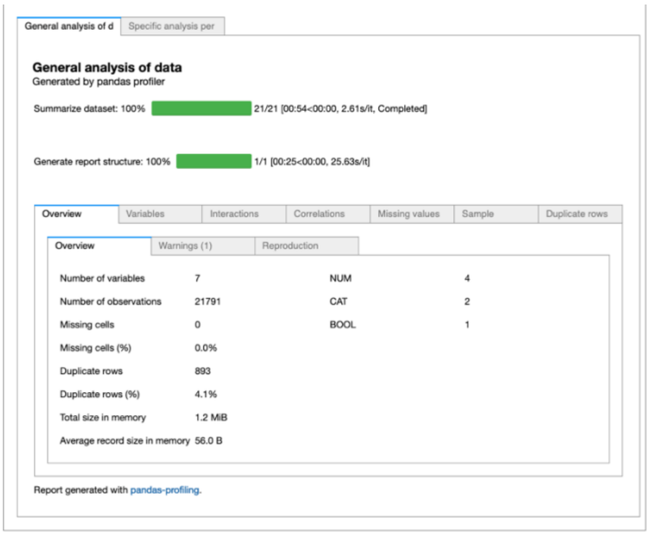
Step 12: Reflect on target distribution.

This step encourages accountability for model fairness.

The step also enables the practitioner to identify that conversations with domain expert or stakeholder is necessary.

The answers selected will form part of a report that the practitioner can use to engage in a discussion with stakeholders.

There is an intentional absence of '*don't know',* or similar options, and a text area where observations may be noted.



General analysis of d | Specific analysis per

General analysis of data
Generated by pandas profiler

Summarize dataset: 100% 21/21 [00:54<00:00, 2.61s/it, Completed]

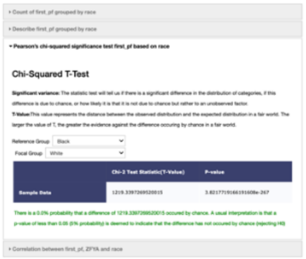
Generate report structure: 100% 1/1 [00:25<00:00, 25.63s/it]

Overview | Variables | Interactions | Correlations | Missing values | Sample | Duplicate rows

Overview | Warnings (1) | Reproduction

Number of variables	7	NUM	4
Number of observations	21791	CAT	2
Missing cells	0	BOOL	1
Missing cells (%)	0.0%		
Duplicate rows	893		
Duplicate rows (%)	4.1%		
Total size in memory	1.2 MiB		
Average record size in memory	56.0 B		

Report generated with pandas-profiling.



▸ Count of first_pf grouped by race

▸ Describe first_pf grouped by race

▾ Pearson's chi-squared significance test first_pf based on race

Chi-Squared T-Test

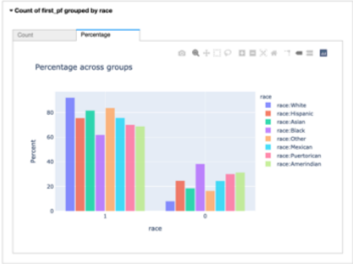
Significant variance: The statistic test will tell us if there is a significant difference in the distribution of categories, if this difference is due to chance, or how likely it is that it is not due to chance but rather to an unobserved factor.
T-Value: This value represents the distance between the observed distribution and the expected distribution in a fair world. The larger the value of T, the greater the evidence against the difference occuring by chance in a fair world.

Reference Group: Black
Focal Group: White

	Chi-2 Test Statistic(T-Value)	P-value
Sample Data	1219.3397269520015	3.8217719166191608e-267

There is a 0.0% probability that a difference of 1219.3397269520015 occured by chance. A usual interpretation is that a p-value of less than 0.05 (5% probability) is deemed to indicate that the difference has not occured by chance (rejecting H0)

▸ Correlation between first_pf, ZFYA and race



▾ Count of first_pf grouped by race

Step 13: Input feature analysis.

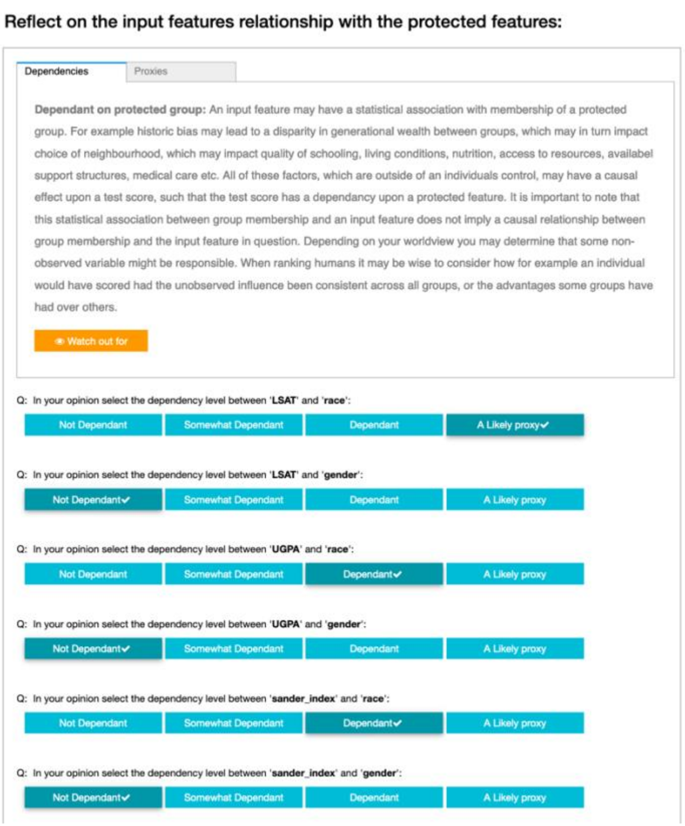
This step contains extensive visualization functionality, including:

General data visualization between input features and output.

Protected feature specific visualisation. For brevity all screen shots will not be included here. The protected feature visualisation is very similar to that for target visualisation.

If the input feature is categorical a *count* and *per centage* visualisation is supported.

If the input feature is categorical the statistical significance test supported is the Chi-Square T-test.



Reflect on the input features relationship with the protected features:

Dependencies | Proxies

Dependant on protected group: An input feature may have a statistical association with membership of a protected group. For example historic bias may lead to a disparity in generational wealth between groups, which may in turn impact choice of neighbourhood, which may impact quality of schooling, living conditions, nutrition, access to resources, availabel support structures, medical care etc. All of these factors, which are outside of an individuals control, may have a causal effect upon a test score, such that the test score has a dependancy upon a protected feature. It is important to note that this statistical association between group membership and an input feature does not imply a causal relationship between group membership and the input feature in question. Depending on your worldview you may determine that some non-observed variable might be responsible. When ranking humans it may be wise to consider how for example an individual would have scored had the unobserved influence been consistent across all groups, or the advantages some groups have had over others.

Q: In your opinion select the dependency level between '**LSAT**' and '**race**':
Not Dependant | Somewhat Dependant | Dependant | A Likely proxy✓

Q: In your opinion select the dependency level between '**LSAT**' and '**gender**':
Not Dependant✓ | Somewhat Dependant | Dependant | A Likely proxy

Q: In your opinion select the dependency level between '**UGPA**' and '**race**':
Not Dependant | Somewhat Dependant | Dependant✓ | A Likely proxy

Q: In your opinion select the dependency level between '**UGPA**' and '**gender**':
Not Dependant✓ | Somewhat Dependant | Dependant | A Likely proxy

Q: In your opinion select the dependency level between '**sander_index**' and '**race**':
Not Dependant | Somewhat Dependant | Dependant✓ | A Likely proxy

Q: In your opinion select the dependency level between '**sander_index**' and '**gender**':
Not Dependant✓ | Somewhat Dependant | Dependant | A Likely proxy

Step 14: Reflect on the input feature.

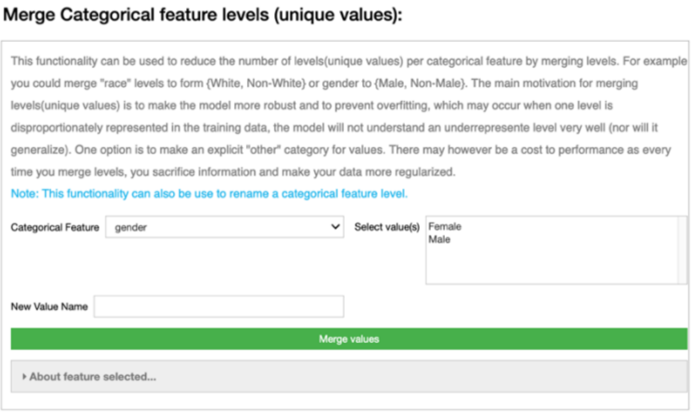
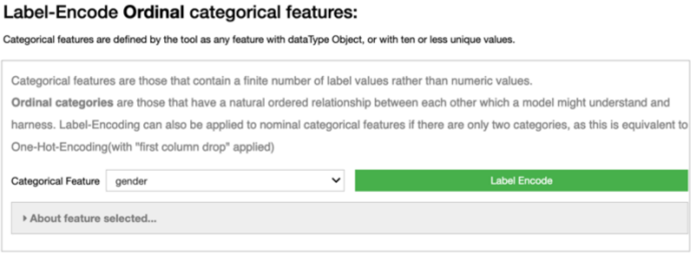
This step encourages accountability for model fairness, and facilitates the uncovering of historical bias and measurement bias introduced through the use of proxies for example.

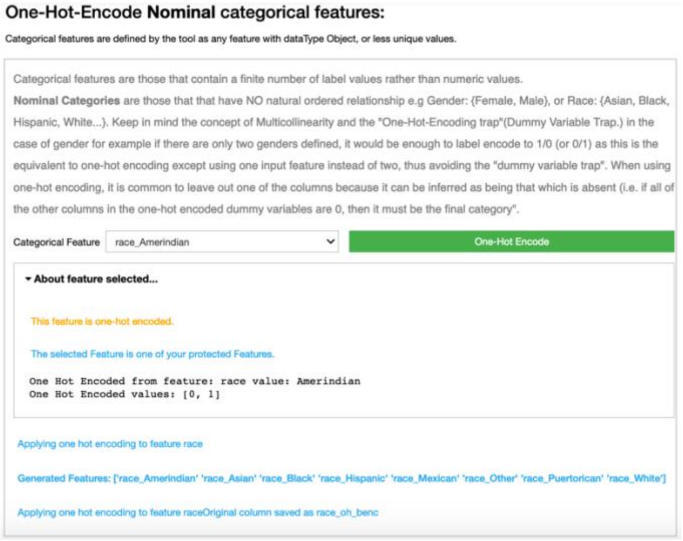
The step also enables the practitioner to identify when collaboration with domain expert or stakeholder may be necessary.

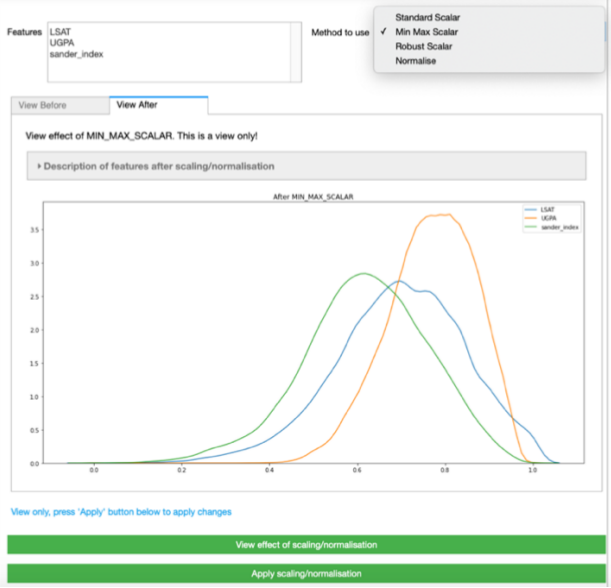
The answers selected will form part of a report that the practitioner can use to engage in a discussion with other stakeholders.

The tool also provides a description of dependencies and proxy features.

There is an intentional absence of ‘*don’t know’*, or similar options, and a text area where observations may be noted.

 **Merge Categorical feature levels (unique values):** This functionality can be used to reduce the number of levels(unique values) per categorical feature by merging levels. For example you could merge "race" levels to form {White, Non-White} or gender to {Male, Non-Male}. The main motivation for merging levels(unique values) is to make the model more robust and to prevent overfitting, which may occur when one level is disproportionately represented in the training data, the model will not understand an underrepresente level very well (nor will it generalize). One option is to make an explicit "other" category for values. There may however be a cost to performance as every time you merge levels, you sacrifice information and make your data more regularized. Note: This functionality can also be use to rename a categorical feature level. Categorical Feature: gender Select value(s): Female, Male New Value Name: 	**Step 15: Merge values.** This is basic feature engineering functionality which facilitated the requirement to merge categorical values within a categorical feature. For example, it could be used to merge cities into states to avoid an excessive number of features after one-hot encoding or similar encoding strategies.
 **Label-Encode Ordinal categorical features:** Categorical features are defined by the tool as any feature with dataType Object, or with ten or less unique values. Categorical features are those that contain a finite number of label values rather than numeric values. **Ordinal categories** are those that have a natural ordered relationship between each other which a model might understand and harness. Label-Encoding can also be applied to nominal categorical features if there are only two categories, as this is equivalent to One-Hot-Encoding(with "first column drop" applied) Categorical Feature: gender 	**Step 16: Label encoding.** This is basic feature engineering functionality which facilitated the requirement to label encode categorical features. There is also a description of the purpose of one hot encoding which may be useful for learners.


One-Hot-Encode Nominal categorical features:
Categorical features are defined by the tool as any feature with dataType Object, or less unique values.
Categorical features are those that contain a finite number of label values rather than numeric values.
Nominal Categories are those that that have NO natural ordered relationship e.g Gender: {Female, Male}, or Race: {Asian, Black, Hispanic, White...}. Keep in mind the concept of Multicollinearity and the "One-Hot-Encoding trap"(Dummy Variable Trap.) in the case of gender for example if there are only two genders defined, it would be enough to label encode to 1/0 (or 0/1) as this is the equivalent to one-hot encoding except using one input feature instead of two, thus avoiding the "dummy variable trap". When using one-hot encoding, it is common to leave out one of the columns because it can be inferred as being that which is absent (i.e. if all of the other columns in the one-hot encoded dummy variables are 0, then it must be the final category".
Categorical Feature: race_Amerindian
 	
This feature is one-hot encoded.
The selected Feature is one of your protected Features.
`One Hot Encoded from feature: race value: Amerindian`
`One Hot Encoded values: [0, 1]`
Applying one hot encoding to feature race
Generated Features: ['race_Amerindian' 'race_Asian' 'race_Black' 'race_Hispanic' 'race_Mexican' 'race_Other' 'race_Puertorican' 'race_White']
Applying one hot encoding to feature raceOriginal column saved as race_oh_benc
	Step 17: One-Hot encode.
	
	This is basic feature engineering functionality which facilitated the requirement to one-hot encode categorical features.
	
	There is also a description of the purpose of one hot encoding which may be useful for learners.

	**Step 18: Scaling or Normalisation.** This is also a basic feature engineering functionality which facilitated the requirement to one-hot encode categorical features. There is also a description of the purpose of one hot encoding which may be useful for learners.

Features: LSAT, UGPA, sander_index

Method to use: Standard Scalar, ✓ Min Max Scalar, Robust Scalar, Normalise

View Before | View After

View effect of MIN_MAX_SCALAR. This is a view only!

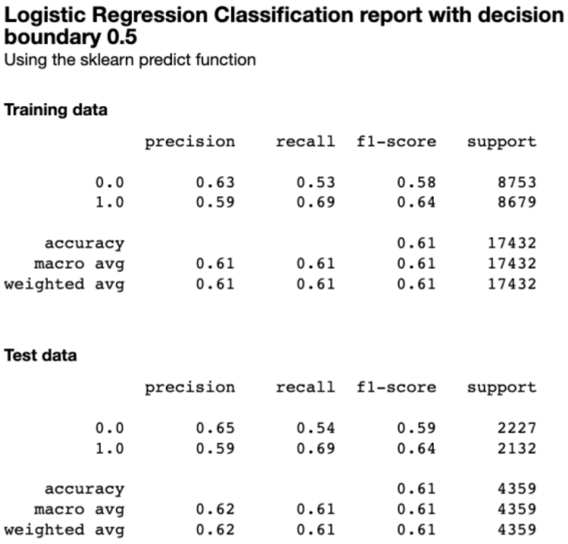
▸ Description of features after scaling/normalisation

View only, press 'Apply' button below to apply changes

View effect of scaling/normalisation
Apply scaling/normalisation

5.4. Fairness and explainability in the model

A logistic regression model is trained with the transformed data. The tool provides a graphical user interface such that the decision boundary can be set for the purpose of converting the continuous numeric output, ZFYA in this case, into a binary format to be used as the logistic regression target. For the purpose of the use case the '*sklearn*' LogisticRegression algorithm was used, values between -3.35 and 0.096 were converted to 0 and values between 0.096 and 3.48 will be converted to 1. No attempt was made to optimise the output as the purpose has been to demonstrate the framework functionality.



Logistic Regression Classification report with decision boundary 0.5

Using the sklearn predict function

Training data

```
              precision    recall  f1-score   support

         0.0       0.63      0.53      0.58      8753
         1.0       0.59      0.69      0.64      8679

    accuracy                           0.61     17432
   macro avg       0.61      0.61      0.61     17432
weighted avg       0.61      0.61      0.61     17432
```

Test data

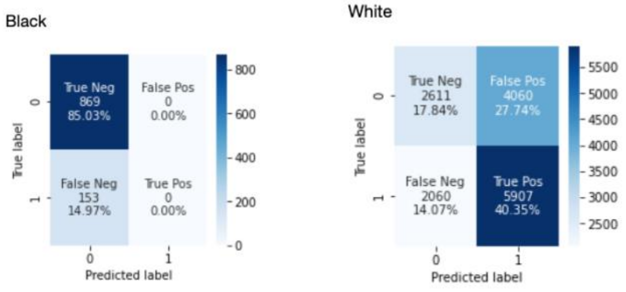
```
              precision    recall  f1-score   support

         0.0       0.65      0.54      0.59      2227
         1.0       0.59      0.69      0.64      2132

    accuracy                           0.61      4359
   macro avg       0.62      0.61      0.61      4359
weighted avg       0.62      0.61      0.61      4359
```

Figure 6: Logistic Regression model accuracy

5.4.1. Model Fairness

To invoke the group fairness analysis functionality it is necessary to import the helper and use the function below, this section is focused on the detection of evaluation bias

```
from machnamh import helper
mh.view_aequitas_fairness_metrics(X_train,   y_train,   y_train_pred,
data_summary)
```



Accuracy: 0.85

Precision/Positive predictive value(PPV): nan

Recall/True Positive Rate/Sensitivity: 0.0

F1 Score=: nan

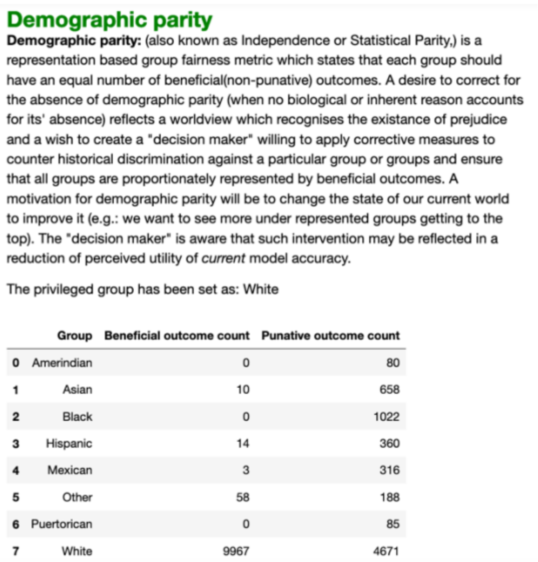
Accuracy: 0.58

Precision/Positive predictive value(PPV): 0.59

Recall/True Positive Rate/Sensitivity: 0.74

F1 Score=: 0.66

Confusion matrix:
An accuracy confusion matrix is generated for each group for each protected feature. For brevity two are shown. Note that although one-hot encoding has been applied the labels 'Black' and 'White' are displayed here. A feature of the framework is to persist these descriptions to helps with interpretation of the results. Notable here from a fairness perspective is that although false negatives are comparable between black and white, false positives (obtaining a position not deserved) is 0 for Black candidates and 27.74% for White candidates. So while accuracy appears much better for the black candidate group it comes with a social disadvantage and at the expense of identifying positive outcomes.



Demographic parity

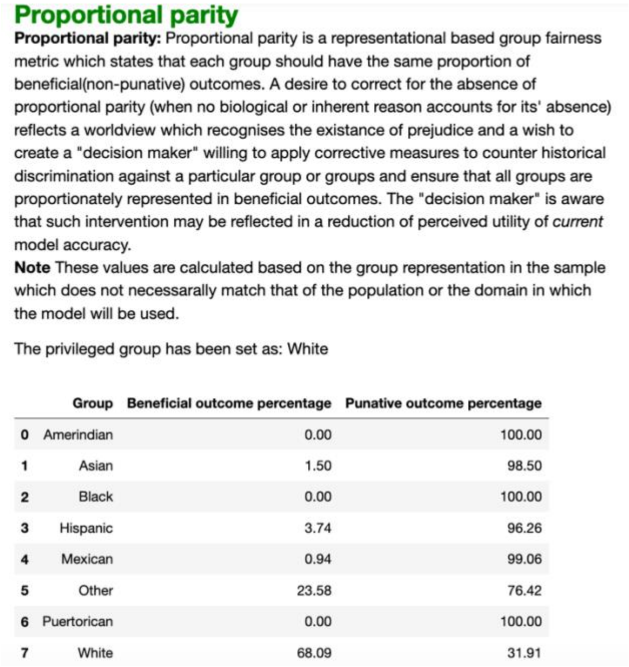
Demographic parity: (also known as Independence or Statistical Parity,) is a representation based group fairness metric which states that each group should have an equal number of beneficial(non-punative) outcomes. A desire to correct for the absence of demographic parity (when no biological or inherent reason accounts for its' absence) reflects a worldview which recognises the existance of prejudice and a wish to create a "decision maker" willing to apply corrective measures to counter historical discrimination against a particular group or groups and ensure that all groups are proportionately represented by beneficial outcomes. A motivation for demographic parity will be to change the state of our current world to improve it (e.g.: we want to see more under represented groups getting to the top). The "decision maker" is aware that such intervention may be reflected in a reduction of perceived utility of *current* model accuracy.

The privileged group has been set as: White

	Group	Beneficial outcome count	Punative outcome count
0	Amerindian	0	80
1	Asian	10	658
2	Black	0	1022
3	Hispanic	14	360
4	Mexican	3	316
5	Other	58	188
6	Puertorican	0	85
7	White	9967	4671

Demographic parity

It is very clear from the results that there is no demographic parity.
The framework takes care of labelling outcomes as beneficial or not based on the options selections during the analysis of the data.
The framework also takes care of reverse hot-encoding the values to display the racial groups with the easily interpreter
A description of the metric is provided to help with the interpretation of the results for the practitioner and as a point of reference for colaborative discussions with stakeholders and/or domain experts.



Proportional parity

Proportional parity: Proportional parity is a representational based group fairness metric which states that each group should have the same proportion of beneficial(non-punative) outcomes. A desire to correct for the absence of proportional parity (when no biological or inherent reason accounts for its' absence) reflects a worldview which recognises the existance of prejudice and a wish to create a "decision maker" willing to apply corrective measures to counter historical discrimination against a particular group or groups and ensure that all groups are proportionately represented in beneficial outcomes. The "decision maker" is aware that such intervention may be reflected in a reduction of perceived utility of *current* model accuracy.
Note These values are calculated based on the group representation in the sample which does not necessarally match that of the population or the domain in which the model will be used.

The privileged group has been set as: White

	Group	Beneficial outcome percentage	Punative outcome percentage
0	Amerindian	0.00	100.00
1	Asian	1.50	98.50
2	Black	0.00	100.00
3	Hispanic	3.74	96.26
4	Mexican	0.94	99.06
5	Other	23.58	76.42
6	Puertorican	0.00	100.00
7	White	68.09	31.91

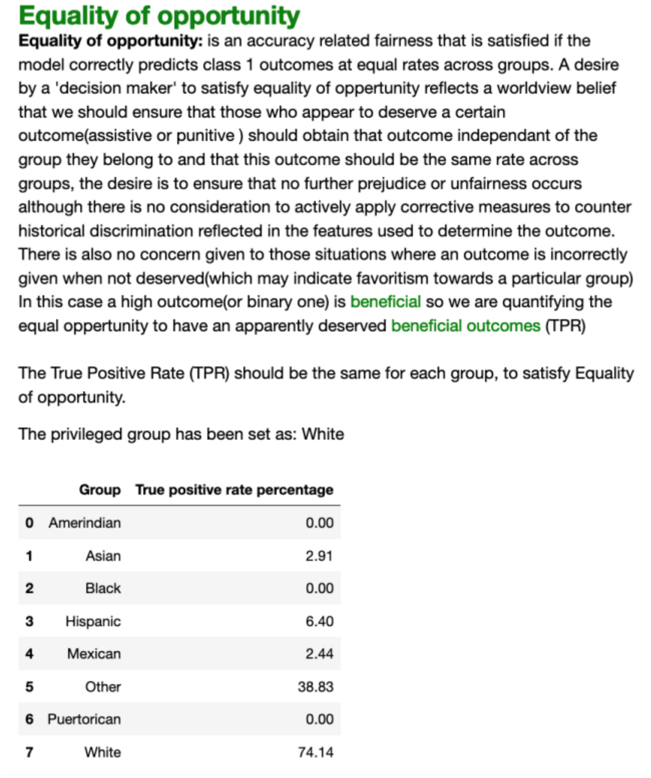
Proportional parity

It is very clear from the results that there is no proportional parity.

It should be noted that the framework also provides a reminder as to the privileged group selected in the data analysis phase.

Here we can see that the White group have by far the largest proportion of positive outcomes.

A description is also provided for the purpose of contet, understanding and discussion.



Equality of opportunity

Equality of opportunity: is an accuracy related fairness that is satisfied if the model correctly predicts class 1 outcomes at equal rates across groups. A desire by a 'decision maker' to satisfy equality of oppertunity reflects a worldview belief that we should ensure that those who appear to deserve a certain outcome(assistive or punitive) should obtain that outcome independant of the group they belong to and that this outcome should be the same rate across groups, the desire is to ensure that no further prejudice or unfairness occurs although there is no consideration to actively apply corrective measures to counter historical discrimination reflected in the features used to determine the outcome. There is also no concern given to those situations where an outcome is incorrectly given when not deserved(which may indicate favoritism towards a particular group) In this case a high outcome(or binary one) is beneficial so we are quantifying the equal oppertunity to have an apparently deserved beneficial outcomes (TPR)

The True Positive Rate (TPR) should be the same for each group, to satisfy Equality of opportunity.

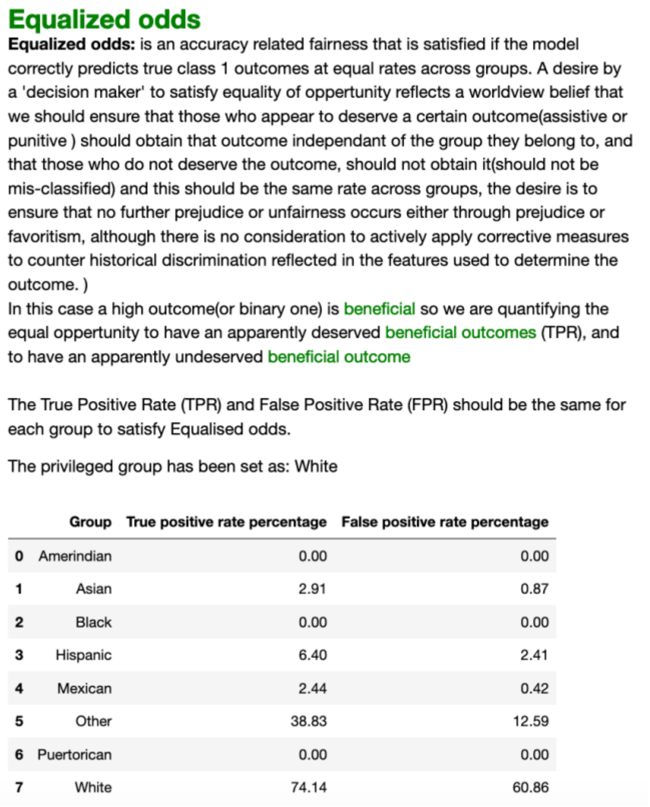
The privileged group has been set as: White

	Group	True positive rate percentage
0	Amerindian	0.00
1	Asian	2.91
2	Black	0.00
3	Hispanic	6.40
4	Mexican	2.44
5	Other	38.83
6	Puertorican	0.00
7	White	74.14

Equality of opportunity

In the use case we can see that the True positive rate is highest for White candidates.

The true positive rate per centage is most accurate for the white group.



Equalized odds

Equalized odds: is an accuracy related fairness that is satisfied if the model correctly predicts true class 1 outcomes at equal rates across groups. A desire by a 'decision maker' to satisfy equality of oppertunity reflects a worldview belief that we should ensure that those who appear to deserve a certain outcome(assistive or punitive) should obtain that outcome independant of the group they belong to, and that those who do not deserve the outcome, should not obtain it(should not be mis-classified) and this should be the same rate across groups, the desire is to ensure that no further prejudice or unfairness occurs either through prejudice or favoritism, although there is no consideration to actively apply corrective measures to counter historical discrimination reflected in the features used to determine the outcome.)
In this case a high outcome(or binary one) is beneficial so we are quantifying the equal oppertunity to have an apparently deserved beneficial outcomes (TPR), and to have an apparently undeserved beneficial outcome

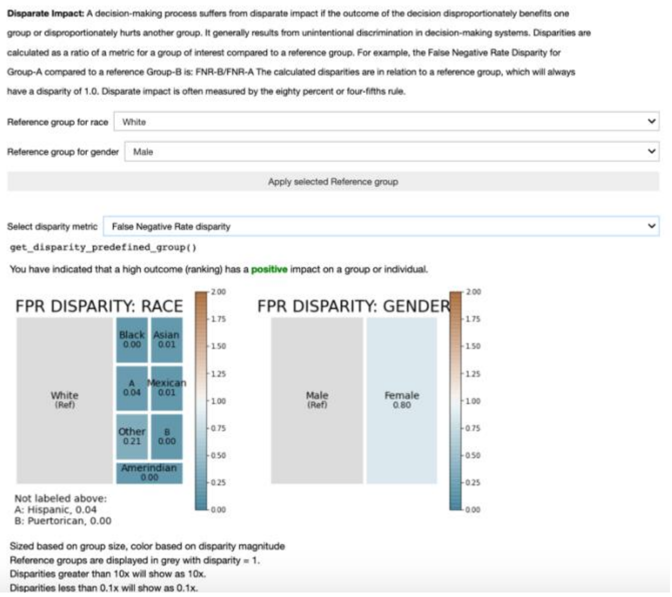
The True Positive Rate (TPR) and False Positive Rate (FPR) should be the same for each group to satisfy Equalised odds.

The privileged group has been set as: White

	Group	True positive rate percentage	False positive rate percentage
0	Amerindian	0.00	0.00
1	Asian	2.91	0.87
2	Black	0.00	0.00
3	Hispanic	6.40	2.41
4	Mexican	2.44	0.42
5	Other	38.83	12.59
6	Puertorican	0.00	0.00
7	White	74.14	60.86

Equalized odds

The framework reminds the practitioner that in this context that true positive rates and false positive rates are beneficial



Disparate Impact: A decision-making process suffers from disparate impact if the outcome of the decision disproportionately benefits one group or disproportionately hurts another group. It generally results from unintentional discrimination in decision-making systems. Disparities are calculated as a ratio of a metric for a group of interest compared to a reference group. For example, the False Negative Rate Disparity for Group-A compared to a reference Group-B is: FNR-B/FNR-A The calculated disparities are in relation to a reference group, which will always have a disparity of 1.0. Disparate impact is often measured by the eighty percent or four-fifths rule.

Reference group for race: White

Reference group for gender: Male

Select disparity metric: False Negative Rate disparity

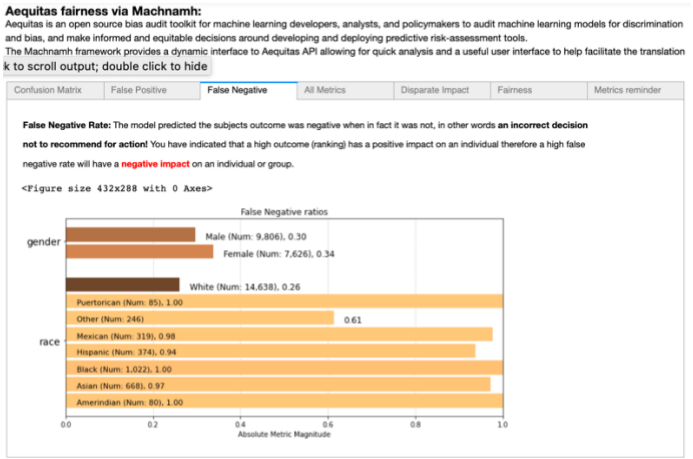
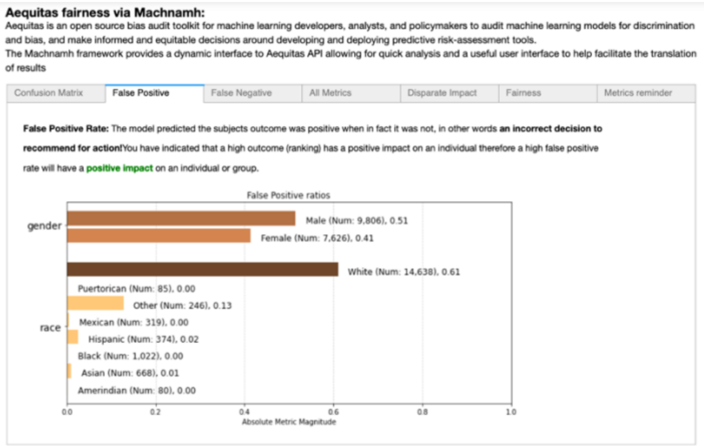
```
get_disparity_predefined_group()
```

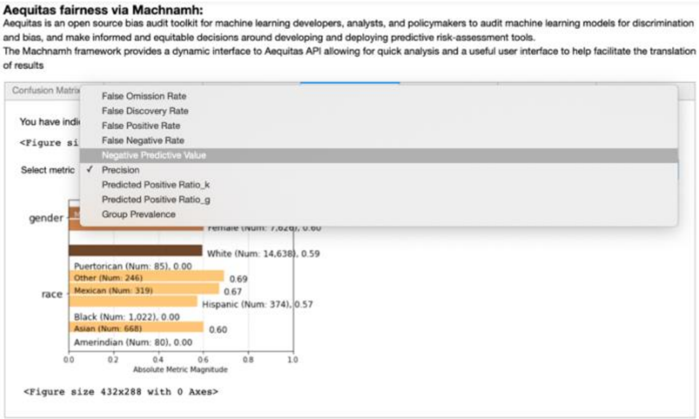
You have indicated that a high outcome (ranking) has a positive impact on a group or individual.

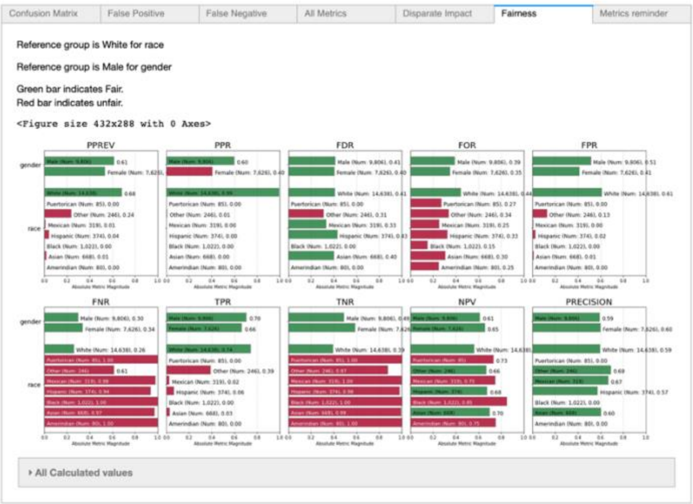
Sized based on group size, color based on disparity magnitude
Reference groups are displayed in grey with disparity = 1.
Disparities greater than 10x will show as 10x.
Disparities less than 0.1x will show as 0.1x.

Disparate Impact

The tool defaults to the privileged group as defined in the data analysis phase. A description of disparate impact is provided in addition to a reminder that a high outcome has a positive effect. The tool also provides some additional text to help interpret the Aequitsa API response.

 **Aequitas fairness via Machnamh:** Aequitas is an open source bias audit toolkit for machine learning developers, analysts, and policymakers to audit machine learning models for discrimination and bias, and make informed and equitable decisions around developing and deploying predictive risk-assessment tools. The Machnamh framework provides a dynamic interface to Aequitas API allowing for quick analysis and a useful user interface to help facilitate the translation k to scroll output; double click to hide Confusion Matrix False Positive False Negative All Metrics Disparate Impact Fairness Metrics reminder **False Negative Rate:** The model predicted the subjects outcome was negative when in fact it was not, in other words **an incorrect decision not to recommend for action!** You have indicated that a high outcome (ranking) has a positive impact on an individual therefore a high false negative rate will have a **negative impact** on an individual or group. `<Figure size 432x288 with 0 Axes>` False Negative ratios gender Male (Num: 9,806), 0.30 Female (Num: 7,626), 0.34 White (Num: 14,638), 0.26 Puertorican (Num: 85), 1.00 Other (Num: 246) 0.61 race Mexican (Num: 319), 0.98 Hispanic (Num: 374), 0.94 Black (Num: 1,022), 1.00 Asian (Num: 668), 0.97 Amerindian (Num: 80), 1.00 0.0 0.2 0.4 0.6 0.8 1.0 Absolute Metric Magnitude	**False Negatives** A graphical representation of the false negatives is provided. The framework also reminds the practitioner that in this implementation, based on the selections provided during the analysis that false negatives will have particularly negative consequences.
 **Aequitas fairness via Machnamh:** Aequitas is an open source bias audit toolkit for machine learning developers, analysts, and policymakers to audit machine learning models for discrimination and bias, and make informed and equitable decisions around developing and deploying predictive risk-assessment tools. The Machnamh framework provides a dynamic interface to Aequitas API allowing for quick analysis and a useful user interface to help facilitate the translation of results Confusion Matrix False Positive False Negative All Metrics Disparate Impact Fairness Metrics reminder **False Positive Rate:** The model predicted the subjects outcome was positive when in fact it was not, in other words **an incorrect decision to recommend for action!**You have indicated that a high outcome (ranking) has a positive impact on an individual therefore a high false positive rate will have a **positive impact** on an individual or group. False Positive ratios gender Male (Num: 9,806), 0.51 Female (Num: 7,626), 0.41 White (Num: 14,638), 0.61 Puertorican (Num: 85), 0.00 Other (Num: 246), 0.13 race Mexican (Num: 319), 0.00 Hispanic (Num: 374), 0.02 Black (Num: 1,022), 0.00 Asian (Num: 668), 0.01 Amerindian (Num: 80), 0.00 0.0 0.2 0.4 0.6 0.8 1.0 Absolute Metric Magnitude	**False Positives** A graphical representation of the false positives is provided. The framework also reminds the practitioner that in this implementation, based on the selections provided during the analysis. that false positives will have beneficial consequences.


Aequitas fairness via Machnamh:
Aequitas is an open source bias audit toolkit for machine learning developers, analysts, and policymakers to audit machine learning models for discrimination and bias, and make informed and equitable decisions around developing and deploying predictive risk-assessment tools.
The Machnamh framework provides a dynamic interface to Aequitas API allowing for quick analysis and a useful user interface to help facilitate the translation of results
 Confusion Matrix
You have indi
`<Figure si`
Select metric
False Omission Rate
False Discovery Rate
False Positive Rate
False Negative Rate
Negative Predictive Value
✓ Precision
Predicted Positive Ratio_k
Predicted Positive Ratio_g
Group Prevalence
gender
Female (Num: 7,626), 0.60
White (Num: 14,638), 0.59
Puertorican (Num: 85), 0.00
Other (Num: 246)
0.69
race
Mexican (Num: 319)
0.67
Hispanic (Num: 374), 0.57
Black (Num: 1,022), 0.00
Asian (Num: 668)
0.60
Amerindian (Num: 80), 0.00
0.0
0.2
0.4
0.6
0.8
1.0
Absolute Metric Magnitude
`<Figure size 432x288 with 0 Axes>`
	All fairness measurements.
	
	The practitioner can review all of the measurements supported by the Aequitas api



Reference group is White for race

Reference group is Male for gender

Green bar indicates Fair.
Red bar indicates unfair.

```
<Figure size 432x288 with 0 Axes>
```

Fairness summary

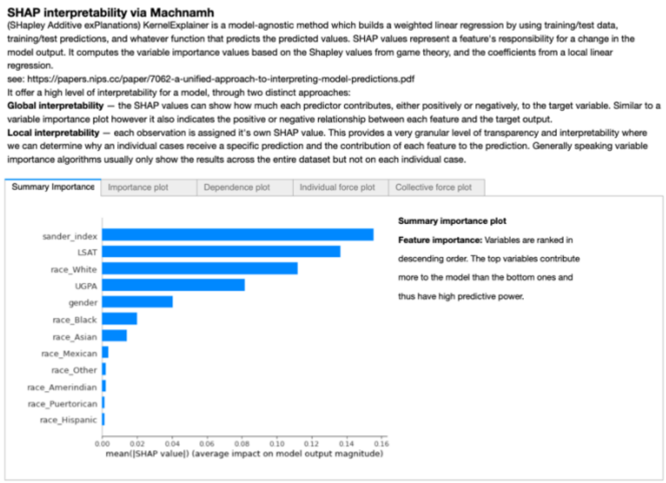
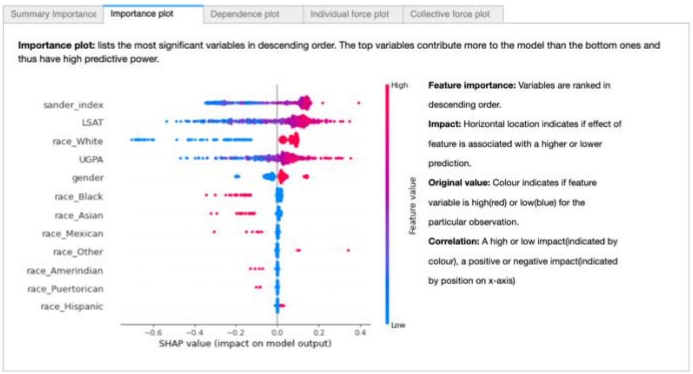
A snapshot of fairness across all available measurements, where green indicates fair and red indicates unfair.

Unfairness may also be flagged where representation in the data is low.

5.4.2. Model Explainability

To invoke the explainability functionality it is necessary to import the helper and run the commands below:

```
from machnamh import helper
explainer,shap_values,x=mh.run_shap_and_serialise_response(
                                                    X_train[training_features_list],
                                                    logistic_reg_model.predict,
                                                    500,
                                                    save_to_path = './output/')

mh.shap_analysis(shap_values, explainer, x, data_summary)
```

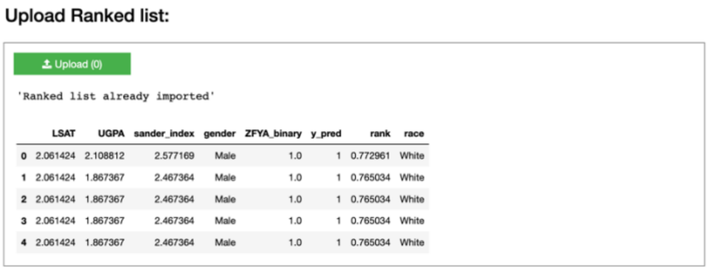
 **SHAP interpretability via Machnamh** (SHapley Additive exPlanations) KernelExplainer is a model-agnostic method which builds a weighted linear regression by using training/test data, training/test predictions, and whatever function that predicts the predicted values. SHAP values represent a feature's responsibility for a change in the model output. It computes the variable importance values based on the Shapley values from game theory, and the coefficients from a local linear regression. see: https://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf It offer a high level of interpretability for a model, through two distinct approaches: **Global interpretability** — the SHAP values can show how much each predictor contributes, either positively or negatively, to the target variable. Similar to a variable importance plot however it also indicates the positive or negative relationship between each feature and the target output. **Local interpretability** — each observation is assigned it's own SHAP value. This provides a very granular level of transparency and interpretability where we can determine why an individual cases receive a specific prediction and the contribution of each feature to the prediction. Generally speaking variable importance algorithms usually only show the results across the entire dataset but not on each individual case. **Summary importance plot** **Feature importance:** Variables are ranked in descending order. The top variables contribute more to the model than the bottom ones and thus have high predictive power. Summary Importance Importance plot Dependence plot Individual force plot Collective force plot sander_index LSAT race_White UGPA gender race_Black race_Asian race_Mexican race_Other race_Amerindian race_Puertorican race_Hispanic 0.00 0.02 0.04 0.06 0.08 0.10 0.12 0.14 0.16 mean(|SHAP value|) (average impact on model output magnitude)	**Summary Importance plot:** The importance plot shows the importance the model gives to each feature. Variables are ranked in descending order. The top variables contribute more to the model than the bottom ones and thus have high predictive power. In our use case we left race and gender in the training data, we can see from the output that *Sanders_index* and *LSAT* are given the most importance and that race 'white' actually has more importance than *UGPA.*
 **Importance plot:** lists the most significant variables in descending order. The top variables contribute more to the model than the bottom ones and thus have high predictive power. **Feature importance:** Variables are ranked in descending order. **Impact:** Horizontal location indicates if effect of feature is associated with a higher or lower prediction. **Original value:** Colour indicates if feature variable is high(red) or low(blue) for the particular observation. **Correlation:** A high or low impact(indicated by colour), a positive or negative impact(indicated by position on x-axis) Summary Importance Importance plot Dependence plot Individual force plot Collective force plot sander_index LSAT race_White UGPA gender race_Black race_Asian race_Mexican race_Other race_Amerindian race_Puertorican race_Hispanic −0.6 −0.4 −0.2 0.0 0.2 0.4 SHAP value (impact on model output) High Feature value Low	**Importance plot:** The importance plot provides more information than the summary plot in this case we can see the direction of the importance. While race 'white' =1(red) has a positive influence on the results we can see that the next most influential racial group 'black' = 1(red) has a negative impact.

5.5. Fairness in Ranking

To invoke the fairness in ranking user interface it is necessary to import the helper and run the commands below.

```
from machnamh import helper
ranked_fairness = review_ranked_fairness_UI(data_summary,
                                              X,
                                              y,
                                              y_prob_1,
                                              y_pred)

ranked_fairness.render()
```

For the Fairness in ranking aspect of the use case, although an option exists to reduce the list to the top-k number or top-p per cent, for any given context we will use the full list for the purpose of demonstration.



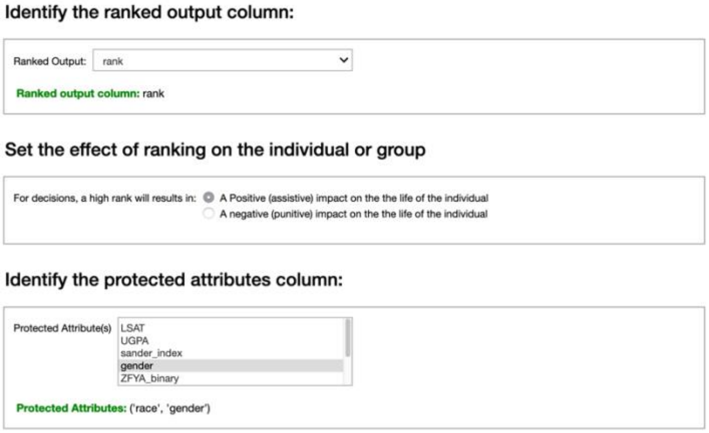
Upload Ranked list:

```
'Ranked list already imported'
```

	LSAT	UGPA	sander_index	gender	ZFYA_binary	y_pred	rank	race
0	2.061424	2.108812	2.577169	Male	1.0	1	0.772961	White
1	2.061424	1.867367	2.467364	Male	1.0	1	0.765034	White
2	2.061424	1.867367	2.467364	Male	1.0	1	0.765034	White
3	2.061424	1.867367	2.467364	Male	1.0	1	0.765034	White
4	2.061424	1.867367	2.467364	Male	1.0	1	0.765034	White

Step 1: Upload the ranked list.
It is possible to call the rank fairness function by identifyin the various components of the model in the function call, in addition to the *data_summary class* from the first step in the analysis which contains all of the gathered information such as the protected features, impact of ranking and the ranked column.

Or any .csv file may be uploaded in which case these values may be selected.



Identify the ranked output column:

Ranked Output: rank

Ranked output column: rank

Set the effect of ranking on the individual or group

For decisions, a high rank will results in: ● A Positive (assistive) impact on the the life of the individual
○ A negative (punitive) impact on the the life of the individual

Identify the protected attributes column:

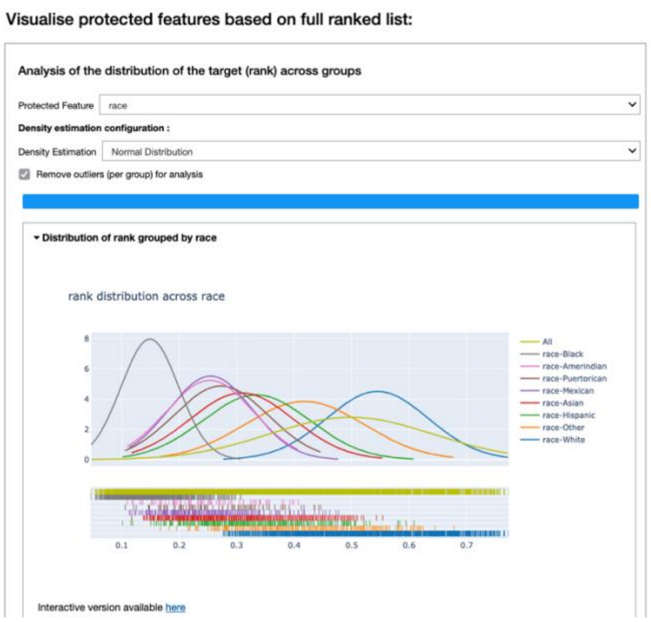
Protected Attribute(s) LSAT
UGPA
sander_index
gender
ZFYA_binary

Protected Attributes: ('race', 'gender')

Step 2: Identify the ranked column.
This will be automatically selected if the function is called with the ‘data_summary’ output from the data analysis phase.

Step 3: Identify the effect of the ranking.
Will a high ranking result in an assistive or punative outcome for the individual?

Step 4: Identify the protected features.
This will also be preselected or can be manually set here.



Visualise protected features based on full ranked list:

Analysis of the distribution of the target (rank) across groups

Protected Feature: race

Density estimation configuration :

Density Estimation: Normal Distribution

☑ Remove outliers (per group) for analysis

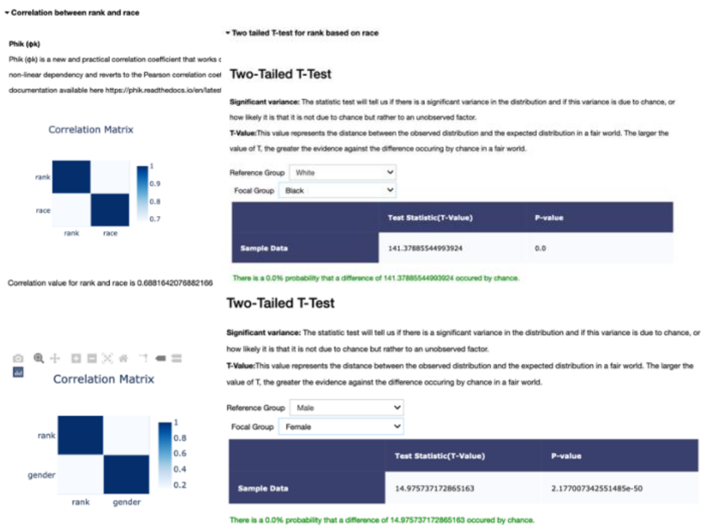
▾ **Distribution of rank grouped by race**

Interactive version available here

Step 5a: Review the distribution of the full ranked list.

This is similar functionality to that of the targer analysis in the first phase during the analysis of the data.

Here a significant dfference in the ditribution of the ranking scores across race can be visualised.



▾ **Correlation between rank and race**

Phik (φk)

Phik (φk) is a new and practical correlation coefficient that works c non-linear dependency and reverts to the Pearson correlation coe documentation available here https://phik.readthedocs.io/en/latest

Correlation value for rank and race is 0.6881642076882166

▾ **Two tailed T-test for rank based on race**

Two-Tailed T-Test

Significant variance: The statistic test will tell us if there is a significant variance in the distribution and if this variance is due to chance, or how likely it is that it is not due to chance but rather to an unobserved factor.

T-Value: This value represents the distance between the observed distribution and the expected distribution in a fair world. The larger the value of T, the greater the evidence against the difference occuring by chance in a fair world.

Reference Group: White

Focal Group: Black

	Test Statistic(T-Value)	P-value
Sample Data	141.37885544993924	0.0

There is a 0.0% probability that a difference of 141.37885544993924 occured by chance.

Two-Tailed T-Test

Significant variance: The statistic test will tell us if there is a significant variance in the distribution and if this variance is due to chance, or how likely it is that it is not due to chance but rather to an unobserved factor.

T-Value: This value represents the distance between the observed distribution and the expected distribution in a fair world. The larger the value of T, the greater the evidence against the difference occuring by chance in a fair world.

Reference Group: Male

Focal Group: Female

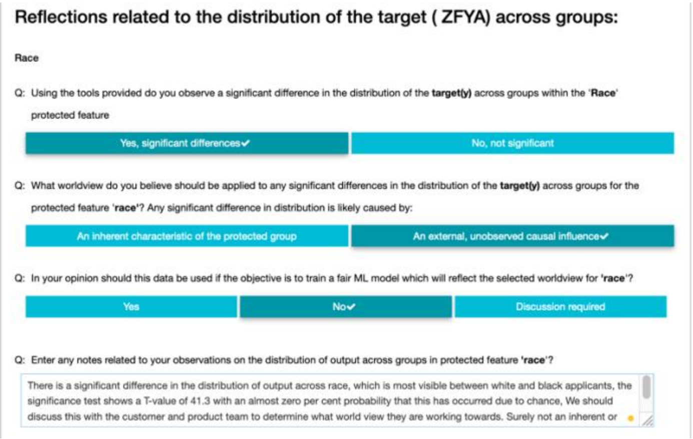
	Test Statistic(T-Value)	P-value
Sample Data	14.975737172865163	2.177007342551485e-50

There is a 0.0% probability that a difference of 14.975737172865163 occured by chance.

Step 5b: Correlation
Here we can see that the correlation between ranking and race is .648, and between ranking and gender it is .16

Step 5c: Significance Test
In terms of ranking, when the significance test is viewed based on the distribution of Black candidates and White candidates the T-value is 141.37 with a 0% possibiltiy that this result has happened by chance.

T-value between Male and Female is 15, with an almost 0% probability that the difference has occurred by chance.



Reflections related to the distribution of the target (ZFYA) across groups:

Race

Q: Using the tools provided do you observe a significant difference in the distribution of the **target(y)** across groups within the '**Race**' protected feature

[✓ Yes, significant differences✔] No, not significant

Q: What worldview do you believe should be applied to any significant differences in the distribution of the **target(y)** across groups for the protected feature '**race**'? Any significant difference in distribution is likely caused by:

An inherent characteristic of the protected group [✓ An external, unobserved causal influence✔]

Q: In your opinion should this data be used if the objective is to train a fair ML model which will reflect the selected worldview for '**race**'?

Yes [✓ No✔] Discussion required

Q: Enter any notes related to your observations on the distribution of output across groups in protected feature '**race**'?

There is a significant difference in the distribution of output across race, which is most visible between white and black applicants, the significance test shows a T-value of 41.3 with an almost zero per cent probability that this has occurred due to chance, We should discuss this with the customer and product team to determine what world view they are working towards. Surely not an inherent or

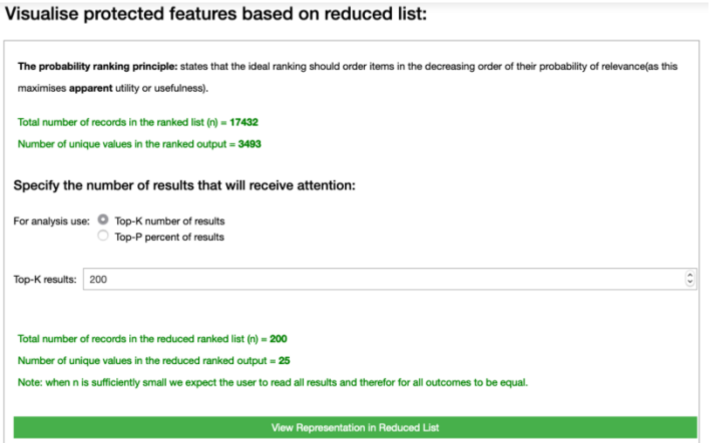
Step 6: Reflection on the distribution of the full ranking

This step encourages accountability for model fairness.

The step also enables the practitioner to identify that conversations with domain expert or stakeholder is necessary.

The answers selected will form part of a report that the practitioner can use to engage in a discussion with stakeholders.

There is an intentional absence of '*don't know'*, or similar options, and a text area where observations may be noted.



Visualise protected features based on reduced list:

The probability ranking principle: states that the ideal ranking should order items in the decreasing order of their probability of relevance(as this maximises **apparent** utility or usefulness).

Total number of records in the ranked list (n) = **17432**

Number of unique values in the ranked output = **3493**

Specify the number of results that will receive attention:

For analysis use: ● Top-K number of results
○ Top-P percent of results

Top-K results: 200

Total number of records in the reduced ranked list (n) = **200**

Number of unique values in the reduced ranked output = **25**

Note: when n is sufficiently small we expect the user to read all results and therefor for all outcomes to be equal.

Step 7a: Consider the probability ranking principle.

This step in the analysis prompts the practitioner to reduce the list to the approximate number of users who will receive attantion. This number will vary depending on the context in which the application is being used.

Step 7b: Reduce the ranked list
As not all results will be based on how many candidates will receive attention. For the purpose of the Use Case we have not reduced the list.

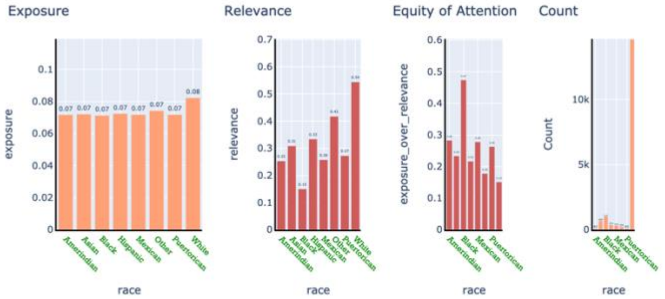


Fig.a Standard log discount(with average)[42][40]

Step 8: Detect Inequality of attention:

Using a standard logarithmic discount as a measure of exposure drop off where: exposure = 1/math.log(j+2, 2). j is the position in the list. Outliers and low group counts can cause mis-interpretation.

Fig.a: Calculate exposure for each position in the ranked list as discussed *by Joachims' and Singh* in *Fairness of Exposure in Rankings.[42]*. Compare the average exposure per group against the average relevance per group, as discussed in the same paper. Review 'Group Equity of

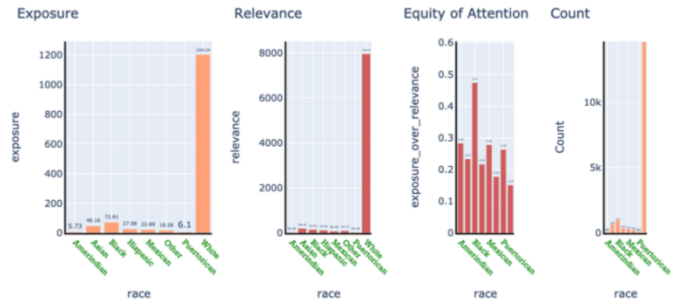


Fig b. Standard log discount(with sum)[42][40][39]

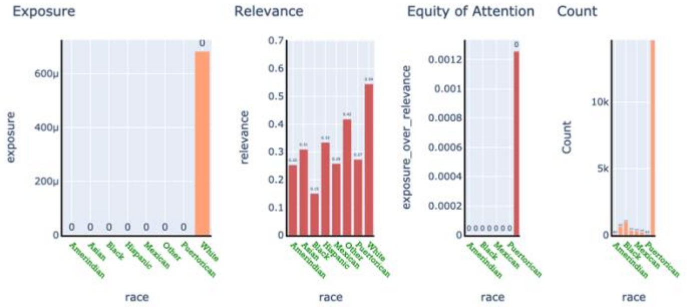
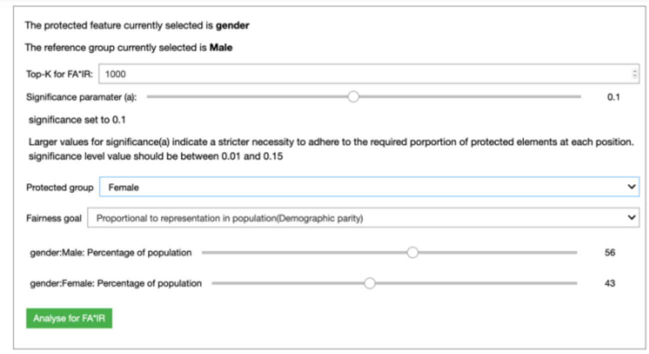


Fig c. Standard log discount(with Exposure is set to p up to a position k*)[42][40]*

attention' as discussed by *Biega et al* in *Equity of Attention: Amortizing Individual Fairness in Rankings [40]*

Fig.b: Rather than comparing the average exposure, the sum of the exposure is used as per *Sapiezynski et al* in *Quantifying the Impact of User Attention on Fair Group Representation in Ranked Lists.*

Fig.c: Exposure is set to p up to a position k, and to 0 when the position is lower than k. Currently p is set to 1 and k is set to 10 and group average value is used.[40]



The protected feature currently selected is **gender**

The reference group currently selected is **Male**

Top-K for FA*IR: 1000

Significance paramater (a): 0.1

significance set to 0.1

Larger values for significance(a) indicate a stricter necessity to adhere to the required porportion of protected elements at each position. significance level value should be between 0.01 and 0.15

Protected group: Female

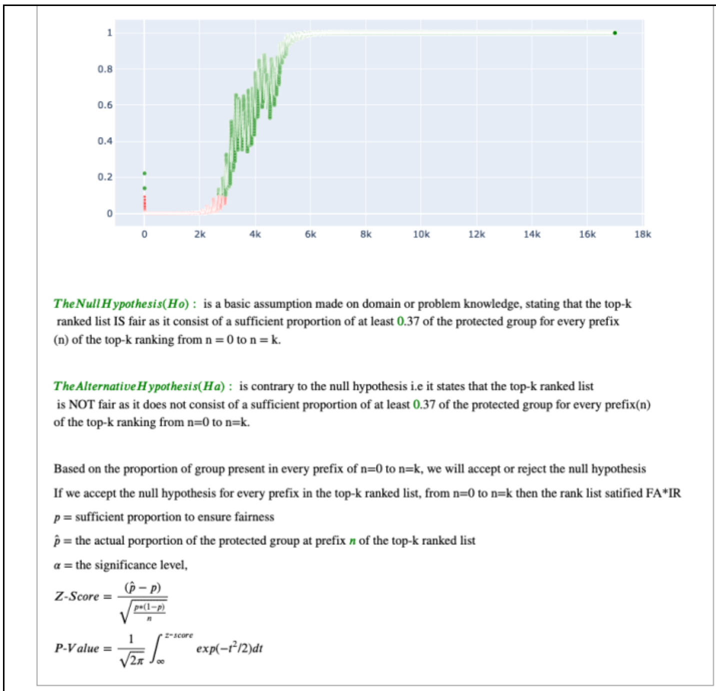
Fairness goal: Proportional to representation in population(Demographic parity)

gender:Male: Percentage of population 56

gender:Female: Percentage of population 43

Step 9. FA*IR analysis
The FA*IR analysis is based on the assumption that the top-k ranked list is fair if it consists of a sufficient proportion of at least *p* of the protected group for every prefix(n) of the top-k ranking from n = 0 to n = k.

The framework prompts for a fairness goal and sets the proportion to the default value of the goal. E.g a demographic parity selection will pick up the proportions from the dataset. However as this is not always reflective it can be modified.



TheNullHypothesis(Ho) : is a basic assumption made on domain or problem knowledge, stating that the top-k ranked list IS fair as it consist of a sufficient proportion of at least 0.37 of the protected group for every prefix (n) of the top-k ranking from n = 0 to n = k.

TheAlternativeHypothesis(Ha) : is contrary to the null hypothesis i.e it states that the top-k ranked list is NOT fair as it does not consist of a sufficient proportion of at least 0.37 of the protected group for every prefix(n) of the top-k ranking from n=0 to n=k.

Based on the proportion of group present in every prefix of n=0 to n=k, we will accept or reject the null hypothesis

If we accept the null hypothesis for every prefix in the top-k ranked list, from n=0 to n=k then the rank list satified FA*IR

p = sufficient proportion to ensure fairness

$\hat{p}$ = the actual porportion of the protected group at prefix n of the top-k ranked list

α = the significance level,

$$Z\text{-}Score = \frac{(\hat{p} - p)}{\sqrt{\frac{p*(1-p)}{n}}}$$

$$P\text{-}Value = \frac{1}{\sqrt{2\pi}} \int_{\infty}^{z\text{-}score} exp(-t^2/2)dt$$

In the use case while the ranked list was fair for White candidates it was at no point fair for Black candidates regardless of the size of K, with significance level set to 0.1 and targeted minimum proportion =18%

For purpose of demonstration the FA*IR analysis was ran for Female candidate (as shown)

There is also an explanation of interpretation.

The output is green when the list is fair for the group in question, and red otherwise. In this case the list is fair at the point where 2632 candidates will receive attention, and remains so for larger values of k.

5.6. Reflections on dataset

It is clear looking at the distribution of the features and the target across groups that disparities exist. The intuition is to consider exam results as being largely objective. However, the framework provides evidence that this is not the case. As the largest disparity has been observed between White candidates and Black Candidates it is these groups that have been the focus of the use case.

A statistically significant difference of T-value **40.6** per the two-tailed T-test.	
White Min: -3.3	Black Min: -3.35
White Max: 3.48	Black Max: 2.6
White Mean: 0.21	**Black Mean: -0.822**
A correlation of **.28** was also observed between race and ZFYA	

Figure 7. Result of analysis for ZFYA

A statistically significant difference of T-value **32.2** per the two-tailed T-test.	
White Min: 1.9	Black Min: 1.8
White Max: 4.2	Black Max: 4
White Mean: 3.3	**Black Mean: 2.89**
A correlation of **.20** was also observed between race and UGPA	
A correlation of **.28** was also observed between race and ZFYA	

Figure 8. Result of analysis for UGPA

A statistically significant difference of T-value **57.7** as per the two-tailed T-test.	
White Min: 17	Black Min: 11
White Max: 48	Black Max: 47
White Mean: 37.5	**Black Mean: 29.4**
A correlation of **.36** was also observed between race and LSAT	
A correlation of **.28** was also observed between race and ZFYA	

Figure 9. Result of analysis for LSAT

A statistically significant difference of T-value 61.5 per the two-tailed T-test.	
White Min: .44	Black Min: .38
White Max: 1	Black Max: .91
White Mean: .77	Black Mean: .64
A correlation of **.38** was also observed between race and Sander_index	
A correlation of **.28** was also observed between race and ZFYA	

Figure 10. Result of analysis for Sander_index

In addition to the initial findings within the data we can also see from the results outlined in the use case that disparities exist across group fairness and also across ranking fairness.

Grade point average (GPA) and First Year Average (FYA) are based on the academic achievements of students through class participation and teacher/professor exam and project grading. Significant disparities exist for which a worldview must be selected. It goes without saying that there are no inherent or biological reasons which would account for such disparities across racial groups, In the biological and social sciences, there is a clear consensus that race is a social construct and not a biological attribute. There is however evidence to suggest that certain student groups are disadvantaged [65-68], not solely due to socioeconomic factors but also through the existence of both implicit teacher bias and stereotype threat (where individuals feel themselves at risk of conforming to stereotypes about their social group, the pressure of which affects performance). In the US and elsewhere, students of colour are significantly more likely to attend low-income schools with less qualified teachers, fewer resources, larger classes sizes, and lower long-term expectations

The LSAT is a standardised test, as significant disparities were also visible between the privileged group (White) and the protected group (Black). No inherent or biological reasons would account for such disparities across racial groups. In the biological and social sciences, there is a clear consensus that race is a social construct and not a biological attribute. Conversations invoked by use of this framework might encourage practitioners and stakeholders to investigate why such disparities exist. Good GPA's alone do not ensure entry to the majority of law school which places huge importance on the LSAT. Debate has continued over many decades concerning the fairness of standardised tests and educational disparities in the U.S. [54-64]. Much discussion has revolved around the very origins of the LSAT itself, introduced in the early 1920s with the support of influential eugenicists with outspoken racist ideologies(Luis Terman, one of the forefathers of the original IQ test believed that IQ was highly heritable, and Carl Brigham the primary developer of the SAT also supported the "native intelligence" hypothesis). Research suggests that standardised tests are in general biased towards the success of the group who have historically held the most power in any society. Such tests have been developed and written in terms of the understandings of that privileged class based upon the world and the culture that facilitates this privilege. LSAT disparities often result in minority students attending lower-ranked law schools, often leading to higher debt, lower employment rates, and lower income levels. A cheating scandal surrounding standardised tests recently implicated several college administrators and celebrities [62]. The result of one such case was a fourteen-day prison sentence and a 30,000 USD fine (based on admittance of paying 15,000USD toward answers alteration). Another case in 2011 resulted in a Black woman and her father being charged with felonies related to the theft of public education and a 10-year jail sentence (suspended down to 10 days, however, with the implications that a criminal record brings).[64] In this case the mother had enrolled her children in a school using their grandfathers address with the objective of obtaining an education at a school that met all 26 educational standards as opposed to the 4 in her own neighbourhood school. This might also lead one to question the role of socioeconomics in education and standardised test results. A quick google search also indicates that BAR preparation and LSAT preparation courses are generally offered at a cost of 2,000 to 4000 USD (or 150 USD per hour for private tuition). It could be asked if this already place those with socioeconomic difficulties at a more significant disadvantage than those who can afford to take such preparation courses.

Sander index: Is a combination of the LSAT and GPA scores into a single weighted average, as we can see from the results the disparity in distribution increases with this combination.

As can be seen from the use case reflection and research are required when creating applications that rank humans, this particularly true when the model uses measurements of complex human characteristics such as intelligence, resilience or potential which have been converted, by one means or another, to abstract quantities. The creation of a predictive model in the context of law school education can have severe consequences for individuals as well as for groups and entire societies. As a practitioner or engineer it is often easy to get lost in the process of delivering a solution with the goal of optimising a decision process to reflect what has been accurate in the past and to lose site of the potential consequences to individuals and to society. It is very clear that this data should not be used to create a model that might in any way influence decisions related to university admission and that the entire system should be interrogated. In September 2020, in response to a lawsuit filed in December 2019, a Florida judge has ruled[1] that the University of California school system is blocked from using standardised tests as an entry criterion. The law suit[2] was filed by Public Counsel on behalf of a group of students, educators and advocates, which argued that the SAT and ACT tests discriminate against minority and low-income applicants. The judge's ruling cited the disadvantage that the test placed on disabled students during the covid-19 pandemic.

[1] http://www.publiccounsel.org/tools/assets/files/1489.pdf

[2] http://www.publiccounsel.org/tools/assets/files/1250.pdf

Chapter 6

Conclusion

This work differs from much of the prior work in that it considers the worldview as the means through which the final mathematical definitions of fairness should be selected, rather than using a worldview to characterise how biased the data or model is. The framework guides reflection related to observations in the data and a consideration to the context and demography under which the resulting application will be used. The objective is to provide an analytical framework to facilitate the identify of a worldview that the organisation wishes to reflect with respect to a particular protected group and a particular context, by encouraging collaboration between stakeholders which may necessitate social research and interaction with experts in the social domain in question.

6.1. Discussion

There are multiple ways to define fairness, to detect fairness and to mitigate unfairness. However, the key to ethical AI is in identifying the values that the producers of such systems are striving to reflect in any model, and transparency regarding the degree to which the model meets these value goals. Companies are very vocal about declaring commitments to diversity and fairness. It would be hard to imagine any company making a declaration in support of *individual fairness* or *fairness by unawareness* in a 'real-world' context where no biological or inherent factors account for discrepancies. However, when it comes to automated decision-making systems, there is still a possibility to hide behind the mathematics of fairness and the language often co-opted from the legal profession. The proposed framework gives organisations the tools necessary to apply critical thinking to the creation of a model and to facilitate alignment of objectives and ideals of fairness framed through a worldview of each protected feature in each implementation.

6.2. Future work

The *Machnamh* framework allows for the simultaneous review of multiple protected features, however does not provide an intersectional analysis (e.g. black, female). There is also a need for input from other fields of study in order to re-asses and review the natural language being used to describe the various concepts of fairness and of worldviews. It is hoped that this work can be used as a starting point for a framework which will support and invoke conversations around fairness, quantified in a mathematical and statistical form yet framed through the use of natural language that can be more easily interpreted, explained, discussed and agreed upon.

The framework is currently solely focused on unfairness detection and the conversation around the objectives of fairness mitigation; however, it could be enhanced to lead the discussions through the various fairness mitigation strategies aligned with the selected worldview or further sub-categorisations of each world view to correctly focus the objective (e.g. Demographic parity and Equalised odds). Farness in ranking could in particular be expanded further

There are also many additional features and functionalities that could be added to the framework to increase adoption, such as enhancing the data preparation functionality to include binning and date extraction. In addition, there are many more libraries and unfairness detection strategies which could be added to further facilitate the detection of disparities, such as the use of adversarial models for unfairness detection.

1. Bibliography

[1] Holstein, K., Vaughan, J. W., Daumé, H., Dudík, M. & Wallach, H. Improving fairness in machine learning systems: What do industry practitioners need? in *Conference on Human Factors in Computing Systems - Proceedings* (Association for Computing Machinery, 2019). doi:10.1145/3290605.3300830

[2] Dwork, C., Hardt, M., Pitassi, T., Reingold, O. & Zemel, R. Fairness through awareness. in *ITCS 2012 - Innovations in Theoretical Computer Science Conference* 214–226 (2012). doi:10.1145/2090236.2090255

[3] Verma, S., & Rubin, J. (2018). Fairness definitions explained. In *Proceedings - International Conference on Software Engineering* (pp. 1–7). IEEE Computer Society. https://doi.org/10.1145/3194770.3194776

[4] Arneson, R. J. Equality and equal opportunity for welfare. *Philosophical Studies* **56,** 77–93 (1989).

[5] Miller, W. in *The Social History of Crime and Punishment in America: An Encyclopaedia* (SAGE Publications, Inc., 2014). doi:10.4135/9781452218427.n590

[6] Hillier, A. E. Redlining and the Home Owners' Loan Corporation. *Journal of Urban History* **29,** 394–420 (2003).

[7] Messner, C. & Wänke, M. Unconscious information processing reduces information overload and increases product satisfaction. *Journal of Consumer Psychology* **21,** 9–13 (2011).

[8] Kintsch, W. Recognition and free recall of organized lists. *Journal of Experimental Psychology* 78 (3p1):481 (1968)

[9] Bruce, D. & Fagan, R. L. More on the recognition and free recall of organized lists. *Journal of Experimental Psychology* **85,** 153–154 (1970).

[10] Keane, M. T., O'Brien, M. & Smyth, B. Are people biased in their use of search engines? *Communications of the ACM* **51,** 49–52 (2008).

[11] Clark, B. B., Robert, C. & Hampton, S. A. The Technology Effect: How Perceptions of Technology Drive Excessive Optimism. *Journal of Business and Psychology* **31,** 87–102 (2016)

[12] Kahneman, D. Thinking fast, thinking slow. *Interpretation, Tavistock, London* 499 (2011).

[13] Lebrecht, S., Pierce, L. J., Tarr, M. J. & Tanaka, J. W. Perceptual other-race training reduces implicit racial bias. *PLoS ONE* **4,** (2009).

[14] Brosch, T., Bar-David, E. & Phelps, E. A. Implicit Race Bias Decreases the Similarity of Neural Representations of Black and White Faces. *Psychological Science* **24,** 160–166 (2013).

[15] Russell, J. A., Brock, S. & Rudisill, M. E. Recognizing the Impact of Bias in Faculty Recruitment, Retention, and Advancement Processes. *Kinesiology Review* **8,** 291–295 (2019)

[16] Bartlett, Robert & Morse, Adair & Stanton, Richard & Wallace, Nancy. (2017). Consumer Lending Discrimination in the FinTech Era. SSRN Electronic Journal. 10.2139/ssrn.3063448

[17] https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G, retrieved 1st February 2020.

[18] https://www.wired.com/story/the-apple-card-didnt-see-genderand-thats-the-problem/, retrieved 1st February 2020.

[19] Cathy O’Neil.Weapons of math destruction: How big data in-creases inequality and threatens democracy. Broadway Books, 2016.

[20] Noble, S. U. in *Algorithms of Opression: How Search Engines Reinforce Racism* 15–63 (2018)

[21] Gov.uk: Ethnicity facts and figures, Stop and Search (2020). https://www.ethnicity-facts-figures.service.gov.uk/crime-justice-and-the-law/policing/stop-and-search/latest

[22] FullFact.org: Stop and search in England and Wales (2019) https://fullfact.org/crime/stop-and-search-england-and-wales/

[23] The Guardian: Met police 'disproportionately' use stop and search powers on black people. https://www.theguardian.com/law/2019/jan/26/met-police-disproportionately-use-stop-and-search-powers-on-black-people

[24] Runnymede Trust report: Race and racism in English secondary schools (2020) https://www.runnymedetrust.org/projects-and-publications/education/racism-in-secondary-schools.html

[25] Devlin, M. “The Power and Potential of Diversity and Inclusion: A Compendium Based on the TIAA Institute’s 2016 Higher Education leadership Conference.” TIAA Institute. March 2017. Page 17

[26] LinkedIn: study (2017): https://news.linkedin.com/2017/6/eighty-percent-of-professionals-consider-networking-important-to-career-success

[27] ProPublica (2016): https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

[28] Schwirtz, Michael, and Michael Winerip. “Kalief Browder, Held at Rikers Island for 3 Years Without Trial, Commits Suicide.” The New York Times, The New York Times, 8 June 2015, www.nytimes.com/2015/06/09/nyregion/kalief-browder-held-atrikers-island-for-3-years-without-trial-commits-suicide.html.

[29] NorthPoint(2016) https://www.equivant.com/response-to-propublica-demonstrating-accuracy-equity-and-predictive-parity/

[30] Report-UN-Sentencing:https://www.sentencingproject.org/publications/un-report-on-racial-disparities/

[31] Gordon, F. (2019). Virginia Eubanks (2018) Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor. New York: Picador, St Martin’s Press. *Law, Technology and Humans*, 162–164. https://doi.org/10.5204/lthj.v1i0.1386

[32] Binns, R. (2020). On the apparent conflict between individual and group fairness. In *FAT* 2020 - Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (pp. 514–524). Association for Computing Machinery, Inc. https://doi.org/10.1145/3351095.3372864

[33] Sorelle A Friedler, Carlos Scheidegger, and Suresh Venkatasubramanian. On the (im) possibility of fairness. *arXiv preprint arXiv:1609.07236*, September 2016.

[34] Kusner, M., Loftus, J., Russell, C. & Silva, R. Counterfactual fairness. in *Advances in Neural Information Processing Systems* 2017-December**,** 4067–4077 (Neural information processing systems foundation, 2017).

[35] Lipton, Z. C., & Steinhardt, J. (2019). Troubling trends in machine-learning scholarship. *Queue*, *17*(1). https://doi.org/10.1145/3317287.3328534

[36] A framework has been suggested by Suresh and Guttag in 'A Framework for Understanding Unintended Consequences of Machine Learning' A

[37] Jackson, J. W. Explaining intersectionality through description, counterfactual thinking, and mediation analysis. *Social Psychiatry and Psychiatric Epidemiology* **52,** 785–793 (2017)

[38] Larson, J., Mattu, S., Kirchner, L., & Angwin, J. (2016). How We Analyzed the COMPAS Recidivism Algorithm. *ProPublica*, 2016. Retrieved from https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm

[39] Sapiezynski, P., Zeng, W., Robertson, R. E., Mislove, A. & Wilson, C. Quantifying the impact of user attention on fair group representation in ranked lists. in *The Web Conference 2019 - Companion of the World Wide Web Conference, WWW 2019* 553–562 (Association for Computing Machinery, Inc, 2019). doi:10.1145/3308560.3317595

[40] Biega, A. J., Gummadi, K. P. & Weikum, G. Equity of attention: Amortizing individual fairness in rankings. in *41st International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2018* 405–414 (Association for Computing Machinery, Inc, 2018). doi:10.1145/3209978.3210063

[41] Zehlike, M. et al. FA∗IR: A fair top-k ranking algorithm. in International Conference on Information and Knowledge Management, Proceedings Part F131841, 1569–1578 (Association for Computing Machinery, 2017)

[42] Singh, A. & Joachims, T. Fairness of exposure in rankings. in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 2219–2228 (Association for Computing Machinery, 2018). doi:10.1145/3219819.3220088

[43] Ninareh M, Morstatter F, Saxena N, Lerman K and Galstyan A. A Survey on Bias and Fairness in Machine Learning(2019) arXiv:1908.09635

[44] Mosca, I., & Wright, R. E. (2020). The long-term consequences of the irish marriage bar. *Economic and Social Review*, *51*(1), 1–34.

[45] Bellamy, R. K. E., Mojsilovic, A., Nagar, S., Ramamurthy, K. N., Richards, J., Saha, D., … Mehta, S. (2019). AI Fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias. *IBM Journal of Research and Development*, *63*(4–5). https://doi.org/10.1147/JRD.2019.2942287

[46] Shepherd, George B., No African-American Lawyers Allowed: The Inefficient Racism of the Aba's Accreditation of Law Schools (March 5, 2001). Emory University School of Law Working Paper, Available at SSRN: https://ssrn.com/abstract=263211 or http://dx.doi.org/10.2139/ssrn.263211

[47] Subotnik, D. (2013). Does Testing = Race Discrimination? Ricci, the Bar Exam, the LSAT, and the Challenge to Learning. *U. Mass. L. Rev, 8,* 332.

[48] Eubanks, Virginia. *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor.* New York: Picador , St Martin's Press, 2018

[49] UCLA (2020) Hollywood Diversity Report, https://socialsciences.ucla.edu/wp-content/uploads/2020/02/UCLA-Hollywood-Diversity-Report-2020-Film-2-6-2020.pdf

[50] PWC(2020).Women In Work Index 2020: https://www.pwc.co.uk/services/economics/insights/women-in-work-index.html

[51] Bruce, S., Schuman, H., Steeh, C., & Bobo, L. (Revised edition 1998)

[52]). Racial Attitudes in America: Trends and Interpretations. *The British Journal of Sociology, 38*(4), 584. https://doi.org/10.2307/590918

[53] Yeom. S, Tschantz.M(2018) Discriminative but Not Discriminatory: A Comparison of Fairness Definitions under Different Worldviews *arXiv*..http://arxiv.org/abs/1808.08619v5

[54] Reeves, R. V., & Halikias, D. (2017). Race gaps in SAT scores highlight inequality and hinder upward mobility. The Brookings Institute. Retrieved from https://www.brookings.edu/research/race-gaps-in-sat-scores-highlight-inequality-and-hinder-upward-mobility/

[55] A. Allensworth, E. M., & Clark, K. (2020). High School GPAs and ACT Scores as Predictors of College Completion: Examining Assumptions About Consistency Across High Schools. Educational Researcher, 49(3), 198–211. https://doi.org/10.3102/0013189X20902110

[56] Race gaps in SAT scores highlight inequality and hinder upward mobility: https://www.brookings.edu/research/race-gaps-in-sat-scores-highlight-inequality-and-hinder-upward-mobility/

[57] (Il)logical Reasoning: The LSAT's Troubling History of Exclusion https://brownpoliticalreview.org/2019/12/illogical-reasoning-the-lsats-troubling-history-of-exclusion/

[58] Leslie, B. (2000). Nicholas Lemann: "The Big Test: The Secret History of the American Meritocracy." American Studies in Scandinavia, 32(2), 97–101. https://doi.org/10.22439/asca.v32i2.2772

[59] Taylor, B. P. (2014). Testing Wars in the Public Schools: A Forgotten History. Journal of American History, 100(4), 1197–1197. https://doi.org/10.1093/jahist/jau035

[60] Subotnik, D. (2013). Does Testing = Race Discrimination?: Ricci, the Bar Exam, the LSAT, and the Challenge to Learning. U. Mass. L. Rev, 8, 332.

[61] Warne, R. T. (2019). An Evaluation (and Vindication?) of Lewis Terman: What the Father of Gifted Education Can Teach the 21st Century. Gifted Child Quarterly, 63(1), 3–21. https://doi.org/10.1177/0016986218799433

[62] How the largest college admissions scandal ever let wealthy parents cheat the system. https://www.latimes.com/local/lanow/la-me-college-admissions-fraud-scheme-20190313-story.html

[63] Factors Affecting Bar Passage Among Law Students: The REAL Connection between Race and Bar Passage. The African American Attorney Network.

https://aaattorneynetwork.com/factors-affecting-bar-passage-among-law-students-the-real-connection-between-race-and-bar-passage/

[64] Her Only Crime Was Helping Her Kids. Kelley Williams-Bolar, like Felicity Huffman, was punished for trying to get her children a better education. https://www.theatlantic.com/ideas/archive/2019/09/her-only-crime-was-helping-her-kid/597979/

[65] Cohen, G. L., Garcia, J., Apfel, N., & Master, A. (2006). Reducing the racial achievement gap: A social-psychological intervention. Science, 313(5791), 1307–1310. https://doi.org/10.1126/science.1128317

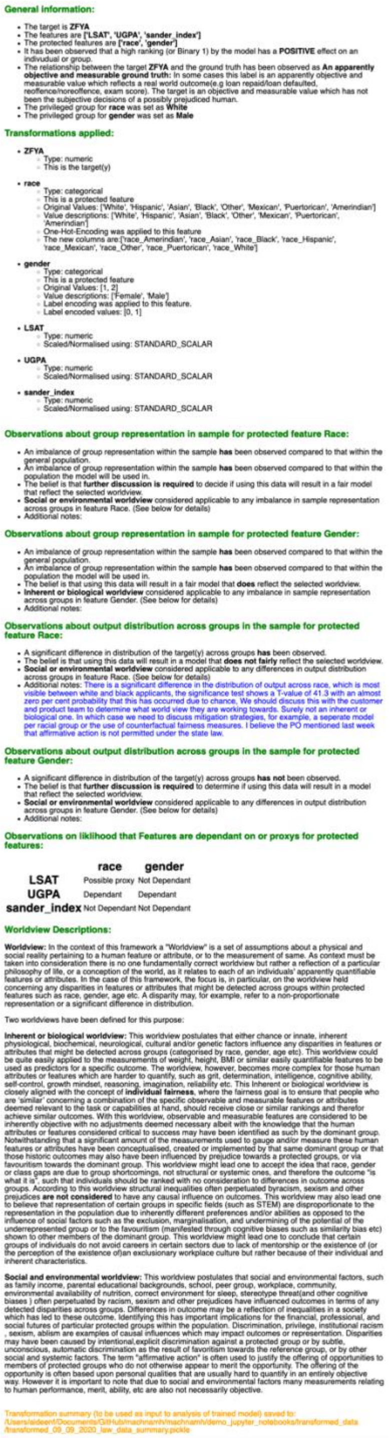
[66] P, M. H., Christina, M., Rosann, T., Ray, W., Dan, F., Sara, M., … University, A. (2014). Opportunity and Equity: Enrolment and Outcomes of Black and Latino Males in Boston Public Schools. Annenberg Institute for School Reform at Brown University. Retrieved from http://search.ebscohost.com/login.aspx?direct=true&db=eric&AN=ED553658&site=ehost-live

[67] Bonefeld, M., & Dickhäuser, O. (2018). (Biased) Grading of Students' performance: Students' names, performance level, and implicit attitudes. Frontiers in Psychology, 9(MAY). https://doi.org/10.3389/fpsyg.2018.00481

[68] U.S. Department of Education Office for Civil Rights CIVIL RIGHTS DATA COLLECTION Data Snapshot: Early Childhood Education, Issue Brief No. 2 (March 2014) https://www2.ed.gov/about/offices/list/ocr/docs/crdc-early-learning-snapshot.pdf

2. Appendices

Appendix A: Sample report generated from the framework



General information:

- The target is **ZFYA**
- The features are **['LSAT', 'UGPA', 'sander_index']**
- The protected features are **['race', 'gender']**
- It has been observed that a high ranking (or Binary 1) by the model has a **POSITIVE** effect on an indivudual or group.
- The relationship between the target **ZFYA** and the ground truth has been observed as **An apparently objective and measurable ground truth:** In some cases this label is an apparently objective and measurable value which reflects a real world outcome(e.g loan repaid/loan defaulted, reoffence/noreoffence, exam score). The target is an objective and measurable value which has not been the subjective decisions of a possibly prejudiced human.
- The privileged group for **race** was set as **White**
- The privileged group for **gender** was set as **Male**

Transformations applied:

- **ZFYA**
  - Type: numeric
  - This is the target(y)
- **race**
  - Type: categorical
  - This is a protected feature
  - Original Values: ['White', 'Hispanic', 'Asian', 'Black', 'Other', 'Mexican', 'Puertorican', 'Amerindian']
  - Value descriptions: ['White', 'Hispanic', 'Asian', 'Black', 'Other', 'Mexican', 'Puertorican', 'Amerindian']
  - One-Hot-Encoding was applied to this feature
  - The new columns are:['race_Amerindian', 'race_Asian', 'race_Black', 'race_Hispanic', 'race_Mexican', 'race_Other', 'race_Puertorican', 'race_White']
- **gender**
  - Type: categorical
  - This is a protected feature
  - Original Values: [1, 2]
  - Value descriptions: ['Female', 'Male']
  - Label encoding was applied to this feature.
  - Label encoded values: [0, 1]
- **LSAT**
  - Type: numeric
  - Scaled/Normalised using: STANDARD_SCALAR
- **UGPA**
  - Type: numeric
  - Scaled/Normalised using: STANDARD_SCALAR
- **sander_index**
  - Type: numeric
  - Scaled/Normalised using: STANDARD_SCALAR

Observations about group representation in sample for protected feature Race:

- An imbalance of group representation within the sample **has** been observed compared to that within the general population.
- An imbalance of group representation within the sample **has** been observed compared to that within the population the model will be used in.
- The belief is that **further discussion is required** to decide if using this data will result in a fair model that reflect the selected worldview.
- **Social or environmental worldview** considered applicable to any imbalance in sample representation across groups in feature Race. (See below for details)
- Additional notes:

Observations about group representation in sample for protected feature Gender:

- An imbalance of group representation within the sample **has** been observed compared to that within the general population.
- An imbalance of group representation within the sample **has** been observed compared to that within the population the model will be used in.
- The belief is that using this data will result in a fair model that **does** reflect the selected worldview.
- **Inherent or biological worldview** considered applicable to any imbalance in sample representation across groups in feature Gender. (See below for details)
- Additional notes:

Observations about output distribution across groups in the sample for protected feature Race:

- A significant difference in distribution of the target(y) across groups **has** been observed.
- The belief is that using this data will result in a model that **does not fairly** reflect the selected worldview.
- **Social or environmental worldview** considered applicable to any differences in output distribution across groups in feature Race. (See below for details)
- Additional notes: There is a significant difference in the distribution of output across race, which is most visible between white and black applicants, the significance test shows a T-value of 41.3 with an almost zero per cent probability that this has occurred due to chance, We should discuss this with the customer and product team to determine what world view they are working towards. Surely not an inherent or biological one. In which case we need to discuss mitigation strategies, for example, a seperate model per racial group or the use of counterfactual fairness measures. I believe the PO mentioned last week that affirmative action is not permitted under the state law.

Observations about output distribution across groups in the sample for protected feature Gender:

- A significant difference in distribution of the target(y) across groups **has not** been observed.
- The belief is that **further discussion is required** to determine if using this data will result in a model that reflect the selected worldview.
- **Social or environmental worldview** considered applicable to any differences in output distribution across groups in feature Gender. (See below for details)
- Additional notes:

Observations on liklihood that Features are dependant on or proxys for protected features:

	race	gender
LSAT	Possible proxy	Not Dependant
UGPA	Dependant	Dependant
sander_index	Not Dependant	Not Dependant

Worldview Descriptions:

Worldview: In the context of this framework a "Worldview" is a set of assumptions about a physical and social reality pertaining to a human feature or attribute, or to the measurement of same. As context must be taken into consideration there is no one fundamentally correct worldview but rather a reflection of a particular philosophy of life, or a conception of the world, as it relates to each of an individuals' apparently quantifiable features or attributes. In the case of this framework, the focus is, in particular, on the worldview held concerning any disparities in features or attributes that might be detected across groups within protected features such as race, gender, age etc. A disparity may, for example, refer to a non-proportionate representation or a significant difference in distribution.

Two worldviews have been defined for this purpose:

Inherent or biological worldview: This worldview postulates that either chance or innate, inherent physiological, biochemical, neurological, cultural and/or genetic factors influence any disparities in features or attributes that might be detected across groups (categorised by race, gender, age etc). This worldview could be quite easily applied to the measurements of weight, height, BMI or similar easily quantifiable features to be used as predictors for a specific outcome. The worldview, however, becomes more complex for those human attributes or features which are harder to quantify, such as grit, determination, intelligence, cognitive ability, self-control, growth mindset, reasoning, imagination, reliability etc. This Inherent or biological worldview is closely aligned with the concept of **individual fairness**, where the fairness goal is to ensure that people who are 'similar' concerning a combination of the specific observable and measurable features or attributes deemed relevant to the task or capabilities at hand, should receive close or similar rankings and therefor achieve similar outcomes. With this worldview, observable and measurable features are considered to be inherently objective with no adjustments deemed necessary albeit with the knowledge that the human attributes or features considered critical to success may have been identified as such by the dominant group. Notwithstanding that a significant amount of the measurements used to gauge and/or measure these human features or attributes have been conceptualised, created or implemented by that same dominant group or that those historic outcomes may also have been influenced by prejudice towards a protected groups, or via favouritism towards the dominant group. This worldview might lead one to accept the idea that race, gender or class gaps are due to group shortcomings, not structural or systemic ones, and therefore the outcome "is what it is", such that individuals should be ranked with no consideration to differences in outcome across groups. According to this worldview structural inequalities often perpetuated byracism, sexism and other prejudices **are not considered** to have any causal influence on outcomes. This worldview may also lead one to believe that representation of certain groups in specific fields (such as STEM) are disproportionate to the representation in the population due to inherently different preferences and/or abilities as opposed to the influence of social factors such as the exclusion, marginalisation, and undermining of the potential of the underrepresented group or to the favouritism (manifested through cognitive biases such as similarity bias etc) shown to other members of the dominant group. This worldview might lead one to conclude that certain groups of individuals do not avoid careers in certain sectors due to lack of mentorship or the existence of (or the perception of the existence of)an exclusionary workplace culture but rather because of their individual and inherent characteristics.

Social and environmental worldview: This worldview postulates that social and environmental factors, such as family income, parental educational backgrounds, school, peer group, workplace, community, environmental availability of nutrition, correct environment for sleep, stereotype threat(and other cognitive biases) often perpetuated by racism, sexism and other prejudices have influenced outcomes in terms of any detected disparities across groups. Differences in outcome may be a reflection of inequalities in a society which has led to these outcome. Identifying this has important implications for the financial, professional, and social futures of particular protected groups within the population. Discrimination, privilege, institutional racism , sexism, ablism are examples of causal influences which may impact outcomes or representation. Disparities may have been caused by intentional,explicit discrimination against a protected group or by subtle, unconscious, automatic discrimination as the result of favoritism towards the reference group, or by other social and systemic factors. The term "affirmative action" is often used to justify the offering of opportunities to members of protected groups who do not otherwise appear to merit the opportunity. The offering of the opportunity is often based upon personal qualities that are usually hard to quantify in an entirely objective way. However it is important to note that due to social and environmental factors many measurements relating to human performance, merit, ability, etc are also not necessarily objective.

Transformation summary (to be used as input to analysis of trained model) saved to: /Users/aideent/Documents/GitHub/machnamh/machnamh/demo_jupyter_notebooks/transformed_data /transformed_09_09_2020_law_data_summary.pickle