Introduction to Ethics in Data Science

including Recidivism Case Study via Python

Mike Ludkovski (based on materials by Alex Franks & Netasha Pizano)

Overview of Today's Lecture

1. Defining ethics
2. Ethical issues related to computers and data
3. Case study: COMPAS recidivism

“The subject matter of ethics is supplied by two intimately related terms, right and good” (Sharp 1928).

Defining Ethics

- Definition:
  - The judgment upon actions which are categorized as right or wrong
  - The societal use of terms good and bad
- Ethics derived from the Greek word ethos - spirit of the community.
- Ethicists try to answer "What actions are regarded as right or wrong by the members of it's society?"

Establishment of Ethics in Science

- The study of ethics dates back to work of Socrates, Plato, and Aristotle
    - Hippocratic Oath
        - Established several medical principles that are still significant today, including consent.
- Ethics in experimental science was a global concern after WWII
    - Nuremberg Code (1947)
        - A set of ethical research principles for human experimentation as a result of the Nazi Nuremberg trials.

Advancements in computer and data communications technology have resulted in the need to reevaluate the application of ethical principles.

- quote from (Weiss 1990) “The XXII self-assessment: The ethics of computing”

Ethical Issues Related to Computers and Data

Many of the daily work-related responsibilities of data scientists rely on making an ethical choice

Data privacy Social justice Transparency Consent Accuracy Accountability

- Data privacy legislation differs across jurisdictions.
- *Data privacy* in the California Consumer Protection Act is the protection of personal information (e.g., Name, email address, purchase and browsing history, location and employment data, IP address, sensitive personal information)

Ethical Issues Related to Computers and Data: Example

- **Uber** was fined millions by the Netherlands government after a watchdog found that the U.S. company was transferring data about drivers from the Netherlands to the U.S.
- **Target** was sued after a father was able to determine that his daughter was pregnant. A predictive algorithm determined that she was likely pregnant given the items she purchased and sent coupons for expecting mothers.

Ethical Issues Related to Computers and Data

- Greater public access of computers raised ethical concerns
- Ethical issues related to computers (Weiss 1990):
  - Ownership of assets/material (e.g., programs, code, visual material)
  - The degree that programs, software, and computer users are held responsible for outcomes

Legal Frameworks which Regulate Data Science

- There is variation across jurisdictions in the laws relating to privacy protection and permissible data usage.
- Two fundamental common themes in the **EU** and **US** legislation which are fundamental to data protection :
  - Anti-discrimination rights
  - Personal data protection rights

Legal Frameworks which Regulate Data Science

- Anti-discrimination Rights in the U.S. - The Civil Rights Act of 1964 prohibits discrimination based on color, race, sex, religion, or nationality, - The Disabilities Act of 1990 prohibits discrimination based on disability
- Personal data protection in the U.S.
  - In the United States, the Fair Information Practice Principles (1973) delineate principles that agencies use when evaluating information systems, processes, programs, and activities that affect individual privacy
- Since 2017, an increase in proposed laws and regulations related to AI have been proposed - The US's Blueprint for an AI Bill of Rights

Data Ethics Principles

- To provide broad guidance,various organizations (e.g., American Statistical Association, (American Statistical Association 2022)) and government agencies have proposed numerous data ethics principles.
- These principles vary in number, scope, and practicality
- Different corporations may chose to uphold the principles that meet and fit the company's culture due to the lack of federal legislation

Data Ethics Principles (OECD)

- The Organization for Economics Co-operation and Development’s (OECD) (OECD 2002) guidelines in the Protection of Privacy and Trans border Flows of Personal Data are one of the most supported set of principles by U.S. government agencies.

1. Collection Limitation Principle
2. Data Quality Principle
3. Purpose Specification Principle
4. Use Limitation Principle
5. Security Safeguards Principle
6. Openness Principle
7. Individual Participation Principle
8. Accountability principles

Data Ethics Principles (OECD)

1. Collection Limitation Principle: There should be limits to the collection of personal data and any such data should be obtained by lawful and fair means and, where appropriate, with the knowledge or consent of the data subject.
2. Data Quality Principle: Personal data should be relevant to the purposes for which they are to be used, and, to the extent necessary for those purposes, should be accurate, complete and kept up-to-date.

Data Ethics Principles (OECD)

3. Purpose Specification Principle: The purposes for which personal data are collected should be specified no later than at the time of data collection and the subsequent use limited to the fulfillment of those purposes or such others as are not incompatible with those purposes and as are specified on each occasion of change of purpose.
4. Personal data should not be disclosed, made available or otherwise used for purposes other than those specified in accordance with Principle 3, except:

a. with the consent of the data subject; or b) by the authority of law.

Data Ethics Principles (OECD)

5. Security Safeguards Principle: Personal data should be protected by reasonable security safeguards against such risks as loss or unauthorized access, destruction, use, modification or disclosure of data.
6. Openness Principle: There should be a general policy of openness about developments, practices and policies with respect to personal data. Means should be readily available of establishing the existence and nature of personal data, and the main purposes of their use, as well as the identity and usual residence of the data controller.

Data Ethics Principles (OECD)

7. Individual Participation Principle: An individual should have the right to obtain from a data controller, or otherwise, confirmation of whether or not the data controller has data relating to them; and to have communicated them that data relating to them.
8. Accountability Principles: A data controller should be accountable for complying with measures which give effect to the principles stated above.

Group Discussion

How does data science play a role in this example? What dilemmas must data scientists consider?

Discussion (4 minutes)

In 2017, Meta announced that they analyze and flag user-generated content related to suicidal intent or risk (e.g., post and comments). Later, Meta announced that the company used the suicide predictions to notify local authorities and that over 100 wellness checks were conducted by law-enforcement. Meta also announced that the suicide detection program would expand across all other countries. In 2018, Meta revealed that they also scan the content of users’ private messages for suicide risk.

04:00

Reset Start

Ethical Practices and Considerations

To meet data ethic principles and engage in socially-just data science, practitioners (e.g. data scientists, researchers) can practice the following:

Honesty Understanding Communication Reporting Protection Cite

Report

- Remain honest about level of competence related skill, content/subject, potential bias, and timeliness.

Ethical Practices and Considerations

Algorithms affect the lives of individuals, sometimes in profound ways

Question

What are example of algorithms that you've encountered in your life which have influenced your life or your behavior?

Question

What are the ethical practices and considerations that you think the creator of algorithm(s) did or did not take into account?

Case Study

Criminal Risk Assessment Algorithms

- Machine learning is increasingly used to to inform decisions about individuals in the criminal justice system
  - Setting bond amounts
  - Length of sentence
- One major component of these decisions is trying to evaluate a defendant's risk of future crime
- *Recidivism*: the tendency of a convicted criminal to re-offend.

The COMPAS recidivism algorithm

- In 2016, ProPublica published an a story about one of the most used algorithms called 'COMPAS'
- **COMPAS** stands for Correctional Offender Management Profiling for Alternative Sanctions, and was mostly used to assess the risk of a pretrial release
- In their article, they conducted a rigorous analysis of possible bias by analyzing data and COMPAS predictions from more than 10,000 criminal defendants in Broward County, Florida

Fairness & Transparency

- Fairness: ideally, COMPAS would have the same impact on all demographic subgroups. Probably not possible!
- Transparency: COMPAS is a *closed-source* algorithm. The public is not allowed to see how the algorithm works.

Question

What information about an individual do you think is “fair” to include in an algorithm which predicts recidivism? What information would be “unfair”?

COMPAS data

The risk-scores are derived come from answers to a 137 question survey and the defendants criminal record.

Predictors used by COMPAS include:

- Prior arrests and convictions (and if any friends had priors)
- Address, GPA, wealth
- If the defendant’s parents separated

Important

Purportedly, race is not used as a predictor.

Recidivism Data

There are 6172 observations in the data we’ll analyze. 44.6% of individuals received a high risk score.

Age Group	Proportion
25 - 45	0.57
Greater than 45	0.21
Less than 25	0.22

Race	Proportion
African-American	0.51
Asian	0.01
Caucasian	0.34
Hispanic	0.08
Native American	0
Other	0.06

Predictions

- The COMPAS algorithm is *closed-source* but we can learn about the algorithm by **fitting our own model** to the COMPAS predictions.
- Predict COMPAS score (“high” vs “low” risk) using data about the defendants collected by ProPublica.
- Data includes age, race, and prior counts
- Compare the predictions for different inputs to the model

Predictions

```
Optimization terminated successfully.
         Current function value: 0.511094
         Iterations 6
                           Logit Regression Results
==============================================================================
Dep. Variable:                      y   No. Observations:                 6172
Model:                          Logit   Df Residuals:                     6162
Method:                           MLE   Df Model:                            9
Date:                Fri, 13 Sep 2024   Pseudo R-squ.:                  0.2563
Time:                        17:55:24   Log-Likelihood:                -3154.5
converged:                       True   LL-Null:                       -4241.7
Covariance Type:            nonrobust   LLR p-value:                     0.000
===============================================================================================
                                  coef    std err          z      P>|z|      [0.025      0.975]
-----------------------------------------------------------------------------------------------
const                          -0.7715      0.083     -9.333      0.000      -0.934      -0.609
priors_count                    0.3013      0.011     27.374      0.000       0.280       0.323
gender_factor_Male             -0.1436      0.078     -1.840      0.066      -0.297       0.009
age_factor_Greater than 45     -1.4504      0.099    -14.688      0.000      -1.644      -1.257
age_factor_Less than 25         1.4264      0.074     19.212      0.000       1.281       1.572
race_factor_Asian              -0.8241      0.479     -1.722      0.085      -1.762       0.114
race_factor_Caucasian          -0.4884      0.068     -7.140      0.000      -0.622      -0.354
```

Note

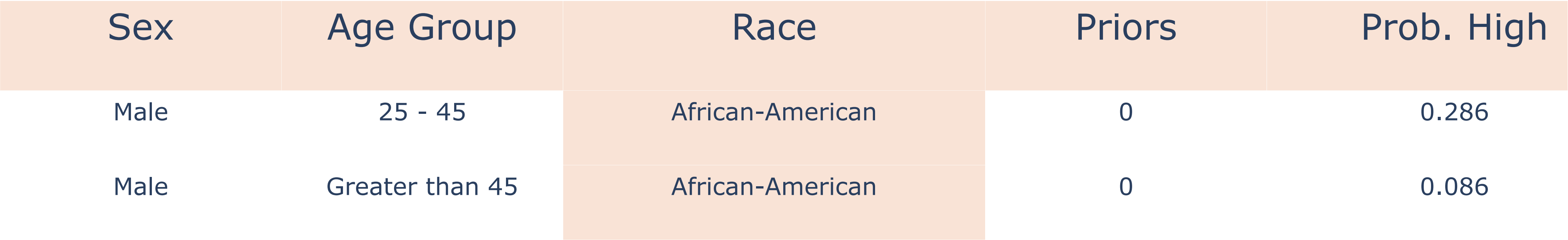
The classifier we used here is called a “logistic regression”. You’ll learn much more about this in later statistics courses.

Exploring Predictions

- Use our classifier to predict the probability that COMPAS would give an individual a high score
- Looks at inputs which are identical except in one characteristic
- Compare predictions for different ages and different races

The importance of age

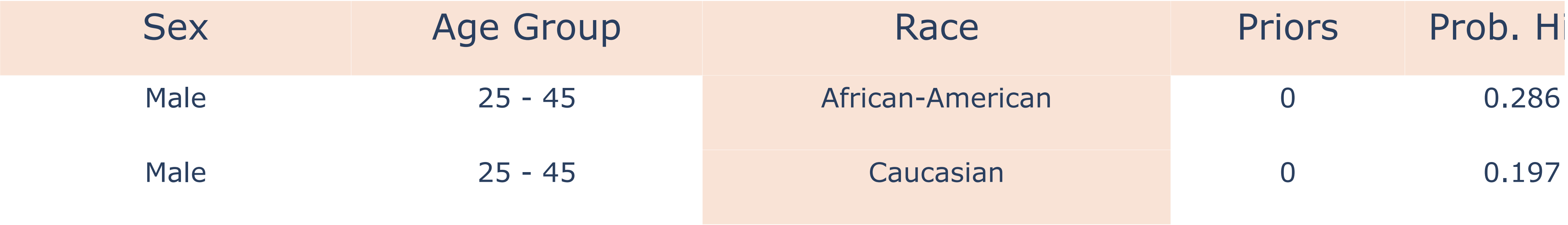
Compare predictions for two hypothetical individuals identical in all characteristics except age.



Sex	Age Group	Race	Priors	Prob. High
Male	25 - 45	African-American	0	0.286
Male	Greater than 45	African-American	0	0.086

The importance of race

Compare predictions for two hypothetical individuals identical in all characteristics except race.



Sex	Age Group	Race	Priors	Prob. Hi
Male	25 - 45	African-American	0	0.286
Male	25 - 45	Caucasian	0	0.197

Important

But “race” is purportedly **not an input** to COMPAS!

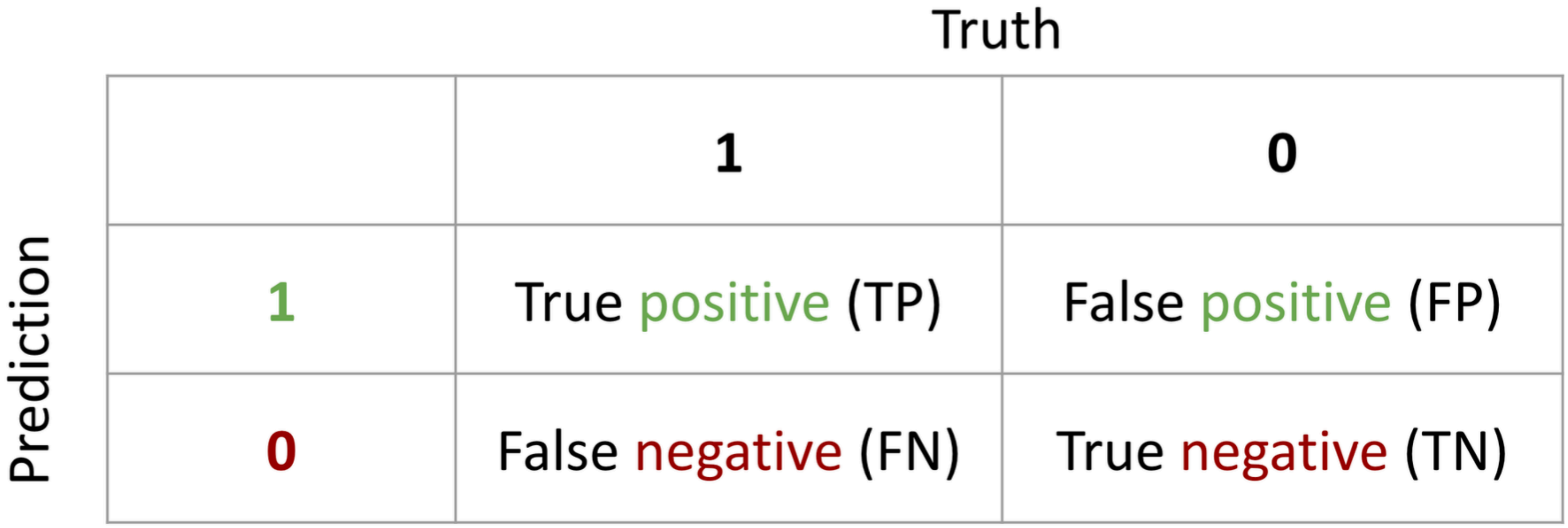
Proxy Variables

Just because `race` isn't an input to the algorithm does not mean the algorithm makes the same predictions for all race groups!

Question

How might the COMPAS algorithm implicitly factor racial information into its predictions even though `race` is not an input?

Evaluating Mis-classifications



Truth

Prediction

	1	0
1	True positive (TP)	False positive (FP)
0	False negative (FN)	True negative (TN)

- Accurary:TP + FP / n
- False positive rate (FPR): FP / (FP + TN)
- False negative rate (FNR): FN / (FN + TP)

Examining FPR and FNR

- ProPublica provides a binary variable named `two_year_recid` which indicates whether an individual committed a new crime within two years of the screening or not.
- We can compare the COMPAS risk predictions to the `two_year_recid` variable which we'll assume is our "ground truth" to compute FPR and FNR
- There is a tradeoff between FPR and FNR!

Examining FPR and FNR

```python
# Create a contingency table
print(pd.crosstab(compas_scores['score_factor'], compas_scores['two_year_recid']))
```

```
two_year_recid     0     1
score_factor
HighScore       1018  1733
LowScore        2345  1076
```

- FPR: 1018 / (2345 + 1018) = 0.30
- FNR: 1076 / (1733+1076) = 0.38
- Accuracy: (2345 + 1733) / n = 0.66

Question

What is the meaning of FPR and FNR in this context? Is one worse than the other?

FPR and FNR for African-Americans

```python
# Filter the data for African-American race
filtered_data = compas_scores[compas_scores['race'] == 'African-American']

# Create a contingency table for score_factor and two_year_recid
print(pd.crosstab(filtered_data['score_factor'], filtered_data['two_year_recid']))
```

```
two_year_recid    0     1
score_factor
HighScore        641  1188
LowScore         873   473
```

FPR = $641/(873+641) = 0.57$

FNR =

ACC =

FPR and FNR for Caucasians

```python
# Filter the data for Caucasian race
filtered_data = compas_scores[compas_scores['race'] == 'Caucasian']
print(pd.crosstab(filtered_data['score_factor'], filtered_data['two_year_recid']))
```

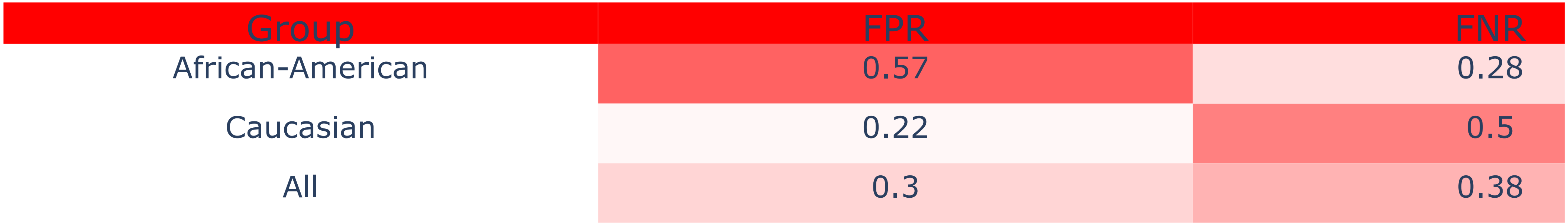
```
two_year_recid    0    1
score_factor
HighScore       282  414
LowScore        999  408
```

FPR =

FNR =

ACC =

FPR and FNR



Group	FPR	FNR
African-American	0.57	0.28
Caucasian	0.22	0.5
All	0.3	0.38

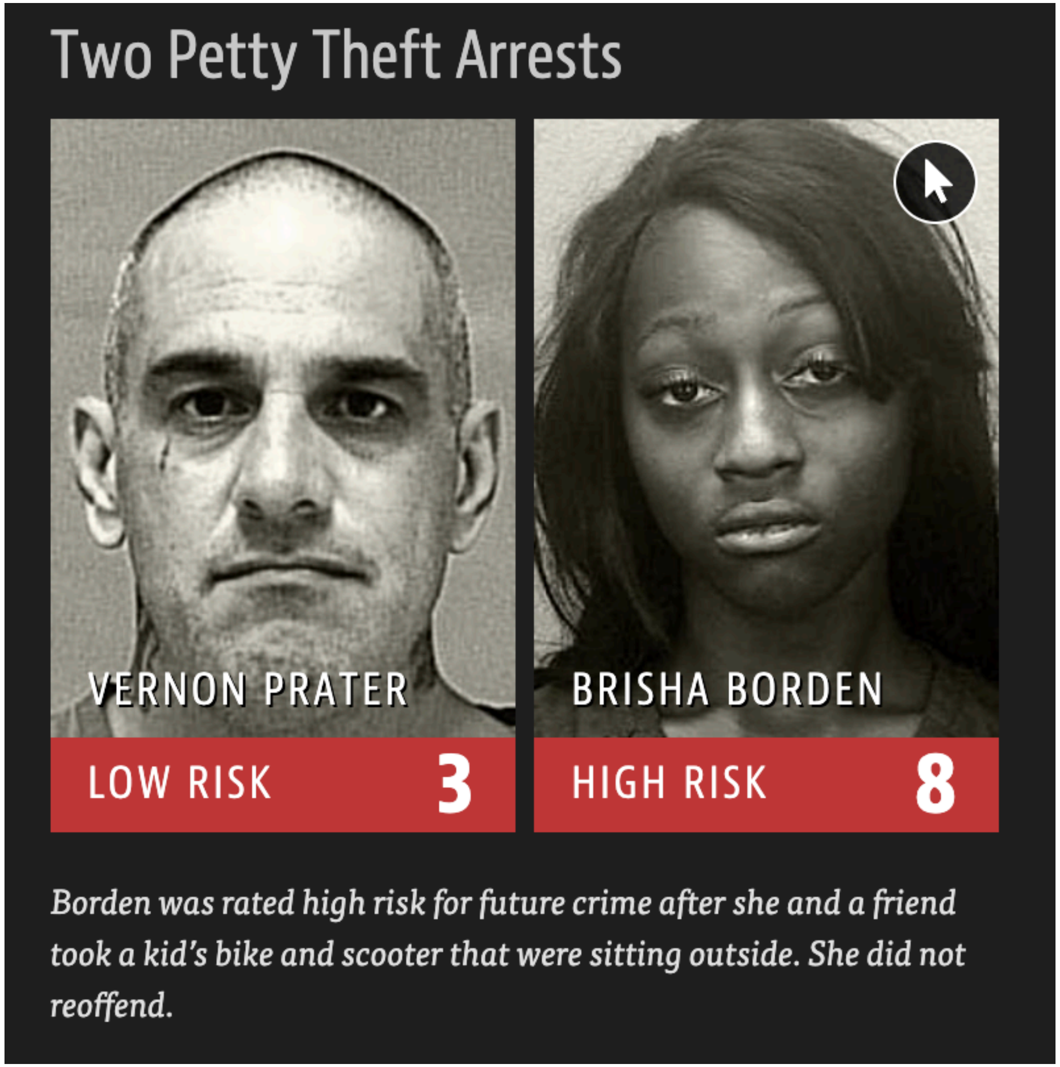
Similar accuracy across the two race groups are attained in very different ways!

Putting a human face on it

ProPublica discusses two different incidents in Broward County:

1. In 2014, an 18-year-old girl, Brisha Borden, and her friend grabbed an unlocked bicycle and scooter and rode them down the street, then abandoning it. Police arrived and arrested the girls for burglary and theft of $80 worth of goods.
2. In 2013, a 41-year-old man, Vernon Prater, was caught shoplifting $86 worth of goods from Home Depot. He had prior convictions for armed robbery and had previously served 5 years in prison.

Putting a human face on it

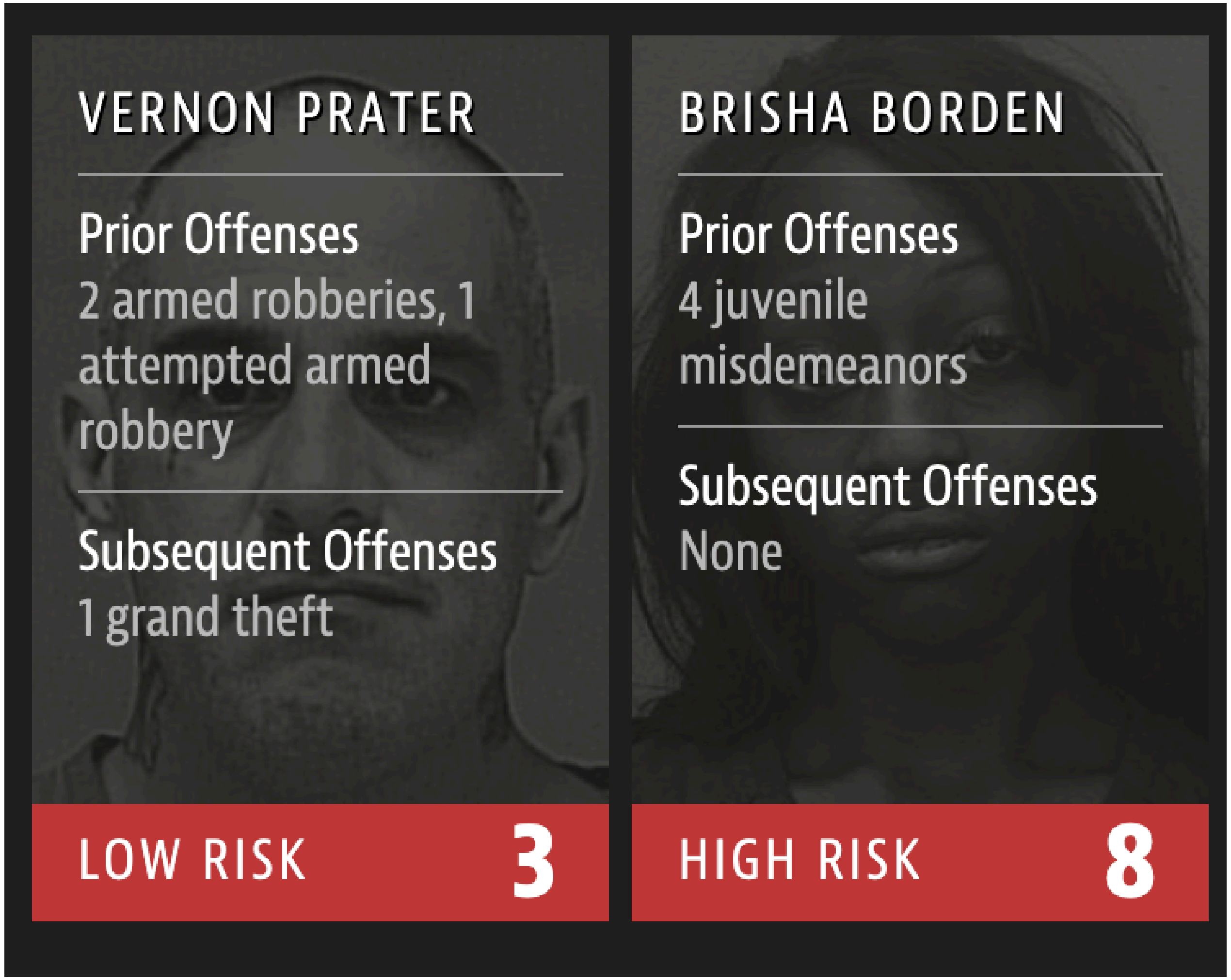


Two Petty Theft Arrests

Borden was rated high risk for future crime after she and a friend took a kid's bike and scooter that were sitting outside. She did not reoffend.

source

Putting a human face on it



VERNON PRATER

Prior Offenses
2 armed robberies, 1 attempted armed robbery

Subsequent Offenses
1 grand theft

LOW RISK 3

BRISHA BORDEN

Prior Offenses
4 juvenile misdemeanors

Subsequent Offenses
None

HIGH RISK 8

source

Summary

- We have highlighted unfairness in the COMPAS algorithm
- Understanding the root causes and devising potential solutions is non-trivial. There are often unintended consequences that data scientists must grapple with.
- Ethical data science and fairness in machine learning is challenging but essential!

Read more

- The ProPublica article, “Machine Bias”,
- Methodology for the analyses used in “Machine Bias”
- Fairness and Algorithmic Decision Making - COMPAS Chapter

References

American Statistical Association. 2022. "Ethical Guidelines for Statistical Practice."

Froomkin, A Michael. 2019. "Big Data: Destroyer of Informed Consent." *BIG DATA*.

OECD. 2002. *OECD Guidelines on the Protection of Privacy and Transborder Flows of Personal Data*. OECD. https://doi.org/10.1787/9789264196391-en.

Sharp, Frank C. 1928. "The Problem of Ethics." In, 3–20. The Century Philosophy Series. New York, NY: The Century Co.

Weiss, Eric A. 1990. "The XXII Self-Assessment: The Ethics of Computing." *Communications of the ACM* 33 (11): 110–32. https://doi.org/10.1145/92755.92780.