




Algorithm to Predict Crime, or Criminal Algorithm?

Pablo Boix - Sarah Jacobs - Nazrath Palliparambil



- 
- Title Slide (DONE)
 - Your Background (DONE)
 - Project Introduction / Background (DONE)
 - Methods (ONGOING)
 - Results (ONGOING)
 - Summary (ONGOING)
 - Conclusions (ONGOING)
 - Questions (DONE - JUST THE SLIDE)



Pablo J. Boix

Pablo has completed his Bachelor of Telecommunications Engineering at Universidad ORT Uruguay, also got his MBA at IE Business School in Madrid, Spain.

He has experience managing indirect distribution channels in the Product Identification Industry in Latin America for more than 15 years. He fluently speaks English, Portuguese and Spanish

Also worked as an QMS Auditor for many companies in S. America and Caribbean region (ISO 9001, 14001, 27001, OHSAS 18001)





Sarah Jacobs

Sarah graduated from The College of Charleston with a degree in Corporate and Organizational Communication.

She has been a project manager in the construction industry for the last 10 years. Prior to that her experience included product development, HR work, marketing and accounting.





Nazrath Palliparambil

Nazrath has completed her Bachelor of Technology in Electrical and Electronics Engineering from Calicut University, India.

She has experience maintaining large electrical systems by working in State Electricity Board in her home state of Kerala, India.





Project Introduction / Background

You should have a couple slides of background information about the topic you are covering. For instance, if you are asking and answering questions about the stock market, you should give a little basic information about the stock market. Try to build this up, so start with the most general information and then get more specific with information that directly relates to your data or the questions you will be answering with the data.

COMPAS Recidivism Bias Study

Project Overview:

COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) is an algorithm used in the American criminal justice system for judges and parole officers to make judgements on a defendant's likelihood of reoffending - recidivism.

We wanted to check if there was an ideal combination of variables that would produce less bias in the algorithm.

The **COMPAS** system is owned by a for-profit company and is therefore proprietary. The data that we used came from a Kaggle project and included intake records on over 18,000 people arrested in Broward County Florida. This information came from freedom of information requests filed by ProPublica (FOIA Act) and some data pre-processing done by them.



What are we looking for?

- What variables most accurately predict high COMPAS scores?
(NEED TO CHANGE FOR THE FINAL ONE THIS IS THE ORIGINAL QUESTION)
- What age groups is more likely to get a high COMPAS score?



“The ability to argue with an algorithm requires confronting the base assumption that an algorithmically-derived assessment is objectively true, distant, and fixed.”

“However, it can be difficult to assess the validity of information generated through proprietary algorithms because vendors often claim that their algorithms are trade secrets that cannot be shared”

Anne L. Washington PhD





Methods

Next comes your methods section. This is where you talk about all of the details regarding your data. You want to give a very high level overview of what you did to:

- Gather/find data (DONE)
- Manipulate / wrangle data (DONE)
- Create new variables (DONE)

You also want to paint a picture of what your data is like. Include details such as:

- Important variables and their summary statistics
- Sample size

The methods section should only be a few slides, and **should not** include any code. You are presenting to a wide, non-data science audience, and thus should not go into a lot of detail.

Data Gathering

- Since all the data linked to prisons, risk analysis and recidivism is disperse. Propublica did an intensive gathering and wrangling of this data into manageable csv files.
- Most papers and evaluations done by third parties also rely on ProPublica initial gather process.
- Even the company response to Propublica initial findings were based on their datasets.
- Since any extra data implies going to several sources, with FOIA requests in some cases, we decided to use the available datasets and try to answer our test questions from there.
- Main source data is located on github: <https://github.com/propublica/compas-analysis>
- Alternative data sources from other scholars (at the end of this presentation in acknowledgments slide)



Manipulation and Data Wrangling



- The wrangling techniques used with datasets were among standard practices:
 - Check for duplicate records
 - Add, Delete and Merge Columns.
 - Dummy coding of variables
 - Create ID variables
 - Joins (while using Tableau for graphics)
 - Create/Link categorical variables from scales or ordinal values.

Relevant Variables



- COMPAS Score (in either flavour: by decile, categorical and/or raw)
- 2 yr Recidivism
- Race
- Age

“Missing” variables:

- Input variables to “black box” COMPAS algorithm. (137 questions)

TOOLS & METHODS:



- Python:
 - Machine Learning
- R
 - Regressions, Graphics
- Tableau
 - Graphics



Results—

The results section will be the meat of your presentation. You should divide your results section into parts by evaluation question, so that you can easily signpost things. In the results section, you will go over any of the exploratory findings you have you want discuss, as well as the answers to each evaluation question. Ensure that you provide LOTS of beautiful visuals to go along with your findings.

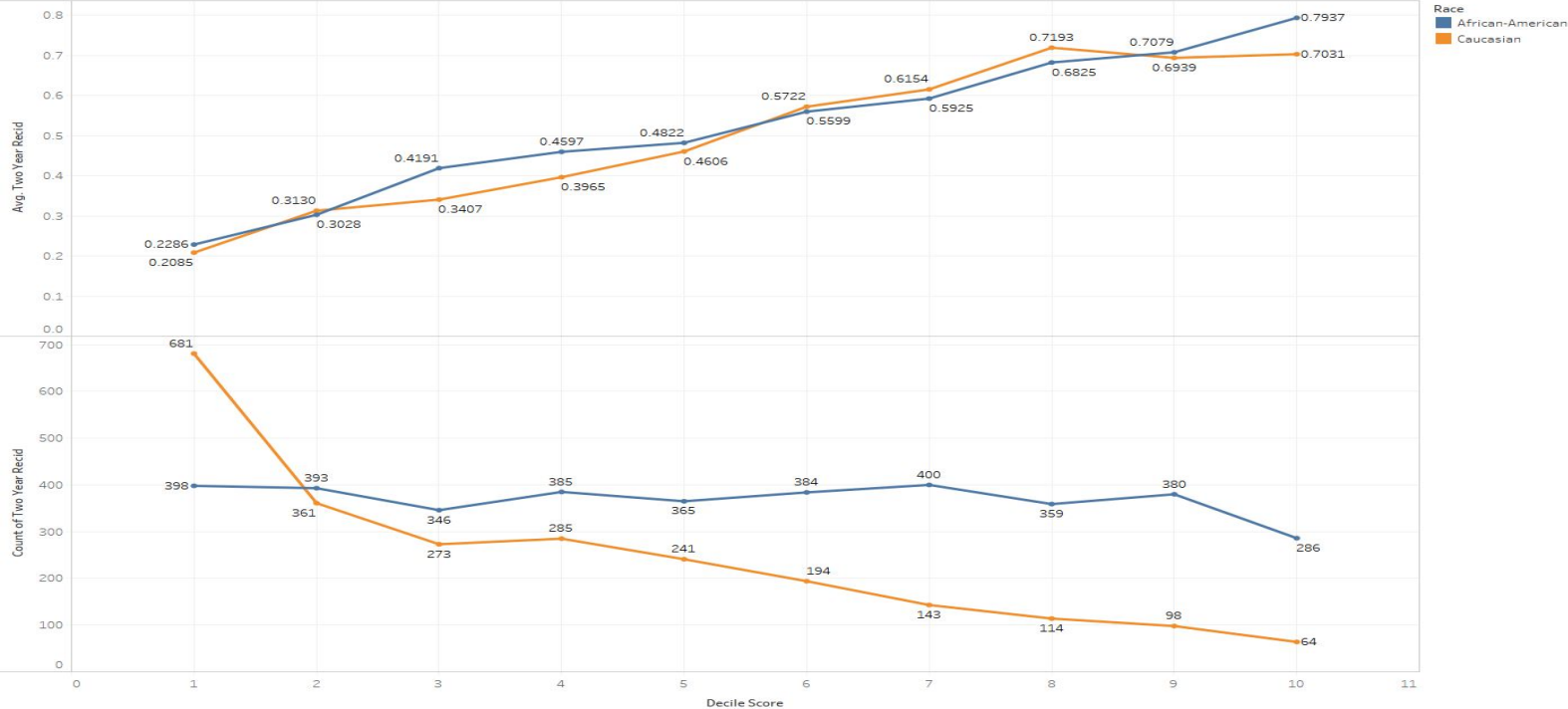
Results:



Recidivism - Count - Decile Score

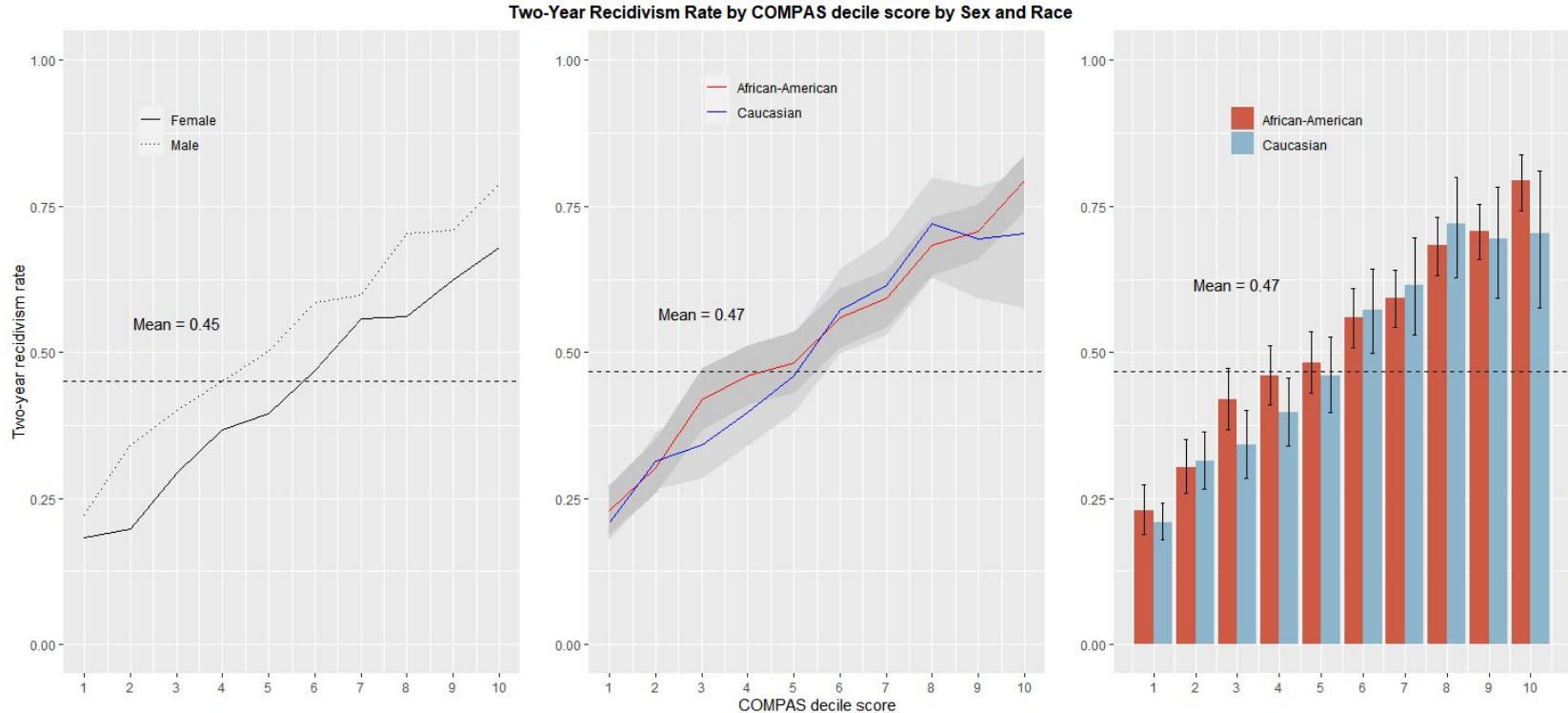


Recidivism and Count vs Decile Score



The trends of average of Two Year Recid and count of Two Year Recid for Decile Score. Color shows details about Race. Details are shown for Race. The view is filtered on Race, which keeps African-American and Caucasian.

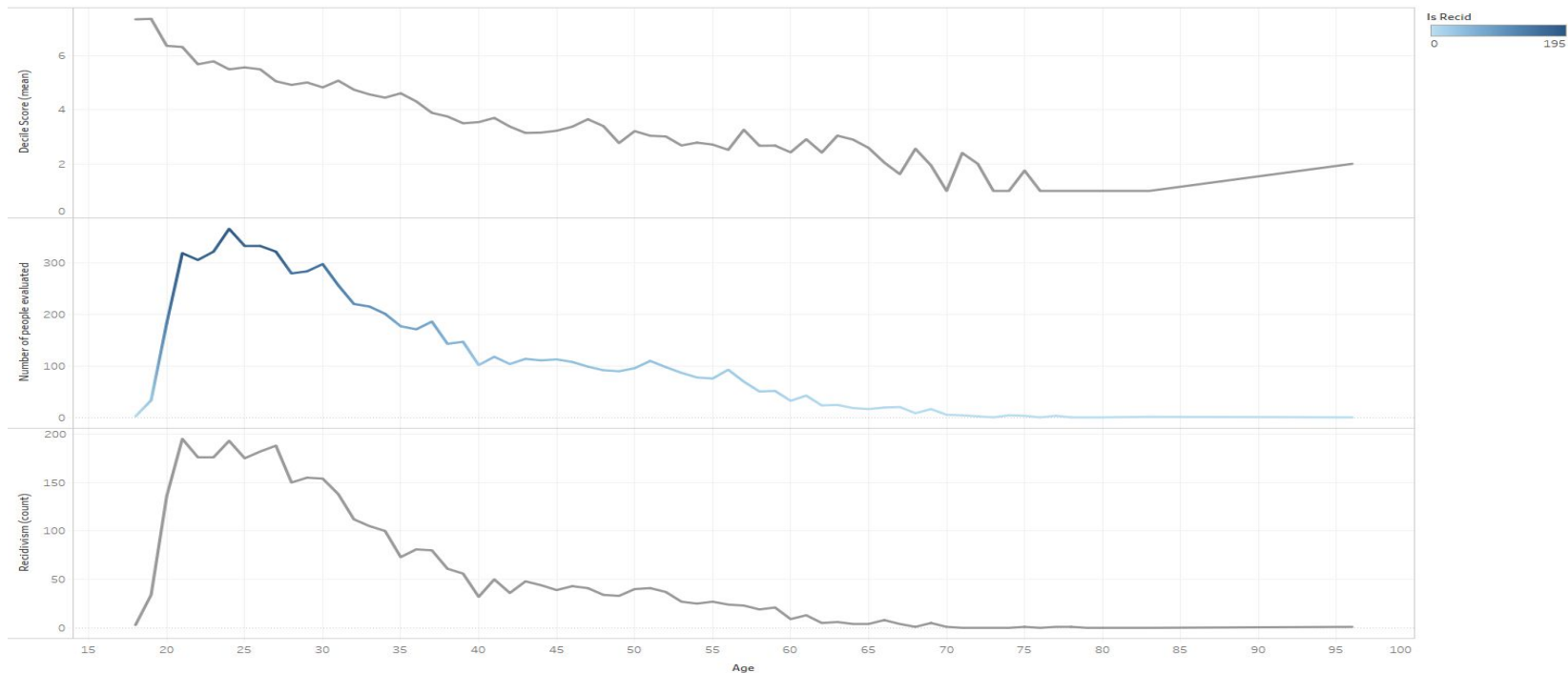
Recidivism Rate vs COMPAS Score



COMPAS Score and Recidivism by Age



Score and Recidivism by Age



The trends of average of Decile Score, count of Decile Score and sum of Is Recid for Age. For pane Count of Decile Score: Color shows sum of Is Recid.

Scores in Percentage →

Below 25

High - 21.16%

Low - 45.54%

Medium - 33.3%



Between 25 & 45

High - 10.72%

Low - 68.43%

Medium - 20.85%



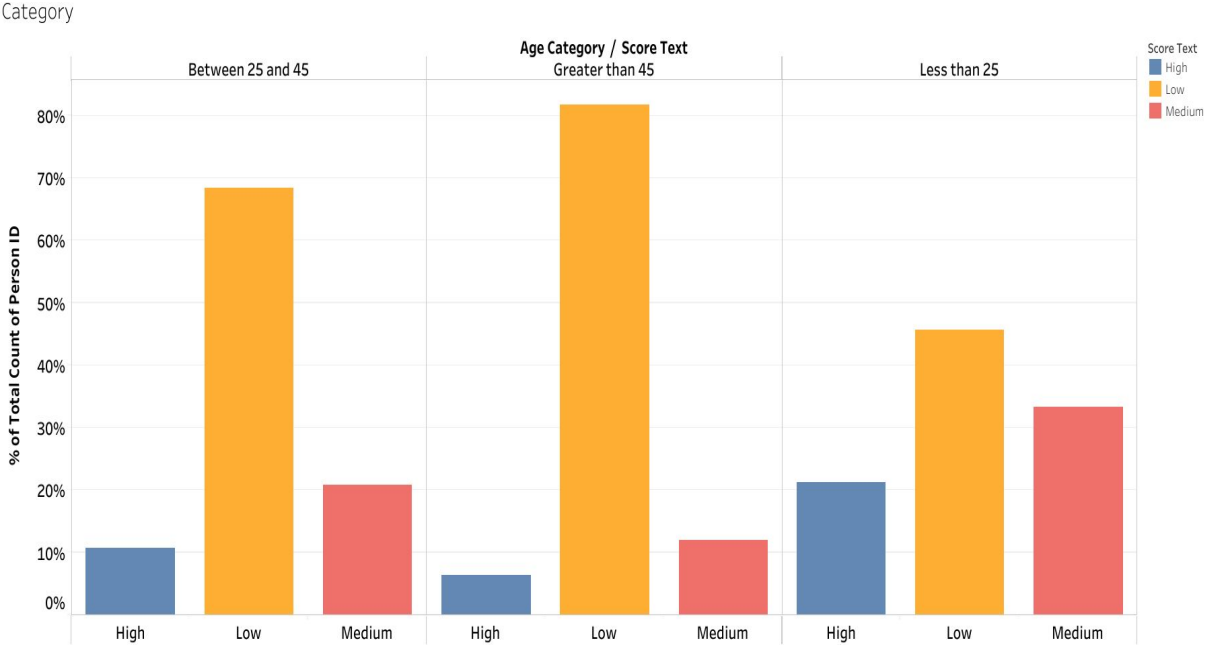
Greater than 45

High - 6.38%

Low - 81.69%

Medium - 11.93%

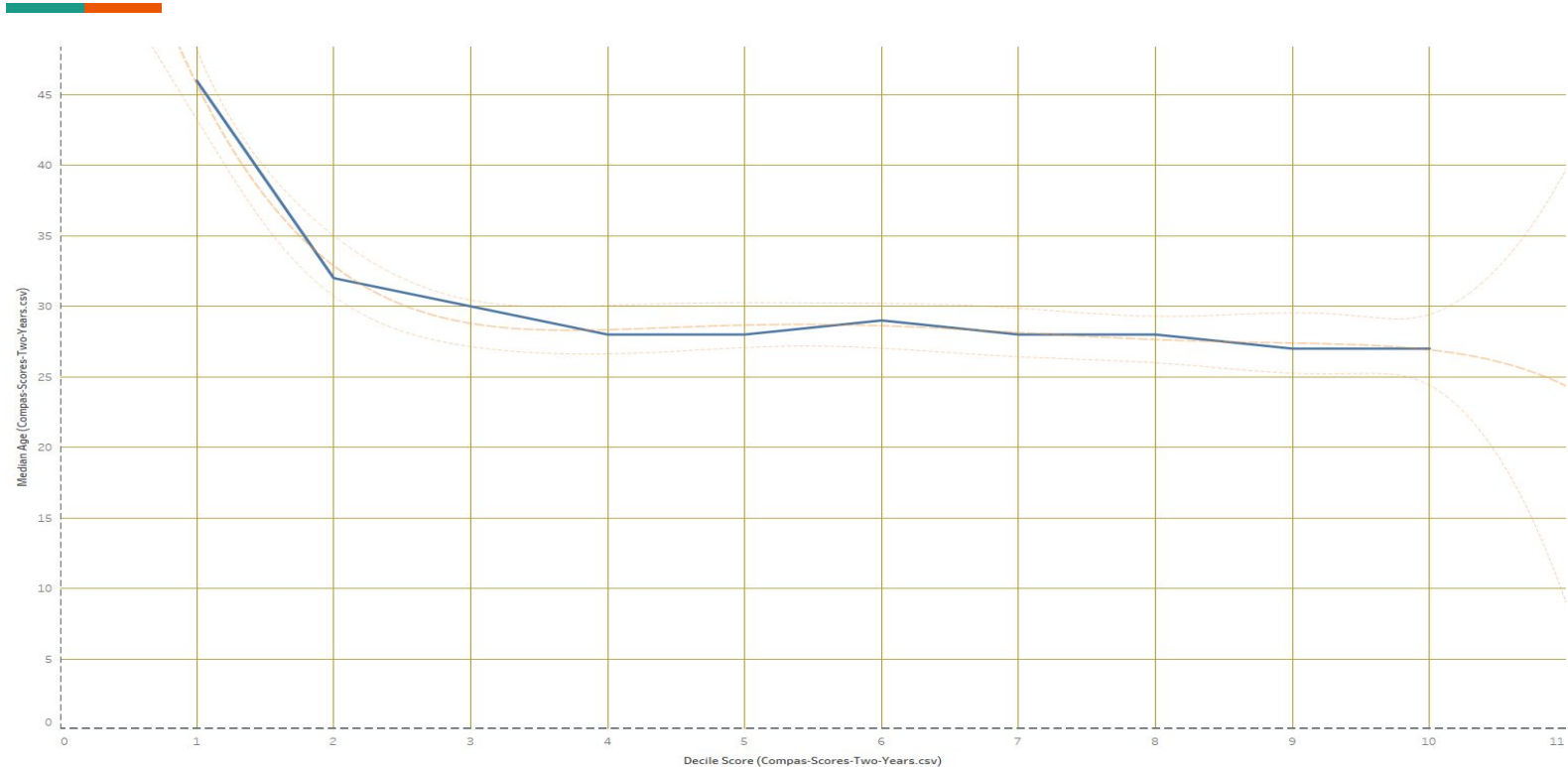
Compas Score By Age Category



Below 25 - High 21.16%

Greater than 45 - Low 81.69%

Median Age by Decile Score



The trend of median of Age (Compas-Scores-Two-Years.Csv) for Decile Score (Compas-Scores-Two-Years.Csv).



Results

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt labore dolore magna aliqua. Lorem ipsum dolor sit amet consectetur adipiscing elit.

- 01 | Consectetur adipiscing elit
- 02 | Sed do eiusmod tempor incididunt ut labore
- 03 | Magna aliqua lorem ipsum dolor sit amet
- 04 | Eiusmod tempor incididunt ut labore et
- 05 | Dolore magna aliqua





Summary

The summary should be JUST ONE SLIDE,
and it should contain a summary of the
entirety of your results section. It should be in
layman's terms, quick and dirty.



Conclusions

Your conclusion section should also only be one slide long. In it, you should have some bullet points containing information about:

- How do your findings impact the world at large?
- What's important about this work?
- Big picture information

Conclusions



- Transparency is a key issue in any analysis.
 - Data Gathering from a disperse number of record is not easy (in cases legal action is needed).
 - Algorithms for risk assessment are obscure in the best scenario (black-box)
- Starting a research in many Data Science cases is like entering a “Rabbit Hole”

TO KEEP IN MIND:

- Between May 2016 and December 2017, the Propublica “Machine Bias” article was cited by at least 578 scholars and researchers.
- No clear consensus has ever been achieved as to whether the COMPAS risk system is biased beyond the results evaluated here.
- The Propublica research is still subject of passionate and heated discussions.
- The lack of agreement could be attributed to, among other things::
 - Legal challenges linked to due process (i.e: State vs Loomis)
 - Lack of transparency on what variables the algorithm uses (137 questions serve as IV input).
 - “Chicken or the egg” mentality - in other words the mindset that higher numbers of minorities are in the justice system because they commit more crimes rather than the system targeting them disproportionately.
- Anne L Washington (PhD) grouped all discussion related this subject in three general themes:
 - Mathematical definitions of fairness
 - Explainable interpretation of models
 - Importance of population comparison groups



ETHICAL AND LEGAL OPEN ISSUES



“The question presented for review by the Court was whether the proprietary nature of the COMPAS violated a defendant’s constitutional right to due process because a defendant cannot challenge the algorithm’s accuracy or scientific validity. The United States Supreme Court declined to hear the case...”

“A research led by Jon Kleinberg presents three conditions that could denote fairness: (1) calibration; (2) balancing negative impact; and (3) balancing positive impact. Kleinberg includes mathematical proofs that show that it is not possible to simultaneously have all three conditions at once.”

The factors that are the input for COMPAS Score could have a different influence on on future behaviour on different groups.

Original technical paper by algorithm developer “Northpointe” was based on 2328 cases, propublica original study based on less than 10000...



References, Bibliography and Acknowledgements:



- ❖ EVALUATING THE PREDICTIVE VALIDITY OF THE COMPAS RISK AND NEEDS ASSESSMENT SYSTEM
Tim Brennan, William Dieterich, Beate Ehret - Northpointe Institute for Public Management Inc.
- ❖ Data & original analysis gathered by ProPublica.
Original Data methodology article: [How We Analyzed the COMPAS Recidivism Algorithm](#)
- ❖ [ProPublica Responds to Company's Critique of Machine Bias Story](#)
- ❖ [HOW TO ARGUE WITH AN ALGORITHM: LESSONS FROM THE COMPAS- PROPUBLICA DEBATE - Anne L. Washington](#)
- ❖ [ProPublica's COMPAS Data Revisited -2019 - Matias Barenstein](#)
 - https://github.com/mbarenstein/ProPublica_COMPAS_Data_Revisited
- ❖ [FairML: Auditing Black-Box Predictive Models](#)
- ❖



Questions?





Review DSO104 Data Wrangling and Visualization for help making your presentation stunning!