Lecture 4
Security and Ethical Aspects of Data

Amaya Nogales Gómez
amaya.nogales-gomez@univ-cotedazur.fr

MSc Data Science & Artificial Intelligence
Université Côte d'Azur

October 25, 2021

Tentative Content

1. Introduction
   - Preliminaries and Machine Learning basics
   - Motivation for ethics in Artificial Intelligence
2. Sources of unfairness
   - Bias in data
   - Algorithmic unfairness: Examples
3. Fairness criteria
   - Types of discrimination
   - Definitions of fairness
   - Fairness analysis of datasets
4. Algorithms and methods for fair ML
   - Pre-processing methods
   - In-processing methods
   - Post-processing methods
   - Legal and policy perspectives

Notation

- Dataset (X, A, Y)
- X: non-protected features
- A: protected attribute
- $Y \in \{-1, +1\}$ label, class membership
- $\hat{Y} = f(X, A)$: prediction

Disclaimer: abuse of notation

- feature $\equiv$ attribute $\equiv$ variable $\equiv$ characteristic
- protected $\equiv$ sensitive

Disparate treatment

A decision making process suffers from disparate treatment if its decisions are partly based on the subject's sensitive attribute information.

Disparate impact

A decision making process suffers from disparate impact if its outcomes disproportionately hurt people with certain sensitive attribute values.

Disparate Impact or The $p\%$-rule

$$UNF_{DI} = \min\left(\frac{\mathcal{P}(\hat{Y}=+1|A=1)}{\mathcal{P}(\hat{Y}=+1|A=0)}, \frac{\mathcal{P}(\hat{Y}=+1|A=0)}{\mathcal{P}(\hat{Y}=+1|A=1)}\right) \times 100.$$

Fairness Through Unawareness

It implies that $f(X, A)$ does not use the value of A and is a legal requirement in many domains where processing sensitive information about individuals is forbidden in order to guarantee no disparate treatment. A predictor $\widehat{Y}$ satisfies unawareness if it does not use the protected attribute A, i.e.,

$$\mathcal{P}\left(\widehat{Y}|X, A\right) = \mathcal{P}\left(\widehat{Y}|X\right).$$

Demographic Parity or Statistical Parity

A stronger definition of fairness compared to unawareness is **demographic parity**. A predictor $\widehat{Y}$ satisfies demographic parity with respect to protected attribute A, if $\widehat{Y}$ is independent of A, i.e.,

$$\mathcal{P}\left(\widehat{Y}|A\right) = \mathcal{P}\left(\widehat{Y}\right).$$

or equivalently

$$\mathcal{P}\left(\widehat{Y}|A=0\right) = \mathcal{P}\left(\widehat{Y}|A=1\right).$$

The protected and unprotected groups should receive the same distribution of output values.

Demographic Parity

$$UNF_{DP} = |\mathcal{P}\left(\widehat{Y}=+1|A=0\right) - \mathcal{P}\left(\widehat{Y}=+1|A=1\right)|.$$

Equalized odds

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.

$$\mathcal{P}\left(\widehat{Y} = +1 | A = 0, Y = y\right) = \mathcal{P}\left(\widehat{Y} = +1 | A = 1, Y = y\right), \quad \forall y \in \{-1, +1\}.$$

The protected and unprotected groups should have equal true positive and false positive rates.

Equalized Odds

$$UNF_{EOdds} = |\mathcal{P}\left(\widehat{Y} = +1 | Y = +1, A = 0\right) - \mathcal{P}\left(\widehat{Y} = +1 | Y = +1, A = 1\right)|$$

$$+ |\mathcal{P}\left(\widehat{Y} = +1 | Y = -1, A = 0\right) - \mathcal{P}\left(\widehat{Y} = +1 | Y = -1, A = 1\right)|.$$

Equal Opportunity

We say that a binary predictor $\widehat{Y}$ satisfies **equal opportunity** with respect to A and Y if

$$\mathcal{P}\left(\widehat{Y} = +1 | A = 0, Y = +1\right) = \mathcal{P}\left(\widehat{Y} = +1 | A = 1, Y = +1\right)$$

The protected and unprotected groups should have equal true positive rate.

Equal opportunity

$$UNF_{EOpp} = |\mathcal{P}\left(\widehat{Y} = +1 | Y = +1, A = 0\right) - \mathcal{P}\left(\widehat{Y} = +1 | Y = +1, A = 1\right)|.$$

Predictive equality

We say that a binary predictor $\widehat{Y}$ satisfies **equal opportunity** with respect to A and Y if

$$\mathcal{P}\left(\widehat{Y} = +1 | A = 0, Y = -1\right) = \mathcal{P}\left(\widehat{Y} = +1 | A = 1, Y = -1\right)$$

The protected and unprotected groups should have equal false positive rate.

Predictive Equality

$$UNF_{PE} = |\mathcal{P}\left(\widehat{Y} = +1 | Y = -1, A = 0\right) - \mathcal{P}\left(\widehat{Y} = +1 | Y = -1, A = 1\right)|.$$

Predictive Parity

We say that a binary predictor $\widehat{Y}$ satisfies **predictive parity** with respect to A and Y if

$$\mathcal{P}\left(Y = +1 | \widehat{Y} = +1, A = 0\right) = \mathcal{P}\left(Y = +1 | \widehat{Y} = +1, A = 1\right)$$

Predictive parity requires the same positive predictive value (i.e., precision) in both groups.

Predictive parity

$$UNF_{PP} = |\mathcal{P}\left(Y = +1 | \widehat{Y} = +1, A = 0\right) - \mathcal{P}\left(Y = +1 | \widehat{Y} = +1, A = 1\right)|.$$

This lecture analyses the investigation by Propublica about a commercial tool made by Northpointe, Inc. to assess the criminal defendant's likelihood of becoming a recidivist.
All original materials for this study are publicly available:

- The original story: Machine Bias.
- How they analyzed the algorithm.
- A GitHub Repository.

`https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing`
`https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm`
`https://github.com/propublica/compas-analysis/`

The story

How two different lives interact with the justice system:

This real story helps us understand:

- **How** it affects individuals.
- **What** type of harms it inflicts to individuals.

Brisha Borden

On a spring afternoon in 2014, Brisha Borden was running late to pick up her god-sister from school when she spotted an unlocked kid's bicycle and a scooter. Borden and a friend grabbed the bike and scooter and tried to ride them down the street.
Just as the 18-year-old girls were realizing they were too big for the tiny conveyances - which belonged to a 6-year-old boy - a woman came running after them saying, "That's my kid's stuff". Borden and her friend immediately dropped the bike and scooter and walked away.
But it was too late - a neighbor who witnessed the heist had already called the police. Borden and her friend were arrested and charged with burglary and petty theft for the items, which were valued at a total of 80\$.

Vernon Prater

The previous summer, 41-year-old Vernon Prater was picked up for shoplifting $86.35\$$ worth of tools from a store.
Prater was the more seasoned criminal. He had already been convicted of armed robbery and attempted armed robbery, for which he served five years in prison, in addition to another armed robbery charge.
Borden had a record, too, but it was for misdemeanors committed when she was a juvenile.

When Borden and Prater were booked into jail a computer program gave a score predicting the likelihood of each committing a future crime.

When Borden and Prater were booked into jail a computer program gave a score predicting the likelihood of each committing a future crime.

Borden

high risk

Vernon

low risk

When Borden and Prater were booked into jail a computer program gave a score predicting the likelihood of each committing a future crime.

Borden	Vernon
high risk	low risk

Two years later, we know the computer got it **exactly backward**.

When Borden and Prater were booked into jail a computer program gave a score predicting the likelihood of each committing a future crime.

Borden
high risk
has not been charged with any new crimes

Vernon
low risk
is serving an eight-year prison term

Two years later, we know the computer got it **exactly backward**.

When Borden and Prater were booked into jail a computer program gave a score predicting the likelihood of each committing a future crime.

Borden
high risk
has not been charged with any new crimes

Vernon
low risk
is serving an eight-year prison term

Two years later, we know the computer got it **exactly backward**.

Borden is **black**
Prater is **white**

COMPAS algorithm: recividism prediction

Prediction Fails Differently for Black Defendants

	WHITE	AFRICAN AMERICAN
Labeled Higher Risk, But Didn't Re-Offend	23.5%	44.9%
Labeled Lower Risk, Yet Did Re-Offend	47.7%	28.0%

Overall, Northpointe's assessment tool correctly predicts recidivism 61 percent of the time. But blacks are almost twice as likely as whites to be labeled a higher risk but not actually re-offend. It makes the opposite mistake among whites: They are much more likely than blacks to be labeled lower risk but go on to commit other crimes. (Source: ProPublica analysis of data from Broward County, Fla.)

Were the risk scores reasonable?

In this case, COMPAS risk scores were **incorrect**.

What harms came from these incorrect decisions?

- Borden: the score influencing the judge's decisions for setting bail. The impact of COMPAS's poor decision resulted in the girl spending the night in jail.
- Other departments use COMPAS scores in trial and sentencing. Similar poor algorithmic decisions elsewhere may impact the amount of time spent in prison, future job prospects, and the right to vote.

Unfortunately...

What explains such a discrepancy in COMPAS scores?

The COMPAS algorithm is a **black box** proprietary algorithm that we can only indirectly investigate.

The COMPAS dataset

An acronym for Correctional Offender Management Profiling for Alternative Sanctions

- An assistive software used to predict recidivism risk.
- Helpful in ways that it provides scores from 1 (being lowest risk) to 10 (being highest risk).
- And a categorical feature: high risk, medium risk or low risk of recidivism.
- For simplicity: medium risk and high risk of recidivism vs. low risk of recidivism.
- The original input dataset used for prediction of recidivism contains 137 variables.
- Race is not an explicit feature considered by the model.

ProPublica

ProPublica obtained two years worth of COMPAS scores from the Broward County Sheriff's Office in Florida, through a public records request, the **Freedom of Information Act**.

- ProPublica is an independent, nonprofit newsroom that produces investigative journalism with moral force.

Aim

To expose abuses of power and betrayals of the public trust by government, business, and other institutions, using the moral force of investigative journalism to spur reform through the sustained spotlighting of wrongdoing.

`https://www.propublica.org/` `https://www.foia.gov/`

How they acquired the data

- Through the public records request, ProPublica obtained COMPAS scores and for all 18,610 people who were scored in 2013 and 2014.
- Three COMPAS scores: "Risk of Recidivism", "Risk of Violent recividism" and "Risk of Failure to Appear".
- Scores range from 1 to 10. Scores 1 to 4 were labeled by COMPAS as "Low"; 5 to 7 were labeled "Medium"; and 8 to 10 were labeled "High".
- Starting with the database of COMPAS scores, they built a profile of each person's criminal history, matching the criminal records to the COMPAS records.
- To determine race, they used the race classifications used officially, which identifies defendants as African-American, Caucasian, Hispanic, Asian and Native American.

The data

The data from which the risk-scores are derived come from a combination of answers to a 137 question survey and the defendant's criminal record. These variables include:

- Prior arrests and convictions
- Address of the defendant
- Whether the defendant a suspected gang member
- If the defendant's parents separated
- If friends/acquaintances of the defendant were ever arrested
- Whether drugs are available in the defendants neighborhood
- How often the defendant has moved residences
- The defendants high school GPA
- How much money the defendant has
- How often the defendant feels bored or sad

`https://www.documentcloud.org/documents/`

Risk Assessment

PERSON

Name:	Offender #:	DOB:
[redacted]	[redacted]	[redacted]

	Gender:	Marital Status:	Agency:
[redacted]	Male	Single	DAI

ASSESSMENT INFORMATION

Case Identifier:	Scale Set:	Screener:	Screening Date:
[redacted]	Wisconsin Core - Community Language	[redacted]	[redacted]

Current Charges

☐ Homicide	☑ Weapons	☑ Assault	☐ Arson
☐ Robbery	☐ Burglary	☐ Property/Larceny	☐ Fraud
☐ Drug Trafficking/Sales	☐ Drug Possession/Use	☐ DUI/OUIL	☑ Other
☐ Sex Offense with Force	☐ Sex Offense w/o Force		

1. Do any current offenses involve family violence?
   ☑ No ☐ Yes
2. Which offense category represents the most serious current offense?
   ☐ Misdemeanor ☐ Non-violent Felony ☑ Violent Felony
3. Was this person on probation or parole at the time of the current offense?
   ☑ Probation ☐ Parole ☐ Both ☐ Neither
4. Based on the screener's observations, is this person a suspected or admitted gang member?
   ☐ No ☑ Yes
5. Number of pending charges or holds?
   ☑ 0 ☐ 1 ☐ 2 ☐ 3 ☐ 4+
6. Is the current top charge felony property or fraud?
   ☑ No ☐ Yes

Criminal History

Exclude the current case for these questions.

7. How many times has this person been arrested before as an adult or juvenile (criminal arrests only)?
   5

- ProPublica also conducted public records research to determine which defendants re-offended in the two years following their COMPAS screening.
- They were able to follow up on approximately half the defendants.
- This dataset contains a field two_year_recid that is 1 if the defendant re-offended within two years of screening and 0 otherwise. Following the notation of the course, this represents the label y.
- We will concern ourselves with comparing the black and white populations, as in the article.
- Similarly, we will consider a COMPAS score of either 'Medium' or 'High' to be a prediction that the defendant will re-offend within two years.

Outcome

- The true outcome being modeled by COMPAS is whether a defendant will commit another crime upon early release from custody.
- This true outcome is unobservable and requires simplification in a number of ways:
  - A time-frame must be set for observing whether someone re-offends (Northpointe sets two years).
  - To be observed re-offending, the police must come into contact with, arrest, and charge the defendant.
  - There will be defendants in the training set that are incorrectly labeled as a 'non-re-offender', only because they committed a crime that was not pursued.
  - There will **likely be bias** in this mislabeling of re-offenders due to **studied** police behavior favoring white communities over black communities.

How the score is used in practice

- The COMPAS score is used at the pretrial detention, trial, sentencing, and parole steps of the justice system.
- Given that the risk-assessment is supposed to model 'likelihood of re-offending' of a certain type, the developers of COMPAS only recommend using the algorithm to judge decisions like early release with access to social services.
- However, once the score became available to the criminal justice system as a whole, it became used in wholly inappropriate ways.
- For example, whether someone may commit another crime in the future has no bearing on whether they did or did not commit the current crime in question.

- The score itself does **not** actually make the decision.
- It's another piece of information that judges and juries use when making more holistic decisions.
- However, as the output of this model is nothing more than an integer, it does not explain to the decision maker how to weight this information.

Basic statistical descriptive analysis

The first step in a statistical descriptive analysis of a sample of data: obtain **tables** or other visualization outputs that allow **summarize** and **order** the data, helping its posterior analysis.

- Let us consider a sample composed of n individuals, for which we will observe variable X, having n data: $x_1, x_2, ..., x_n$.
- Let $x_1, \ldots, x_k$ the k different **values** observed.

Absolute frequency

The frequency (or absolute frequency) of an event x_i is the number n_i of times the observation occurred in an experiment.

$$\sum_{i=1}^{k} n_i = n$$

Relative frequency

The relative frequency of x_i, denoted by f_i, is the proportion of occurrences observed for this event, i.e.,

$$f_i = \frac{n_i}{n}, \qquad 1 \leq i \leq k$$

$$\sum_{i=1}^{k} f_i = 1$$

Frequency distribution

A frequency distribution is a table (*frequency table*) or graph (bar plot or histogram) that displays the frequency of the events in a sample. Each entry in the table contains the frequency of the occurrences of values within a particular group or interval.

x_i	n_i	N_i	f_i	F_i
x_1	n_1	N_1	f_1	F_1
x_2	n_2	N_2	f_2	F_2
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_k	n_k	N_k	f_k	F_k
	n		1	

Example

The COMPAS score of recividism calculated for the 15 defendants on a given day was:

$$4,\ 3,\ 7,\ 5,\ 6,\ 4,\ 5,\ 4,\ 5,\ 6,\ 7,\ 7,\ 3,\ 4,\ 5$$

Example

The COMPAS score of recividism calculated for the 15 defendants on a given day was:

$$4,\ 3,\ 7,\ 5,\ 6,\ 4,\ 5,\ 4,\ 5,\ 6,\ 7,\ 7,\ 3,\ 4,\ 5$$

The frequency table for this sample is:

x_i	n_i	N_i	f_i	F_i
3	2	2	0.133	0.133
4	4	6	0.266	0.4
5	4	10	0.266	0.666
6	2	12	0.133	0.8
7	3	15	0.2	1
	15		1	

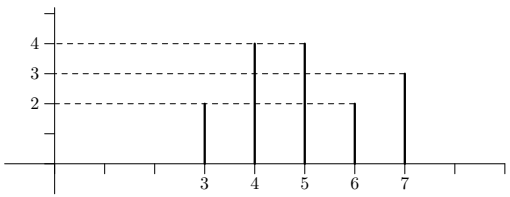
Graphical representations

- Variable
  - Discrete
    - Bar chart
    - Frequency polygon
    - Cumulative curve
  - Continuous
    - Histogram
    - Frequency polygon
    - Cumulative curve
- Attribute
  - Bar chart
  - Pie chart
  - Pictogram

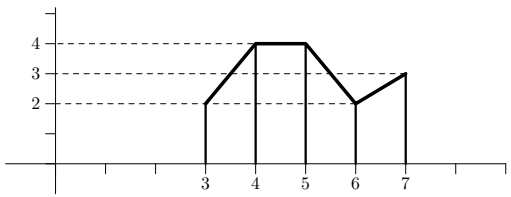
Example: discrete variable

Let us consider again the variable X="defendant's COMPAS risk score".

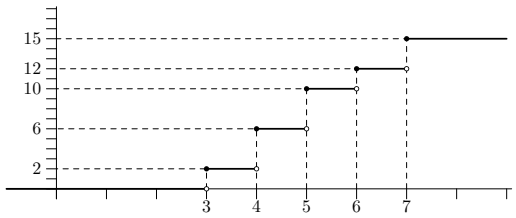
x_i	n_i	N_i	f_i	F_i
3	2	2	0.133	0.133
4	4	6	0.266	0.4
5	4	10	0.266	0.666
6	2	12	0.133	0.8
7	3	15	0.2	1
	15		1	



Bar chart



Frequency polygon

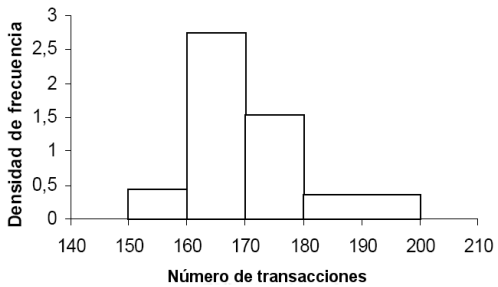


Cumulative curve

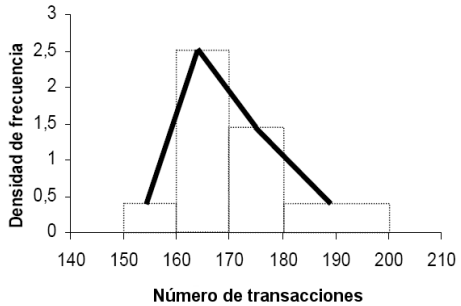
Example: continuous variable

Let us consider the variable X="Height in cm", observed in $n = 50$ defendants.

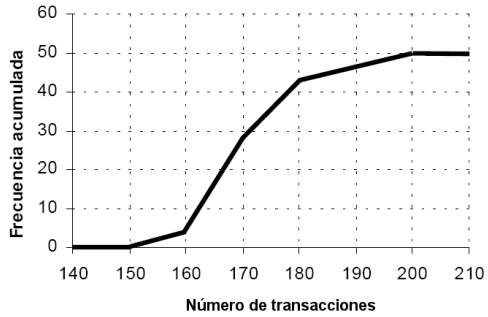
$(L_{i-1}, L_i]$	n_i	a_i	h_i	N_i
$(150, 160]$	4	10	0.4	4
$(160, 170]$	25	10	2.5	29
$(170, 180]$	14	10	1.4	43
$(180, 200]$	7	20	0.35	50



Histogram



Frequency polygon

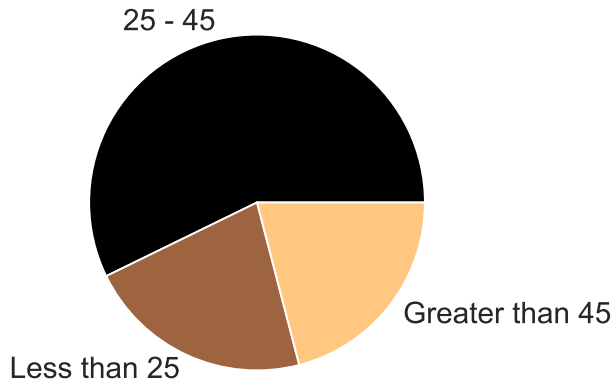


Cumulative curve

Pie chart

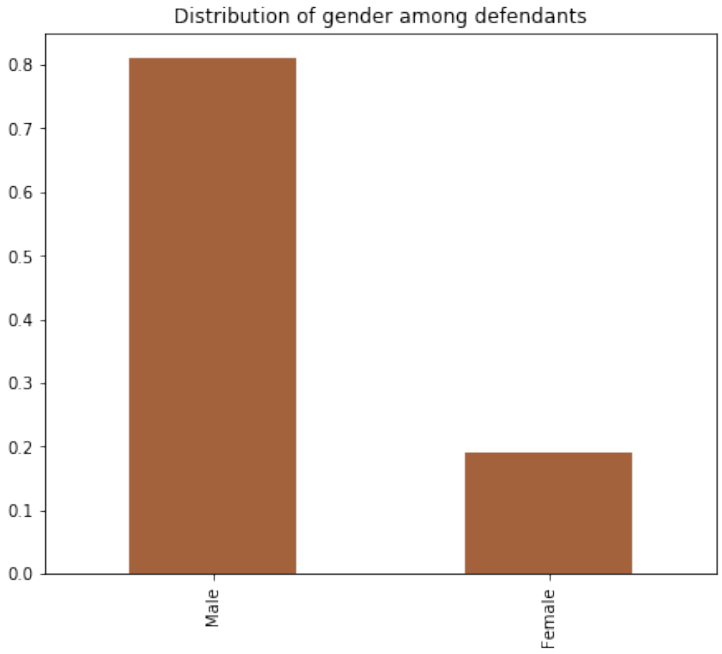
Within a circle, each category is asigned a sector proportional to its frequency.

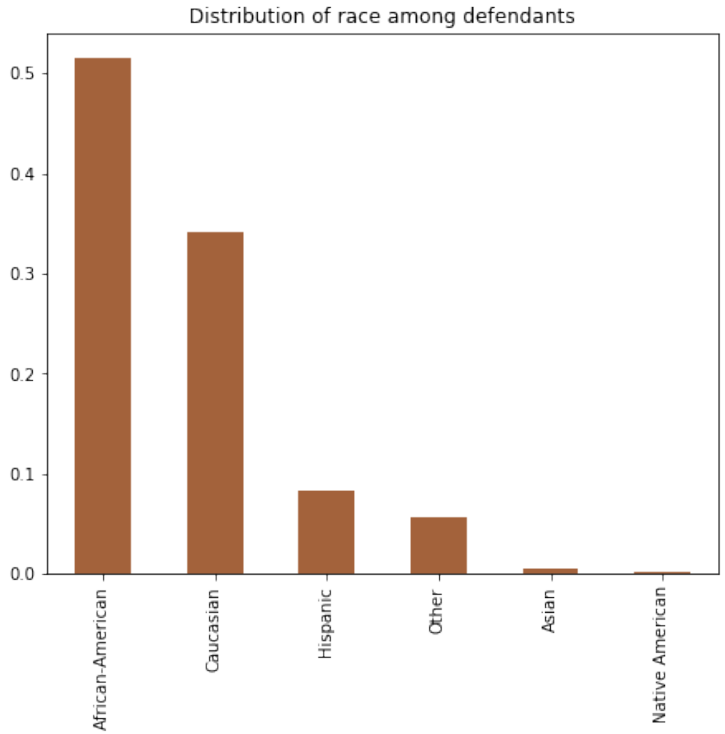
Defendant's age	n_i	f_i	$f_i \times 360$
Less than 25	1347	0.22	79^o
Between 25 and 45	3532	0.57	206^o
Greater than 45	1293	0.21	75^o
	6172	1	360^o



Analysis of Fairness

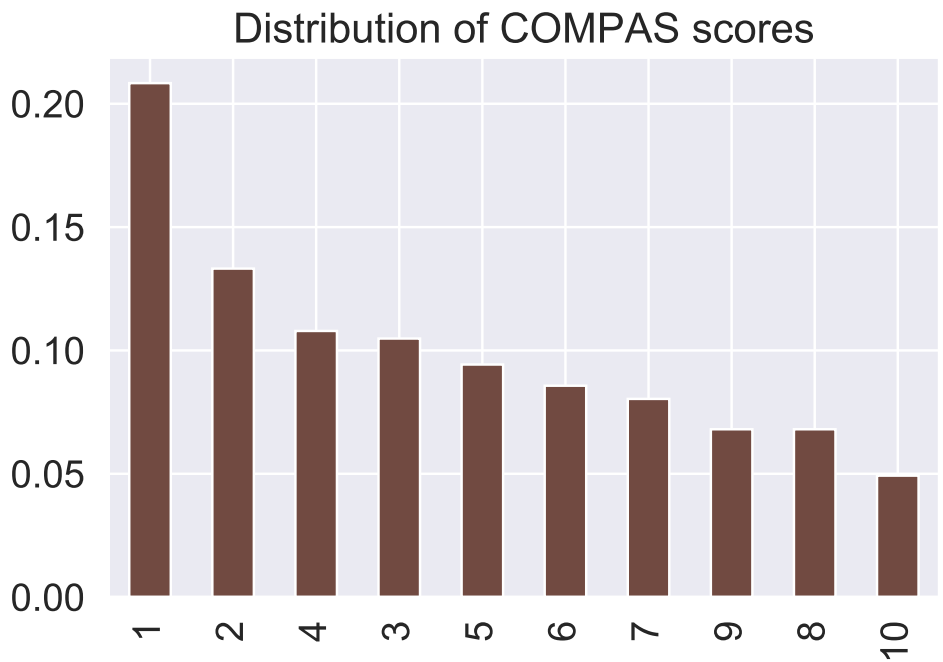
- We will now investigate the COMPAS algorithm on the data collected by ProPublica.
- We will analyze the COMPAS scores for "Risk of Recidivism". An equivalent analysis could be made for "Risk of Violent Recidivism".
- We will analyse: distribution of gender, races, distribution of the COMPAS decile scores for different groups, fairness analysis.





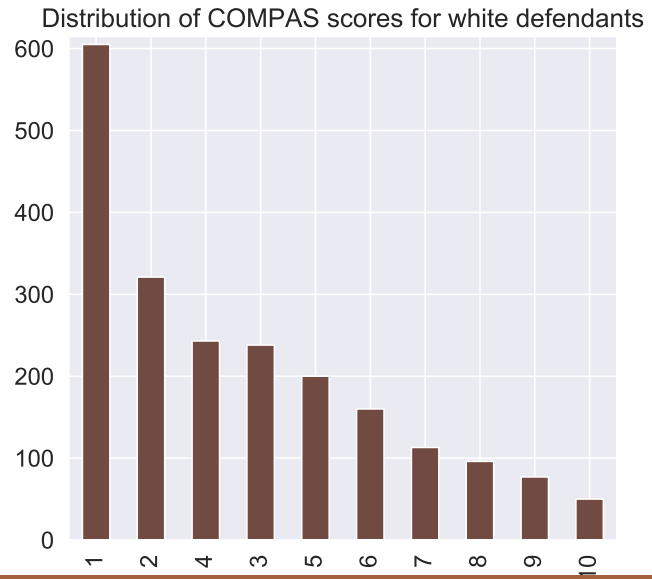
Almost 80% of defendants are classified as male, while the white and black defendants comprise of approximately 85% of the total population of defendants.

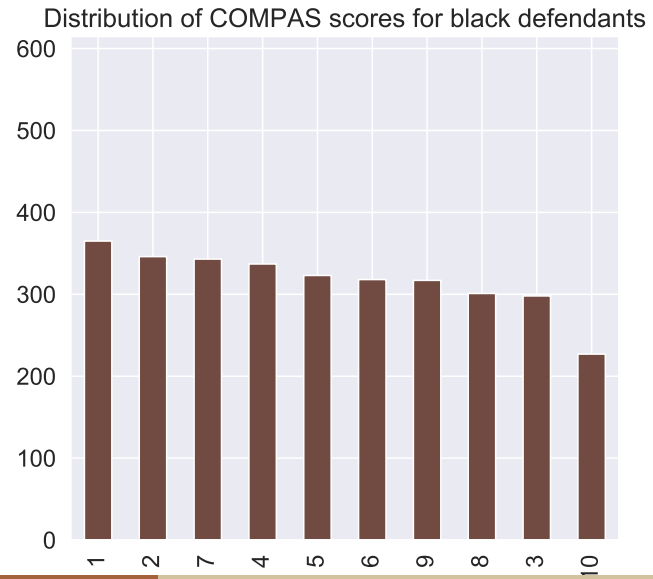
- We analyze the COMPAS scores for "Risk of Recidivism".
- We plot the distribution of the COMPAS decile scores.
- We plot the distribution of these scores for all 6,172 defendants who had not been arrested for a new offense or who had recidivated within two years.



Comparing the COMPAS score based on race

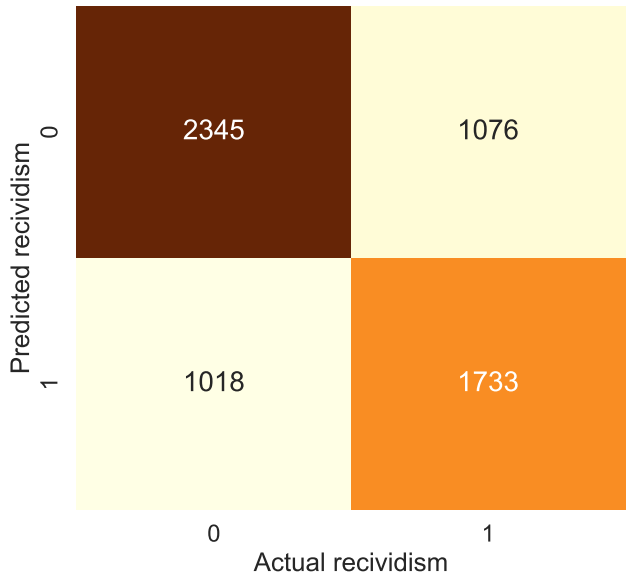
- There is a qualitative difference in the distributions among the Black and white defendants.
- Scores for white defendants were skewed toward lower-risk categories
- Scores for black defendants were evenly distributed across scores.
- These observations do not prove any demographic or behavioral bias.

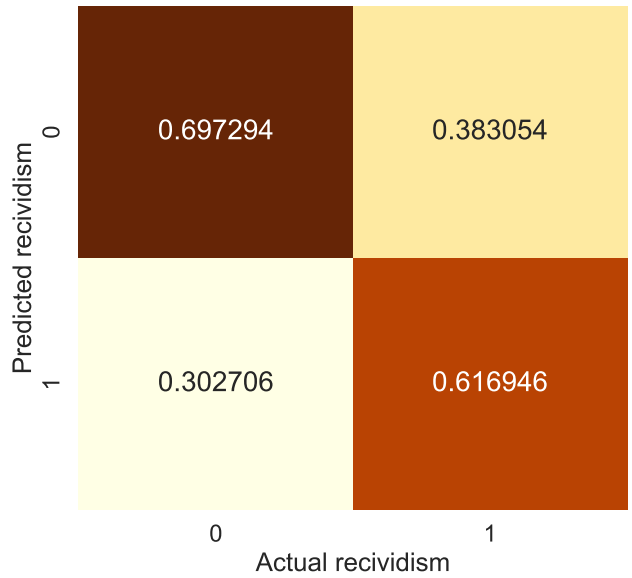




- The COMPAS algorithm, on the dataset as a whole, is relatively balanced.
- The positive class (actual recividists) represent a 46% of the dataset, which slightly coincides with the predicted recividists (45%).
- A 34% of the population experienced an incorrect decision, roughly balanced between false positives and false negatives.

	Acc(%)	FNR	FPR
All	66.0	0.38	0.30

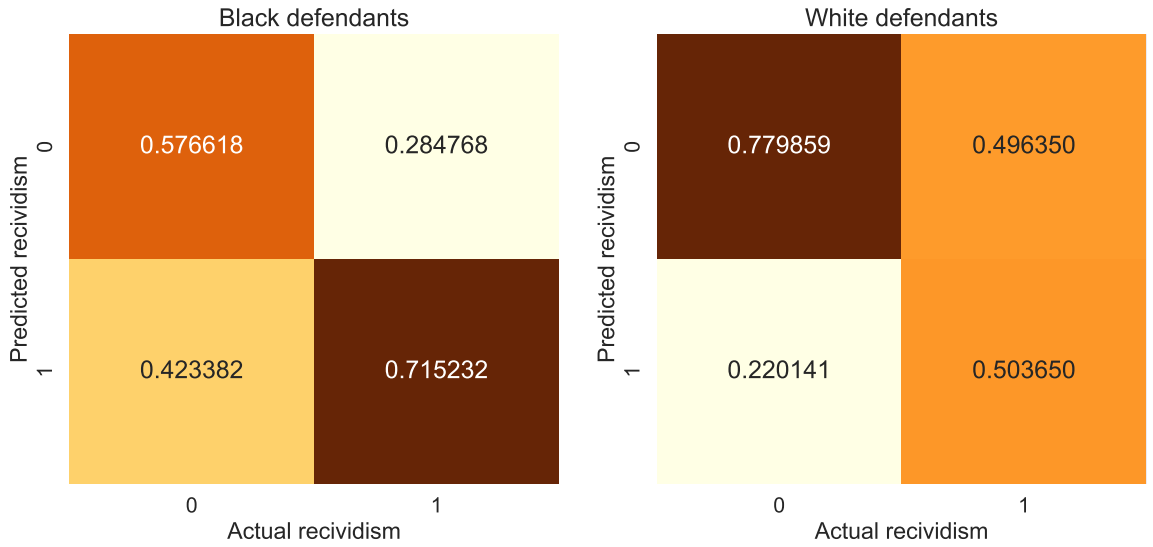




If we look at black and white populations separately:

- A greater proportion of black defendants experience an incorrect "will re-offend" prediction than white defendants.
- A greater proportion of white defendants experience an incorrect "won't re-offend" prediction than black defendants.

	Acc(%)	FNR	FPR
All	66.0	0.38	0.30
Black	64.9	0.28	0.42
White	67.2	0.49	0.22



Some interesting reading

- https://www.propublica.org/
- https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm
- https://github.com/propublica/compas-analysis/
- https://www.foia.gov/
- https://www.documentcloud.org/documents/2702103-Sample-Risk-Assessment-COMPAS-CORE.html