Methods to Achieve Algorithmic Fairness in Artificial Intelligence

Megha Nanda (ID: 19240177)
Msc. in Artificial Intelligence
National University of Ireland Galway

Supervisors

Heike Felzmann
Enda Barrett

In partial fulfillment of the requirements for the degree of

Artificial Intelligence - Online

August 2021

DECLARATION I, Megha Nanda, do hereby declare that this thesis entitled 'Methods to Achieve Algorithmic Fairness in Artificial Intelligence' is a bonafide record of research work done by me for the award of MSc. in Artificial Intelligence-Online (PROGRAMME) from National University of Ireland Galway. It has not been previously submitted, in part or whole, to any university or institution for any degree, diploma, or other qualification.

Acknowledgements

The subject of this thesis is very close to my heart as I strongly feel about understanding bias and mitigating the same in various aspects of life. This work combines the aspects of Law, Ethics and Computer science which motivated me to pick up this topic. Being a technology professional, this is one way that I believe we can encounter the bias in our day-to-day work and ensure we contribute our bit in making a better and fair world. It seems like a small step in technology domain as this is a very recent and challenging subject which has received lot of attention in last 5 years, but as it is said “Small steps lead to big changes”.

I would thank my family for tremendous support during my master’s programme and helping me to pursue my studies along with full time career.I would also like to thank my supervisor Heike Felzmann for guiding me and teaching the subject on ‘Ethics in Artificial Intelligence’ in the course that helped me in various aspects of this thesis as well. Many thanks to my second supervisor Enda Barrett for his guidance, discussion and challenging this thesis by discussing key questions such as profitability vs morality which motivated me to cover not only the technological aspects but discuss about Business Ethics.

Abstract

Growing data and use of machine learning models have opened an interesting area of research on the sociological, ethical and legal aspect of the discrimination in computer science literature. As per US law, the bias is coded into system through disparate impact, which has different outcomes for different set of groups.In consumer lending, AI is used to approve loan applications based on the credit worthiness of the applicant.There are concerns raised that data driven technologies can lead to gender bias in its outcomes.It is hard to determine the impact of bias or rather estimate the algorithmic fairness as it requires a deeper understanding of how algorithms make decisions.There are more than 20 definitions of fairness proposed in the literature that attempts to assess the model fairness. In this research, we uncover some of these methods of testing and correcting algorithmic bias through pre-processing, in processing and post processing techniques.We also showcase the results of fairness methods on Home Credit dataset.

Keywords: Fairness in Artificial Intelligence, Fairness in Machine Learning, Responsible Artificial intelligence, Bias, Discrimination,Explainable Artificial intelligence.

Contents

List of Figures

List of Tables

Chapter 1

Background

Machine learning has diverse range of applications in today's era. It has increased the efficiency of the processes and is used now a days to filter loan applications (Mahoney and Mohen,2007), college admissions (Ragab et al.,2014), make bail and parole decisions (Brennan et al.,2009) and, skin cancer identification(Lau and Al-Jumaily,2009).Such a major influence in a short span of time motivates the need to investigate both positive and negative effects of technology as AI as a cutting-edge technology assigns certain autonomy to the machines.

Key advantages of Artificial intelligence include time saving across activities, enabling multi-tasking, execution of complex tasks, 24*7 operational time, highly scalable across industry and divisions and it enables faster and smarter decision making by augmenting the capabilities of humans. In critical domains like healthcare, AI is used in various practices such as drug development, patient monitoring, diagnosing health care issues, treatment development etc. [Bartoletti, 2019]. In financial services it allows to create more personalised products and services, enhancing the cybersecurity and reducing the risk. AI applications in many other areas such as virtual assistance like Siri, Alexa helps in automating day to day tasks,it helps in Agriculture to monitor the enhancement the crop yield using drones, in automotive to design autonomous vehicles, in E-commerce to manage automated warehousing and in supply chain, it creates a significant value and impact by increasing the productivity and reducing the risks.

However, there are widespread concerns with machines making life changing decisions as some of the outcomes may not align with our ethical norms and values. This impact known as algorithmic bias can be described as unfair prediction of outcomes in favour of protected class based on gender, religion, race or colour (Friedman and Nissenbaum, 1996). It creates unintentional discrimination in the predictive outcomes (Bigman and Gray, 2018). It has been shown that data driven methods contains implicit human bias that leads to discrimination and introduce new biases (Kim,2016). Such biases have been proven in natural language processing applications where algorithms relate men to technical occupations and women to domestic roles(Bolukbasi et al.,2016) and in online advertising where google showed advertisements that person arrested is more often black than white (Sweeney,2013).As a result, it is vital to understand what biases are present, identify its sources, quantitatively measure it and understand what steps can be taken to mitigate the bias. Discriminatory problems created by machine learning and how to mitigate it, is defined as the algorithmic fairness.

In the area of customer lending, AI is used extensively in loan approval processes to make it more efficient, and all the information of the loan applicant is used predict their repayment abilities. There are concerns around the use of AI in credit lending if algorithms trained on the datasets that reflect disparity between men and women could leads to lending bias against one of the groups. Having access to loans can increase the ability of an individual to improve his standard of living as it helps them to buy new house, support education and plan for any contingencies. When it comes to Gender, discrimination is quite prominent. Women face disparities in financial inclusion. Even if datasets are free of bias, sensitive variables are not used in the model building process, it is key to assess the fairness of the model results. We would test this in later sections using home credit dataset. Ensuring fairness is key for the economic growth as gender equality and economic growth are positively correlated.

Algorithmic fairness is a relatively new area of research due to rapid technological advancement in last 2-3 decades. It has been explored in white house report (Mitchell et al.,2021), popular books (Eubanks,2018), and inspired multiple software packages. Also, annual conferences are conducted to bring together experts from various domains to come together, characterise and address fairness, transparency, and accountability principles in AI system. One of such conferences is FATML and these events are required as it is a challenging area that requires discussion and agreements on addressing the complex issues. For example: one of the complex issue is multiple features in machine learning that could potentially reflect relation to the protected features such as gender or race. One such field is address which is highly correlated to race and even if race is not used as a parameter in the models, the models could reflect racial bias.

To study the bias in models, in 2018 IBM launched aif360 toolkit(Bellamy et al.2018) which is a python library to mitigate bias and improve model fairness. This toolkit is used in the later sections for home credit dataset to test and mitigate bias. Facebook and google have also published the set of tools post that(Bakalar et al.,2021). As a result of rapid growth in this field, there are inconsistent terminology and notations. In this study, we focus on some of the key metrics from IBM's fairness toolkit for bias identification and methods to achieve fairness.

1.1 Sources of Bias

Before we dive into identifying bias and mitigation, it is fundamental to understand various ways in which bias is encoded into machine learning models. These include the following sources:

1. Defining the problem statement:Before starting to build the models, it is key to define what the model is going to predict. For example: if we are predicting the creditworthiness of a customer, then the term 'creditworthiness needs to be defined in quantitative terms. The objective of why we need to assess the creditworthiness also needs to be clearly specified like we want to assess creditworthiness to minimise default rate or to maximise the margins. Understanding what are we trying to achieve helps us in assessment and model implementation.An ill defined problem statement could focus only on business outcomes lacking social responsibility.

2. Data collection: It is very often that data collected already contains human bias. We would agree that humans or society is not fair or unbiased in its own and the same data is recorded or saved in the data warehouses. For example: decisions regarding bail or parole considers the probability of criminal committing another crime in next few periods once he is released. But the data captured is not on the crimes committed but the on the ones who got arrested and the arrest rate could be higher for minority populations due to the higher rate of policing. Since models or predictions are based on the data that contains bias, they are bound to incorporate the bias in outcomes. (Rothwell,2014)

3. Sample Bias: Sample bias could occur when data is too big, and sample

is taken which doesn't accurately represent the population on which the final decisions are being made(Huang et al.,2006). Underprivileged population can be put at a disadvantage within the sensitive attribute due to sampling bias.For example: if image classification models attempt to identify people across multiple cultures and countries and model is trained for only sample population which excludes some countries then it would not be able to predict accurately
4. Minimising average error fits the majority population: For multiple population distribution in the dataset, the methodology of minimising the average error will lead the model to fit majority population.For example: In the prediction of suitable job candidate, the majority population is men and minority population is women. It is natural that recruitment rate is higher for men than women. If we train the model to minimise error, it will fit majority of male population. In most of the problems, like predicting recidivism the predictions that are made by the algorithms are based on the data or actions that had taken in the past. Learning theory emphasis that we need to deep dive, explore more and gather more data(Chen et al.,2018).
5.Feature Selection:The bias could be introduced during the data preparation and feature selection stage. There are few aspects of customer data that is prohibited to be used. There could be certain features that could be highly correlated to sensitive features so correlation needs to be validated before the feature selection. These features could significantly determine model's accuracy, but it is equally important to assess if features selected will not create any bias.

To assess the fairness of the algorithm or to find the bias, 20+ definitions of fairness have been proposed. However, what is lacking is the consensus on the most efficient methods for identifying gender inequality in lending. Also, these methods only show the bias without identifying the key sources of bias.

1.2 Why algorithmic bias assessment is important

Discrimination has been researched from last 50 years across various domains such as economic, statistical, and legal. However, in computer science literature, the study has emerged in last few years. Algorithms are often considered as neutral and unbiased as they are mathematical computations. However, as we saw above there are various sources from which bias can infiltrate into algorithms. They can project higher authority than humans in some cases. Also, it becomes easier to replace human ignorance, mistakes, or their responsibility by putting all of it on algorithms. In 2019, apple card algorithm was accused of gender bias where it gave a higher credit limit to male than female despite having a higher credit score. It highlights the importance of investigating the moral loss especially in high stake environment where the repercussions are high.

In such cases, responsibility and the retribution gap increase in case of any harm. Retribution gap is gap between human desire for punishment and absence of subject for blame. Danaher [Danaher,2016] stated 3 implications of this retribution gap 1) There exists higher risk of moral scapegoating 2) The confidence in law and justice declines 3) Non retributive approaches to crime are supported.

1.3 Legal Framework

In Europe, the European Convention on Human rights (ECHR) protect the human rights from discrimination caused by AI and the political freedom. The article 14 of ECHR states

> *"the enjoyment of the rights and freedoms set forth in this Convention shall be*

secured without discrimination on any ground such as sex, race, color, language, religion, political or other opinion, national or social origin, association with a national minority, property, birth or other status."

The discrimination is classified as direct and Indirect discrimination. The direct discrimination is called disparate treatment where there is direct intent for discrimination. The Indirect discrimination is called 'Disparate Impact' where it is unintentional, but it leads to disproportionate impact on one of the protected classes. The civil rights Act of 1964 "prohibits discrimination on the basis of race, color, religion, sex or national origin". The Fair Housing Act of 1968 prohibits discrimination

1.4 Why it is hard to fix Bias?

There are various reasons due to which it is hard to identify and fix the bias. Key ones include:
Sources: It is hard to identify the source of bias so that it can be removed. In case of Amazon where the system was penalising the female candidates, the system was programmed to ignore words related to 'women'. However, later it was realised that system was still picking the implicit words related to gender like 'executed' which had high correlation with men than women(Kodiyan,2019)
Algorithm Design: the algorithms are designed to maximise accuracy and not to ensure fairness. Moreover, the data preparation and validation where we try to test the model on validation dataset, has the same bias as the training dataset.
Lack of social context: The mental framework with which data scientists frame the problem may not be perfect way to solve social problems. For example: data scientists might make a model that can be implemented in multiple scenarios

like default model. However, the system designed in one country could be very different from another country as the definitions of fairness changes

Definition of fairness: It is difficult to fix bias without defining what is fair or how the absence of bias looks like. Fairness has lot of definitions, however for algorithmic fairness, it needs to be defined in mathematical terms. Does it mean same proportion of males or females should get high default scores? Or same level of risk is assigned regardless of gender? Once we define what the fairness means, we can then test it and try to achieve the same.

Fairness vs Profitability Trade off: When it comes to financial lending, making the models fair could mean declining profitability by giving credit only to selected individuals. The discussion of business ethics is quite subjective and balancing the profitability and fairness could be challenging (Bowie,1998). Following the course of legality is the choice for most people, however there are loopholes and abiding legal norms may not always lead to the path of 'rightness'. For example: companies set up operations in countries that are tax havens to increase profitability, but it comes at the cost of other human beings and environment.

Profitability objective is the backbone of every organisation; however, a greater onus is also placed on the shoulders of these companies to create fair and just society. There are some of the actions suggested by business experts to guide through the path of ethical behaviour such as creating a mission statement of values that are known widely to all the individuals. Owners need to ensure successful adherence at regular interval, and they should themselves set an example by following ethical practices. A company driven by values also drives profitability by earning customer's respect and loyalty (Ferrell,2004)

Fairness vs accuracy Trade off: The standard wisdom is that there is a trade off between accuracy and achieving fairness. The focus should be comparing the societal costs and notion of 'fairness'. In the COMPAS algorithm research, it is stated that there is "an inherent tension between minimizing violent crime and satisfying common notions of fairness." [Corbett-Davies et al.,2017]. This implies that we need to compromise on public safety by satisfying fairness definitions and releasing the persons of colour who could potentially be high risk criminals. And it could also result in disproportionate impact of higher crimes in whites or African American community creating another type of fairness cost.

The researchers must try to reduce the disparity without affecting the accuracy first mainly where the trade off is higher. Sometimes the investigation on the dataset, different algorithm, investigation of bugs could also help to maximise the overall accuracy. One other option is to analyse the distribution across key sensitive features and ensure data is balanced. Buolamwini's facial detection experiments(Buolamwini and Gebru,2018) is one of the good examples where the under-representation of women and people with dark skin was transformed so that they are better represented in the entire dataset to ensure neural networks learn from the fair datasets. For a perfectly balanced data and two popular fairness metrics equality of opportunity and demographic parity, there is no trade off between accuracy and fairness. In nutshell, companies need to evaluate the risk tolerance where the trade off are critical to equity issues or if any external support from key stakeholders or human decision makers are required to make the decision.

1.5 Definition of Fairness

One of the major questions before trying to combat bias is how to define algorithmic fairness? If algorithms make similar decisions for similar individuals, then is it considered fair? Or if algorithms make similar decisions across different groups (e.g. males vs females) then it is fair? Or is there any another definition to be used? Fairness is broadly classified into individual and group fairness (Gajane and Pechenizkiy,2017).

Individual Fairness: Broadly, individual fairness focus on defining constraints for specific individuals such that "similar individuals are treated similarly" (Dwork et al.,2012) where the similarity metric is defined based on specific task which is determined on case by case basis. There are various types of individual fairness such as disparate impact, disparate treatment, unawareness and counterfactual fairness (Barocas and Selbst,2016; Russell et al.,2017). Even though these approaches are more meaningful, they require strong underlying assumptions.

Group Fairness: The group fairness focuses on defining a specific protected demographic group and then work towards statistical parity across all these groups. The most popular measures include positive classification rate also called statistical parity (Calders and Verwer,2010; Kamishina et al, 2011), false positive and false negative rates also called equalised odds (Chouldechova,2017; Kleinberg et al, 2016) and positive predicted value (Chouldechova, 2017; Kleinberg et al., 2016) also called equalised calibration. They key advantage of this approach is that it's easy to implement without making lot of assumptions. However, since it focuses on one group fairness which is defined in the beginning, it does not promise fairness on one or more sub-groups defined over the protected attribute such as race or gender. This is called "fairness gerrymandering" (Kearns et al.,2018). It also discusses that fairness over subgroups is computationally intensive and hence difficult.

In some research papers, it is assumed that Group and Individual fairness metrics are conflicting. If let's say a model has a gender bias and for interview process, it gives highly likely predictions to men than women. Post applying fairness methods, assume a male applicant is not invited to an interview. A male applicant complains that he was not invited despite the similar qualifications. Now the dilemma is that should employer adjust the model to ensure equal or more qualified men are also invited? Or should it continue with interviewing females to justify fairness.

Group fairness is a distributive justice measure of fairness. It ensures statistical parity in outcomes across gender or race or other protected features. They are based on independence, separation and sufficiency. However, it is not possible to achieve all at the same time which means some bias is unavoidable. Independence assumes the same distribution of loans across males and females. Separation assumes that sensitive characteristic is statistically independent of predictions. Sufficiency assumes that ground truth to benefit that applicant gets while predictions show injustice.

Individual fairness is also a distributive measure, however, unlike group fairness it does not depend on any criterion. It ensures that people who are similar, they received similar outcomes. It requires to formulate clearly what is similarity and how it can be measured. This notion assigns the responsibility of fairness on the distance metrics.

Chapter 2

Research Question

The research question that this review aims to analyse is ”How can an analyst audit the fairness of the black box predictive models?” Imagine it is the year 2050. NUIG is using predictive ensemble models in the admission process. In the end, couple of applicants who meet the required criterion are rejected. These applicants claim that they have been discriminated on the basis of race. How NUIG would be able to detect bias in its models?

In 2016, analysis from Bloomberg revealed racial bias in the same day home delivery service of Amazon. This paid service excluded some of the areas with predominant black population in the six major cities. In New York, it excluded Bronx which has majority of black and Hispanic Population. Similarly in Atlanta, it covers northern part which is mainly white population and left southern part which has predominantly black population. Amazon confirmed that its decisions of same day delivery are not decided based on the ethnic compositions, but rather based on the population of prime members and proximity to Amazon warehouses. However, no matter what methodology was used, the outcomes demonstrated racial bias. How could amazon analyse what went wrong? How can it test for this racial bias even though they did not use ethnicity in their decision-making process and how can they make it fair?

The objective of this research is to build a framework where machine learning models are built, bias is tested and how to achieve fairness is demonstrated using Home credit dataset. It answers below key questions:

Research Question 1: How we can assess gender bias in the original data and the model outcomes

Research Question 2: How we can debias the data and model outcomes to ensure fair outcomes in lending

Chapter 3

Related Work

3.1 Model Development

This section explains some of the basic terminology and modelling process. In data mining process, one of the standard industry processes is CRISP-DM that stands for Cross Industry Standard Process for data mining (Chapman et al.,1999). It involves 6 stages of developing and deploying the machine learning models.

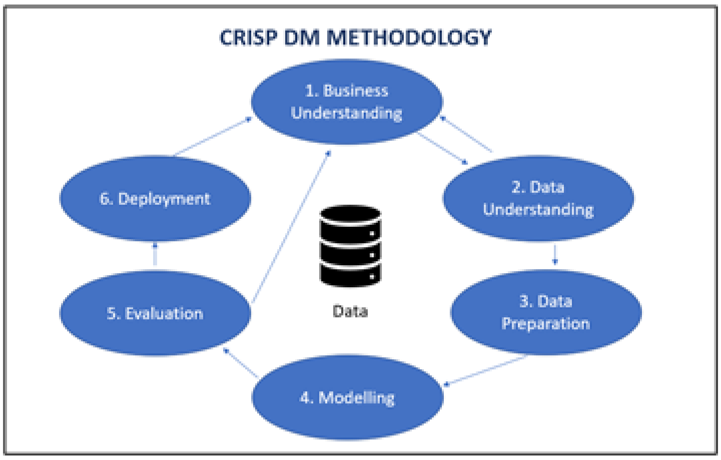


Figure 3.1: CRISP DM Process Diagram. Adapted from Chapma at al, 1999

Various stages of CRISP DM Process is explained below:
1.Business Understanding: This stage involves understanding the business objective. For example: Whether a company wants to reduce cost or increase profits. This helps in the further steps when we investigate data if we have the required metrics and if the objective can be met with the modelling process.
2.Data Understanding: This stage includes data collection, understanding various data dimensions, assess the data quality and creating various hypothesis based on the data availability and quality.
3.Data Preparation: This stage includes various steps such as improving the data quality by missing value treatments, outlier treatments, splitting the data into training and testing dataset, variable selection to make data ready for the modelling process.
4.Modelling: This stage involves model selection which involves choosing the right model technique if let's say we need to build classification model or linear model or use unsupervised learning techniques to get the desired outcomes.
5.Evaluation: This stage includes assessing the robustness of the model and testing if it meets the required business objective outlined in the initial step.
6.Deployment: The insight from the model needs to be structured in a proper way. It needs to ensure that it is used further in the day-to-day decision making which requires implementation and maintenance on regular basis.

When decisions are made about default or no default, approval, or rejection of mortgages or wherever there are binary results, classification models are used. In this scenario, we look at if the loan was paid back or default was made on the loan. We use the raw data and divide it in the 70% training and 30% testing dataset. Training dataset is used for model development whereas the model is tested on the remaining data. The protected or the sensitive feature used in this assessment is 'CODE_GENDER' that takes the values Male and Female. We assess with the data and classification model if there is any statistically significant difference in the model results with respect to the sensitive feature. Bias is identified using key metrics in IBM AIF360 toolkit.

3.2 Why AIF360 Toolkit?

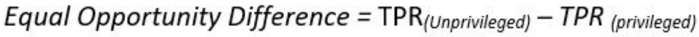
There are many open-source toolkits to assess fairness in AI models but some of them only focus till bias detection and propose no further methods of fairness. Zehlike et al.,2017 describes various metrics such as disparate impact, odds ratio and mean difference. Fair ML(Adebayo et al.,2016) also provides auditing tool for predictive models and assess the impact of features on the predictive ability of the model. FairTest (Tramèr et al.,2015) toolkit also works on identifying the relationship between the output and protected sub group defined by gender, race etc. It reports strong associations as bugs and rank them as well. Aequitas(Stevens et al.,2018) is also another toolkit that has python library and tool on a website where data can be uploaded for bias assessment. It also presents "Fairness Tree" that helps to decide the appropriate fairness metric. Themis(Galhotra et al.,2017) also generates test suites to measure discrimination.
IBMs AIF 360 combines multiple libraries and created one open source toolkit with comprehensive set of bias and fairness metrics. It also shared tutorials with python code available on github to ensure it can be easily used by data scientists with ease. However, it does require datasets to be in a specific format and there are few challenges as examples are shared on different datasets.

3.3 Bias Identification

Identifying the bias is the first step in the journey of fairness. First, we need to define the protected attribute in which we need to identify the bias. When detecting bias, various set of metrics are used to validate the anomalous results. Details of them are as below:

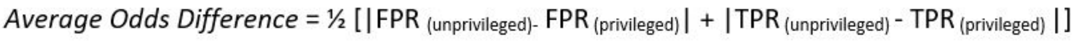
1.Statistical parity Difference This metric measures the difference between privileged and unprivileged classes to get a particular outcome. Smaller the difference, higher is the statistical parity that confirms fairness.

2.Equal Opportunity Difference This metric measures the difference of True positive rates across the two groups: privileged and unprivileged where True positive rate is the probability that actual test will test positive.It needs to be close to zero for fairness.



$$\textit{Equal Opportunity Difference} = \mathrm{TPR}_{(Unprivileged)} - TPR_{(privileged)}$$

3.Average Odds Difference This metric measures the average difference of True positive rate and false positive rate between the privileged and unprivileged group. True positive Rate is the percentage of correctly identified actual positives and False Positive rate is the percentage of incorrectly identifying actual negatives. A value of 0 zero implies the odds difference between groups is negligible and they have equal benefit.



$$\textit{Average Odds Difference} = ½\,[\,|\mathrm{FPR}_{(unprivileged)} - \mathrm{FPR}_{(privileged)}| + |\mathrm{TPR}_{(unprivileged)} - \mathrm{TPR}_{(privileged)}|\,]$$

4.Disparate Impact This fairness metric compares the proportion of customers that receives the positive outcome across unprivileged and privileged groups. For the binary labels the value of disparate impact ranges from 0 to infinity.

-The value less than 1 indicates privileged group has higher proportion of predicted positive outcome than the unprivileged group which is called positive bias.

-A value of one indicates parity.

-The value of more than 1 indicates, unprivileged group has higher proportion of

predicted positive outcomes than the privileged. This is called Negative Bias.
5.Theil Index This metric shows the generalised entropy index with alpha = 1. The generalised entropy index is proposed as a group and individual fairness measure in (Speicher et al.,2018). It is a measure of inequality in a population.It needs to be close to zero for fairness.It estimates how well estimated values are closer to the actual values
6.Euclidean Distance This metric measures the average Euclidean difference between the samples of 2 datasets. Higher the difference, higher the disparity between two groups.
7.Mahalanobis Distance This metric measures the average Mahalanobis difference between the samples of 2 datasets. It measures how many standard deviations away is one sample from another sample. Higher the distance, higher is the inequality between two samples.
8.Manhattan Distance This metric measures the average Manhattan difference between the samples of 2 datasets. The distance metrics are mainly for individual fairness, and they call for similar treatment for similar individuals.
As we see from above metrics, some of them require predicted values and some metrics are calculated on the original dataset. The metrics that require predictions are Equal opportunity difference, average absolute odds difference and Theil Index whereas others do not require the predicted values.

3.4 Debiasing input data and model outcomes

In the literature, the fairness methods are broadly classified into three categories namely: Pre-Processing, In-Processing and Post Processing algorithms (Hajian et al.,2016; Madras et al.,2018; Adebayo et al.2016)

3.4.1 Pre-Processing algorithms

The hypothesis behind pre-processing techniques is that training data itself is biased, which algorithm learns and provides discriminatory outcomes. Hence the data needs to be evaluated in the beginning and data transformation is required to avoid historical bias in the first place. In pre-processing techniques, the input data is modified to quantify the bias in the data. (Calders Verwer, 2010) presents a method to transform data labels with respect to protected attribute to remove discrimination. In their approach, the distribution of protected feature values was modified given each data label. In this proposed method, the model is trained on positive labels and then highly ranked negatively classified items from the protected attribute ,is changed to get fair outcome. The model is then trained on data with more balanced distribution of positive outcome. Similarly (Kamiran and Calders,2009) proves the effectiveness of data transformation on the German credit dataset and achieves the discrimination free outcomes with a small drop in accuracy. One algorithm discussed from (Feldman et al.,2015) relates to the disparate impact. Disparate impact in US labour laws means unintentional practices that puts one group at systematic disadvantage, even though it appears to be neutral. In this process, Feldman talks about modification of each attribute such that the marginal distributions of the parameter with a given sensitive value are equal. (Johndrow and Lum,2019) approaches the same methodology from likelihood-based perspective, that allows to make changes to the mutual independence between protected variable and other attributes and not only just pairwise independence. In summary, pre-processing techniques aim to modify the data to ensure models learn from the “fair” data to mitigate any bias.

3.4.2 In-Processing algorithms

In this technique, the algorithm used for predictions is modified to reduce discrimination. This usually includes adding a regularizer which acts as a penalty for the model's cost function and helps to determine the degree of acceptable bias. This technique assumes that model's cost function is well behaved, and models are completely defined. However, the ensemble models usually do not have the well-defined cost function and hence this method could be difficult to implement in such cases. (Kamishina et al., 2011) has discussed about adding a regularizer which is called "prejudice remover". It aims to reduce the learning behaviour and hence, indirect influence of sensitive variable in the model. In this paper, they added a regularizer to the logistic regression model. The parameter x is added that controls the degree of bias and larger the x, more bias can be reduced. However, there is a bias accuracy tradeoff(Fish et al.,2016) as when the bias is removed the model accuracy is decreased as well. Hence the objective is to strike a balance between bias and accuracy. Another method of in-processing is described by (Pedreshi et al.,2008) and it explains the unfairness in the association rules. It explains that classification rule A, B to C can be split into potentially discriminating and non-potentially discriminating part. A measure called "alpha protection" is defined which is used to tune the desired level of protections against potentially discriminatory part.

3.4.3 Post-Processing algorithms

In this method, the results of the predictive models are manipulated to achieve statistical parity. These techniques aim to balance classification errors across protected sub-groups to achieve equal false negative rates, false positive rates or both (equal odds). (Hardt et al.,2016) has proven that given the data about the target, predictor and protected groups, any learned predictor can be optimally adjusted to remove discrimination. Researchers have argued that these methods are incompatible with individual fairness (Chouldechova, 2017; Corbett-Davies,2017; Kleinberg et al.,2016). (Woodworth et al.,2017) has also demonstrated that under some assumptions, the post processing methods for achieving fairness could be suboptimal. (Kamiran et al.,2012) highlights the limitations of above two methods and proposes two methods namely "Reject Option based classification(ROC)" in which it exploits the low confidence region of probabilistic classifier and relabel instances between deprived and favoured classes to achieve parity. The second proposed method is defined as "Discrimination Aware Ensemble (DAE)" in which it exploits the disagreement region of classifier and relabel again the deprived and favoured classes.

There are 20+ metrics of fairness(refer table below) grouped into individual level measures, group level measures and casual measures. Group measures separates population into privileged and unprivileged group and try to get a statistical measure that is equal across both groups. Individual fairness aims for individuals that are similar to be treated similarly. Causal measures focus on the use of non-discriminatory features to assess the impact of features on prediction outcomes.

In this research paper, we focus mainly on group level measures due to the conflict discussed above between various measures. Such contradictions are also known as Trolley Problem in machine learning[Renda,2018]. Also, examples shared in the IBM toolkit were explained in detailed for the group measures to follow for our dataset.

Definition	Type of Measure	Citation
Statistical Parity	Statistical Measure	Dwork et al.
Conditional Statistical Parity	Statistical Measure	Corbett-Davies et al.
Predictive Parity	Statistical Measure	Chouldechova
False Positive Error Rate Balance	Statistical Measure	Chouldechova
False Negative Error Rate Balance	Statistical Measure	Chouldechova
Equalised Odds	Statistical Measure	Hardt et al.
Conditional Use Accuracy Equality	Statistical Measure	Berk et al.
Overall Accuracy Equality	Statistical Measure	Berk et al.
Treatment Equality	Statistical Measure	Berk et al.
Calibration	Statistical Measure	Chouldechova
Well Calibration	Statistical Measure	Kleinberg et al.
Balance for Positive Class	Statistical Measure	Kleinberg et al.
Balance for Negative Class	Statistical Measure	Kleinberg et al.
Causal Discrimination	Similarity Based Measure	Galhotra et al.
Fairness Through Unawareness	Similarity Based Measure	Russell et al.
Fairness Through awareness	Similarity Based Measure	Dwork et al.
Counterfactual Fairness	Causal Measure	Russell et al.
No unresolved discrimination	Causal Measure	Kilbertus et al.
No proxy discrimination	Causal Measure	Kilbertus et al.
Fair Inference	Causal Measure	Nabi and Shpitser

Table 3.1: Definitions of Fairness. Adapted from Rubin and Verma(2018)

Chapter 4

Methodology and Implementation

4.1 Overview

This section covers the overview of the methodology used in the research process. In the second section, technologies that are used in this process are discussed. The third section covers data exploration, data manipulation, algorithms that are used to build the model and test the model fairness are described and fairness assessment results are shown.

The methodology includes following steps to assess the methods to achieve fairness.

Firstly, supervised machine learning model is built on the credit dataset predicting the probability of loan repayment.

Secondly, Based on the original dataset and on the model outcomes the following measures of fairness would be tested namely, disparate impact, Statistical Parity difference, average odds difference, Equal opportunity difference and Theil Index. These metrics measures the parity among males and females where male

is privileged, and female is unprivileged group.
Thirdly, Following the assessment of bias we would discuss bias mitigation methods such as pre- processing that modifies the features and labels in the training data, In processing methods such as prejudice remover that lead to the fair classifier and post processing methods such as Equalised calibrated odds and assess the outcomes.

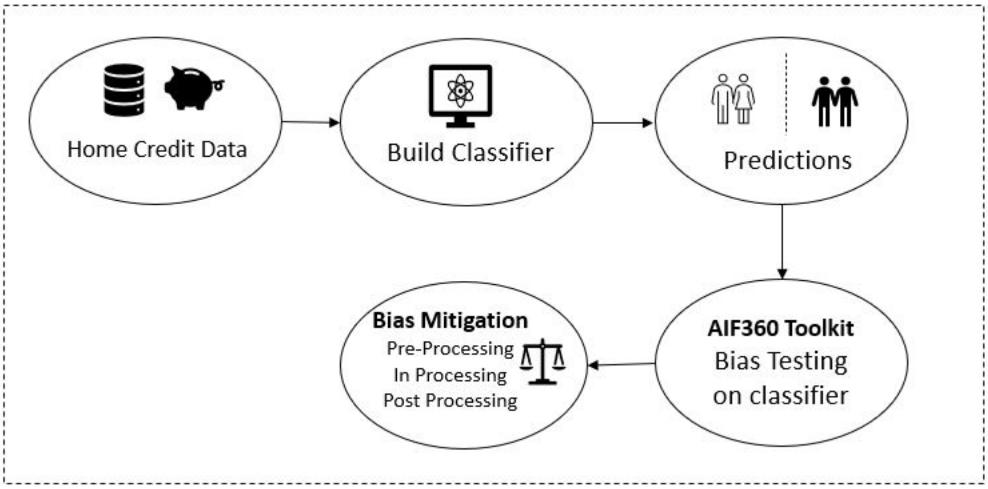


Figure 4.1: Fairness Testing and Implementation Methodology

In testing and mitigation of bias, IBMs AI Fairness 360 toolkit has been used. AI Fairness 360 toolkit contains 70 fairness metrics and 10 bias mitigation algorithms that have been developed by researchers and aim to translate research into actual practice. The metrics relevant to the dataset and used case would be selected to show results. The aif360 package includes:
-Metrics for datasets and models to assess bias
-Explanations of each metric and the benchmark values
-Various methods to mitigate bias in datasets.

The main objective of this library is designed to translate research into actual practice across industries.

4.2 Technologies

This research used Python as a tool for data exploration, model development and fairness testing. Within Python, key libraries used include pandas and numpy for data manipulation, sklearn for machine learning models, and aif360 fairness package is used to import fairness toolbox to estimate the methods of bias and perform fairness treatments.

4.3 Data Processing

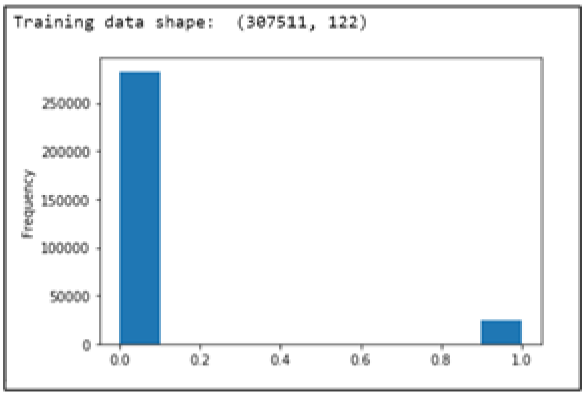
4.3.1 Data Acquisition

The data used in this study is obtained from kaggle website [Credit,2018] where Home Credit Group, Non-banking financial institution shared their dataset to predict the ability of the client to repay the loan. Although multiple datasets across various parameters were shared on kaggle, we used mainly loan application data for building models. The training data also has labels 0: when the loan was repaid and 1: when the loan was not repaid. This also contains the protected variable Gender based on which we defined Male as the privileged group and tested if there is a statistically significant difference between the loan repayment capabilities between males and females based on which they might be treated differently.

Target:
1: the loan was not repaid(defaulters)
0: the loan was repaid

4.3.2 Data Exploration

The training data had 307,511 observations and 122 variables. We analyse the missing percentage of all the features in the data. There were more than 50% of missing values in lot of columns and imputation was required to fit the model on this data while retaining the key information from the data. The distribution of target variable is 92% (282,686) of loans were repaid whereas 8% (24,825) of the loans were not repaid on time as people faced payment difficulties. We explored the reasons for difficulties further.



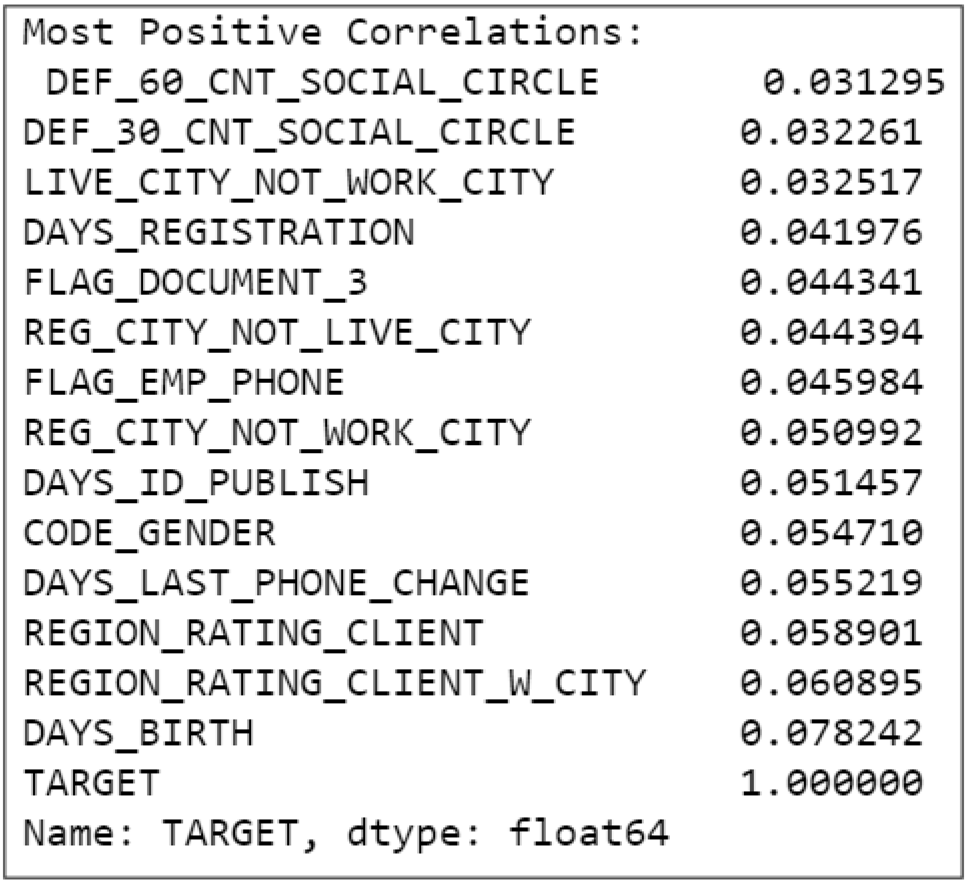
```
Training data shape:  (307511, 122)
```

Figure 4.2: Target Distribution

Anomalies: While doing data exploration data anomalies are identified by looking at the statistics of various columns such as min, max and quartile values. We looked at the 'days-birth' variable that defines the number of days of birth before the date of application. It is a metric of defining age of an individual. There were no outliers in the variable.

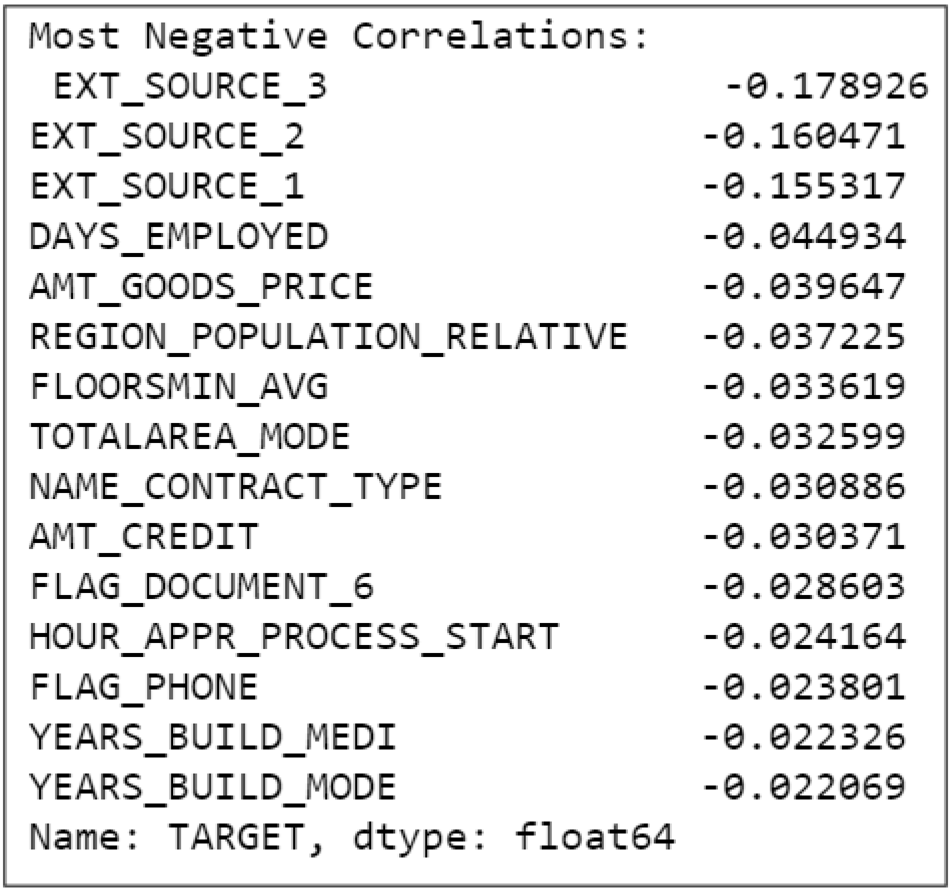
The field 'days-employed' had a maximum value is about 1000 years which seems to be incorrect. The customers with such anomalous values had lower rate of default. These values were replaced to missing so that they could be replaced by missing imputation.

Correlations: The correlation of the target variable with other variables are assessed. The top 10 variables with positive and negative correlations are printed below:



```
Most Positive Correlations:
 DEF_60_CNT_SOCIAL_CIRCLE      0.031295
DEF_30_CNT_SOCIAL_CIRCLE      0.032261
LIVE_CITY_NOT_WORK_CITY       0.032517
DAYS_REGISTRATION             0.041976
FLAG_DOCUMENT_3               0.044341
REG_CITY_NOT_LIVE_CITY        0.044394
FLAG_EMP_PHONE                0.045984
REG_CITY_NOT_WORK_CITY        0.050992
DAYS_ID_PUBLISH               0.051457
CODE_GENDER                   0.054710
DAYS_LAST_PHONE_CHANGE        0.055219
REGION_RATING_CLIENT          0.058901
REGION_RATING_CLIENT_W_CITY   0.060895
DAYS_BIRTH                    0.078242
TARGET                        1.000000
Name: TARGET, dtype: float64
```

Figure 4.3: Variables with Positive Correlations with Target Variable



```
Most Negative Correlations:
 EXT_SOURCE_3                    -0.178926
EXT_SOURCE_2                    -0.160471
EXT_SOURCE_1                    -0.155317
DAYS_EMPLOYED                   -0.044934
AMT_GOODS_PRICE                 -0.039647
REGION_POPULATION_RELATIVE      -0.037225
FLOORSMIN_AVG                   -0.033619
TOTALAREA_MODE                  -0.032599
NAME_CONTRACT_TYPE              -0.030886
AMT_CREDIT                      -0.030371
FLAG_DOCUMENT_6                 -0.028603
HOUR_APPR_PROCESS_START         -0.024164
FLAG_PHONE                      -0.023801
YEARS_BUILD_MEDI                -0.022326
YEARS_BUILD_MODE                -0.022069
Name: TARGET, dtype: float64
```

Figure 4.4: Variables with Negative Correlations with Target Variable

There are 3 variables with strongest negative correlation namely EXT_SOURCE_3, EXT_SOURCE_2, EXT_SOURCE_1. These features are normalised credit score from external sources that represents the credit worthiness of the customer. Higher the score, lower is the probability of default. Most positive correlation is with variable 'DAYS_BIRTH' which is age of the customer at the time of application. This variable has negative values, hence higher the age, less likely he is to default.

Effect of Age on Repayment: It is observed that as the customer gets older, the probability of repayment is higher. When we look at the age distribution across targets and non-targets, we observe that target= 1 distribution is skewed towards the left which is the younger age group which implies younger people have higher probability of default.

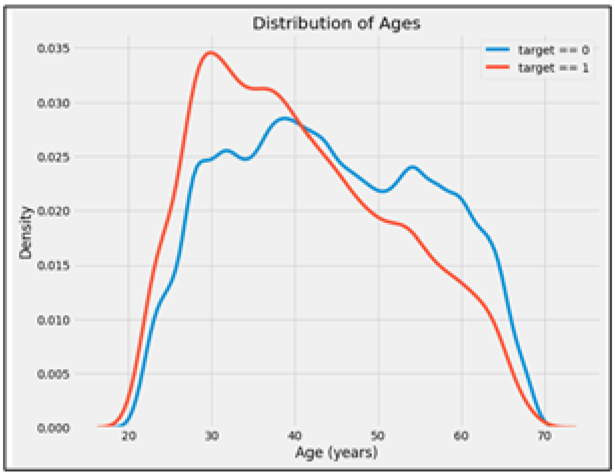


Figure 4.5: Distribution of Age across Target and Non target population

Gender and Repayment: It is observed that Males had a higher default rate of 10% compared to 7% default rate for females.

TARGET	F	M
0	188,278	94,404
1	14,170	10,655
% defaulters(1s)	7%	10%

Table 4.1: Target Distribution across Males and Females

Effect of Family Status on Repayment: By observing the application distribution, we observe 63% of applicants are married, 15% of applicants are single. If we look at defaulters in each of the family status, we observe that percentage of defaults are relatively higher in people under civil marriage and single group.

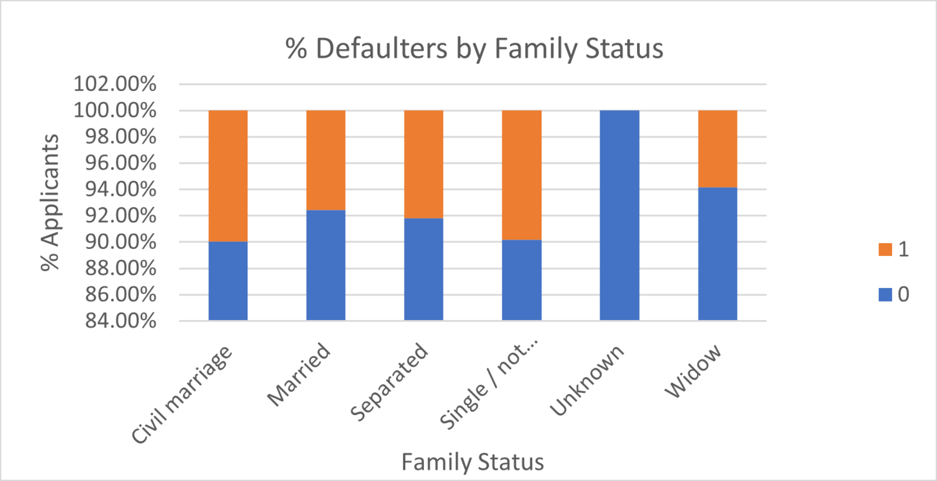


Figure 4.6: Target distribution across Family Status

Missing Value Imputation: Missing data is complex task when analysing the results and it is important to handle the missing values before the model build stage to avoid any bias created by missing data. There are multiple methods of missing value imputation, however in this case we used median value of each column for imputation.

Scaling: We scaled the numeric columns to ensure there is no bias as variables measured in different scales do not contribute equally to the predictions. We used minmaxscaler function from the sklearn library. This ensures each feature is scaled and transformed individually and the final values range between 0 and 1. The transformation is given by

X_standardised = (X- X.min)/(X.max-X.min)

X_scaled = X_standardised * (max-min) +min

4.3.3 Data Manipulation

Class Imbalance: Class imbalance occurs when the total number of targets(1s) is far less than the total number of other classes. In this case, given data had 8% of customers who defaulted which is far less than the 92% of the customers who were able to repay back their loan. This class imbalance poses the problem in machine learning tasks as most of the machine learning algorithms were designed on the assumption of balanced classes and as a result, models do not predict accurately especially for the minority class. The accuracy is no longer a proper measure as it does not differentiate between target and non-target class. To deal with the class imbalance, oversampling is performed to get the equal percentage of targets(defaulters) and non-targets in the data. Post oversampling the data distribution looks as below:

TARGET	Observations	% Observations
0	99,651	50%
1	100,349	50%
Total	200,000	100%

Table 4.2: Target Distribution post oversampling

Categorical Variables: Since in machine learning models, categorical variables cannot be used directly, they need to be encoded into numerical form. There are 2 methods of encoding categorical variable:

Label Encoding: In this method, no new columns are added in the data. The variable is transformed into numeric values assigning a unique value to each category. For example: there is field called 'NAME_CONTRACT_TYPE' that indicates if loan is a cash loan or revolving loan. The cash loan was encoded as 0 and revolving loans were encoded as 1. Similarly, Males were encoded as 1 and females were encoded as 0.

One Hot Encoding: In one hot encoding, a new column is created for each categorical value of a character variable. For example: For 'NAME_CONTRACT_TYPE' , there could be one column for Cash loans taking values 0 or 1 and another column for Revolving loans taking values 0 and 1.

In case of multiple values of a categorical variable, one hot encoding is better as label encoding gives random numeric values which reflects an arbitrary order.

Standard Data Set: The dataset is then converted into Standard dataset as per fairness aif360 module. While defining this we define the target variable name, the protected attribute which is 'Code_Gender' and the privileged class which is Male.

Data Split: we split the features and targets and created 2 datasets with 70% training data(140,000 records) and 30%(60,000 records) validation data.

4.3.4 Model Build

We used logistic regression to model 'Target' as a function of predictors. To evaluate the model performance, we looked at accuracy which came out to be 73% and F1 score is 74%. Below are the model statistics.

Model Accuracy: 73%

F1 Score: 74%

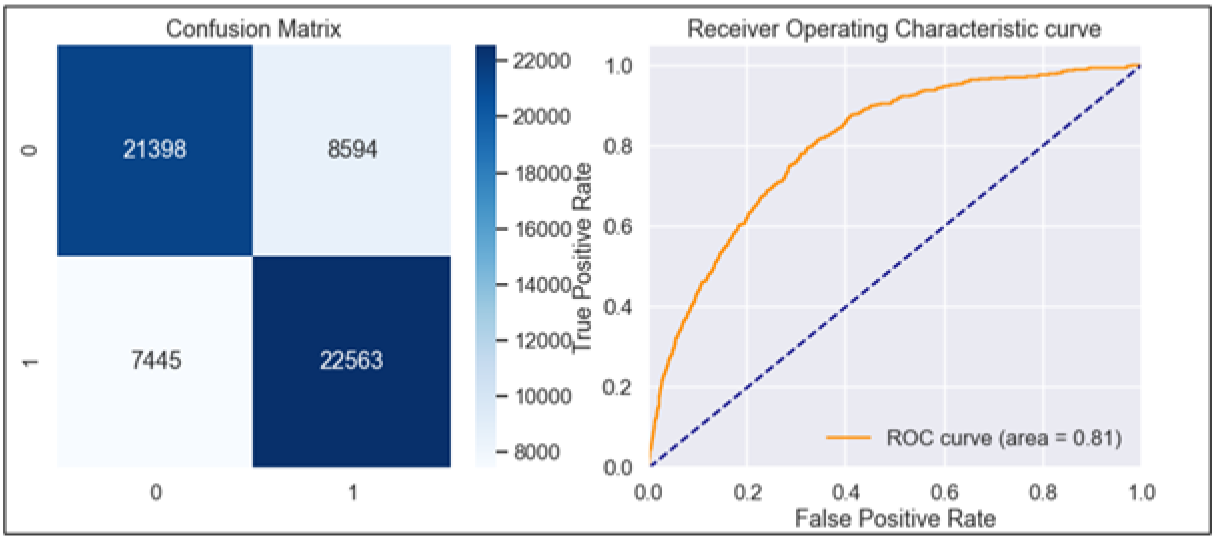


Figure 4.7: Logistic Regression Model Results

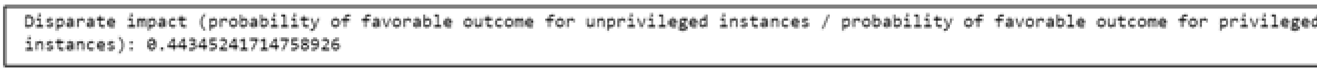
The key predictors included the credit score from the external sources, loan annuity, Age,Credit amount of the loan, duration of employment prior to the loan application, duration of the identity document which is submitted during the loan application, type of loan whether it is cash loan or revolving loan, price of the good for which loan is given and population density of the place where the applicant lives.

To evaluate model performance in terms of separating 2 classes we see the Receiver Operating characteristic(ROC) curve. It shows model performance at various classification thresholds. In this we plot false positive rate on the X axis and True positive rate on the Y axis. The area under the curve(AUC) can be calculated from ROC curve. AUC is used for model comparison and higher the AUC , better is the model. AUC in our case is 0.81. Confusion Matrix is another way of showing the performance of the classification model. It helps to assess the accuracy, recall and other key measures.

4.3.5 Bias Identification

A range of U.S. and European law forbids discrimination against protected classes in a variety of contexts, such as employment, credit, housing, public accommodation, public education, jury selection, use of genetic information, and health care and health insurance. They include various grounds such as gender, marital status, family status, age disability, sexual orientation, race, religion, and membership of the Traveller community.In this case, we assess if there is any bias in default outcomes among males and females.

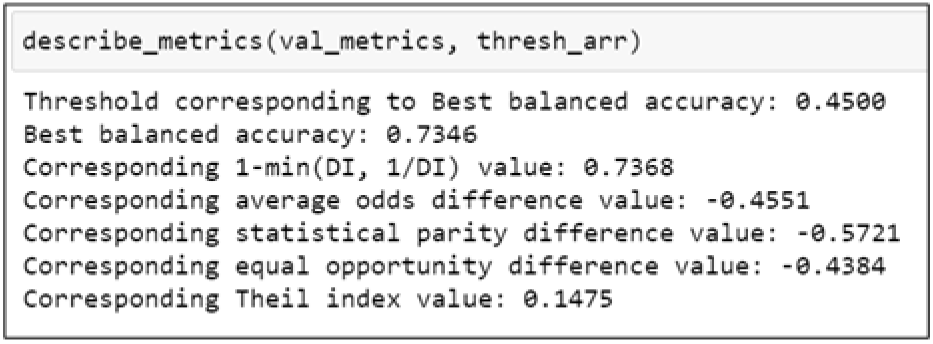
Disparate Impact: To apply the fairness processing and calculation of metrics, we set up the Binary Label dataset object to hold the data. We specified gender as the protected characteristic and male as the privileged group and female as the underprivileged group. AIF fairness comes with the metrics that runs various relevant metrics on the dataset. The disparate impact metric is calculated on the original dataset as below.



```
Disparate impact (probability of favorable outcome for unprivileged instances / probability of favorable outcome for privileged
instances): 0.44345241714758926
```

In this case, the disparate metric of 0.44 indicates positive bias and higher proportion of predicted positive outcomes for privileged group that is males.The ideal value is 1 for parity, however 0.8 can also be considered as a reasonable threshold for parity.

We could also look at the value of 1 - min(disparate impact, 1/disparate impact) and the we want this value to be less than 0.2 to imply fairness We calculate the fairness/bias metrics on the validation data.



```
describe_metrics(val_metrics, thresh_arr)

Threshold corresponding to Best balanced accuracy: 0.4500
Best balanced accuracy: 0.7346
Corresponding 1-min(DI, 1/DI) value: 0.7368
Corresponding average odds difference value: -0.4551
Corresponding statistical parity difference value: -0.5721
Corresponding equal opportunity difference value: -0.4384
Corresponding Theil index value: 0.1475
```

Best Balanced accuracy: This metric shows the model accuracy of around 74%.

1-min(DI,1/DI) value: This metric shows the disparate impact. The ideal value must be less than 0.2 for classifier predictions to be fair, however in this model we see the value is 0.73 implying unfairness

Average Odds Difference: The average odds difference must be close to zero for a classifier to be fair. In this case the value is -0.45.

Statistical Parity / Mean Difference: This represents the difference in mean outcomes between the privileged and underprivileged groups. This metric should be 0 to ensure there is no difference between privileged and underprivileged group. If this is not 0, it means one group has better outcomes than the other one. In this case, mean difference is -0.57 which is negative, it implies that males was getting 57% more positive outcomes in the training data. Males have high probability of default than females which could lead to unfavourable outcomes when granting the credit.

Equal Opportunity Index: This metric explains the difference in recall scores i.e. True Positive rate between unprivileged and privileged groups. Ideal value of 0 indicates equality of opportunity.In above case we can see the value is -0.44 for the equal opportunity index implying the model is unfair.

Theil Index: For models to be fair, the ideal value of Theil index should be 0. In above case we see Theil Index value is 0.14 which implies slight unfairness.

Chapter 5

Implementation of Fairness Methods

In this section, we discuss various ways of mitigating the bias at various steps.

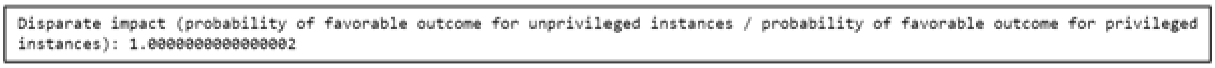
5.1 Pre-processing Algorithm

As discussed above, Pre-processing algorithms transforms the input data to reduce bias. One of the pre processing methods we discuss here is Reweighing.

Reweighing: As we saw in the previous step, there is positive bias in the data which is reflected across multiple metrics. As this is not desirable, bias mitigation is the next step. One of the methods that can be used is called 'Reweighing'. It is a pre-processing algorithm because the mitigation happens before model creation.

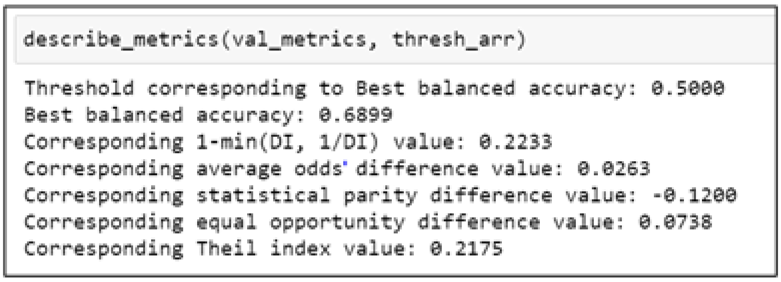
In reweighing, the weights across the group and target combination are given to ensure fairness before classification [Kamiran et al]. This method focuses only on training data and it reweighs the observations. It assigns larger weights to unprivileged group with favorable outcome. The reweighing class is created in aif360 algorithms and the parameters that we need to specify are privileged and unprivileged groups. There are two functions fit and transform that are used here in which Fit computes the weights for reweighing the dataset and transform function converts the dataset into new dataset based on estimated weights.

Post reweighing, we compute the disparate impact metric again on transformed data and result is as below:



```
Disparate impact (probability of favorable outcome for unprivileged instances / probability of favorable outcome for privileged
instances): 1.0000000000000002
```

As we see the disparate impact metric is 1 which means that probability of favourable outcome for privileged and unprivileged group is exactly same. To test for impact on other fairness metrics, we use the reweighted dataset and train the logistic regression model. Below are the results on the transformed data:



```
describe_metrics(val_metrics, thresh_arr)

Threshold corresponding to Best balanced accuracy: 0.5000
Best balanced accuracy: 0.6899
Corresponding 1-min(DI, 1/DI) value: 0.2233
Corresponding average odds difference value: 0.0263
Corresponding statistical parity difference value: -0.1200
Corresponding equal opportunity difference value: 0.0738
Corresponding Theil index value: 0.2175
```

We observe that

1-min(DI,1/DI) value: This metric reduces from 0.73 in original data to 0.22 in transformed data which is desirable.

Average Odds Difference: This metric reduce significantly from -0.45 to 0.02. Ideal value for fairness being 0 , reweighting has improved this metric and shows fair model

Statistical Parity / Mean Difference: This metric reduced from -0.57 to -0.12 post pre-processing the data.

Equal Opportunity Difference: This metric reduced from -0.44 to 0.07

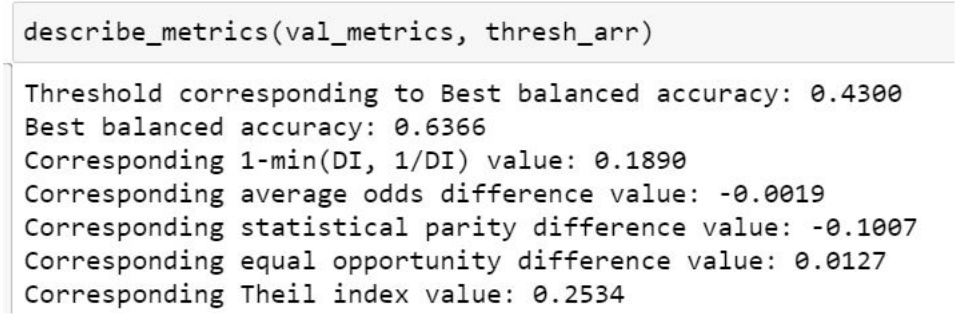
Theil Index: This metric increased from 0.14 to 0.21 reducing the fairness.

5.2 In-Processing Methods:

In In-processing method of reducing bias, the algorithms that are used for training the model are modified to mitigate bias. We have shown ‘Prejudice remover’ as one of the in-processing technique for the home credit data.

Prejudice remover:

This is an in-processing technique where discrimination aware regularisation term is added to the learning objective [Kamishima et al.,2011]. The input for this method includes stating name of the sensitive attribute, the target variable and ‘eta’ which is the fairness penalty parameter. The ‘fit_transform’ function here train the model on the input data and transform the data. The fit function learns the regularised logistic regression model. We validate the Prejudice remover model on validation dataset and below are the fairness metrics post prejudice remover algorithm:



```
describe_metrics(val_metrics, thresh_arr)
```

```
Threshold corresponding to Best balanced accuracy: 0.4300
Best balanced accuracy: 0.6366
Corresponding 1-min(DI, 1/DI) value: 0.1890
Corresponding average odds difference value: -0.0019
Corresponding statistical parity difference value: -0.1007
Corresponding equal opportunity difference value: 0.0127
Corresponding Theil index value: 0.2534
```

We observe that

1-min(DI,1/DI) value: This metric reduces from 0.73 in original data to 0.18 which is desirable.

Average Odds Difference: This metric reduce significantly from -0.45 to -0.0019 . Ideal value for fairness being 0 , inprocessing has improved this metric and shows fair model

Statistical Parity / Mean Difference: This metric reduced from -0.58 to -

0.10 after processing the data.

Equal Opportunity Difference: This metric reduced from -0.44 to 0.012

Theil Index: This metric increased from 0.14 to 0.25 reducing the fairness.

5.3 Post-Processing Methods:

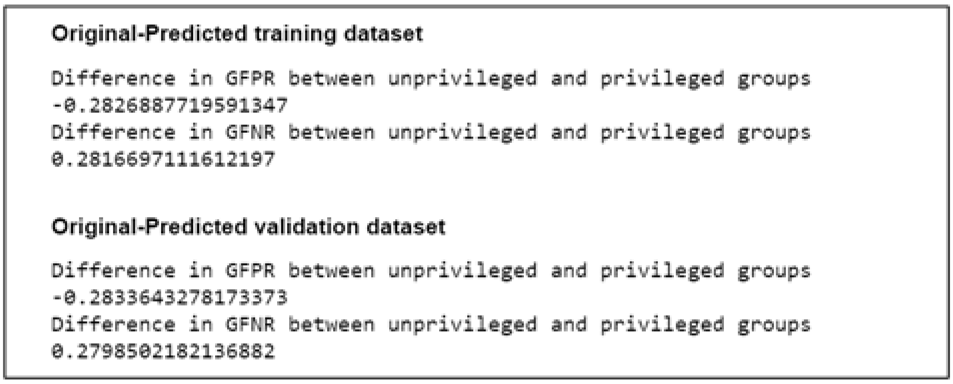
In this approach of building fairness, the results of the previously trained classifier are modified to achieve desired fair results. One of the methods that we used here is Calibrated Equalised Odds Post Processing.

Calibrated Equalised Odds Post Processing: This method calibrates over the classifier score outputs to find probabilities with which we can change the output labels to achieve objective of equalised odds[Pleiss et al.,2017].

The parameters include specifying priviledge and unpriviledged groups, cost constraint which could take values False Positive Rate, False Negative rate or weighted. Seed is an optional parameter to ensure repeatability.

The fit function computes the parameters using true and predicted score for calculating equalised generalised odds while preserving calibration. The predict function in next step, modifies the predicted scores to generate new labels that would satisfy equalised odds constraints.

We observe on our original dataset first the difference between False Positive Rate and False negative rate between two groups. The False positive rate difference is -0.28 and False negative rate difference is 0.27.



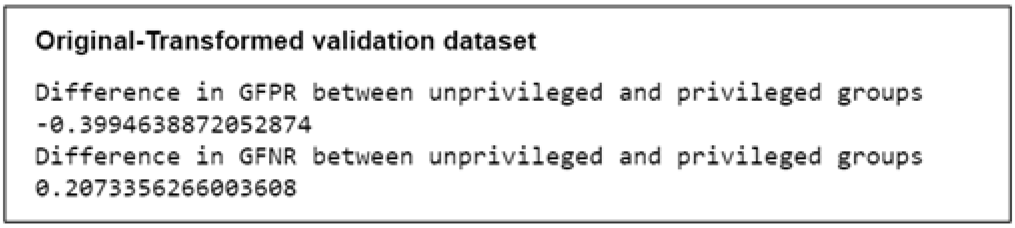
Original-Predicted training dataset

```
Difference in GFPR between unprivileged and privileged groups
-0.2826887719591347
Difference in GFNR between unprivileged and privileged groups
0.2816697111612197
```

Original-Predicted validation dataset

```
Difference in GFPR between unprivileged and privileged groups
-0.2833643278173373
Difference in GFNR between unprivileged and privileged groups
0.2798502182136882
```

Post transforming the validation dataset using odds equalising post processing, we observe that difference in false positive rate reduced to -0.39 and difference in False negative rate reduced to 0.20.



Original-Transformed validation dataset

```
Difference in GFPR between unprivileged and privileged groups
-0.3994638872052874
Difference in GFNR between unprivileged and privileged groups
0.2073356266003608
```

Chapter 6

Results

In this study, we learnt various sources of Bias in machine learning models where it could enter during the collection stage, during the data processing stage, or while building the models.It is important to understand and identify bias as AI is used in various domains in current era for decision making and to ensure it follow legal and ethical norms.It is challenging to quantify bias and mitigate it due to various data sources, complex algorithm design, subjective decisions that goes into building models and define what fairness means. Fairness and profitability objectives are sometimes poles apart and then deeper thoughts and discussion are required on how to balance the same.

6.1 Answers to research questions and key findings

Research Question 1: How we can assess gender bias in the original data and the model outcomes

To answer this question, we need to define the protected attribute first that need to be assessed in the business outcomes. There could be various protected attributes such as gender, race, family status etc. that could lead to discrimination and stakeholder discussions are required to define the key priorities.In this study, we tested if there is bias amongst males and females when predicting the loan default outcomes.

As we discussed above, multiple open source toolkits have been developed such as FAIRML, Fairtest, Aequitas, Themis , AIF360 etc. and we used AIF360 toolkit in this research as it provides comprehensive set of bias and fairness metrics.We explained some of the bias metrics, how they are calculated and their ideal values. We tested for group fairness metrics by dividing the population into privileged and unprivileged groups where male is defined as privileged group. We analysed the key metrics in the original dataset such as disparate impact, Average Odds Value, Statistical Parity Difference, Equal Opportunity Difference and Theil Index. As some of these metrics require model outcomes, we used logistic regression modeling technique to build a classification model which had model accuracy of 73

Research Question 2: How we can debias the data and model outcomes to ensure fair outcomes in lending We could see the bias in the data(Refer table 6.1 below) and the model outcomes as the metrics differs from the ideal values.We used reweighting as the pre-processing techniques, prejudice remover as the in-processing technique and Calibrated equalised odds as the post processing technique that significantly improve the metrics and moves them closer to the ideal values. For post-processing we observe that difference between False Positive rate and False Negative Rate between the privileged and unprivileged groups decrease.

Though we made an attempt to highlight 5 metrics of fairness, there are multiple other metrics of fairness as well. It is not simple to define and satisfy all notions of fairness at the same time. There are fairness accuracy trade offs as well.

Metrics	Original	Pre-Processing	Post-Processing	Ideal Value
Disparate Impact	0.44	1	1	1
1-min (DI,1/DI) value	0.73	0.22	0.18	0.2
Average Odds Difference	-0.46	0.02	-0.001	0
Statistical Parity	-0.58	-0.12	-0.10	0
Equal Opportunity Difference	-0.44	-0.07	-0.012	0
Theil Index	0.14	0.21	0.25	0

Table 6.1: Final Results

Fairness is rooted in ethical beliefs and is hard to satisfy through mathematical formula. Hence, human involvement is required to understand and make decisions where there are serious consequences.

Chapter 7

Conclusion

Predictive models are key to decision making in various industries. To ensure they help to make society fairer on ethical and moral grounds, careful deployment of them is of prime importance. It is important to develop tools for validation and fairness testing to identify if model outcomes are leading to any bias. In this project we discussed 3 methods for finding bias and working on them to translate them into fair outcomes.

Having said that, we must also understand that the notion of gender equality has not been achieved in real life.Even though attempts are made to solve the complex issue of inequality, it is highly likely to be found in the data and hence model outcomes

7.1 Role of Ethical Frameworks

What we discussed in this research is the developer's responsibility of ensuring fair outcomes. What is also important is to have ethical frameworks that defines industry standards and guidelines and ensure that they are followed. At least 84 public and private initiatives are working on defining the guidelines for ethical development and governance of AI. There are several international and US based efforts in developing ethical guidelines for the use of AI [Lee et al.,2019]. The key objective of these guidelines is to ensure the use of emerging technologies is in best interest of public. The European commission also presented Ethical guidelines on trustworthy AI which specifies 7 principles namely Human agency and oversight; technical robustness and safety; privacy and data governance; Transparency; Diversity, non-discrimination, and fairness; environmental and societal well-being and accountability[Floridi,2019]. These principles interpret fairness in terms of equal access, equal treatment, and equality in design processes. OECD also specifies key principles which includes human centered values and fairness [Floridi and Cowls,2019]. It states, "AI systems should be designed in a way that respects the rule of law, human rights, democratic values and diversity, and should include appropriate safeguards to ensure a fair and just society".

AI Ethics have converged to set of principle that is like the medical profession. However, there are concerns around its impact on AI development. Brent Mittelstadt(Mittelstadt,2019) states that AI development lacks *"(1) common aims and fiduciary duties, (2) professional history and norms, (3) proven methods to translate principles into practice, and (4) robust legal and professional accountability mechanisms."* Above reasons suggests that we need to still wait and watch to experience the success of these high-level principles.

Link to the Github Repository

https://github.com/mnanda-g/Algorithmic-Fairness

References

Julius A Adebayo et al. *FairML: ToolBox for diagnosing bias in predictive modeling*. PhD thesis, Massachusetts Institute of Technology, 2016. 16, 18

Chloé Bakalar, Renata Barreto, Stevie Bergman, Miranda Bogen, Bobbie Chern, Sam Corbett-Davies, Melissa Hall, Isabel Kloumann, Michelle Lam, Joaquin Quiñonero Candela, et al. Fairness on the ground: Applying algorithmic fairness approaches to production systems. *arXiv preprint arXiv:2103.06172*, 2021. 3

Solon Barocas and Andrew D Selbst. Big data's disparate impact. *Calif. L. Rev.*, 104:671, 2016. 10

Ivana Bartoletti. Ai in healthcare: Ethical and privacy challenges. In *Conference on Artificial Intelligence in Medicine in Europe*, pages 7–10. Springer, 2019. 1

Rachel KE Bellamy, Kuntal Dey, Michael Hind, Samuel C Hoffman, Stephanie Houde, Kalapriya Kannan, Pranay Lohia, Jacquelyn Martino, Sameep Mehta, Aleksandra Mojsilovic, et al. Ai fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias. *arXiv preprint arXiv:1810.01943*, 2018. 3

Richard Berk, Hoda Heidari, Shahin Jabbari, Michael Kearns, and Aaron Roth.

Fairness in criminal justice risk assessments: The state of the art. *Sociological Methods & Research*, 50(1):3–44, 2021. 22

Yochanan E Bigman and Kurt Gray. People are averse to machines making moral decisions. *Cognition*, 181:21–34, 2018. 2

Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in neural information processing systems*, 29:4349–4357, 2016. 2

Norman Bowie. Companies are discovering the value of ethics. *USA Today Magazine (January 1998)*, 22:41, 1998. 8

Tim Brennan, William Dieterich, and Beate Ehret. Evaluating the predictive validity of the compas risk and needs assessment system. *Criminal Justice and Behavior*, 36(1):21–40, 2009. 1

Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, pages 77–91. PMLR, 2018. 9

Toon Calders and Sicco Verwer. Three naive bayes approaches for discrimination-free classification. *Data mining and knowledge discovery*, 21(2):277–292, 2010. 10

Pete Chapman, Julian Clinton, Randy Kerber, Thomas Khabaza, Thomas Reinartz, Colin Shearer, and Rüdiger Wirth. The crisp-dm user guide. In *4th CRISP-DM SIG Workshop in Brussels in March*, volume 1999. sn, 1999. 14

Irene Chen, Fredrik D Johansson, and David Sontag. Why is my classifier discriminatory? *arXiv preprint arXiv:1805.12002*, 2018. 5

Alexandra Chouldechova. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data*, 5(2):153–163, 2017. 10, 22

Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd acm sigkdd international conference on knowledge discovery and data mining*, pages 797–806, 2017. 9, 22

Home Credit. Home credit default risk. *https://www.kaggle.com/c/home-credit-default-risk/data*, 2018. 25

John Danaher. Robots, law and the retribution gap. *Ethics and Information Technology*, 18(4):299–309, 2016. 6

Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pages 214–226, 2012. 10, 22

Virginia Eubanks. *Automating inequality: How high-tech tools profile, police, and punish the poor.* St. Martin's Press, 2018. 3

Michael Feldman, Sorelle A Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. Certifying and removing disparate impact. In *proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, pages 259–268, 2015. 19

OC Ferrell. Business ethics and customer satisfaction. *Academy of Management Executive*, 18(2):126–129, 2004. 8

Benjamin Fish, Jeremy Kun, and Ádám D Lelkes. A confidence-based approach for balancing fairness and accuracy. In *Proceedings of the 2016 SIAM International Conference on Data Mining*, pages 144–152. SIAM, 2016. 20

Luciano Floridi. Establishing the rules for building trustworthy ai. *Nature Machine Intelligence*, 1(6):261–262, 2019. 45

Luciano Floridi and Josh Cowls. A unified framework of five principles for ai in society. *Available at SSRN 3831321*, 2019. 45

Pratik Gajane and Mykola Pechenizkiy. On formalizing fairness in prediction with machine learning. *arXiv preprint arXiv:1710.03184*, 2017. 10

Sainyam Galhotra, Yuriy Brun, and Alexandra Meliou. Fairness testing: testing software for discrimination. In *Proceedings of the 2017 11th Joint Meeting on Foundations of Software Engineering*, pages 498–510, 2017. 16, 22

Sara Hajian, Francesco Bonchi, and Carlos Castillo. Algorithmic bias: From discrimination discovery to fairness-aware data mining. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 2125–2126, 2016. 18

Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29:3315–3323, 2016. 20, 22

Jiayuan Huang, Arthur Gretton, Karsten Borgwardt, Bernhard Schölkopf, and Alex Smola. Correcting sample selection bias by unlabeled data. *Advances in neural information processing systems*, 19:601–608, 2006. 5

James E Johndrow and Kristian Lum. An algorithm for removing sensitive information: application to race-independent recidivism prediction. *The Annals of Applied Statistics*, 13(1):189–220, 2019. 19

Faisal Kamiran and Toon Calders. Classifying without discriminating. In *2009 2nd international conference on computer, control and communication*, pages 1–6. IEEE, 2009. 19

Faisal Kamiran, Asim Karim, and Xiangliang Zhang. Decision theory for discrimination-aware classification. In *2012 IEEE 12th International Conference on Data Mining*, pages 924–929. IEEE, 2012. 21

Toshihiro Kamishima, Shotaro Akaho, and Jun Sakuma. Fairness-aware learning through regularization approach. In *2011 IEEE 11th International Conference on Data Mining Workshops*, pages 643–650. IEEE, 2011. 38

Michael Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. In *International Conference on Machine Learning*, pages 2564–2572. PMLR, 2018. 10

Niki Kilbertus, Mateo Rojas-Carulla, Giambattista Parascandolo, Moritz Hardt, Dominik Janzing, and Bernhard Schölkopf. Avoiding discrimination through causal reasoning. *arXiv preprint arXiv:1706.02744*, 2017. 22

Pauline T Kim. Data-driven discrimination at work. *Wm. & Mary L. Rev.*, 58: 857, 2016. 2

Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. Inherent trade-offs in the fair determination of risk scores. *arXiv preprint arXiv:1609.05807*, 2016. 21, 22

Akhil Alfons Kodiyan. An overview of ethical issues in using ai systems in hiring with a case study of amazon's ai based hiring tool. *Researchgate Preprint*, 2019. 7

Ho Tak Lau and Adel Al-Jumaily. Automatically early detection of skin cancer: Study based on nueral netwok classification. In *2009 International Conference of Soft Computing and Pattern Recognition*, pages 375–380. IEEE, 2009. 1

Nicol Turner Lee, Paul Resnick, and Genie Barton. Algorithmic bias detection and mitigation: Best practices and policies to reduce consumer harms. *Brookings Institute: Washington, DC, USA*, 2019. 45

David Madras, Elliot Creager, Toniann Pitassi, and Richard Zemel. Learning adversarially fair and transferable representations. In *International Conference on Machine Learning*, pages 3384–3393. PMLR, 2018. 18

John F Mahoney and James M Mohen. Method and system for loan origination and underwriting, October 23 2007. US Patent 7,287,008. 1

Shira Mitchell, Eric Potash, Solon Barocas, Alexander D'Amour, and Kristian Lum. Algorithmic fairness: Choices, assumptions, and definitions. *Annual Review of Statistics and Its Application*, 8:141–163, 2021. 3

Brent Mittelstadt. Principles alone cannot guarantee ethical ai. *Nature Machine Intelligence*, 1(11):501–507, 2019. 45

Razieh Nabi and Ilya Shpitser. Fair inference on outcomes. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018. 22

Dino Pedreshi, Salvatore Ruggieri, and Franco Turini. Discrimination-aware data mining. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 560–568, 2008. 20

Geoff Pleiss, Manish Raghavan, Felix Wu, Jon Kleinberg, and Kilian Q Weinberger. On fairness and calibration. *arXiv preprint arXiv:1709.02012*, 2017. 39

Abdul Hamid M Ragab, Amin Y Noaman, Abdullah S Al-Ghamdi, and Ayman I Madbouly. A comparative analysis of classification algorithms for students college enrollment approval using data mining. In *Proceedings of the 2014 Workshop on Interaction Design in Educational Environments*, pages 106–113, 2014. 1

Andrea Renda. Ethics, algorithms and self-driving cars–a csi of the 'trolley problem'. *CEPS Policy Insight*, (2018/02), 2018. 21

Jonathan Rothwell. How the war on drugs damages black social mobility. *Brookings Institution*, 2014. 4

Chris Russell, M Kusner, C Loftus, and Ricardo Silva. When worlds collide: integrating different counterfactual assumptions in fairness. In *Advances in neural information processing systems*, volume 30. NIPS Proceedings, 2017. 10, 22

Till Speicher, Hoda Heidari, Nina Grgic-Hlaca, Krishna P Gummadi, Adish Singla, Adrian Weller, and Muhammad Bilal Zafar. A unified approach to quantifying algorithmic unfairness: Measuring individual &group unfairness via inequality indices. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2239–2248, 2018. 18

A Stevens, A Anisfeld, B Kuester, J London, P Saleiro, and R Ghani. Aequitas: Bias and fairness audit. *Center for Data Science and Public Policy, The University of Chicago*, 2018. 16

Latanya Sweeney. Discrimination in online ad delivery. *Communications of the ACM*, 56(5):44–54, 2013. 2

Florian Tramèr, Vaggelis Atlidakis, Roxana Geambasu, Daniel J Hsu, Jean-Pierre Hubaux, Mathias Humbert, Ari Juels, and Huang Lin. Discovering unwarranted associations in data-driven applications with the fairtest testing toolkit. *CoRR, abs/1510.02377*, 2015. 16

Blake Woodworth, Suriya Gunasekar, Mesrob I Ohannessian, and Nathan Srebro. Learning non-discriminatory predictors. In *Conference on Learning Theory*, pages 1920–1953. PMLR, 2017. 21

Meike Zehlike, Francesco Bonchi, Carlos Castillo, Sara Hajian, Mohamed Megahed, and Ricardo Baeza-Yates. Fa* ir: A fair top-k ranking algorithm. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, pages 1569–1578, 2017. 16