Emprical Analysis

Table of Contents

# Discrimination Mitigation

## Introduction

* machine learning models and wide data more prevalent.
* application to personnel
* While nothing new, brings risks of discrimination, unfairness, and other potential ethical problems to the fore
* In this white paper, we discuss our implementation of an algorithm for mitigating unintended discrimination in predictive models. Motivated by the Frisch-Waugh-Lovell theorem, this algorithm can be generically applied to already-trained machine learning models – so long as these models were trained using protected class variables against which the user desires to mitigate discrimination.

### Literature Review

#### ‘fairness’ criteria

The machine-learning ‘fairness’ literature is wide; a large number of fairness criteria have been proposed, but there has been little follow-up on particular methods.

The fairness criteria have many shortcomings: 1. Generically, different fairness criteria cannot be satisfied simultaneously. 2. Fairness criteria often conflate prediction with decision rules, which tend to preclude their application to frameworks in which the decision rules are contingent or unknown.  
3. Algorithms intended to satisfy fairness criteria often require special transformation of the data used to train the model or require customization of the estimation procedure, such as through the loss function.  
Consquently, such fairness criteria cannot be implemented in a model that has already been trained.  
4. Many algorithms intended to satisfy fairness criteria are limited to particular

Below we discuss several fairness criteria with the aim of illustrating the above problems. This discussion is not comprehensive; instead we would point readers to \_\_\_\_\_.

[discussion of several fairness criteria ]

The methodology we propose address statistical discrimination, which we believe to be the most pernicious and also the most rectifiable risk of machine learning models. In essence,

#### statistical discrmination

Classically, economists consider two types of discrmination: taste-based disrimination and statistical discrmination. [cite]

definition taste based

definition statistical

* why this distinction is relevant
* a user without animus might nevertheless engage in statistical discrimination because doing so is beneficial to his objective.
* a model that predicts an outcome relevant to a decision maker however,
* taste-based discrimination is not totally irrelevant, however. That is, taste-based discrimination can generate bias in the target or covariates used to train the model. For example, promotion decisions might be biased against,

classically, statistical discrimination occurs when a decisionmaker overtly uses the predictive content of protected class.

However, excluding protected class from a prediction model is not sufficient to prevent it from discriminating.

### Proxy Discrimination and Frisch-Waugh-Lovell

[discussion written in lyx]

## discussion of algorithm with illustrations

### base case

### consequence - downweighting of proxy predictors

### shortomings

good faith in the decision rule will still be important this methodology does not insulate predictions from the consequences of structural discrimination.

# List of Figures

# List of Tables

# References