Investigating discrimination bias in predictive modelling

Jonathan Rittmo Sara Thiringer

2020-10-22

Abstract

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Curabitur tempor quis tellus in convallis. Etiam mattis est laoreet mi facilisis fringilla. Curabitur sed tortor blandit, consectetur sem quis, posuere urna. Donec at sem turpis. Ut nec lacus vitae nisl pretium varius non a sem. Cras venenatis rhoncus facilisis. Curabitur viverra leo eu dui varius, ac hendrerit tellus maximus. Duis nec ipsum feugiat, egestas tortor et, elementum felis. Donec sagittis nec ante quis convallis. Aenean volutpat sem nec gravida ullamcorper. Ut vestibulum elementum sem. Morbi a malesuada ipsum. Nam placerat neque diam, vel lobortis diam dictum sed. Phasellus id urna ligula. Morbi mattis ex vel posuere mattis. Sed viverra felis et suscipit tempor. Ut at ante vitae est tempus sodales. Fusce condimentum in erat id dignissim. Nam hendrerit quis lorem quis dignissim. Mauris quis arcu accumsan, tincidunt nunc id, scelerisque sapien.

Introduction

As more data and consequently more data-driven decisions have entered the world, the problem with algorithms reinforcing discriminatory structures have received an increasing amount of attention. Several cases showcase how algorithms that were designed to be neutral decision-makers have made discriminatory predictions. Examples include facial recognition being less accurate for people with darker skin [@buolamwini_gender_2018], ads for higher paid jobs being shown more frequently to men [@datta_automated_2015], healthcare predictions underestimating the illness of black people [@obermeyer_dissecting_2019] as well as individual tech company scandals such as the Apple card seemingly granting men a higher credit limit than women and Amazon's automated recruitment tool unrighteously favoring men. However unintentional, these examples show the need for an awareness of discrimination and fairness when collecting data, training models and using predictions for decision-making.

In this project we have looked at methods for deadling with potential discrimination in models. First, we present x ways of measuring fairness in order to be able to evaluate it. In this part we also discuss different methods of dealing with bias. Secondly, we train several statistical models on a dataset were we know that discriminatory bias is present. We then use the R package *fairmodels* to evaluate how the model performed both in terms of accuracy and fairness. Lastly, we use a pre-processing bias mitigation method called impact disparate remover to reduce the presence of bias in the models. We evaluate the success of this tool and discuss it in relation to the different models.

For simplicity, this project focuses solely on classification methods. (maybe present more restrictions)

[Something on terminology... Maybe small appendix.]

Understanding Fairness in Statistical Models

In order to build models that are less discriminatory then previously mentioned examples have managed to be, we first need a way to measure fairness. In recent years, a lot of work have been put into defining fairness in a quantitive way. These different definitions, listed and explained below, implicate somewhat different views on what fairness really is. This ongoing discussion draws on previous social and legal research on equality and fairness.

Quantifying Fairness

Optimising for Fairness

Pre-Processing

Training

Prediction

Classification on the COMPAS Data

Models

We trained five different models on the compas data. [Kanske typ en table?]

Evaluation of Bias in Models

Disparate Impact Remover

Final Discussion