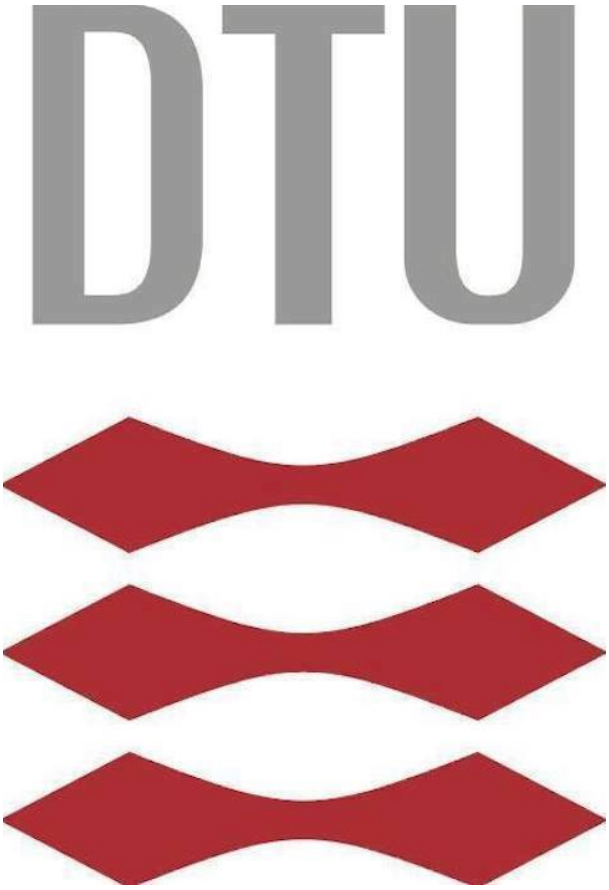
02446 - Project in Artificial Intelligence and Data Science

Fairness in classification



Rasmus Stokholm Bryld, s183898
Matilde Maria De Place, s183960
Sunniva Olsrud Punsvik, s183924

Contents

1 Introduction

Fairness and bias are relatively new concepts in the realm of computer science, and over the last decade, the quality and accuracy of different AI models (especially machine and deep learning) have increased significantly. The outcome of this success has led us to introduce neural networks as part of various decision making tasks. In the U.S. there has been a rise in the use of software for risk assessments to help in court and every stage of the justice process[1], which has received heaps of criticism because this software, which is used in several states already, allegedly is biased against African Americans.

This project will investigate the meaning of bias in data and algorithms, how to detect and examine them, and how biases contribute to risk assessment scores. This project will specifically delve into the Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) dataset, to see if earlier findings of bias are reproducible. Bias in the data will be searched for using permutation test and thereafter implementing a classifier trained on the COMPAS dataset by following the method used in [2], which introduces an equality of opportunity classifier to maximise the accuracy of the predictions, hereby correcting bias in a classifier. Then by judging the importance of different features and the direction they influence the decile score, this will be done by considering some selected data, that is, looking into all defendants who committed misdemeanours and comparing their respective risk assessment scores with their race. By inspecting whether they re-offended or not, one can look for which combination of features are mostly responsible for the decile score (also known as feature selection). Lastly the report will consider the ethical aspects of using such technology and ask the ethical question of why we sentence people to jail and when this should be done.Der vil komme mere præcise forklaringer til hvad vi gør, når vi er helt sikre på det

1.1 State-of-the-art and Recent Studies

The-state-of-the-art in this domain is quite interesting, some papers on fairness in classification often work with the notion of parity; equality of outcome for protected groups i.e. the notion of parity requires the outcome of given classifier to be independent of the protected attribute. However, correcting a classifier in order to achieve for example demographic parity [2] brings about a significant disadvantage, namely accuracy. Additionally, this method raises an important ethical question: which attributes should be treated as protected and which requirements must a classifier obey to ensure fairness? Another state-of-the-art paper which combined computer science, criminology, and statistics concluded that in most non-trivial cases, it is impossible to maximise accuracy simultaneously with fairness because one needs to consider challenging trade-offs[3]. Due to the fact that these are new concepts, the consensus is not apparent.

Recent studies have vast different definitions of fairness, as well as approaches that is different from parity. One approach proposed by Dwork, Hard, Pitassi, Reingold and Zemel in [4]. Their definition of fairness: *"We capture fairness by the principle that any two individuals who are similar with respect to a particular task should be classified similarly"* allows a focus on each individual. They apply statistical parity [1], however, only as a consequence of their concept on fairness constraints, a consequence they actively try to achieve. Another study, by Barrioa, Gamboab, Gordalizaa and Loubes [5] defines fairness according to Disparate Impact (DI) and Balanced Error Rate (BER) [2] The study aspire to correct bias in data by modifying values of

[1]Definition of statistical parity according to [4]: *Statistical parity is the property that the demographics of those receiving positive (or negative) classifications are identical to the demographics of the population as a whole.* Additionally, statistic parity is, by [4] referred to as "group fairness"

[2]Definition of DI according to [6]: *"Disparate Impact is a metric to evaluate fairness. It compares the proportion of individuals that receive a positive output for two groups: an unprivileged group and a privileged group."*

the attributes and hence ensure that predictions of the target do not depend on the defined protected attribute. The study argues, from at practical view, that e.g. companies will find that blurring data rather than changing the model is more applicable. The study aim to obtain statistical parity by the means of modifying data.
Der kommer mere til recent studies her

2 Data Description

This project takes advantages of two, out of several, data sets obtained by ProPublica: `compas-scores.csv` and `compas-scores-two-years.csv`[7]. `compas-scores.csv` consists of 47 attributes and 11,757 observations, all with COMPAS scores accessed at pretrial. `compas-scores-two-years.csv` is a subset of `compas-scores.csv`, consisting of 7,214 observations 53 attributes. One of the additional attributes being a binary variable with information of two years recidivism. Hereto, one shall know Northpoints definition of recidivism as well as the interpretation made by Propublica: *"finger-printable arrest involving a charge and a filing for any uniform crime reporting (UCR) code. We interpreted that to mean a criminal offence that resulted in a jail booking and took place after the crime for which the person was COMPAS scored."*[7].

2.1 How the Data is Acquired

Before diving into a description of the remaining attributes, an elaboration of ProPublicas data aggregation will be stated. From Broward County Sheriff's Office in Florida, ProPulica received and downloaded COMPAS scores of 18,610 people obtained in the period 2013 to 2014, as well as jail records from January 2013 o April 2016. They only kept the 11,757 COMPAS scores assessed at pretrial stage. From Broward County Clerk's Office website, April 1, 2016, ProPublica obtained 80,000 public criminal records and from the Florida Department of Corrections website, Propublica downloaded incarceration records aw well. In order to build each relevant profile, Propublica matched COMPAS scores with criminal history using date of birth and first and last name. A persons race was defined according to Broward County Sheriffs Office classification[7].

2.2 Description of Attributes

This section introduces attributes if the aforementioned data sets `compas-scores.csv` and `compas-scores-two-years.csv` respectively. For simplicity, all attributes are divided into 5 categories. Hereto one shall notice that `compas-scores.csv` only contains four of the five categories. This report will refer to the categories as their names.

1. *Personal information of defendant*
2. *Case of which COMPAS scores are assessed*
3. *Recidivism*: Information of whether or not recidivism occurs. Remaining attributes have content if recidivism occurs and NaNs if not.
4. *Violent recidivism*: Information of whether or not violent recidivism occurs. Remaining attributes have content if recidivism occurs and NaNs if not.
5. *2 years recidivism*: Whether or not recidivism occurs within 2 years.

Once again for simplicity, attributes of *Recidivism* and *Violent recidivism* with definition equivalent a attributes of *Case of which COMPAS scores are assessed* are not included in the respective table but listed below with their equivalent attribute. One ought to notice that attributes staring with r_ belongs to *Recidivism* and v_ and vr_ belongs to *Violent recidivism.*

- r_case_number and vr_case_number are equivalent to c_case_number
- r_days_from_arrest equivalent to c_days_from_arrest
- r_offense_date and vr_offense_date equivalent to c_offense_date
- r_jail_in equivalent to c_jail_in
- r_jail_out equivalent to c_jail_out
- r_charge_desc and vr_charge_desc are equivalent to c_charge_desc

Before moving on to an elaboration of each attribute, one shall notice that the data sets of interest consists attributes with unclear definition. As a result, most definitions in table 1 are found by merging the knowledge gained by interpreting the data sets and reading [8] and [7]. Hence definitions in 1 shall be viewed as defined by this report. In addition, the unclear definitions is why the attributes are describes in details.

Attributes in category *"Personal information of defendant"*

Attribute	Description of attribute
name, first, last	Name of defendant (full and divided intro first and last name)
sex	Male or Female
age	Age of defendant in years
age_cat	'Greater than 45', '25 - 45', 'Less than 25'
race	'Other', 'African-American', 'Caucasian', 'Hispanic', 'Native American', 'Asian'
dob	Date of birth of defendant (year-month-date)

Attributes in category *"Case of which COMPAS scores are assessed"*

days_b_screening_arrest	Days between c_jail_in and (compas_screening_date) (integer)
c_jail_in	Jail time begins (year-month-date, hour:minute:second)
c_jail_out	Jail time begins (year-month-date, hour:minute:second)
c_case_number	Case number of case of relevant case
c_offense_date / c_arrest_date	Offence/arrest date of relevant case (only one completed)
c_days_from_compas	Days between compas_screening_date and c_offense_date or c_arrest_date
c_charge_degree	Classification of relevant crime (misdemeanor or felony)[9]
c_charge_desc	Describtion of crime

Attributes in category *"Recidivism"*

is_recid	Binary, 1 if recidivism occurs 0 otherwise
r_charge_degree	Possible values: (F3), (M1), (F2), (M2), (MO3), (F1), (F6),(F7), (CO3), (F5) [3]
decile_score.1 and decile_score	COMPAS score "Risk of Recidivism", integers from 1-10
score_text	Categorial "Risk of Recidivism". Low: 1-4 Medium: 5-7 High: 8-10
screening_date	Date of which "Risk of Recidivism" is assessed

Attributes in category *"Violent recidivism"*

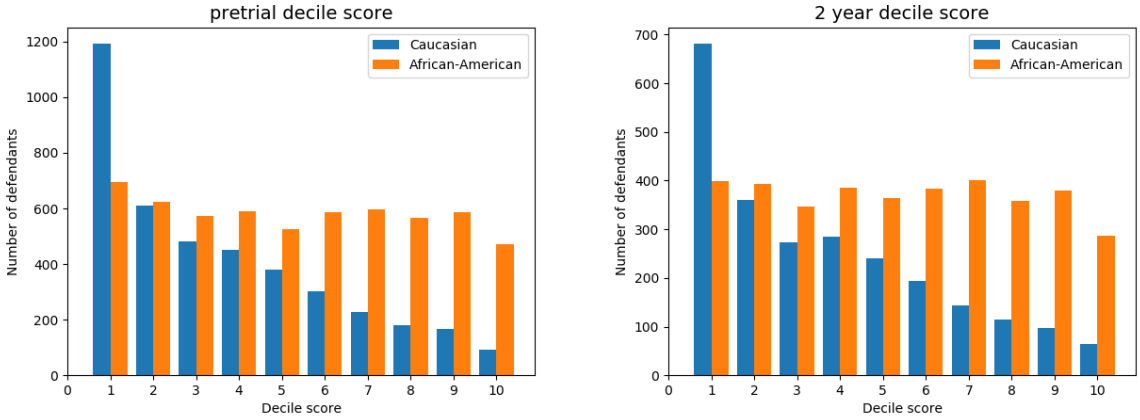
is_violent_recid	Binary, 1 if violent recidivism occurs 0 otherwise
vr_charge_degree	Possible values: (F3), (F2), (F1), (M1), (MO3), (M2), (F6), (F7), (F5) [4]
v_decile_score	COMPAS score "Risk of Violence", integers from 1-10
v_score_text	Categorial "Risk of Violence". Low: 1-4 Medium: 5-7 High: 8-10
v_screening_date	Date of which "Risk of Recidivism" is assessed

Attributes in category *"2 years recidivism"*

two_years_recid	Binary, 1 if recidivism occurs within two years, 0 otherwise

Table 1

2.3 Distribution



(a) (b)

Figure 1

3 Method

The first point in question is to comprehend the concept of bias in data and algorithms, to search for these biases in the COMPAS dataset and precisely what makes the algorithm biased. One ought to make a distinct definition of what bias entails in this context, to have a clear-cut interpretation surrounding bias and fairness for consistency reasons, as well as it is important to distinguish that bias in data may not be the same as bias in a classifier. The literature on this is quite foggy and some papers have different standards as to what bias means, where one typically designates a mathematical definition when it comes to classifiers, however, first one needs to tackle what bias means in the domain of data analysis. Second point in question is how to detect bias and consequently reduce them.

3.1 Bias in Data

This paper will use the following definition for bias in data: *"Bias is taken to mean interference in the outcomes of research by predetermined ideas, prejudice or influence in a certain direction. Data can be biased but so can the people who analyse the data. When data is biased, we mean that the sample is not representative of the entire population"*[10]. To search for bias in data, one can do a permutation test, which is a re-sampling method. The idea behind this is to build a sampling distribution (without assuming) by re-sampling the observed data. Without replacement, the data is rearranged and thereby assigning different outcomes (from the actually observed outcomes) to different values. The null hypothesis H_0 states that there is no difference between recidivism scores between races. der skal skrives her hvad man opnår med en permutations test i denne kontekst (i dette projekt)

3.2 Bias in Algorithms

To obtain a definition of bias in classification, this paper will use the definition used in [2], which states two criterion what bias is in supervised learning (notice that the cited paper uses the terms "discriminatory" and "non-discriminatory" and this paper will use "biased" and "unbiased"). The two criterion suggested in [2] are *equalised odds and equal opportunity*. This method is a probabilistic approach to define bias, and quantifiable able to reduce them. The goal here is to predict risk assessment scores, based on the available features in the COMPAS dataset,

whilst assuring the prediction is as unbiased as possible with respect to a protected attribute, which this case in point is "race". A protected attribute is an issue of fairness and this will be described in the next section. Equalised odds is given as **definition 2.1**[2]:

Definition 2.1 (Equalized odds). We say that a predictor $\widehat{Y}$ satisfies *equalized odds* with respect to protected attribute A and outcome Y, if $\widehat{Y}$ and A are independent conditional on Y.

Figure 2: Equalised Odds Definition

and equal opportunity is given as **definition 2.2**[2]:

Definition 2.2 (Equal opportunity). We say that a binary predictor $\widehat{Y}$ satisfies *equal opportunity* with respect to A and Y if $\Pr\{\widehat{Y} = 1 \mid A = 0, Y = 1\} = \Pr\{\widehat{Y} = 1 \mid A = 1, Y = 1\}$.

Figure 3: Equal Opportunity Definition

Der skal uddybes en del om metoden her

3.3 The Concept of Fairness

3.4 Constructing a Classifier

The classifier will consist of a feedforward artificial neural network (ANN) and will be trained on the `compas-scores.csv` dataset with $decile_score.1$ as target value, which is an integer value from 1-10.

3.5 Correcting the classifier

3.6 Evaluating bias in classifiers

To investigate bias in the three classifiers (COMPAS, the standard ANN and the corrected ANN) three confusion matrices are created. According to ProPublica [1] COMPAS was biased against Black people in the amount of people who were given given a high risk assessment but did in fact not re-offend. This can be described mathematically in a confusion matrix (See figure...) where the bias will appear as Black people having a higher rate of false negatives.
False Positives (FP), False Negatives (FN), True Positives (TP) and True Negatives (TN)

4 Results

5 Discussion

5.1 Bias in Data

The problem which occurs when investigating bias in data, is how to search for it? By looking at the current demographics is a good staring point; the racial makeup of the U.S. population compared to the U.S.s' adult correctional population. The U.S. is renowned for having the largest incarcerated population in the world, accumulating 21% of the world's prisoners[11]. African Americans accounts for 12.7% of the U.S. population[12], but comprises 34% of the correctional population[11] (when adding Hispanics, the correctional populations accumulates to 54%).

6 Conclusion

References

[1] J. Angwin, L. Kirchner, J. Larson, and S. Mattu, "Machine bias." https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing, May 2016. Accessed on 20.02.2020.

[2] M. Hardt, E. Price, and N. Srebro, "Equality of opportunity in supervised learning," *NIPS 2016*, Oct 2016. Accessed on 14.03.2020.

[3] R. Berk, H. Heidari, S. Jabbari, M. Kearns, and A. Roth, "Fairness in criminal justice risk assessments: The state of the art," *SAGE Journals*, July 2018. Accessed on 14.03.2020.

[4] M. Hardt, C. Dwork, O. Pitassi, Toniann Reingold, and R. Zemel, "Fairness through awareness," Nov 2011. Accessed on 19.03.2020.

[5] E. del Barrioa, F. Gamboab, P. Gordalizaa, and J.-M. Loubes, "Obtaining fairness using optimal transport theory," July 2019. Accessed on 19.03.2020.

[6] S. Ronaghan, "Ai fairness — explanation of disparate impact remover." https://towardsdatascience.com/ai-fairness-explanation-of-disparate-impact-remover-ce0da59451f1 Apr 2019. Accessed on 19.03.2020.

[7] J. Angwin, L. Kirchner, J. Larson, and S. Mattu, "How we analyzed the compas recidivism algorithm." https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm, May 2016. Accessed on 14.03.2020.

[8] Missing, "Compas analysis.ipynb." https://github.com/propublica/compas-analysis/blob/master/Compas%20Analysis.ipynb, Missing Missing. Accessed on 14.03.2020.

[9] Missing, "Pennsylvania crime classification." https://www.davidcohenlawfirm.com/pennsylvania-crime-classification?fbclid=IwAR1tGFrdZOZT7qeBFw2X1jlonmyKJxs7w5d1bTo5FBUJe-pC2I0x-ext2YM, Missing Missing. Accessed on 14.03.2020.

[10] S. d. Vries, "5 types of bias in data analytics." https://cmotions.nl/en/5-typen-bias-data-analytics/, Nov 2017. Accessed on 14.03.2020.

[11] NAACP, "Criminal justice fact sheet." https://www.naacp.org/criminal-justice-fact-sheet/, 2019. Accessed on 14.03.2020.

[12] U. S. C. Bureau, "The united states census quick search." https://www.census.gov/quickfacts/fact/table/US/PST045218, 2019. Accessed on 14.03.2020.