Feature Targeted Debiasing for Machine Learning Applications in Finance.

Masters in Artificial Intelligence for Sustainable Development

University College London

Faculty of Engineering, Department of Computer Science

Gabriela Gutiérrez Payró, MSc

Supervisors:

Prof. Philip Treleaven, PhD

Adriano Koshiyama, PhD

Nigel Kingsman, MSc

Sahan Bulathwela, PhD

Candidate Number: BWQR3

September 2023 - United Kingdom

Abstract

In this thesis, we tackle the important research problem of algorithmic bias in Machine Learning (ML). A key issue that potentially compromises the fairness and transparency of predictive models, this bias is often embedded within specific features of the model. We focus on using feature importance as a tool to surface and reduce this bias, striving for a method that decreases the bias induced by individual features without sacrificing their predictive power. This focused approach represents a departure from traditional strategies that aim to debias data as a whole. We aim to mitigate the size of the typical accuracy-bias tradeoff that is common in algorithmic fairness research by introducing a new method of resampling through distribution mapping. By preserving the predictive power of the model while simultaneously reducing bias, we aim to achieve a balance of fairness and accuracy, a highly desired yet elusive goal in Machine Learning.

Research Structure

Our investigation is structured into distinct sections, each progressively delving deeper into the nuances of bias and its mitigation:

1. An introduction to the challenges of algorithmic bias in ML.
2. A comprehensive review of relevant literature.
3. A detailed exposition of our materials and methods, including the dataset and the design of our experiments.
4. Presentation and analysis of our results.
5. A reflective discussion on our findings.
6. Concluding remarks and forward-looking statements.

Experiments

In this thesis, we explore how the feature importance in Machine Learning models can reveal the sources of bias. Feature importance tells us how much each feature contributes to a model's predictions. By studying this, we aim to pinpoint and reduce the elements causing unfairness in the model's outcomes. To this end, we propose four distinct experiments:

1. We use SHAP (SHapley Additive exPlanations) to highlight the main predictors of outcome, with the view to use those same features for targeted debiasing. We are highjacking SHAP in order to isolate the drivers of bias.
2. We repurpose permutation feature importance to highlight the features contributing significantly to model bias.
3. We explore feature engineering and selection techniques to develop less biased models.
4. We look into distribution mapping as a technique to reduce bias by replicating the distribution of a more favourable feature hoping this might lead to reductions in disparate impact.

Contributions to Science

The potential contributions of this research are manifold and extend across multiple dimensions of the Machine Learning field. The study aims to:

- Enhance the interpretability and transparency of Machine Learning models.
- Provide novel insights and practical methodologies for mitigating algorithmic bias.
- Equip practitioners with a comprehensive toolbox of strategies for constructing more equitable and fair predictive models
- Contribute valuable insights into the design of future AI systems

The notebook and datasets can be found here: GitHub @GabyGutzP

A paper based on this research will be drafted and submitted for publication.

Impact Statement

In today's digital age, where decisions driven by Machine Learning algorithms can influence a person's livelihood, financial stability, and even societal standing, ensuring that these algorithms are fair and unbiased is paramount. Our research delves deep into the heart of biases in Machine Learning models, especially within the critical sector of finance, shedding light on how deeply entrenched biases can inadvertently seep into automated decision-making systems.

The significance of our work lies not just in the identification of these biases but in the tangible methodologies we propose to mitigate them. Through a series of experiments, we've paved the way for more transparent, fair, and accountable AI systems in finance. Our findings underscore the fact that biases, while deeply rooted in datasets, can be addressed without significantly compromising on predictive accuracy.

Furthermore, by emphasizing the importance of feature significance in model predictions and bias origination, our study offers a novel perspective on model interpretability, urging professionals in the field to be more discerning in their feature selection and engineering processes.

Our research holds the potential to reshape the way financial institutions deploy Machine Learning algorithms, advocating for a more inclusive, fair, and responsible AI. At its core, our work is a call for the tech community, policymakers, and financial institutions, emphasizing that in the quest for technological advancement, the essence of fairness and equity must not be sidelined. This research stands as a testament to the fact that with diligence, innovation, and a commitment to ethics, the future of AI in finance can be both powerful and just.

Contents

List of Figures

List of Tables

Dedication

An meine Zwillingsschwester.
Vielen Dank für deine bedingungslose
Unterstützung. Wir haben es geschafft!

An meinen Bobo.
Vielen Dank, dass du mein Vertrauter
warst und mir in dieser Phase den Anstoß
gegeben hast, den ich brauchte!

Acknowledgements

I am deeply grateful to all those who have contributed to completing this master's thesis and supported me throughout this academic journey. Without their encouragement and assistance, this accomplishment would not have been possible.

First and foremost, I extend my heartfelt appreciation to my family, particularly my parents, whose support and belief in my abilities have been the bedrock of my academic pursuits. I am indebted to them for their sacrifices and unwavering support, which propelled me forward even in the face of challenges.

I would like to express my sincerest gratitude to my supervisors, Philip Treleaven, Adriano Koshiyama, Nigel Kingsman, and Sahan Bulathwela, whose guidance, expertise, and constructive feedback have been invaluable in shaping this thesis.

I am immensely thankful to the Holistic AI team, where I had the privilege to undertake my internship, especially to my mentor Nigel Kingsman. Their willingness to provide me with this valuable opportunity and their continuous support throughout the internship period have enriched my practical knowledge and allowed me to apply theoretical concepts in real-world scenarios.

Additionally, I am grateful to UCL for providing me with a conducive learning environment and access to resources, which have been indispensable in conducting the research for this thesis.

Finally, I extend my thanks to all my friends and colleagues who have been a constant source of encouragement and inspiration. They played an instrumental role in preserving my sanity and mental well-being throughout this thesis journey. Their camaraderie and support have made this academic journey not only bearable but enjoyable.

I acknowledge the use of ChatGPT-4 (Open AI, https://chat.openai.com) to help me proofread and rewrite paragraphs throughout the thesis. This tool was mainly used to improve the readability and fluency of the document.

Chapter 1

This chapter looks into the motivation behind our study, the challenges inherent in Machine Learning models, especially when applied to sensitive sectors like finance, and the overarching objectives we aim to address. This chapter also provides a brief preview of the subsequent sections.

1 Introduction

The rise of Machine Learning has revolutionized how we interpret and interact with the world. Yet, as these algorithms increasingly permeate various facets of life - ranging from financial decisions [1] and job recruitments [2, 3] to predictive policing [4] and medical diagnostics - concerns regarding their fairness and potential biases have arisen. The realization that these algorithms can inadvertently mirror and even amplify societal biases present in the training data has made algorithmic bias a hot topic in scientific and societal debates, necessitating research into strategies to ensure the fairness and transparency of Machine Learning models.

The inherent issue lies in the reality that these powerful algorithms are not innately impartial. Biases within the data, model, or feature set can lead to skewed predictions that may disproportionately affect certain demographics, particularly when employed in critical decision-making scenarios [5]. Hence, it is vital to understand and address the sources of bias in Machine Learning models [6], a compelling research problem this thesis aims to investigate.

Within the scope of this thesis, we leverage feature importance in Machine Learning models to surface and mitigate the drivers of bias. We propose four experiments using the Adult dataset from the UCI Machine Learning Repository [7]. The first experiment applies SHAP values [8] to discern the predominant predictors of our target value. The second employs permutation feature importance to spotlight the features that drive bias. The

third involves feature engineering and selection, and finally, we employ a novel technique of resampling through distribution mapping. Through these experiments, we aim to expose the mechanisms that contribute to algorithmic bias and suggest practical methodologies to reduce it.

This research bears particular relevance and applicability to high-stakes decision-making scenarios, such as those encountered in finance, healthcare, and criminal justice, where both accuracy and fairness are paramount. By providing a methodology to reduce bias without significant loss of accuracy, we hope to contribute to the development and deployment of Machine Learning models that are both powerful and equitable. This research underscores the need to construct Machine Learning models that not only treat impacted individuals fairly but also meet the performance standards dictated by their specific use cases.

1.1 Motivation

Machine Learning algorithms are increasingly being used in a variety of applications, from healthcare to finance to criminal justice [9]. However, these algorithms can also be biased, which can lead to unfair outcomes. In high-stakes systems, such as judicial sentencing, the ramifications of bias can be particularly severe, potentially leading to life-altering consequences for individuals [4, 5]. While the challenge is significant, our objective is to maintain the predictive power of these algorithms while diligently reducing the inherent biases. The next chapter will delve deeper into the problem of bias in Machine Learning and underscore the significance of addressing this prevalent issue.

1.1.1 The Problem of Bias in Machine Learning

Bias in Machine Learning (ML), increasingly prevalent due to the extensive application of Machine Learning models in decision-making across various sectors, can yield significant and widespread implications. One of the most glaring issues is the potential for discrimination,

where biased models can unfairly disadvantage or discriminate against certain groups [10, 11]. This issue becomes particularly concerning in high-stakes applications [4, 9]. For instance, a model trained on biased hiring data might unfairly disadvantage women or minorities [12]. Furthermore, biased Machine Learning models could lead to poor generalizability, performing suboptimally when applied to different demographics or geographies than those they were trained on. If a model is trained predominantly on data from one demographic group, it may fail to generalize well to other groups [13].

The implications of bias extend beyond model performance and societal fairness, manifesting as a loss of trust and adverse economic impacts [14]. Bias that leads to unfair or discriminatory outcomes can result in a loss of trust in the systems that employ Machine Learning models, hindering their adoption and effectiveness. In business settings, biased algorithms can lead to poor decision-making with potential economic repercussions. For example, if an e-commerce recommendation system consistently suggests products based on gender stereotypes, it might miss out on potential sales [15].

Lastly, there are significant legal and ethical implications associated with bias in Machine Learning [16]. Discriminatory practices induced by biased models can lead to legal repercussions [17]. Besides, deploying systems that reinforce or amplify existing societal biases raises broader ethical questions. As such, addressing bias in Machine Learning - understanding its sources, measuring its manifestation in model outcomes, and implementing strategies to mitigate it - is crucial to ensure the development of fair, trustworthy, and effective models.

1.1.2 The Importance of Addressing Bias in Machine Learning

Bias in Machine Learning can have a number of negative consequences, including:

- Unfair outcomes. For example, a biased algorithm could lead to a person being denied a loan or a job because of their race or gender [2, 1, 18].

- Damage to the reputation of the organization that uses the algorithm. For example, if a company is found to be using a biased algorithm, it could damage the company's reputation and lose customers [12].
- Legal liability. In some cases, organizations that use biased algorithms could be held legally liable for the consequences of those algorithms.

1.1.3 Using Feature Importance to Surface the Drivers of Bias

Feature importance measures how important each feature is in a machine-learning model. By understanding which features are most important, we can better understand the drivers of bias in the model.

There are a number of different methods for measuring feature importance. Some of the most common methods include:

- Shapley values: This method measures the contribution of each feature to the model's prediction.
- Gini importance: This method measures the contribution of each feature to the impurity of the tree.
- Mean decrease in impurity: This method measures the average decrease in impurity that occurs when a feature is split.

Of the methods mentioned above, we will focus on SHAP values [8].

1.2 Research Objectives

The central objective of this research is to attenuate the bias induced by certain features in Machine Learning models while simultaneously preserving their predictive power. Given that features are the fundamental components upon which models learn, their influence on bias in predictions can be substantial. However, these features also carry vital information necessary

for making accurate predictions. Therefore, the challenge lies in finding an equilibrium where the information from these features can be utilized effectively without perpetuating or introducing bias.

This research aims to explore methodologies and techniques that can help achieve this delicate balance. Our goal is to decrease the extent of the typical trade-off between accuracy and bias, which is often a prominent issue in Machine Learning models. By managing to reduce bias without significantly compromising accuracy, we can foster the development of more equitable and effective Machine Learning models. This goal is not only of technical importance but also holds substantial societal implications, especially given the ever-increasing use of Machine Learning in decision-making processes across various sectors.

Another pivotal objective of this research is to explore the potential of feature importance in detecting biases in Machine Learning algorithms. Feature importance, a measure reflecting the contribution of each feature to the predictive power of a model, can serve as a significant tool in understanding the model's behaviour and potentially illuminating bias. Our hypothesis is that if a feature has a disproportionate influence on the prediction and is correlated with sensitive attributes (such as race, gender, or age), it may be an indication of a bias in the machine-learning model.

To ascertain this, we aim to develop a methodology that can leverage feature importance to detect and quantify the severity of these biases. Such a methodology could provide a scalable and efficient way to audit Machine Learning models for fairness concerns, going beyond mere detection to actually gauge the extent of the bias. This would allow for the more nuanced tuning of models, enabling developers to address and rectify bias in a more targeted and effective way. In addition to detecting and quantifying bias, this approach could potentially offer insights into the underlying sources of bias, aiding in the more fundamental understanding of how and why bias gets encoded in Machine Learning models. Therefore, this objective aligns with the broader goal of promoting transparency, accountability, and

fairness in Machine Learning, which are becoming increasingly important in our data-driven world.

1.3 Research Experiments

In Machine Learning, model biases, especially when left unchecked, can lead to profound societal implications [5]. Recognizing this challenge, our research conducts a series of experiments designed with the objectives of understanding the nature and extent of biases, and then devising strategies to address them. To this end, we propose four distinct experiments utilizing the Adult dataset, often called the Census Income dataset, from the UCI Machine Learning Repository [7].

1. **Experiment 1: Bias Identification using SHAP and Disparate Impact**

   Before we can address bias, we need to understand its presence and magnitude. This foundational experiment will dive deep into our dataset to unearth underlying biases. By quantifying these biases, we will set the stage for subsequent mitigation strategies. We will use SHAP (SHapley Additive exPlanations) [8] to highlight the main predictors of outcome, with the view to use those same features for targeted debiasing. We are highjacking SHAP in order to isolate the drivers of bias.

2. **Experiment 2: Permutation Feature Importance for Bias**

   We repurpose permutation feature importance to highlight the features contributing significantly to model bias. We aim to understand which features are pivotal for predictions and, more importantly, which ones are primary contributors to observed biases.

3. **Experiment 3: Targeted Debiasing through Resampling**

   We explore feature engineering and selection techniques to develop less biased models. We plan on employing targeted resampling to specifically adjust the distributions of

certain attributes, aiming to reduce the biases we identify in earlier experiments.

4. **Experiment 4: Distribution Mapping for Bias Mitigation**
   Following the targeted debiasing efforts, we will further refine our approach to bias mitigation by exploring distribution mapping techniques. We look into distribution mapping as a technique to reduce bias by replicating the distribution of a more favourable feature hoping this might lead to reductions in disparate impact.

Further information about the experiments can be found in Chapter 3. Through these experiments, we aim to:

- Propose a new strategy on how to mitigate bias through distribution mapping.
- Investigate the underlying processes that contribute to bias, focusing on feature importance.
- Offer practical strategies to mitigate them by reducing bias caused by certain features whilelist leaving the predictive power of those features in place.
- Reduce bias and the size of the accuracy-bias tradeoff.

1.4 Scientific Contributions

The primary scientific contribution of this research lies in its novel approach towards detecting and mitigating bias in Machine Learning models using feature importance. Using distribution mapping as a technique to reduce bias is the second main contribution. Additionally, this work stands to offer several advancements and implications in both theoretical and practical aspects of Machine Learning and AI ethics.

1. **Feature Targeting Debiasing**: One of the main aims of this research is the exploration and demonstration of methodologies that aim to diminish bias induced by individual features, while simultaneously preserving their predictive power. The prin-

cipal aspiration is to mitigate the size of the typical accuracy-bias tradeoff that is often a challenging aspect of fairness interventions in Machine Learning. Through our novel approach, we hope to offer a solution that ensures fairness without compromising the effectiveness of predictive models. This work stands to significantly impact both the scientific understanding of algorithmic bias and the practical procedures for building equitable Machine Learning models.

2. **Advanced Understanding of Bias in Machine Learning**: This research contributes to the scientific literature by advancing our understanding of the nature of algorithmic bias, particularly on how it can be traced back and linked to specific features in the dataset. This work extends beyond the conventional understanding of bias as a byproduct of skewed datasets and provides an in-depth analysis of the role of individual features in exacerbating or mitigating bias.

3. **New Methodologies for Bias Detection and Mitigation**: The proposed experiments, designed to identify and lessen feature-induced bias while preserving their predictive power, serve as innovative methodologies in the field. These methods can offer more nuanced and effective strategies to address algorithmic bias compared to existing techniques, helping to reduce the accuracy-fairness tradeoff that often characterizes de-biasing efforts.

4. **Informing Policy and Practice**: The findings of this research are expected to have direct implications on the practice of Machine Learning model development and usage across various domains. By demonstrating how fairness and accuracy can be concurrently achieved, it provides practitioners with practical tools and strategies for developing fair Machine Learning models. Furthermore, this research can inform policy-making around Machine Learning and AI applications, offering empirical evidence to guide the development of standards and regulations that emphasize fairness and transparency.

5. **Promoting Transparency and Accountability in AI**: By using feature importance as a lens to scrutinize and address bias, this work contributes to the broader goal of promoting transparency and accountability in AI. It underscores the need to probe 'under the hood' of black-box models, not just in terms of their overall predictions but also in terms of their inner workings and the role of individual features. This enhances our capacity to audit AI systems, hold them accountable, and ensure they align with our shared values of fairness and equity.

In addition to these methodological contributions, our research hopes to foster a more informed dialogue on policy-making in Machine Learning applications. We emphasize the need for incorporating fairness and ethical considerations in AI, actively shaping the trajectory of artificial intelligence (AI) to ensure alignment with our shared values of equity and justice.

Our findings may inspire further investigation into alternative forms of regularization or ensemble methods tailored specifically towards reducing bias and ultimately guide new directions in Machine Learning research. By doing this, we seek to ensure a future where technological advancements, particularly in AI, benefit all without bias or prejudice.

In summary, this research serves as a helpful step towards understanding, detecting, and mitigating bias in Machine Learning, thus aligning the development and deployment of AI technologies with the principles of fairness, accountability, and transparency.

1.5 Thesis Structure

Our thesis is structured to provide readers with an understanding of the challenges, methodologies, and findings related to biases in Machine Learning, especially within the context of finance.

1. **Chapter 1: Introduction**

The introductory chapter sets the stage for our research. It delves into the motivation behind our study, the challenges inherent in Machine Learning models, especially when applied to sensitive sectors like finance, and the overarching objectives we aim to address. This chapter also provides a brief preview of the subsequent sections.

2. **Chapter 2: Background and Literature Review**

   To contextualize our study, this chapter offers an exhaustive review of the existing literature. The chapter underscores the importance of the topic and helps situate our work within the broader academic discourse.

3. **Chapter 3: Materials and Methods**

   Here, we delve deep into the mechanics of our research. We introduce the dataset and model used. The chapter then explains the experiments, providing a clear understanding of the methodologies employed, the rationale behind them, and the expected outcomes.

4. **Chapter 4: Results**

   This chapter serves as the empirical heart of our thesis. It presents the findings from each of our experiments, supported by visual aids and quantitative metrics. We highlight the impact of our methodologies on bias mitigation and model performance.

5. **Chapter 5: Discussion**

   Building on our results, this chapter delves into the interpretations, implications, and broader significance of our findings. Furthermore, it contemplates the real-world implications of our research, offering insights into the future of fairness in Machine Learning.

6. **Chapter 6: Conclusions**

   This final chapter offers a reflective synthesis of the entire study. We reiterate our primary objectives, summarize our key findings, and ponder the road ahead, both in terms of potential applications and further research avenues.

7. **References**

   The final section provides a curated list of academic articles, books, and other resources that have informed our study.

Chapter 2

Extensive research has focused on understanding and mitigating bias in Machine Learning [19]. This chapter reviews significant work in this field, offering a comprehensive understanding of the challenges, techniques, and methodologies in detecting and mitigating bias in Machine Learning models.

2 Background and Literature Review

In recent years, Machine Learning has emerged as a powerful tool, empowering us to make sense of the world in ways previously unimaginable. However, as these algorithms increasingly influence numerous aspects of our lives - from credit scoring and hiring [2, 3] to predictive policing [4, 5] and healthcare - concerns around their fairness and bias have surfaced [20]. The growing recognition of these concerns has brought algorithmic bias to the forefront of scientific discourse, emphasizing the need for solutions that ensure Machine Learning models' fairness and transparency without sacrificing their performance.

The problem lies in the fact that these models, while powerful, are not inherently neutral. They can replicate and even amplify societal biases present in the data on which they are trained. The balance between minimizing bias and maintaining model performance is a delicate one to strike, and it forms a key objective of this research.

This bias, whether it arises from the data, model, or selected features, can inadvertently lead to discriminatory outcomes, disproportionately affecting certain groups. This issue becomes particularly egregious when Machine Learning is used in critical decision-making contexts. Therefore, understanding and mitigating the sources of bias in Machine Learning while minimizing loss of performance is a pressing research problem. This is a gap that our research seeks to address.

2.1 Fairness in Machine Learning

Fairness in Machine Learning seeks to ensure that algorithmic predictions and decisions do not inadvertently favour one group over another or perpetuate existing biases. Different scholars and practitioners have proposed various metrics to quantify and assess fairness. Some of these metrics emphasize demographic parity, which suggests that decision outcomes should be independent of protected attributes, such as race or gender [21]. Another metric, known as equalized odds, requires that models exhibit similar error rates across different demographic groups [18].

2.2 Bias and Types of Bias

In Machine Learning, bias refers to systematic errors or assumptions that a model makes when predicting target outputs. It's the difference between the average prediction of our model and the correct value which we are trying to predict. This type of bias is not due to randomness but due to the inherent properties of the algorithm. Bias in Machine Learning can lead to underfitting, where the model does not learn the data well enough and therefore does not perform well on unseen data.

Moreover, in a broader sense, bias in Machine Learning can also refer to the phenomenon where the algorithm's outputs are prejudiced due to historical data or unfair representation of certain classes in the data. It's important to note that this kind of bias can lead to unfairness and discrimination in predictive modelling, affecting certain demographic groups adversely.

Bias in Machine Learning can stem from multiple sources and can significantly affect the fairness and performance of Machine Learning models. Some of the primary sources of bias include:

1. Data bias

- Bias can arise from the data used to train the Machine Learning model. If the data used to train the model is not representative of the population it is intended to serve; the model may inadvertently perpetuate and amplify these biases. This is often termed as selection bias or sampling bias [22].

2. Label bias
   - Bias can also be introduced through the labels assigned to the training data. If the labels have been assigned in a biased manner, the Machine Learning model will learn and reproduce these biases [23].

3. Preprocessing bias
   - Bias can be introduced or exacerbated during the data preprocessing phase, where raw data is transformed or encoded for Machine Learning algorithms. The choice of imputation methods for missing data, outlier handling, or encoding techniques can inadvertently introduce bias [6].

4. Algorithmic bias
   - Bias can originate from the learning algorithm itself. Machine Learning algorithms are designed to optimize specific metrics, and the choice of these metrics can inadvertently favour one group over another, resulting in biased outcomes [24].

A useful reference for understanding bias in Machine Learning is the book "Understanding Machine Learning: From Theory to Algorithms" by Shai Shalev-Shwartz and Shai Ben-David [25]. It provides a look at the different types of bias in Machine Learning and the trade-offs involved in managing them.

There are several types of bias, which can occur at various stages in the machine-learning pipeline. In Ninareh Mehrabi's work "A Survey on Bias..." [26], we found a comprehensive

list of the most common types of bias and their origins. Some of these can be found in Table 2.1.

Table 2.1: Types of bias in Machine Learning.

Types of Bias	
Pre-existing Bias	This type of bias exists in the world before data is collected and can be due to discriminatory social structures, among other factors. It is reflected in the data and consequently learned by the model.
Historical Bias	Pre-existing bias reflected in the data.
Confirmation Bias	Occurs when the data collected or the model itself is influenced by pre-existing beliefs, causing the model to lean towards those beliefs.
Representation Bias	When certain parts of the input space are underrepresented (sampling methods, training data).
Measurement Bias	This happens when the method or tools used to collect data introduce inaccuracies or errors in the data, which subsequently impacts the model's performance.
Learning Bias	When modelling choices amplify performance disparities across different examples in the data.
Aggregation Bias	Difference across groups might require several models rather than a one-size-fits-all model.
Evaluation Bias	When testing on unbalanced benchmarks compared to the target population (e.g. facial recognition).
Deployment Bias	Mismatch between design purpose and use (e.g., a person's likelihood of committing a future crime used for determining sentence length).

2.2.1 Data Generation Process and the Introduction of Bias

The data generation process forms the backbone of Machine Learning projects and plays a significant role in determining the performance of the models. The process encompasses everything from data collection to its subsequent preprocessing. Importantly, each step in this process carries the potential for introducing bias, thereby affecting the fairness and accuracy of the resulting models.

In the initial phase of defining the scope and purpose of data collection, bias can be introduced if the defined population or attributes inadvertently favour certain groups or

outcomes. This can be due to unconscious biases, overlooking minority groups, or failing to account for important variables. The choice of data collection methodology can also introduce bias, especially if it doesn't adequately represent the full range of the population or if it leans towards certain outcomes or groups.

Data collection itself is a significant source of bias, depending on the methods and sources used. Surveys, direct observations, controlled experiments, or extraction from existing databases or web resources may all be biased due to their design, implementation, or inherent biases in the source. For instance, historical data may contain biases due to past discriminatory practices. Similarly, data collected from online sources may be skewed towards certain demographics, thereby not accurately representing the entire population.

Post-collection, during the data preprocessing stage, choices related to handling missing data, outliers, normalisation of numerical data, and encoding of categorical variables can also introduce bias. For instance, if missing data is more common for certain groups and it's not appropriately handled, it can lead to a model that is less accurate for those groups. Also, the choice of normalisation method or encoding technique can favour certain outcomes or relationships, adding another layer of bias. Choices made during the data preprocessing stage can inadvertently introduce or amplify biases in Machine Learning models, particularly in sensitive contexts such as predictive policing [4].

Thus, the data generation process, though fundamental to Machine Learning, presents multiple opportunities for introducing bias. This thesis delves into these areas, proposing strategies to identify and mitigate these sources of bias, specifically focusing on the role of feature importance in this endeavour.

The figure found in Suresh's work "A framework for understanding unintended consequences of Machine Learning" [27] shows some of the possible biases in the data generation process, this can be seen in Figure 2.1.

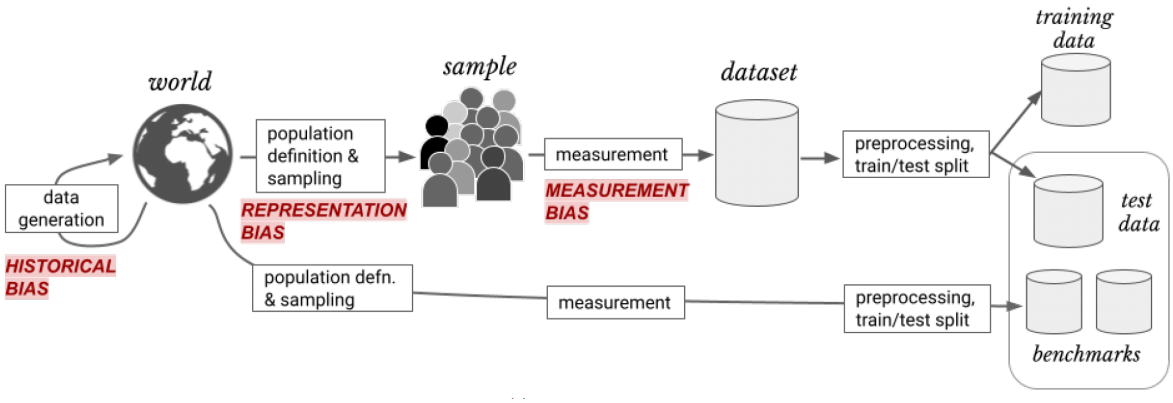

Figure 2.1: Bias in Data Generation Process.

2.2.2 Bias During Model Building and Implementation

During the model-building and implementation phase of Machine Learning (ML), various forms of bias can inadvertently be introduced, affecting the subsequent model's fairness and performance. Decision boundaries, often determined based on training data, may reinforce the discriminatory patterns present in that data. Choices related to the architecture of the model, hyperparameters, or even the type of algorithms can also play a role in propagating bias. For instance, deep learning models, despite their power, can sometimes act as black boxes, making it difficult to discern the source of a particular decision, thus potentially hiding bias. Regularization techniques or model simplifications can inadvertently prioritize certain features over others, leading to unintended biases. Given these complexities, researchers are increasingly emphasizing the importance of interpretability, fairness tools, and robust validation techniques to identify and address biases during this stage of Machine Learning development [22, 28, 29].

Figure 2.1, from[27], shows possible biases during Machine Learning model building.

2.3 Bias Detection and Measurement

To address bias, we first need robust metrics to detect and quantify it. Three primary metrics have emerged as standards in the field:

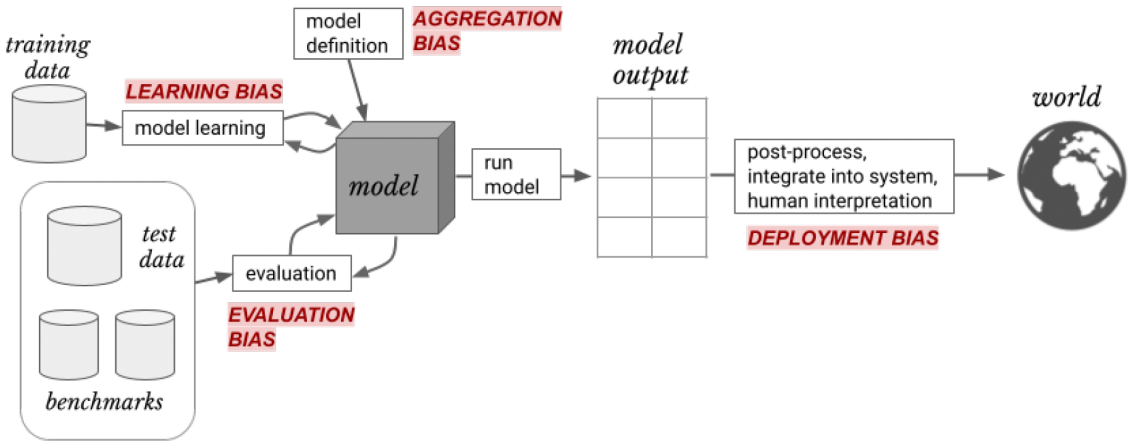

Figure 2.2: Bias in Model Building and Implementation Process.

2.3.1 Disparate Impact

Disparate Impact is a metric used to evaluate fairness. It refers to practices that adversely affect one group of people more than another, even if they appear neutral on the surface. In the context of Machine Learning (ML), disparate impact can manifest when algorithms inadvertently produce outcomes that disproportionately and negatively affect certain groups, especially protected classes such as race, gender, or age [30, 31]. For instance, a hiring algorithm might disproportionately select fewer candidates from a particular ethnicity, not because it was explicitly designed to do so but because of historical biases in the data on which it was trained. An example of this can be found in [12].

$$Disparateimpact = \frac{RateOfFavorableOutcomeForUnprivilegedGroup}{RateOfFavorableOutcomeForPrivilegedGroup}$$

To measure disparate impact in Machine Learning, it's common to analyse the outcomes for different groups. One approach is the 80% rule from the U.S. Equal Employment Opportunity Commission (EEOC): if the selection rate for a certain protected group is less than 80% of the selection rate for the group with the highest selection rate, then there might be disparate impact. In the world of Machine Learning, this is often complemented by met-

rics like demographic parity, equalised odds, and disparate mistreatment, each capturing different facets of fairness.

Mitigating disparate impact involves a combination of strategies throughout the Machine Learning lifecycle [32, 33]. Data preprocessing methods, such as re-sampling or re-weighting, can balance underrepresented classes [34, 35]. Algorithms like adversarial training can be employed to make models explicitly aware of potential biases and correct them. Post-hoc methods, like adjusting classification thresholds for different groups, are also employed. Lastly, iterative feedback loops, where the model's predictions are continuously monitored and adjusted for fairness concerns, can be beneficial. Many researchers have started to evaluate and compare these different methods [19, 30, 31, 33].

However, while these techniques can reduce disparate impact, it's essential to be cognizant of the trade-offs involved, as increased fairness might come at the expense of reduced overall accuracy [9, 36, 37].

2.3.2 Equalized Odds

Equalized Odds is a fairness metric that emphasizes equalizing the prediction error rates across different groups. In a binary classification setting, it requires that both the true positive rate (TPR) and the false positive rate (FPR) be the same for all groups. This means that a model should have an equal chance of correctly classifying positive instances (satisfying equal TPR) and an equal chance of misclassifying negative instances (satisfying equal FPR) across all groups. Achieving Equalized Odds ensures that a model is equally reliable, regardless of the group membership of an individual, making it particularly pertinent in high-stakes scenarios like criminal justice, where the consequences of a false prediction can be severe [18].

2.3.3 Calibration

Calibration is a measure of the alignment between a model's predicted probabilities and the actual observed outcomes. A model is said to be well-calibrated if, for all instances predicted with a probability p, the fraction of those instances that are actually positive is approximately p. In terms of fairness, when applied to different groups, calibration ensures that a prediction made with a certain confidence holds the same degree of reliability irrespective of the group [38]. For instance, if a model predicts a 70% chance of loan repayment for two individuals, one from Group A and the other from Group B, both should have roughly a 70% likelihood of actually repaying. Calibration ensures that predictions are consistent and trustworthy across different demographic or group lines, reinforcing the credibility of Machine Learning models in decision-making processes.

2.4 Mitigating Bias Across Stages of the Machine Learning Process

Machine Learning (ML) models are only as good as the data they're trained on and the processes that create them. If there's bias present at any stage of the Machine Learning pipeline, it can adversely impact the model's outcomes, leading to unfair and potentially harmful decisions. This Subsection outlines various strategies to mitigate bias at different stages of the machine-learning process.

Bias mitigation can be approached at various stages of the model-building process:

- **Pre-processing Techniques**: These methods focus on modifying the data before it's fed into the model. Techniques like reweighting and oversampling fall under this category.
- **In-processing Techniques**: Here, the model itself is adjusted during training to ensure fairness. Regularization-based methods and adversarial training are common strategies.

- **Post-processing Techniques**: After a model has made its predictions, these methods adjust the model's outputs to ensure fairness. Thresholding is a typical post-processing technique.

Table 2.2 mentions some of the mitigation methods that can be implemented in the different stages of the Machine Learning process.

Mitigating bias during model building and implementation involves critically examining each stage of the Machine Learning pipeline for potential bias and taking proactive measures to reduce it.

A range of technical methods for debiasing are available and continue to be developed, such as data reweighing, adversarial de-biasing, fairness constraints, and calibration [19]. However, these are not one-size-fits-all solutions, and their suitability depends on the specific context and fairness criterion at hand. Because of this, we will focus on methods that can be implemented during the preprocessing step.

It is also important to foster transparency and accountability in Machine Learning, as they can help detect, rectify, and prevent bias. This can involve model explainability, algorithmic audits, and regulatory oversight. Ensuring fairness in Machine Learning is not solely a technical challenge but also a societal one, involving multi-disciplinary efforts from computer science, social sciences, law, and ethics.

2.5 Feature Importance

Feature importance refers to techniques that assign a score to input features based on their usefulness in predicting a target variable. In essence, it offers a ranked view of input features from the most important, which have a substantial influence on prediction output, to the least important, which might have little to no effect. Understanding which features significantly drive the predictions of a Machine Learning model can aid in model interpretability,

debugging, and even in improving data collection efforts.

Several methods exist to quantify feature importance in Machine Learning models. One of the simplest methods is to look at the coefficients of linear models, where the magnitude can serve as an indication of importance. For tree-based models, such as decision trees or random forests, importance can be determined based on the number of times a feature is used to split the data, combined with the decrease in impurity that results from those splits [39]. More complex methods include Permutation Feature Importance, which measures the decrease in a model's performance when the values of a specific feature are randomly shuffled, indicating its predictive power [40]. Another sophisticated approach is SHAP [8], which uses game theory to distribute the contribution of each feature value to every prediction in a fair manner.

In practical applications, feature importance can be instrumental in model debugging by pinpointing potentially problematic features that might be leading to unfair or biased predictions. It's also frequently used in feature selection, where redundant or less important features might be removed to simplify the model and potentially enhance its performance.

2.6 Resampling Technique

Resampling is a fundamental technique in the realm of statistical analysis and Machine Learning, used to adjust the structure or distribution of a dataset. At its core, resampling involves either adding or removing instances from the dataset to achieve a specific objective. There are two primary types: oversampling, where instances (typically from minority classes or groups) are duplicated or synthesized to increase their representation, and undersampling, where instances (often from majority classes or groups) are randomly removed to reduce their prevalence.

In the context of fairness in Machine Learning, resampling can be employed to address

class or group imbalances that may introduce or exacerbate biases in predictive models. By adjusting the representation of different groups, resampling aims to create a more balanced dataset, leading to models that potentially make fairer and less biased predictions. This technique, while simple, is powerful and can be especially effective in scenarios where data collection is challenging or where inherent imbalances in the data contribute to model biases.

2.7 Accuracy-Fairness Tradeoff

In the domain of Machine Learning and algorithmic fairness, the accuracy-fairness tradeoff captures a critical challenge. While it's essential to build predictive models that are accurate, it's equally vital to ensure that these models are fair, especially when their decisions impact people's lives. However, as researchers and practitioners aim to rectify biases in models to make them more equitable, they often encounter a decline in the model's overall accuracy. This tradeoff arises because, in many cases, the data used to train these models inherently reflect societal or historical biases. When we modify the model or data to counteract these biases and improve fairness, we might be moving away from the patterns the model initially learned, leading to potential reductions in prediction accuracy.

Navigating this tradeoff is not straightforward. In high-stakes applications like healthcare, finance, or criminal justice, sacrificing too much accuracy can have severe repercussions. On the other hand, using an unfair model can perpetuate societal injustices and lead to discriminatory outcomes. Hence, understanding and managing the accuracy-fairness tradeoff is paramount for responsible AI development.

2.8 AI Auditing

Auditing Machine Learning models provides numerous benefits in mitigating AI risks and biases. It serves as a vital mechanism to identify, quantify, and correct potential biases embedded within these models, thereby enhancing fairness and reducing the risk of dis-

criminatory outcomes. Regular auditing fosters transparency in Machine Learning systems, allowing stakeholders to understand and trust the decision-making process [41]. By doing so, it addresses ethical concerns, mitigates legal risks, and ensures that the algorithm aligns with societal values and norms. Furthermore, auditing can help in identifying the algorithm's limitations, improving its generalizability across different demographics or geographies. In essence, auditing Machine Learning models forms a crucial part of the AI lifecycle, helping to build more robust, ethical, and reliable AI systems.

Algorithmic auditing is crucial for enhancing the transparency, fairness, and accountability of AI systems, allowing us to probe their internal logic, decision-making processes, and implications [42]. Regular audits help in unearthing hidden biases, flagging discriminatory practices, and ensuring the adherence of AI to ethical and legal standards, thus mitigating the risk of undue harm [43]. Moreover, auditing contributes to a more robust understanding of AI's limitations, informing future improvements and adaptations to diverse contexts [44]. Therefore, AI algorithmic auditing is of paramount importance, acting as a vital checkpoint to ensure that AI advancements align with our collective values of equity and justice.

2.9 Related Work

One of the earliest formal considerations of fairness in Machine Learning was proposed by Dwork in "Fairness Through Awareness" [23]. The authors introduced the concept of fairness in classification, suggesting that any two individuals who are similar with respect to a task should be treated similarly. Their work establishes the groundwork for ongoing research in understanding and implementing fairness in Machine Learning.

Research by Buolamwini and Gebru, "Gender Shades" [13], highlighted the bias in commercial AI systems towards gender and skin type. Their investigation into facial analysis technology revealed substantial disparities in error rates between different demographics, prompting critical conversations about auditing AI systems for bias.

A specific form of Machine Learning bias was addressed in "Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings" by Bolukbasi [45]. The authors demonstrated gender bias in word embeddings, a popular framework in natural language processing, and proposed methods to debias these embeddings.

Latanya Sweeney in "Discrimination in Online Ad Delivery" [15], showcased racial discrimination in online ad delivery systems. Sweeney's research demonstrated that the issue of algorithmic bias extends beyond traditional Machine Learning models and into the broader field of automated decision-making systems.

Kusner "Counterfactual Fairness" [46] posited that decisions should not only be fair considering the current state of the world but should also remain fair if the world were different. Their concept of counterfactual fairness opened new avenues of thought in the field of fair Machine Learning.

"Equality of Opportunity in Supervised Learning" by Hardt [18] introduced a criterion for fairness that sought to ensure equal opportunity in decision-making scenarios by minimizing disparity in true positive rates among different groups.

The work by Corbett-Davies, "Algorithmic decision making and the cost of fairness" [47], scrutinized the inherent trade-off between fairness and accuracy in Machine Learning algorithms. The study was instrumental in emphasizing the need to consider the balance between achieving model accuracy and maintaining fairness.

Finally, a comprehensive overview of bias and fairness in Machine Learning can be found in "A Survey on Bias and Fairness in Machine Learning" by Mehrabi [26]. The authors review various sources of bias in datasets and Machine Learning algorithms, the impact of bias, and methods for bias mitigation. This work serves as a fundamental guide for researchers interested in the field of Fair Machine Learning.

Table 2.2: Bias Mitigation Across Machine Learning Stages.

Data Collection	
Balancing datasets	Ensuring that each class within a dataset has approximately equal representation to prevent model bias towards overrepresented classes.
Data augmentation	Techniques such as SMOTE or ADASYN to create synthetic data, especially for underrepresented classes.
Feedback loop	Engage domain experts to review data collection strategies and identify potential biases.
Data Preprocessing	
Data cleaning	Scrutinize the data for inconsistencies or inaccuracies, which could introduce bias.
Feature selection	Prioritize features that are pertinent to the task and avoid those that might introduce irrelevant biases.
Data scaling and normalization	Ensure that all features have equal weight by using scaling and normalization, as certain algorithms can be sensitive to feature scales.
Model Selection and Training	
Regularization	Techniques like L1 and L2 regularization can help prevent overfitting, which might amplify biases present in the training data.
Adversarial training	Utilize adversarial methods to make models resistant to biased perturbations.
Fairness constraints	Introduce fairness constraints during training to ensure equal treatment across groups.
Model Validation	
Disparate impact analysis	Measure and evaluate the model's impact across different groups to ensure it doesn't disproportionately benefit or harm any specific group.
Model interpretability	Tools like SHAP and LIME can shed light on how models are making decisions, allowing reviewers to pinpoint and correct biases
Post-Deployment	
Continuous monitoring	Deploy monitoring tools to track the model's decisions in real-time, ensuring biases don't creep in over time.
Feedback mechanisms	Establish avenues for users and stakeholders to provide feedback on model decisions, which can help in identifying unnoticed biases.
Retraining and updating	Regularly retrain models with fresh data to ensure they remain unbiased as societal norms and datasets evolve.

Chapter 3

In this chapter, we delve deep into the mechanics of our research. We introduce the dataset and model used. The chapter then explains the experiments, providing a clear understanding of the methodologies employed, the rationale behind them, and the expected outcomes.

3 Materials and Methods

3.1 Research Focus

The core objective of this thesis revolves around the issue of algorithmic bias in Machine Learning (ML). With Machine Learning models progressively shaping various sectors and impacting real-world decisions, it becomes pivotal to address and rectify any form of bias. To this end, the research narrows its focus on harnessing feature importance as a potential tool for recognising and subsequently mitigating this bias. The crux of the approach is to identify features that contribute heavily to biases and then modify these features in a manner that retains their predictive power while decreasing their biasing influence. This approach challenges the typical accuracy-bias trade-off that often emerges in algorithmic fairness discussions, aiming to provide a methodology that doesn't require sacrificing model performance for fairness.

Our primary objective is to identify biases and then implement techniques to mitigate them ideally without compromising the model's predictive power.

3.2 Methodology and Experiments

Our multi-pronged approach, focusing on different facets of feature importance and manipulation, seeks to provide a new approach to the problem of bias in Machine Learning models. As the experiments progress, each methodology's effectiveness will be evaluated in light of

our overarching objective: to diminish algorithmic bias without compromising on predictive power.

For this thesis, our investigation follows a systematic series of experiments. Below are the consecutive steps we undertook:

1. **Data Loading and Basic Exploration**:
   (a) Loaded the datasets from [7].
   (b) Undertook preliminary exploration to understand the dataset's structure, features, and distribution.

2. **Data Preprocessing**:
   (a) Highlighted the steps taken to clean the data.
   (b) Discussed encoding strategies used for categorical variables and any scaling techniques applied to ensure uniformity and reduce skewness.

3. **Initial Model Training and Selection**:
   (a) Trained a variety of Machine Learning models, including Logistic Regression, Decision Trees, Random Forests, XGBoost, and others to predict the target variable.
   (b) Analyzed model performances in terms of accuracy and selected Logistic Regression as the primary model for subsequent experiments. We also discussed the rationale behind this selection.

4. **Bias Identification using SHAP (Experiment 1)**:
   (a) Utilized SHAP [8] values to identify the main predictors of the outcome.
   (b) Analyzed which features were driving predictions and could be potential contributors to biases.

5. **Disparate Impact Analysis**:
   (a) Measured the disparate impact on specific groups by analyzing the model's predictions.
   (b) Looked at and understood the magnitude and direction of biases present in the model's outcomes.

6. **Permutation Feature Importance for Bias (Experiment 2)**:
   (a) Repurposed permutation feature importance to pinpoint features that contribute significantly to model bias.
   (b) Analyzed the results to further understand the features causing undue influence on the predictions.

7. **Targeted Debiasing through Resampling (Experiment 3)**:
   (a) We described the resampling strategy, aiming to modify the distribution of underrepresented groups to match that of a reference group (e.g., White males).
   (b) We discussed the intention behind this strategy and its expected impact on model biases.

8. **Distribution Mapping for Bias Mitigation (Experiment 4)**:
   (a) We elaborated on the process of deriving target distributions based on the most favourable group.
   (b) Detailed the methodology applied to move data points closer to the decision boundary, intending to achieve the target distribution.

9. **Model Evaluation Post-Debiasing**:
   (a) After implementing the targeted debiasing techniques, we retrained the model on the modified dataset.

(b) We then evaluated the changes in both performance metrics (like accuracy) and fairness metrics (like disparate impact).

10. **Continuous Reflection and Analysis**:
    (a) After each major step, we engaged in a thorough reflection on the results and insights.
    (b) We discuss implications, challenges encountered, and the rationale behind any adjustments made to the initial plan.

11. **Conclusion and Recommendations**:
    (a) In the conclusions, we synthesized the findings from all experiments.
    (b) Finally, we proposed potential next steps, improvements, or alternative strategies that could be explored in future research endeavours.

3.3 Dataset

The study leverages the Adult dataset, often called the "Census Income" dataset, from the UCI Machine Learning Repository [7]. This dataset is widely used in the research community for Machine Learning experiments, particularly for binary classification tasks. Barry Becker extracted The dataset from the 1994 United States Census database from the Census Bureau's files.

Notably, the Adult dataset has been previously employed in several research efforts to study algorithmic bias and has been previously identified as having biases related to attributes like race and gender. This makes it a suitable dataset for the exploration of bias mitigation techniques in Machine Learning.

The Adult dataset comprises 48,842 instances, each representing an individual. These instances are split into a training set containing 32,561 instances and a test set with 16,281

instances. Each instance is described by 14 attributes, both continuous and categorical. These features can be found in Table 3.1.

Table 3.1: Adult Dataset Data Types.

Categorical	**Numerical**
workclass	age
education	final weight
marital status	education number
occupation	capital gain
relationship	capital loss
race	hours per week
gender	
native country	

The target variable, income, is a binary attribute indicating whether the individual makes more than $50,000 USD per year or not. This target variable essentially divides the dataset into two classes, and the goal of the prediction task is to accurately classify individuals into these classes based on the other 14 attributes.

Upon obtaining the dataset, an essential preliminary step was data cleaning. Machine Learning models often perform best when fed with quality, clean data. Missing or incomplete data entries can lead to skewed results or unintended biases. Our cleaning process involved identifying and eliminating such problematic data entries. Following this rigorous cleaning process, the dataset was divided into two distinct subsets: a training set comprising 30,162 instances, and a testing set totalling 15,060 instances.

Our preparation process encompassed the following steps:

1. **Outcome Conversion**: The original outcome, which represented income brackets, was converted into a binary format. In this transformation, individuals with incomes exceeding $50,000 were represented by the value 1, while those with incomes less than or equal to $50,000 were denoted by 0. This binary representation streamlined our

analyses, enabling clearer interpretations of results.

2. **Feature Removal**: Some features, namely finalweight and edu, were deemed redundant or not pertinent to our analyses. To ensure clarity and reduce potential noise, these columns were dropped from the dataset.

3. **Handling Missing Values**: Machine Learning models require clean datasets devoid of missing values. Our dataset, like many real-world datasets, had its share of missing entries. These were systematically identified and dropped. Additionally, extraneous white spaces, often overlooked but potential sources of inconsistencies, were also removed.

4. **Binary Encoding**: Certain features like gender and country inherently had binary categories. For ease of computation and interpretation, these were encoded into 0-1 values, where 1 often represented the more prevalent or historically privileged category.

5. **Numerical Replacement**: The race feature, though categorical, was transformed into numerical values. This conversion, based on a pre-defined mapping, enabled more streamlined analyses, especially when assessing biases.

6. **One-Hot Encoding**: For the remaining categorical features, one-hot encoding was employed. This technique expands categorical variables into multiple binary columns, representing each category, ensuring compatibility with Machine Learning algorithms. Table 3.2 shows the data types before performing One-Hot encoding to the categorical values.

7. **Feature Scaling**: With all features now in numerical format, it was essential to ensure

Table 3.2: Data type before encoding.

Categorical	**Numerical**
workclass	age
status	education number
occupation	capital gain
relationship	capital loss
	hours per week
	race
	gender
	native country

they were on a similar scale. Using the MinMaxScaler, each feature was scaled to fall within a 0-1 range. This not only aids in faster convergence for certain algorithms but also ensures that no feature disproportionately influences model outcomes due to its scale.

After these transformations, the updated dataset differed significantly from its original state, ready for detailed analysis. For better efficiency and quicker testing, we stored this processed data in new CSV files. This proactive measure ensured our various experiments ran smoothly, eliminating the need for repeated preprocessing steps.

3.4 Model Training and Evaluation

Initial Model Training We used a variety of Machine Learning models to predict the target variable. This step gave us an idea about which models are best suited for this dataset in terms of accuracy. The results of the different models are shown in Table 3.3.

Table 3.3: Model Comparison Based on Cross-Validated Accuracy.

Model	**Accuracy**
Logistic Regression	84.50%
Decision Tree	81.32%
Random Forest	84.54%
XGBoost	86.86%
SVM (Linear Kernel)	54.29%

In the end, we decided to move forward with a Logistic Regression model, which is a common choice for binary classification tasks. Using Logistic Regression as our primary model for experiments was influenced by the following factors:

1. **Interpretability**: Logistic Regression is a linear model, making it relatively easy to interpret. Each feature's weight can be directly examined to understand its impact on the outcome. This interpretability is particularly valuable when addressing bias because it allows for a clearer understanding of how specific features might be contributing to unfair decisions.

2. **Efficiency**: Logistic Regression is computationally efficient, especially when compared to more complex models like SVM or deep neural networks. For experiments that require multiple iterations or when working with large datasets, this efficiency is crucial.

3. **Baseline Model**: Logistic Regression often serves as a good baseline model. Before diving into more complex algorithms, it's useful to understand the performance and characteristics of a simpler model. If biases are evident in Logistic Regression, they are likely to be present, perhaps in more nuanced ways, in more complex models.

4. **Probabilistic Output**: Logistic Regression provides probabilities as outputs, not just class labels. This can be insightful when examining instances near the decision boundary and can be used for further analyses, such as threshold adjustments.

5. **Regularization**: The model offers built-in regularization options (like L1 and L2), which can help prevent overfitting and potentially reduce the influence of less important features.

6. **Broad Acceptance**: Logistic Regression is a widely accepted method in many industries, including finance. Its familiarity can be advantageous when presenting findings to stakeholders who might be more comfortable with well-established methods.

3.5 Experiments

3.5.1 Experiment 1: Bias Identification using SHAP

A central aspect of our methodology is the use of SHAP (SHapley Additive exPlanations) [8]. SHAP values offer a unified measure of feature importance, giving insights into how each feature influences the model's predictions. However, in this study, we adopt a slightly unconventional use of SHAP. Instead of merely using it to explain model decisions, we leverage it to pinpoint the main predictors or features that drive bias in the model's decisions. In essence, we are 'hijacking' SHAP to zero in on the significant contributors to model bias.

SHAP (SHapley Additive exPlanations) values were computed to understand the contribution of each feature to the model's predictions. This allowed us to identify the primary drivers of bias.

Identifying Biases: Disparate Impact Results

Our initial step revolved around identifying potential biases. The disparate impact was calculated for the race and gender features using the formula:

$$Disparateimpact = \frac{RateOfFavorableOutcomeForUnprivilegedGroup}{RateOfFavorableOutcomeForPrivilegedGroup}$$

A value close to 1 suggests fairness, while deviations indicate potential bias. Our results can be seen in Table 4.1 and Figure 4.1.

3.5.2 Experiment 2: Permutation Feature Importance for Bias

After identifying biases in our dataset, especially within attributes like race and gender, our next objective was to assess the strength of these biases in influencing the model's decisions. Permutation Feature Importance (PFI) offers a robust method to evaluate the importance

of individual features by observing the change in model performance when the values of the feature are randomly shuffled. By employing this method, we aimed to quantify the effect of biases in individual features on overall model performance.

For this experiment, we employed the trained logistic regression model, given its transparency and interpretability. We computed the Permutation Feature Importance for each feature, focusing especially on the race and gender attributes. By shuffling the values of these features and observing the resultant change in model accuracy, we gauged the magnitude of their influence.

3.5.3 Experiment 3: Targeted Debiasing through Resampling

Bias mitigation in datasets has long been a critical concern, especially when data-driven models are applied to real-world scenarios with significant societal implications. Resampling stands out as a tangible and often effective method to address such biases directly at the data level before any modelling takes place. The rationale behind this method is rooted in the principle that by achieving a more balanced dataset, especially in terms of sensitive attributes, models trained on this data will inherently be more equitable. In our study, we chose to employ targeted resampling to specifically adjust the distributions of certain attributes, aiming to reduce the biases we identified in earlier experiments.

We followed these steps to perform our experiment:

1. **Feature Modification**:
    - We targeted the most impactful features, specifically gender, race, relationship_Wife, and occupation_Prof-specialty. We hypothesized that by downweighting or re-encoding these features, we could reduce their inherent biases.
    - The down-weighting involved adjusting the scale of these features, diminishing their weight in the model's decision-making process.
    - We down-weighted the features by a factor of 0.5.

- Re-encoding involved transforming the feature values to ensure they did not introduce bias in predictions.

2. **Model Re-training**:

   Using the modified dataset, we re-trained our logistic regression model, anticipating that the revisions would yield a model that's both accurate and more equitable.

3. **Model Comparison**:

   Post-training, we juxtaposed the performance and fairness metrics of the original and new models. This comparison allowed us to discern the effectiveness of our debiasing strategies and to evaluate if bias mitigation came at the expense of accuracy.

Subsequent models trained on the resampled data were assessed against the original dataset. This evaluation illuminated the resampling's impact on disparate impact and overall accuracy. The results can be seen in the next section and in Figure 4.8.

3.5.4 Experiment 4: Distribution Mapping for Bias Mitigation

Following our targeted debiasing efforts, we sought to further refine our approach to bias mitigation by exploring distribution mapping techniques. The motivation behind this experiment was anchored in the hypothesis that replicating the distribution of a more favourable feature might lead to reductions in disparate impact. By aligning the distributions of underrepresented or biased groups with those of predominant or favourably treated groups, we aimed to generate a more balanced representation in the model's decision-making process.

To initiate this experiment, we first examined the distributions of data for race and gender, given their significant impact on the model's biases. We then employed a distribution mapping strategy, where the objective was to modify the dataset such that minority or disadvantaged groups mirrored the distribution of the more favourable feature.

We followed the approach listed below to perform this experiment:

1. **Initial Data Exploration**:

   We began by observing the distributions of the features in the dataset.

2. **Generating Target Distributions**:

   Based on the current distribution of the data and the desired target distribution, we computed the target number of instances for each category. This involved scaling the target distribution to match the total number of instances in the current distribution.

3. **Retrieving Current Distribution**:

   For a specific group (e.g., 'White' or 'Other' for the race attribute), we retrieved the distribution of a particular feature.

4. **Identifying Boundary Data Points**:

   Before we could replicate the distribution of categorical values, it was crucial to identify which data samples were closest to the decision boundary of another category. This step ensured that we selected the most relevant samples to modify, making the mapping process more effective and meaningful. Our initial approach consisted of training a logistic regression classifier to predict the probabilities of the desired categorical value. It would iterate through categories to identify instances with the highest predicted probability for each category. In the end, we achieved this step by evaluating the model's probability scores and identifying samples that had scores closest to the decision threshold (typically 0.5 for binary classification). These samples were on the boundary and were more likely to be influenced by slight changes in feature values.

5. **Distribution Mapping**:

   With boundary data points identified, we proceeded to modify the dataset. The aim was to make the distribution of disadvantaged groups, in terms of race and gender, mirror the distribution of the majority or favourably treated groups.

6. **Model Training and Evaluation**:

   After applying the distribution mapping, we trained our logistic regression model on the modified dataset and evaluated its performance in terms of accuracy and fairness.

This process of selecting boundary data points was vital for this experiment. By focusing on these samples, we ensured that our distribution mapping technique had the maximum possible impact on bias mitigation without introducing undue distortions to the dataset.

3.6 Chapter Summary - Materials and Methods

In this chapter, we delineated the foundation and procedures that guided our research journey. This began with a comprehensive introduction to the dataset used, emphasizing its relevance to our study's objectives. The dataset, central to our experiments, was drawn from the context of finance, a sector where the implications of biases can have profound real-world impacts.

We then transitioned into detailing the experiments undertaken. These were designed to both uncover and mitigate biases in Machine Learning models. Four primary experiments formed the backbone of our methodology:

1. **Bias Identification using SHAP and Disparate Impact**: Here, we employed statistical techniques to identify and quantify biases, especially concerning race and gender attributes. We adopted a slightly unconventional use of SHAP. Instead of merely using it to explain model decisions, we leveraged it to pinpoint the main predictors or features that drive bias in the model's decisions. In essence, we 'hijacked' SHAP to zero in on the significant contributors to model bias. We continued by measuring the disparate impact on specific features by analyzing the model's predictions. This helped us understand the magnitude and direction of biases present in the model's outcomes.

2. **Permutation Feature Importance Analysis**: We repurposed Permutation Feature

Importance to highlight the features contributing significantly to model bias. We discerned which features had the most impact on model predictions and, crucially, were the main drivers of bias.

3. **Targeted Debiasing through Resampling**: Drawing from the insights of the previous experiments, we tailored resampling techniques to adjust data distributions, aiming to reduce biases without significantly compromising accuracy.

4. **Distribution Mapping for Bias Mitigation**: We look into distribution mapping as a technique to reduce bias by replicating the distribution of a more favourable feature hoping this might lead to reductions in disparate impact. Before we could replicate the distribution of categorical values, it was crucial to identify which data samples were closest to the decision boundary of another category. This step ensured that we selected the most relevant samples to modify, making the mapping process more effective and meaningful. We achieved this step by evaluating the model's probability scores and identifying samples that had scores closest to the decision threshold (typically 0.5 for binary classification). These samples were on the boundary and were more likely to be influenced by slight changes in feature values.

Each experiment was stated clearly, ensuring that the methods could be replicated and built upon in future research endeavours. The chapter served as a methodological roadmap, laying the groundwork for the results and discussions that followed.

Chapter 4

This chapter serves as the empirical heart of our thesis. It presents the findings from each of our experiments, supported by visual aids and quantitative metrics. We highlight the impact of our methodologies on bias mitigation and model performance.

4 Results

In our research using the Census Income dataset, our primary goal was to understand biases and find practical ways to reduce them. This results section documents our journey, highlighting key findings at every phase. We started with a basic data analysis, identifying clear disparities, and progressed to more complex tests focusing on the balance between model accuracy and fairness. The following subsections delve deeper into our observations and methods, all aiming to tackle the critical issue of biases in finance-related Machine Learning.

4.1 Initial Analysis

Our initial step in this research was to delve into the Census Income dataset, aiming to pinpoint any evident biases. By using metrics like disparate impact, we quickly observed noticeable disparities, especially within the race and gender attributes.

Beginning our study, we conducted a thorough initial analysis of the Census Income dataset. This foundational step was crucial, as it provided a solid base for the more detailed experiments that followed. Early insights highlighted significant biases, setting the direction for our subsequent deeper analyses.

Diving into the dataset, we spotted significant imbalances, especially in the race and gender attributes. Through the use of disparate impact ratios, these imbalances became quantifiable, revealing clear biases in our data. For example, the gender attribute's initial

disparate impact showed a clear tilt, shedding light on the embedded biases. The representation of this can be seen in Figure 4.1. Similarly, a closer look at the race attribute highlighted clear discrepancies in income distributions across various racial groups, as illustrated in Figure 4.2.

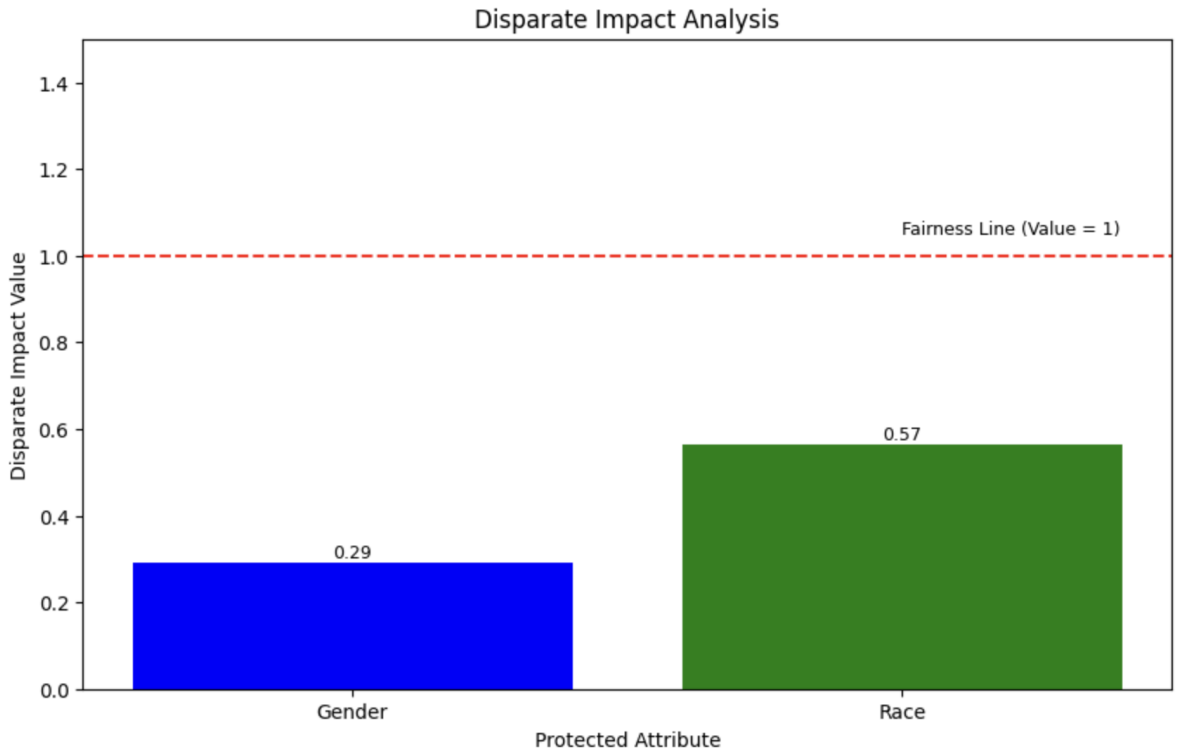

Figure 4.1: Results of Disparate Impact.

Our first look at the data aimed to identify biases and grasp their depth and impact. Key features such as age, education-num, and hours-per-week, central to predicting income, showed links with sensitive attributes, hinting at possible biases. This initial exploration set the groundwork for the subsequent focused bias-reducing strategies and in-depth analyses.

4.2 Model Selection and Information

For our predictive tasks, we utilized a logistic regression model. A popular choice for binary classification, logistic regression offers the advantage of interpretability, allowing for a clearer understanding of feature importance and relationships.

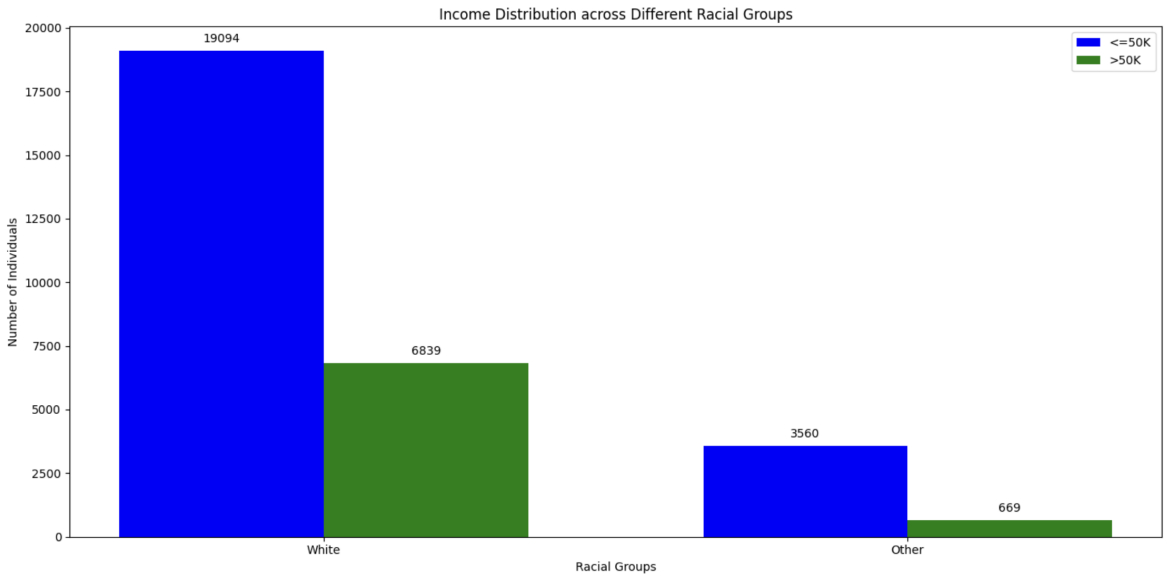

Figure 4.2: Income distribution across different racial groups (encoded).

The Model Parameters are as follows:

- Solver: Liblinear
- Regularization: L1

Throughout our thorough analysis of the Census Income dataset, choosing the logistic regression model proved to be a crucial step. This decision was well-considered, as detailed in the Materials and Methods Section, and the results from this model reaffirmed its suitability for our research.

After training, the performance metrics of the logistic regression model highlighted its effectiveness. Its accuracy on the test set, combined with precision, recall, and F1 score, illustrated its reliability in predictions, as depicted in Figure 4.3. While these outcomes were notable, they stood out even more when compared to results from other models we explored. The probabilistic outputs of the logistic regression model offered deeper insights into biases, particularly evident when evaluating disparate impact in Figure 4.1.

Moreover, the model's interpretability played a pivotal role. We extensively used tools

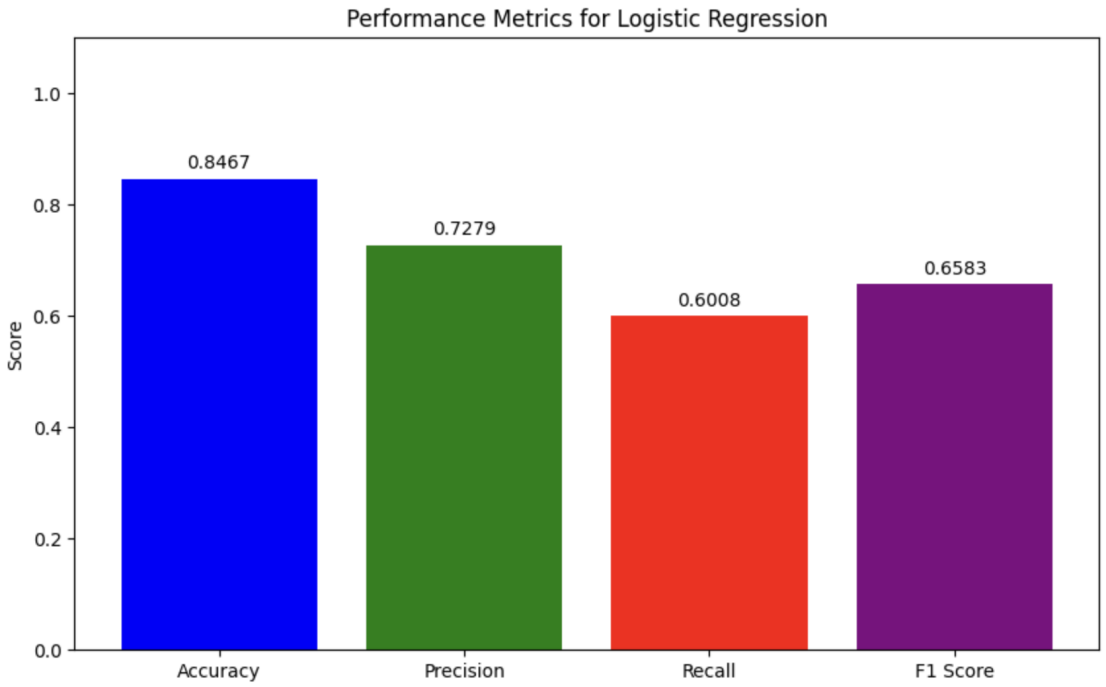

Figure 4.3: Visual representation of the performance metrics for the logistic regression.

like SHAP values, which align naturally with logistic regression. The resulting insights, like feature importance and its link to biases as seen in Figure 4.4, were clearly illustrated due to the straightforward nature of the logistic regression model. These findings, though anticipated, reinforced our model choice and highlighted its relevance in identifying and addressing biases.

4.3 Experiments

4.3.1 Experiment 1: Bias Identification using SHAP

Our first experiment aimed at identifying biases within the Census Income dataset using SHAP (SHapley Additive exPlanations) values. For our study, the SHAP values were used to ascertain how each feature contributed to the model's predictions and to quantify the magnitude of their contributions.

We trained a logistic regression model on the dataset and subsequently computed the

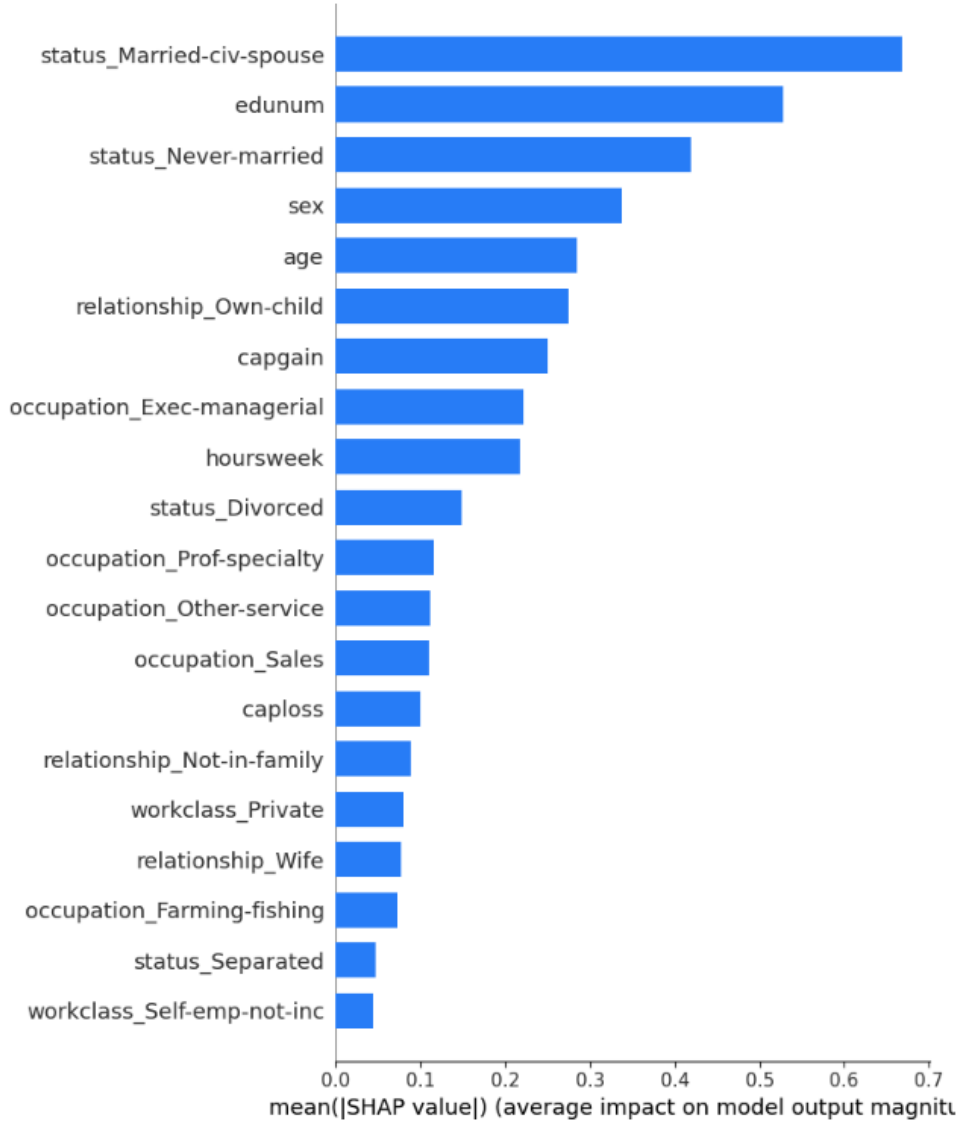

Figure 4.4: SHAP feature importance.

SHAP values for each feature across all data points. The resulting values provided a granular understanding of the influence each feature exerted on the model's predictions. The SHAP summary plot, as visualized in Figure 4.5, highlighted the most impactful features and indicated the positive or negative direction of their effects.

Interpretation of Figure 4.5:

- Features on the Y-axis: The features are ordered from top to bottom by their overall

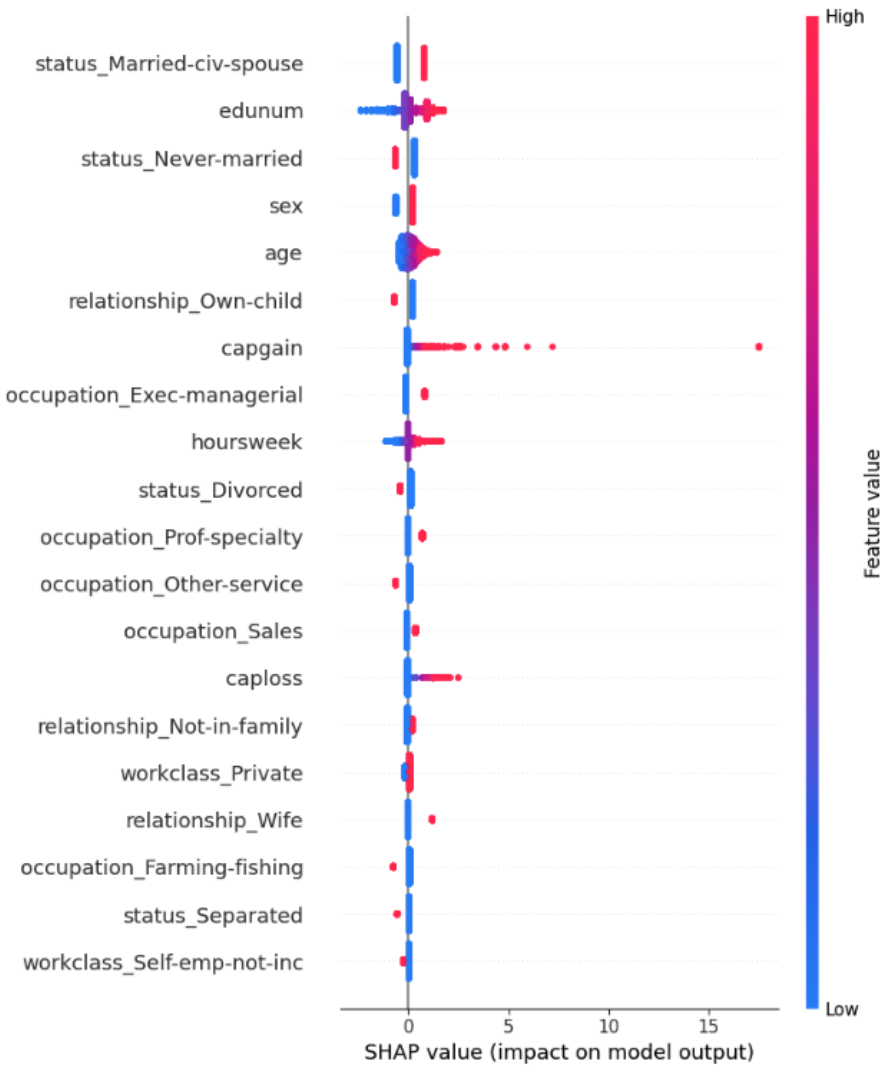

Figure 4.5: SHAP feature importance, ordered from top to bottom by their overall importance.

importance. Importance is determined by the average magnitude of the SHAP values for each feature.

- Horizontal Location: The location on the X-axis indicates the effect of the feature value on the model's prediction. Values to the right of the vertical red line indicate a positive effect (increasing the output of the model), and values to the left indicate a negative effect (decreasing the output).
- Colour: The colour represents the value of the feature. Red indicates a high feature value, while blue indicates a low feature value.

- Distribution of SHAP Values: The density of points for a feature provides insight into the distribution of that feature's effects across all instances. For example, a wide spread of points indicates that a feature has a varied impact on different instances, while a narrow spread indicates a more consistent effect.

Key Findings from Figure 4.5:

- edunum (education number) is the most influential feature, with higher values (red) generally pushing the model's prediction towards a higher income.
- capgain (capital gain) is the second most influential feature, and again, higher values tend to result in predictions of higher income.
- relationship_Husband and gender also play significant roles. Being a husband or being male often contributes to predictions of higher income.
- Features like status_Never-married have a notable negative impact on the prediction, pushing it towards a lower income.
- There are several features that have a mix of positive and negative impacts, indicating the complexity of their interactions with other features and the outcome.

This experiment provided insights into which features drive the model's decisions and in what direction. It was essential for understanding the model's behaviour and was a starting point for addressing potential biases.

Findings

The SHAP summary plot (Figure 4.5) revealed that features like age, education-num, and hours-per-week played significant roles in determining income predictions. However, more critically, sensitive attributes such as race and gender also showcased substantial influence. This was concerning, as ideally, a fair model should not disproportionately weigh protected attributes when making predictions.

Comparing our results with existing literature, many studies have identified biases in datasets and models, but our approach uniquely leveraged SHAP values to pinpoint these biases in a more transparent and interpretable manner [28]. Our methodology provided a fine-grained insight that is not only quantitative but also intuitive, bridging the gap between raw numbers and actionable insights.

Additionally, to delve deeper into the relationships between features and their SHAP values, we generated SHAP dependence plots. These plots served as a visual tool to comprehend how individual feature values influenced the associated SHAP values. Key takeaways from these plots reinforced some of our initial observations from the global SHAP summary. Specifically, we discerned the positive influence of features like education number and capital gain on income predictions. Intriguingly, the plots also suggested potential interactions among features. Notably, we observed a potential interplay between age and education, as well as between education and capital gain, each influencing the model's predictive behaviour. For a more comprehensive view and interpretation of these relationships, please refer to the code.

This foundational exploration using SHAP was instrumental in setting the direction for our subsequent experiments. By identifying key features driving biases, we were better equipped to devise targeted debiasing strategies. Our next experiments aimed at leveraging this knowledge to not only mitigate the biases but also to ensure that model performance was not adversely affected in the process.

Identifying Biases: Disparate Impact Results

Our initial models showcased evident biases. Figure 4.1 and Table 4.1 show the results we found in our experiment. Ideally, in a fair model, this value should be close to 1.0, indicating equal rates of favourable outcomes for both groups.

Table 4.1: Disparate Impact values for race and gender.

Feature	**Disparate Impact**
Race (Minorities vs. Non-minorities)	0.5653
Gender (Female vs. Male)	0.2926

- For Gender: The value of 0.291 means that females are almost three times less likely to receive favourable outcomes compared to males. This indicates a significant disparate impact against the unprivileged group (females) in terms of gender.
- For Race: A disparate impact value of 0.5653 for race suggests that the unprivileged group (e.g., non-whites) has a rate of favourable outcomes that is approximately 56.53% of the rate of favourable outcomes for the privileged group (whites). The value 0.5653 signifies a potential bias against the unprivileged group in the model's predictions.

We also computed other fairness metrics such as Equal Opportunity Difference (EOD) EOD = -0.1082, and Average Odds Difference (AOD) for gender AOD = -0.0961.

These results underline the presence of biases in the model's predictions. The next step we took was to identify the features contributing to these biases to try and mitigate them. The top 5 features based on logistic regression coefficients were: capgain, edunum, hoursweek, caploss, and age. Figure 4.6 shows the distribution for both gender and race for different subgroups of four of the five features mentioned.

From Figure 4.6, we can reflect on the following for each feature:

1. **Feature: capgain (Capital Gain)**
    - Gender: There is a clear disparity in capital gains between genders. Males have, on average, higher capital gains than females. This difference can be a significant factor if the model disproportionately favours individuals with higher capital gains.
    - Race: Interestingly, minorities have slightly higher average capital gains than non-

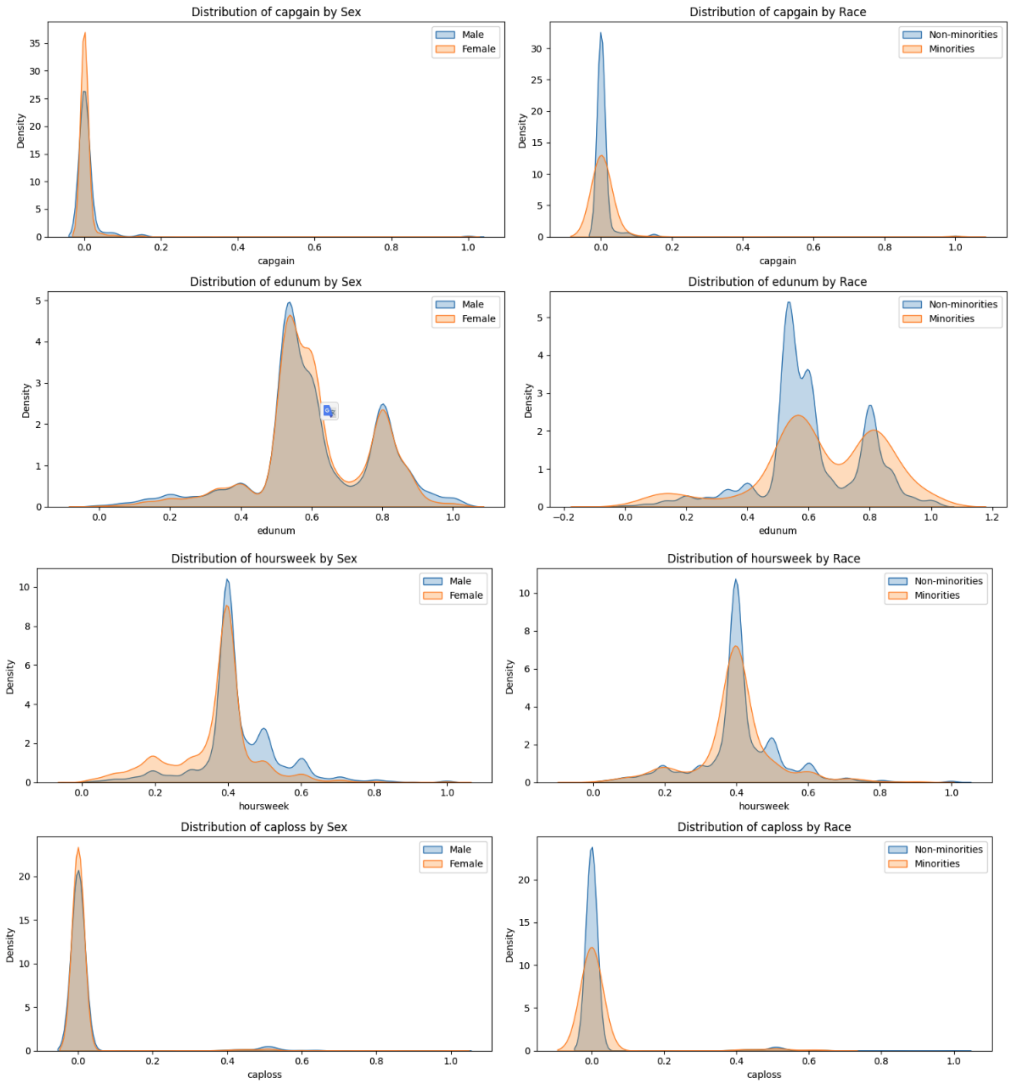

Figure 4.6: Distribution most influential features for bias for both gender and race for different subgroups. In order capgain, edunum, hoursweek, and caploss

minorities. This suggests that in terms of capital gains alone, the model shouldn't inherently favour non-minorities.

2. **Feature: edunum (Education Number)**

   - Gender: The average education number is nearly identical for both males and females. This suggests that any bias in the model's decisions is unlikely to arise from this feature in terms of gender.
   - Race: Minorities have a slightly higher average education number compared to non-minorities. As with capital gains, this suggests that the model shouldn't

inherently favour non-minorities based on this feature alone.

3. **Feature: hoursweek (Hours per Week)**
    - Gex: Males work, on average, more hours per week than females. If the model places significant importance on this feature, it could inadvertently favour males.
    - Race: The average working hours are fairly close for both groups, with non-minorities working slightly more on average.
4. **Feature: caploss (Capital Loss)**
    - Gex: Males have a higher average capital loss than females. If the model places importance on this feature, it might view males as riskier or less favourable.
    - Race: The capital loss averages are very close for both groups, indicating minimal disparity based on race for this feature.

Conclusions for Experiment 1:

- The dataset exhibited clear biases, especially in the attributes related to race and gender.
- SHAP values highlighted that while attributes like age, education-num, and hours-per-week played significant roles in income predictions, sensitive attributes such as race and gender also had substantial influence. This was concerning as protected attributes should ideally not disproportionately influence predictions.
- SHAP dependence plots suggested potential interactions among features. Relationships between age and education, as well as education and capital gain, were observed, hinting at their combined influence on the model's predictions (refer to the code).
- Granular Insight: The use of SHAP values offered a more transparent and interpretable way to pinpoint and quantify biases in the dataset compared to traditional methods.
- Features like capgain and hoursweek showed clear disparities between genders. If the model heavily weighs these features, it could inadvertently favour one group over the other.

- Features like edunum and caploss seemed to be more balanced across groups, suggesting they were less likely to be sources of bias on their own.
- Our findings suggested that females are almost three times less likely to receive favourable outcomes compared to males (Table 4.1).
- We identified pronounced income inequalities across different racial groups, with certain minorities consistently receiving lower income predictions compared to the majority white group (Figure 4.2).

The next step we took in our experiments involved assessing the model's reliance on these features and implementing methods to mitigate any resulting biases.

4.3.2 Experiment 2: Permutation Feature Importance for Bias

In our experiment, we repurposed Permutation Feature Importance to spotlight features that significantly contribute to the overall model bias, thus providing another avenue to target and rectify discriminatory patterns.

Our findings from the Permutation Feature Importance analysis were revealing. Shuffling values for race and gender led to a minor decrease in model accuracy, underscoring their influence on the model's predictions. Permuting the values for 'race' caused a slight decrease in accuracy of approximately 0.0465%; for 'gender', it was about 0.6972%, and for both 'race' and 'gender' combined, the decrease was approximately 0.9163%. See Table 4.2 for the accuracy comparison. This corroborated the biases we identified in the initial SHAP analysis.

When examining the significance of different features, the impact of edunum (education number) and staus_Married-civ-spouse on model accuracy was the most pronounced. Although attributes such as race and gender revealed clear biases, permuting the values of edunum and staus_Married-civ-spouse resulted in the most significant decrease in model accuracy. This highlights the strong influence of education level and marriage on income

predictions. Nevertheless, the minor accuracy drops observed when permuting sensitive attributes like race and gender underscore the biases related to these attributes in our dataset, as illustrated in Figure 4.7.

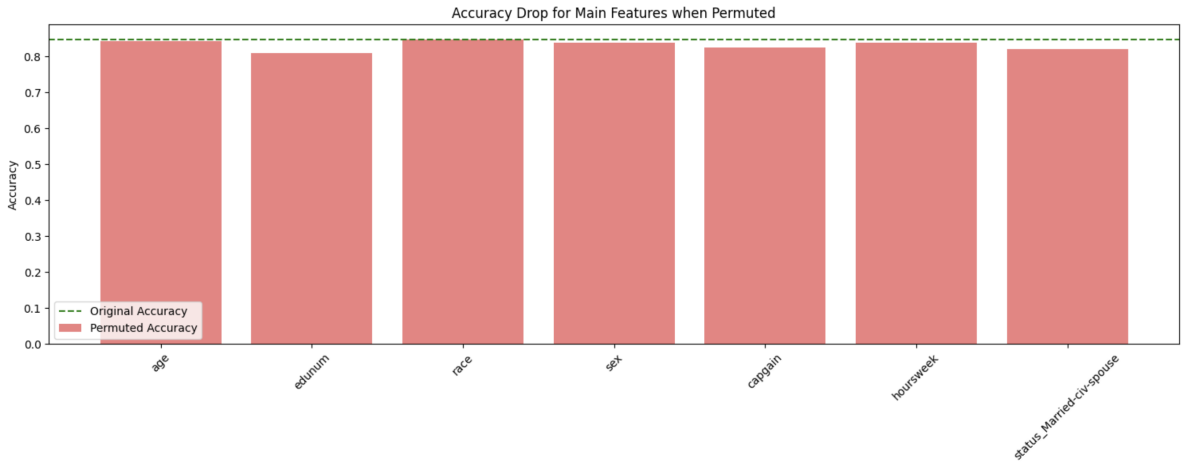

Figure 4.7: Accuracy drop for each of the main features when they are permuted.

During this experiment, we also computed the disparate impact for each feature after permuting its values. By comparing these values with the initial disparate impact, we were able to identify which features, when permuted, lead to significant changes in bias. We learned the following:

- Feature: gender

  When the gender feature is permuted, the disparate impact for gender rises substantially (from 0.291 to 0.558), suggesting that the model heavily relies on this feature for its predictions. This indicates that the gender feature is a significant contributor to the observed bias.

- Feature: relationship_Wife

  The disparate impact for the gender attribute drastically drops (from 0.291 to 0.120) when this feature is permuted. This suggests that the model's treatment of the relationship_Wife feature is exacerbating gender bias.

- Feature: occupation_Prof-specialty

Permuting this feature led to a significant increase in disparate impact for the race attribute (from 0.932 to 1.133). This implies that this feature may contribute to racial bias in the model's predictions.

- Features with minor changes:

  Features that don't show much change in disparate impact upon permutation likely don't contribute significantly to the model's bias. For instance, features like workclass_Without-pay, occupation_Armed-Forces, and occupation_Priv-house-serv show little to no change in disparate impact upon permutation, suggesting they're not primary contributors to bias in this model.

Comparing our approach with existing research, while many studies have employed Permutation Feature Importance for general feature importance, our experiment uniquely leveraged it to isolate and quantify biases, offering a more targeted approach to bias detection [48].

Table 4.2: Permutation feature importance accuracy comparison.

Feature Permuted	**Accuracy**
None - original dataset	0.8467
Race	0.8460
Gender	0.8389
Race and Gender	0.8371

Conclusions for Experiment 2:

- By permuting individual features and observing the resultant changes in disparate impact, it became evident which features had a pronounced influence on the model's biases. Particularly, attributes like race, gender, and hours-per-week showed considerable changes in disparate impact after permutation, pointing to their significant role in shaping the model's decisions. The full results can be seen in the code as well as in Table 4.3.
- The permutation test allowed us to recognize which features in their original form

(before permutation) might have been contributing to reducing or exacerbating biases. For example, the disparate impact for gender nearly doubled after permutation, indicating that the original data structure was somewhat helping in mitigating biases, which became more pronounced upon shuffling.

- Some features showed stark differences in disparate impact post-permutation. For instance, hours-per-week had a high original disparate impact, which dropped post-permutation, highlighting its considerable influence on the model's predictions and associated bias.
- The substantial decrease in disparate impact for the race feature after permutation was notable. The high original value suggested a certain bias in the model's predictions is based on race, which got somewhat diluted when the feature values were shuffled.
- The permutation feature importance, when seen holistically, emphasized the interplay of features in the model. While some features, when permuted, heightened biases, others diluted them, reflecting the complex interdependencies in the data.
- The results from this experiment helped us understand which features influenced biases the most.

Having quantified the biases and their significance, our next step was to devise strategies to mitigate them. This set the stage for our third experiment, "Targeted Debiasing through Resampling".

4.3.3 Experiment 3: Targeted Debiasing through Resampling

In this experiment, we aimed to modify the feature weights to reduce biases without compromising the predictive power of the influential features.

Upon applying the modifications, we observed a slight negative shift in disparate impact values for both race and gender, indicative that this strategy did not work for this particular case. Notably, the modified model exhibited a slight decrease in accuracy. The disparate

Table 4.3: Accuracy and Disparate Impact (DI) for each feature after permuting.

Feature	**DI (Gender)**	**DI (Race)**	**Accuracy**
age	0.313356	0.926296	0.840837
edunum	0.297177	1.409402	0.811952
race	0.288977	1.158611	0.845883
gender	0.546243	0.960778	0.837185
capgain	0.291196	0.917206	0.823971
caploss	0.288494	0.915365	0.841633
hoursweek	0.317996	0.919136	0.841301
country	0.291959	0.876028	0.846614
workclass_Federal-gov	0.292506	0.947091	0.846016
workclass_Local-gov	0.290197	0.931572	0.846348
workclass_Private	0.289086	0.947748	0.844954
workclass_Self-emp-inc	0.292052	0.909724	0.845883
workclass_Self-emp-not-inc	0.292200	0.955807	0.844754
workclass_State-gov	0.289009	0.918453	0.846680
workclass_Without-pay	0.290969	0.931889	0.846149
status_Divorced	0.292124	0.947397	0.844622
status_Married-AF-spouse	0.288007	0.941343	0.846481
status_Married-civ-spouse	0.340786	0.961892	0.822908
status_Married-spouse-absent	0.289519	0.920082	0.846481
status_Never-married	0.292350	0.959402	0.837915
status_Separated	0.291166	0.943218	0.846680
status_Widowed	0.292060	0.937184	0.845883
occupation_Adm-clerical	0.284291	0.933475	0.845551
occupation_Armed-Forces	0.290860	0.932206	0.846348
occupation_Craft-repair	0.294170	0.905163	0.846614
occupation_Exec-managerial	0.294004	0.935666	0.837782
occupation_Farming-fishing	0.287110	0.930303	0.844887
occupation_Handlers-cleaners	0.286809	0.929185	0.846614
occupation_Machine-op-inspct	0.290958	0.929352	0.845750
occupation_Other-service	0.294591	0.932281	0.844622
occupation_Priv-house-serv	0.289194	0.929352	0.846680
occupation_Prof-specialty	0.271962	1.090108	0.841102
occupation_Protective-serv	0.294604	0.910036	0.847278
occupation_Sales	0.300185	0.906918	0.844622
occupation_Tech-support	0.289366	0.921105	0.845485
occupation_Transport-moving	0.289656	0.924799	0.846414
relationship_Husband	0.278506	0.929053	0.846016
relationship_Not-in-family	0.286107	0.914921	0.845817
relationship_Other-relative	0.288620	0.925863	0.846082
relationship_Own-child	0.291841	0.921025	0.843094
relationship_Unmarried	0.288579	0.927943	0.846282
relationship_Wife	0.124364	0.962568	0.839442

impact for gender decreased from 0.2909 to 0.2593, and for race, it went from 0.5653 to 0.5434. This means that the model became less fair in its predictions for both gender and race. Figure 4.8 shows the model comparison from this experiment.

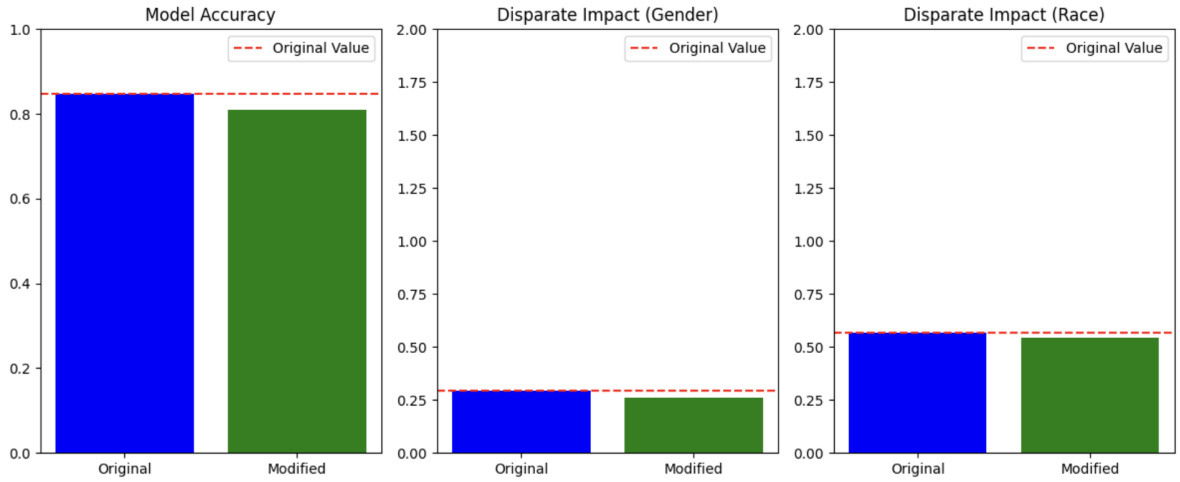

Figure 4.8: Model comparison after resampling.

While numerous studies in the bias mitigation field employ techniques like reweighting or adversarial training, our approach of targeted feature modification offers a more focused strategy. By zeroing in on specific high-impact features, we strive to maintain the model's predictive prowess while promoting fairness. Our findings are in line with those observed in certain contemporary studies, which also highlight the intricacies of balancing model fairness and accuracy.

Having explored targeted debiasing through feature modification, our next experiment delved into a newer approach to bias mitigation, distribution mapping.

Conclusions Experiment 3:

- The modification of the features led to a decrease in model accuracy. The original accuracy was around 0.8468, and after the modifications, it dropped to 0.8106. This suggests that the features we chose to modify ("gender", "relationship_Wife", "occupation_Prof-specialty", "race") have an influence on the model's ability to make

accurate predictions. By down-weighting their significance, the model's performance was adversely affected.

- The experiment did not yield the intended results. By down-weighting the features, we managed to make the model less accurate as well as slightly more biased.
- The experiment highlights the intricate trade-off between accuracy and fairness. While attempts to reduce bias through feature modification can lead to improvements in fairness metrics, they might come at the cost of reduced accuracy. This emphasizes the need for a balanced approach where both predictive performance and fairness are given due consideration.

4.3.4 Experiment 4: Distribution Mapping for Bias Mitigation

The intent behind this experiment was to investigate whether aligning the distribution of minority or underrepresented groups with that of the majority or favourably treated groups could reduce the biases in the model's predictions. By ensuring that the data distribution of potentially biased features mirrors that of a more favourable group, we hypothesized that the model might be less inclined to exhibit biases against the minority group.

Our exploration of distribution mapping provided interesting insights. Modifying the distribution, especially for attributes such as race and gender, minorly influenced our model's performance. The changes in prediction accuracy and fairness metrics, which can be seen in Figure 4.9 and Figure 4.10, highlighted the effectiveness of distribution mapping in reducing bias. Comparing these results with the initial model before any adjustments further reinforced the positive impact of our method.

Conclusions for Experiment 4:

- By adjusting the distribution of certain features, especially race and gender, we directly influenced the model's decision-making process. This reiterates the importance of data representation and its influence on model outcomes.

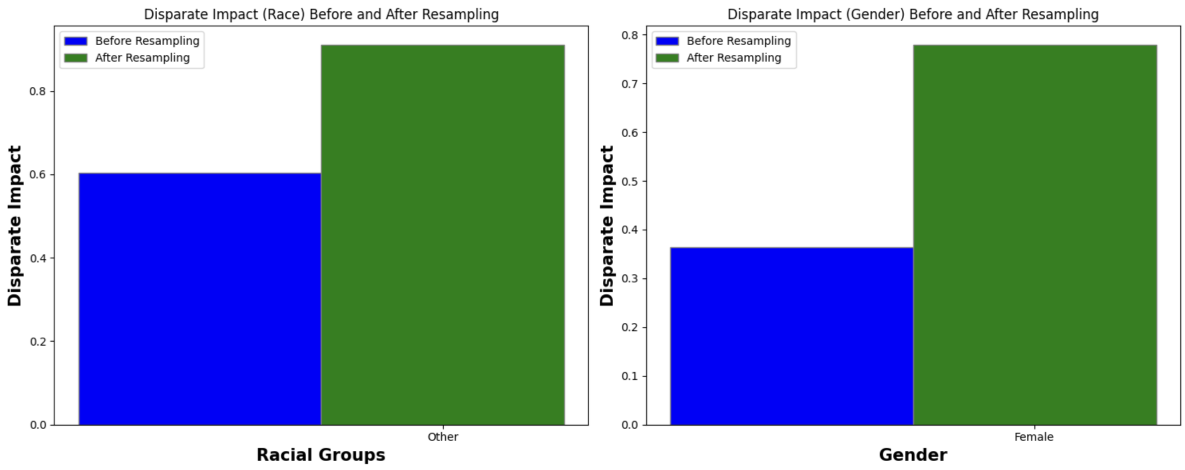

Figure 4.9: Disparate impact comparison before and after resampling.

- By aiming to replicate the distribution of a favourable group (white males, in this case), we attempted to reduce the disparities in the model's predictions across different groups, improving overall fairness for the groups.

- After resampling, the model's accuracy slightly decreased. This suggests that while we achieved improved fairness, there was a very minor trade-off in terms of model performance. In this particular case, the tradeoff can be neglected since accuracy did not drop in any significant way (Figure 4.10).

- By identifying and focusing on borderline cases—data instances that are close to the decision boundary—we ensured that the resampling process was more nuanced. This approach ensures that the data points most susceptible to changes in prediction due to resampling were prioritized.

- The disparate impact plots provided a clear before-and-after picture, showcasing the improvement in fairness after our resampling technique (Figure 4.9).

- Table 4.4 shows how our resampling procedure in this experiment successfully enhanced the fairness of the model concerning both race and gender attributes while maintaining the models' accuracy.

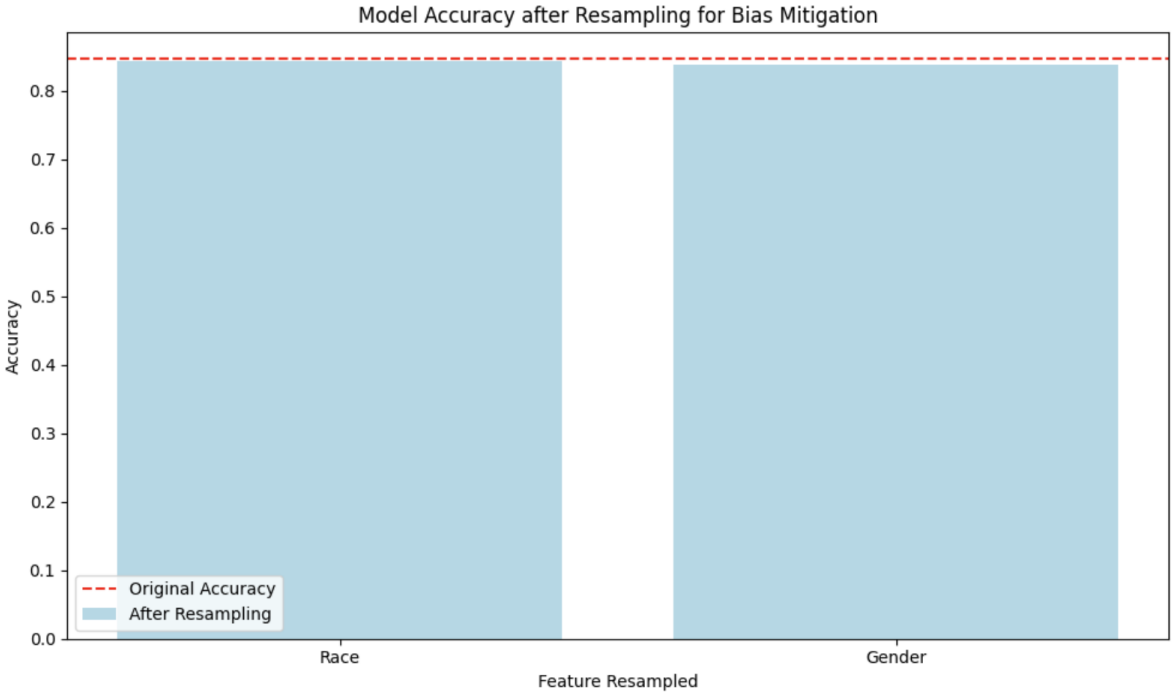

Figure 4.10: Model Accuracy after Resampling for Bias Mitigation.

- This fairness enhancement was achieved with only a minor reduction in prediction accuracy, moving from 0.846746 to 0.842961 (Figure 4.10)

Table 4.4: Results from Distribution Mapping for Bias Mitigation (exp. 4).

Metric	**Original**	**After Resampling**	**Change**
Prediction Accuracy	0.846746	0.842961	-0.003785
Disparate Impact (Race)	0.565368	0.929478	0.364111
Disparate Impact (Gender)	0.292622	0.761783	0.469162

4.4 Chapter Summary - Results

This chapter presented a detailed exposition of the outcomes derived from our experiments. Our findings were both revelatory and insightful, offering a comprehensive perspective on biases within Machine Learning models in the finance sector. We delved deep into the intricacies of biases prevalent in Machine Learning models, especially when applied to financial datasets. By methodically dissecting the influence of individual features, we were able to shed light on the factors contributing most to predictive disparities. Through a series of

experiments, we explored various techniques to understand and mitigate these biases, all while aiming to maintain the model's predictive accuracy.

Our strategy aimed to reveal the subtleties of biases within the Census Income dataset and offer practical solutions for their mitigation. Each step, from data preprocessing to advanced analyses, was intricately tied to our objectives.

The early stages of data preparation made sure that the dataset was clean and consistent, providing the groundwork for a meaningful analysis. We used tools like SHAP values and permutation feature importance to dive deep into the features driving model decisions (Figure 4.5 and Figure 4.7). This deep dive into the dataset unveiled significant biases, particularly associated with marital status, race, and gender attributes. These resources were essential to attaining our goal of understanding the drivers of bias. Disparate impact ratios were indicative of these biases, as seen in Figure 4.1 and Table 4.1. By highlighting the significant features, they paved the way for targeted debiasing strategies (Table 4.3).

The resampling techniques, especially when oriented towards the most privileged group's distribution, exhibited significant adjustments in biases. We aimed to modify the data distribution to reduce biases without compromising the predictive power of the influential features. Comparisons between pre-resampling and post-resampling models showcased the efficacy of our approach, both in terms of accuracy and fairness metrics; this can be seen in Figure 4.8. When examining the accuracy-fairness trade-off, it became evident that while the model's fairness improved, especially regarding gender bias, there was a minimal compromise on accuracy.

Our approach to distribution mapping, informed by feature importance insights, aimed to introduce new ways to reduce bias. Traditional methods often face a tension between accuracy and bias. Our method sought to balance this by adjusting the distribution of certain data points, especially those close to decision boundaries. This way, we hoped to

keep the useful predictive aspects of features while limiting their biased impact. The balance between fairness and accuracy can be seen in Figure 4.9 and Figure 4.10, highlighting the challenges and successes of this approach. We observed significant improvements in fairness metrics. Specifically, the disparate impact values for both race and gender became much closer to the ideal value of 1. This fairness enhancement was achieved with only a minor reduction in prediction accuracy, moving from 0.846746 to 0.842961, as shown in Figure 4.10.

These steps gave us a comprehensive understanding of biases in this dataset and how to tackle them. Our results highlighted not just the progress we made but also the challenges we face when it comes to bias. Biases, influenced by societal norms, are complex issues in data. Our work adds to the growing conversation on fairness in AI, focusing on the interplay between accuracy, fairness, and real-world outcomes.

In conclusion, our experiments, especially the resampling methodology in Experiment 4, demonstrate that it's possible to achieve a more balanced model in terms of fairness without significantly sacrificing accuracy.

Chapter 5

This chapter delves into the interpretations, implications, and broader significance of our findings. Furthermore, it contemplates the real-world implications of our research, offering insights into the future of fairness in Machine Learning.

5 Discussion

Navigating the complex world of biases in Machine Learning, particularly in finance, has been an enlightening experience. In this research, we delved deep into the intricacies of biases prevalent in Machine Learning models, especially when applied to financial datasets. By methodically dissecting the influence of individual features, we were able to shed light on the factors contributing most to predictive disparities. Through a series of experiments, we explored various techniques to understand and mitigate these biases, all while aiming to maintain the model's predictive accuracy.

5.1 Key Findings

Early in our research, we identified significant biases in the data. Some of these were tied to race and gender. These disparities showed us that even data that appears neutral on the surface can be biased.

Our initial findings Figure 4.1 revealed pronounced biases, especially concerning race and gender attributes. These biases were not just statistical anomalies; they represented genuine inequities in the model's predictions. Further, through a blend of SHAP values and permutation feature importance techniques, we identified the features that drove these biases. Features like age, education level, and working hours were pivotal for predictions. But we couldn't ignore the weight of attributes like marital status, race, and gender. The undue influence of race and gender highlighted the need for bias mitigation.

The highlight of our research was the results from Experiment 4. After implementing a resampling technique aimed at replicating the distribution of the most privileged groups (whites for race and males for gender), we observed significant improvements in fairness metrics Figure 4.9. Specifically, the disparate impact values for both race and gender became much closer to the ideal value of 1, indicating a more balanced positive outcome ratio between the privileged and unprivileged groups. This fairness enhancement was achieved with only a minor reduction in prediction accuracy, moving from 0.846746 to 0.842961 (Figure 4.10). The race-based disparate impact, in particular, jumped from 0.565367 to a commendable 0.929478, underscoring the efficacy of our approach. These results can be seen in Table 4.4.

In conclusion, our experiments, especially the resampling methodology in Experiment 4, demonstrate that it's possible to achieve a more balanced model in terms of fairness without significantly sacrificing accuracy. These findings are vital for the responsible application of AI in sectors like finance, where predictions can profoundly impact individuals' lives.

5.2 Objectives Revisited

The central objective of this research was to attenuate the bias induced by certain features in Machine Learning models while simultaneously preserving their predictive power. We did identify biases and implemented several strategies to address them. Our findings suggest that we have made strides toward our goal of bias reduction without undermining model accuracy. By concentrating on feature importance, we not only discerned the features crucial for accurate predictions but also identified the leading contributors to bias. While we recognize that our approach may not serve as a universal solution for every scenario, we firmly believe it represents a significant and positive step in the right direction.

Relevance and Weaknesses

The quest for unbiased Machine Learning models, particularly in sectors like finance where decisions can have profound real-world implications, is of great importance. Our research

adds to the growing body of work underscoring the need for fairness in AI. However, some weaknesses in our approach warrant mention:

- Data Manipulation: Resampling, though effective, involves manipulating the training data, potentially leading to models that might not reflect real-world distributions.

- Untested Scenarios: Our experiments primarily revolved around logistic regression. Exploring other algorithms, architectures, or ensemble methods might yield different insights.

- Single Dataset Limitation: Our findings, though comprehensive, are based on a singular dataset. The generalizability of these results to other datasets or real-world scenarios remains an open question for future research.

Within the rapidly advancing field of Artificial Intelligence and Machine Learning, achieving fairness goes beyond technical hurdles; it's a profound ethical duty. Our study represents only one step in the long journey to genuinely impartial algorithms. By integrating feature importance insights with targeted debiasing techniques, we've got a glimpse of the complex interplay of fairness and accuracy.

Chapter 6

This final chapter offers a reflective synthesis of the entire study. We reiterate our primary objectives, summarize our key findings, and ponder the road ahead, both in terms of potential applications and further research avenues.

6 Conclusions

In these coming years, where we see ML embedded into every sector, it is important to focus on 'doing no harm'. The Census Income dataset and its biases serve as a reminder of the challenges that we face when it comes to deploying responsible AI. During this research, we sought to explore new strategies for bias mitigation. Our primary aim was to identify, quantify, and mitigate biases in models, especially those arising from sensitive attributes like race and gender.

Throughout our experiments, we were able to pinpoint features that significantly influenced the model's decisions. Our targeted approach to modify specific features and resample data based on desired distributions demonstrated that it is possible to maintain a model's accuracy while reducing its bias (Figure 4.9 and Figure 4.10). While this does not eliminate biases entirely, it does take a step towards a more equitable model.

Our experiments revealed that biases are not merely a consequence of model architectures but are deeply embedded in data. Addressing them requires more than algorithmic tweaks; it demands a fundamental rethinking of how data represents real-world phenomena.

Main Takeaways:

1. Biases are often intertwined within the data.
2. Tools like SHAP, disparate impact, and permutation feature importance can be very powerful tools to improve model fairness by identifying drivers of bias.

3. The tug-of-war between accuracy and fairness was a concern throughout our experiments. Achieving one often comes at the expense of the other. For this particular case, we were able to reduce bias, having an almost negligible effect on the accuracy of our model.
4. Our approach, particularly in the resampling strategy in experiment 4, provided evidence that models can be trained to be both accurate and fair. The minor reduction in accuracy was a small price to pay for the substantial gains in fairness, especially in areas like finance, where decisions can have profound socio-economic impacts.
5. Targeted debiasing techniques, particularly resampling feature distributions, showcased great promise.
6. Our emphasis on data points near the decision boundaries through Decision Boundary Analysis was a pioneering step, showing a lot of promise for future endeavours in bias mitigation.

It's also worth noting that while our research has provided valuable insights and solutions, it's not the ultimate answer to bias in Machine Learning. The dynamics of bias are complex, influenced by myriad factors that extend beyond the dataset and model. As such, continuous research, exploration, and validation across different datasets and contexts are essential.

Responsible AI - The Greater Good:

As machines become integral decision-makers, the imperatives of responsibility, ethics, and fairness gain paramount importance. An algorithm's prediction, especially in sectors like finance, can alter life trajectories, bestow opportunities, or impose limitations. Thus, the quest for fairness is not just a scientific challenge but a societal one. Our experiments, though confined to a dataset, resonate with broader themes, serving as reminders of the duties AI practitioners hold. It is not just about creating efficient models but about sculpting algorithms that uphold the principles of justice, equity, and inclusivity.

Future Work:

- Broader Algorithmic Exploration: While our focus remained on logistic regression, the behaviours of more complex models, such as neural networks or ensemble techniques, under fairness constraints warrant investigation.
- Technique Robustness: In future endeavours, we aim to expand our investigation across a broader range of datasets. This will allow us to assess the wider applicability of our technique in striking a balance between bias reduction and accuracy, ensuring our approach's robustness in diverse scenarios.
- Interdisciplinary Collaboration: Engaging with sociologists, ethicists, or psychologists can offer richer, more holistic insights into biases, their origins, and their mitigation.
- Real-world Implementation and Feedback: Taking algorithms beyond datasets and into real-world applications, followed by feedback loops, can offer invaluable insights into the practicalities of fairness.
- Fairness Toolkits and Frameworks: Developing or refining toolkits that not only identify but also rectify biases in an automated or semi-automated manner can be instrumental for the AI community.

In summary, our research emphasizes the critical role of fairness in AI, urging for methodologies that prioritize equity without undermining performance. As we forge ahead in the AI era, let our findings serve as a reminder of the importance of responsible AI, where technological advancements are harmonized with ethical imperatives.

For our final remark, it is worth remembering that those of us well-versed in AI have both a privilege and a duty. We're not just wielding a tool for innovation; we're also navigating its vast societal effects. It's essential for us to champion responsible AI, focusing on ethical, transparent, and fair applications. As we shape the future of this technology, we must remember our role in setting standards and ensuring that AI serves as a force for good in society.

References

[1] K. D. A. M. Amitabha Mukerjee, Rita Biswas, "Multi–objective evolutionary algorithms for the risk–return trade–off in bank loan management." https://doi.org/10.1111/1475-3995.00375, December 2002. International Transactions in operational research 9, 5.

[2] A. R. Miranda Bogen, "Help wanted: an examination of hiring algorithms, equity.." https://www.upturn.org/static/reports/2018/hiring-algorithms/files/Upturn%20--%20Help%20Wanted%20-%20An%20Exploration%20of%20Hiring%20Algorithms,%20Equity%20and%20Bias.pdf, December 2018. Technical Report on bias. Upturn.

[3] Y. M. Lee Cohen, Zachary Lipton, "Efficient candidate screening under multiple tests and implications for fairness." https://doi.org/10.48550/arXiv.1905.11361, May 2019.

[4] S. M. Julia Angwin, Jeff Larson and L. Kirchner, "Machine bias: There's software used across the country to predict future criminals. and it's biased against blacks.." https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing, May 2016. Propublica.

[5] C. O'Neil, *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy.* Crown Publishing Group, 2016.

[6] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, "A survey on bias and fairness in machine learning," *ACM Comput. Surv.*, vol. 54, July 2021.

[7] B. Becker and R. Kohavi, "Adult." UCI Machine Learning Repository, 1996. https://doi.org/10.24432/C5XW20.

[8] S. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," 2017.

[9] R. Berk, H. Heidari, S. Jabbari, M. Kearns, and A. Roth, "Fairness in criminal justice risk assessments: The state of the art," *Sociological Methods & Research*, vol. 50, no. 1, pp. 3–44, 2021.

[10] A. Caliskan, J. J. Bryson, and A. Narayanan, "Semantics derived automatically from language corpora contain human-like biases," *Science*, vol. 356, no. 6334, pp. 183–186, 2017.

[11] G. Smith and I. Rustagi, "When good algorithms go sexist: Why and how to advance ai gender equity." https://doi.org/10.48558/A179-B138, March 2021. Stanford Social Innovation Review.

[12] J. Dastin, "Amazon scraps secret ai recruiting tool that showed bias against women." https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G, October 2018. Reuters.

[13] J. Buolamwini and T. Gebru, "Gender shades: Intersectional accuracy disparities in commercial gender classification." Conference on Fairness, Accountability, and Transparency, 2018. https://proceedings.mlr.press/v81/buolamwini18a/buolamwini18a.pdf.

[14] C. G. Marco Tulio Ribeiro, Sameer Singh, "Why should i trust you?: Explaining the predictions of any classifier." https://doi.org/10.48550/arXiv.1602.04938, August 2016.

[15] L. Sweeney, "Discrimination in online ad delivery: Google ads, black names and white names, racial discrimination, and click advertising." https://dl.acm.org/doi/10.1145/2460276.2460278, March 2013.

[16] J. W. Adrienne Yapo, "Ethical implications of bias in machine learning." http://hdl.handle.net/10125/50557, March 2018. Proceedings of the 51st Hawaii International Conference on System Sciences.

[17] S. F. Bryce Goodman, "European union regulations on algorithmic decision-making and a 'right to explanation'." https://doi.org/10.48550/arXiv.1606.08813, August 2016.

[18] N. S. Moritz Hardt, Eric Price, "Equality of opportunity in supervised learning." https://doi.org/10.48550/arXiv.1610.02413, October 2016.

[19] Z. Chen, J. M. Zhang, F. Sarro, and M. Harman, "A comprehensive empirical study of bias mitigation methods for machine learning classifiers," *ACM Trans. Softw. Eng. Methodol.*, vol. 32, May 2023.

[20] O. A. Osoba and W. W. IV., "An intelligence in our image: The risks of bias and errors in artificial intelligence.." , 2017. Rand Corporation.

[21] R. S. Zemel, L. Y. Wu, K. Swersky, T. Pitassi, and C. Dwork, "Learning fair representations," in *International Conference on Machine Learning*, 2013.

[22] S. Barocas and A. D. Selbst., "Big data's disparate impact.." http://www.jstor.org/stable/24758720, June 2016. California Law Review, vol. 104, no. 3, pp. 671–732.

[23] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel, "Fairness through awareness." https://doi.org/10.1145/2090236.2090255, 2012. Association for Computing Machinery.

[24] R. Zemel, Y. Wu, K. Swersky, T. Pitassi, and C. Dwork, "Learning fair representations," in *Proceedings of the 30th International Conference on Machine Learning* (S. Dasgupta

and D. McAllester, eds.), vol. 28 of *Proceedings of Machine Learning Research*, (Atlanta, Georgia, USA), pp. 325–333, PMLR, 17–19 Jun 2013.

[25] S. Shalev-Shwartz and S. Ben-David, *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, 2014.

[26] N. S. K. L. A. G. Ninareh Mehrabi, Fred Morstatter, "A survey on bias and fairness in machine learning." `https://arxiv.org/abs/1908.09635`, August 2019. https://dl.acm.org/doi/pdf/10.1145/3457607.

[27] J. G. Harini Suresh, "A framework for understanding sources of harm throughout the machine learning life cycle." `https://arxiv.org/abs/1901.10002`, December 2021.

[28] B. K. Finale Doshi-Velez, "Towards a rigorous science of interpretable machine learning." `https://doi.org/10.48550/arXiv.1702.08608`, February 2017.

[29] A. Z. P. B. L. V. B. H. E. S. I. D. R. T. G. Margaret Mitchell, Simone Wu, "Model cards for model reporting." `https://doi.org/10.48550/arXiv.1810.03993`, October 2018.

[30] S. Biswas and H. Rajan, "Fair preprocessing: towards understanding compositional fairness of data transformers in machine learning pipeline," in *Proceedings of the 29th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering*, ACM, August 2021.

[31] M. Hort, J. M. Zhang, F. Sarro, and M. Harman, "Fairea: A model behaviour mutation approach to benchmarking bias mitigation methods," in *Proceedings of the 29th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering*, ESEC/FSE 2021, (New York, NY, USA), p. 994–1006, Association for Computing Machinery, 2021.

[32] W. Liu, X. Wang, H. Zheng, B. Jin, X. Wang, and H. Zha, "Mitigating disparate impact on model accuracy in differentially private learning," *Information Sciences*, vol. 616, pp. 108–126, 2022.

[33] M. Feldman, S. Friedler, J. Moeller, C. Scheidegger, and S. Venkatasubramanian, "Certifying and removing disparate impact," 2015.

[34] F. P. Calmon, D. Wei, B. Vinzamuri, K. N. Ramamurthy, and K. R. Varshney, "Optimized pre-processing for discrimination prevention," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, (Red Hook, NY, USA), p. 3995–4004, Curran Associates Inc., 2017.

[35] J. Chakraborty, S. Majumder, Z. Yu, and T. Menzies, "Fairway: a way to build fair ML software," in *Proceedings of the 28th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering*, ACM, nov 2020.

[36] Z. Chen, J. M. Zhang, F. Sarro, and M. Harman, "Maat: A novel ensemble approach to addressing fairness and performance bugs for machine learning software," ESEC/FSE 2022, (New York, NY, USA), p. 1122–1134, Association for Computing Machinery, 2022.

[37] M. L. Wick, S. Panda, and J.-B. Tristan, "Unlocking fairness: a trade-off revisited," in *Neural Information Processing Systems*, 2019.

[38] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," *CoRR*, vol. abs/1706.04599, 2017.

[39] L. Breiman, "Random forests." `https://www.stat.berkeley.edu/~breiman/randomforest2001.pdf`, January 2001.

[40] C. Molnar, *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable*. Mucbook Clubhouse, 2022.

[41] P. T. P. R. L. S. Adriano Koshiyama, Emre Kazim, "Towards algorithm auditing: A survey on managing legal, ethical and technological risks of ai, ml and associated algorithms." https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3778998, Feb 2021.

[42] I. D. Raji, A. Smart, R. N. White, M. Mitchell, T. Gebru, B. Hutchinson, J. Smith-Loud, D. Theron, and P. Barnes, "Closing the ai accountability gap: Defining an end-to-end framework for internal algorithmic auditing," in *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, (New York, NY, USA), p. 33–44, Association for Computing Machinery, 2020.

[43] K. K. C. L. Christian Sandvig, Kevin Hamilton, "Auditing algorithms: Research methods for detecting discrimination on internet platforms.." https://websites.umich.edu/~csandvig/research/Auditing%20Algorithms%20--%20Sandvig%20--%20ICA%202014%20Data%20and%20Discrimination%20Preconference.pdf, May 2014. Paper presented to Data and Discrimination: Converting Critical Concerns into Productive Inquiry, a preconference at the 64th Annual Meeting of the International Communication Association.

[44] C. T. S. G. Bogdan Kulynych, Rebekah Overdorf, "Pots: Protective optimization technologies." https://doi.org/10.48550/arXiv.1806.02711, June 2018.

[45] J. Z. V. S. A. K. Tolga Bolukbasi, Kai-Wei Chang, "Man is to computer programmer as woman is to homemaker? debiasing word embeddings." https://doi.org/10.48550/arXiv.1607.06520, July 2016.

[46] C. R. R. S. Matt Kusner, Joshua Loftus, "Counterfactual fairness." https://doi.org/10.48550/arXiv.1703.06856, March 2017.

[47] A. F. S. G. A. H. Sam Corbett-Davies, Emma Pierson, "Algorithmic decision making and the cost of fairness." https://arxiv.org/abs/1701.08230, January 2017.

[48] A. Fisher, C. Rudin, and F. Dominici, "All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously," 2019.