Data Science Looks At Discrimination

By: Aditya Mittal
Team: Norm Matloff, Arjun Ashok, Taha Abdullah, Brandon Zarate
University of California, Davis

Agenda

1. Introduction
2. **Part One:** Detecting Discrimination & Adjustment For Confounders
3. **Part Two:** Discovering & Mitigating Bias in Machine Learning
4. Discussion
5. Questions

Introducing dsld (R Package)

- ➔ Broadly aimed at statistics instructors and students, offering a powerful yet user-friendly approach to studying discrimination.
  - ◆ Intended to appeal to students' sense of *social awareness* & increase interest in statistics courses.
  - ◆ Includes an **80 page Quarto book** to serve as a guide of the key statistical principles and their applications.
- ➔ Discrimination remains a critical social issue in the United States and many other countries.
- ➔ **dsld** offers advanced *analytical* and *graphical tools* for detecting and measuring **discrimination** and **bias** related to attributes such as race, gender, age, and marital status.

Part One: Detecting Discrimination

Motivating Example

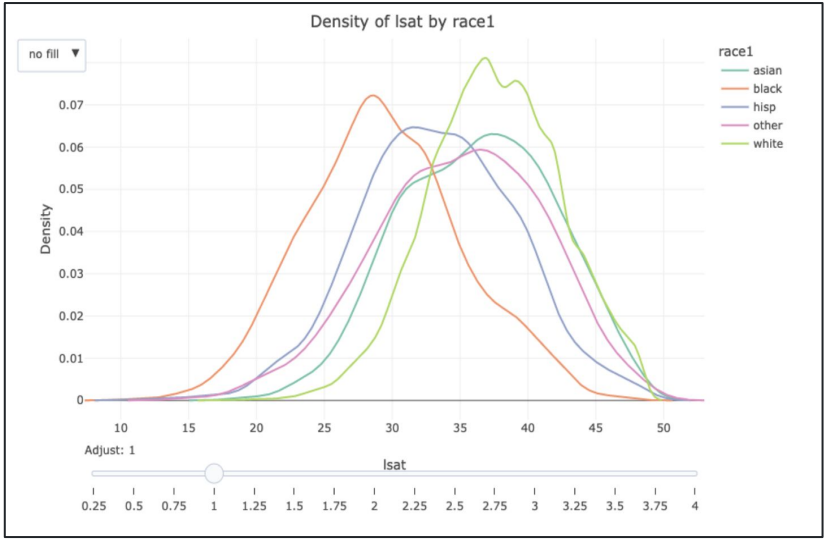
- ➔ Criticism of **standardized testing** for favoring students with more resources.
- ➔ Studies show test discrepancies between Black and White students (Dixon-Roman et al., 2013)
- ➔ Many institutions have removed **SAT** and **GRE** requirements.
- ➔ Reveals importance of examining potential **biases** in standardized testing.

Dataset: Law Schools Admissions

- ➔ Is the LSAT unfair?
- ➔ What are potential **confounding** factors that may affect our analysis?

Graphical Analysis (show applications of various methods provided by dsld)

- ➔ Analyze the the distribution of LSAT scores segmented by race using **dsldDensityByS**.
- ➔ Investigate potential **racial differences** in LSAT scores.
- ➔ Can serve as a starting point for classroom discussions for further analysis.
- ➔ Results may be influenced by effect of **confounding variables.**



Distribution of LSAT scores, segmented by race

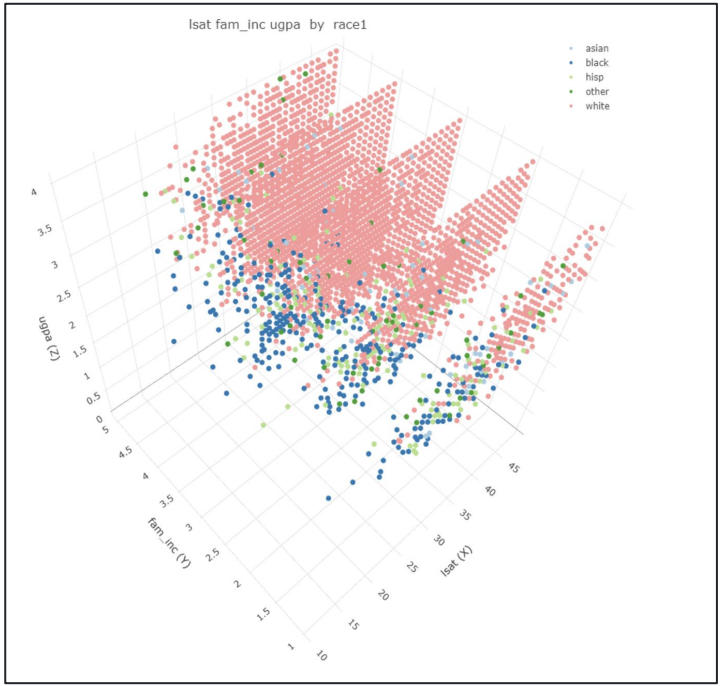
Investigating Confounding Relationships

Investigating confounding relationships among the variables LSAT score, GPA, Family Income, Race, etc.

- Visualize these relationships using **dsldScatterPlot3D**.

- *Lowest family income quintile:* Mostly Black and Latino students; upper levels: Predominantly white students.
- *Lower LSAT scores:* Majority non-white, across all income levels.
- *Undergraduate GPA:* Similar trend to LSAT, but less pronounced.

- Exploratory analysis suggests family income may confound the relationship between race and LSAT score. *Requires further investigation.*



Analysis using **dsldLinear**

Investigate concern that the LSAT and other similar tests are biased against Black and Latino students, and might otherwise have racial issues.

```
$`Sensitive Factor Level Comparisons`
   Factors Compared Estimates Standard Errors
1     asian - black  4.748263       0.1980883
2      asian - hisp  2.001460       0.2035044
3     asian - other  0.868031       0.2625286
4     asian - white -1.247088       0.1546271
5      black - hisp -2.746803       0.1863750
6     black - other -3.880232       0.2515488
7     black - white -5.995351       0.1409991
8      hisp - other -1.133429       0.2562971
9      hisp - white -3.248547       0.1457509
10    other - white -2.115119       0.2194472
```

Pairwise Comparison of estimates of each sensitive levels race in the no-interactions case via **dsldLinear().**

- ➔ Additional arguments required: **Interactions** (boolean), and **StComparisonPts** (Data-frame)
- ➔ In the interactions case, we fit S *different* linear models for each level of race.
- ➔ Racial differences in LSAT scores: Black and white individuals with similar educational backgrounds differ by nearly **6 points**.

Caution needed due to dataset quality and potential hidden confounders, like the quality of undergraduate institutions.

Part Two: Mitigating Bias in Machine Learning

Motivating Example: Compas Algorithm

- **COMPAS** algorithm used to predict recidivism, faced criticism by *ProPublica* for alleged bias against Black defendants.
- Northpointe contested *ProPublica's* findings, while *ProPublica* defends their analysis with statistical evidence.
- The debate highlights the importance of **fairness in machine learning** and teaching fair practices to address biases and promote equitable outcomes.
- **dsld** provides many wrappers from **fairML** and **EDFFair** packages for fair predictive modelling.

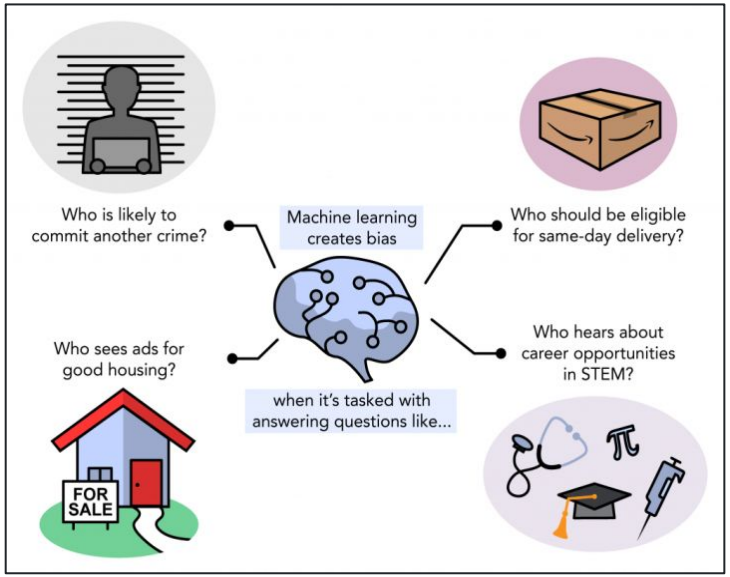
ProPublica. (2016). Machine bias: Risk assessments in criminal sentencing. (Link)

Measuring Fairness

Important to uncover and reduce **biases** in machine learning models to ensure fairness across sensitive groups.

Two main components of fair machine learning:

- ➔ **Measuring unfairness:** How can we measure the level of influence of a sensitive variable *S* on our predictions?
- ➔ **Reducing unfairness:** For a given algorithm, how can we ameliorate its unfairness, yet still maintain an acceptable utility (predictive power) level?



Examples of potential bias in machine learning applications

Fairness vs. Utility Trade-off

Fairness-Utility tradeoff: Inherent tradeoff between fairness and predictive accuracy – prioritizing fairness in an algorithm may lead to decreased accuracy.

- ➔ **Measuring Accuracy:** Measured by the misclassification rate (binary classification) or Mean Absolute Prediction Error (regression).

- ➔ **Measuring Unfairness:** Assess the predicted relation between *Y-S* despite omitting *S* by computing the correlation between predicted Y and S using **Kendall's Tau** correlation (provides value between [-1,1]).

A note on proxy variables: Secondary variables that indirectly infer a protected attribute, potentially introducing bias in decision-making even when the protected attribute is not explicitly used.

COMPAS Example (Introduction)

Goal of **COMPAS** example: Omitting *S* from analyses, possibly due to legal requirements or fairness concerns. We are also concerned about impact of potential *proxy variables*.

- ➔ Correlation between *predicted Y* and *S* which highlights possible fairness concerns and necessitates mitigation strategies.

- ➔ Predict **probability of recidivism** *Y* using **race** as our sensitive variable *S*
- ➔ Use traditional ML algorithms to establish baseline results for fairness vs. utility tradeoffs
  - ◆ Logistic Regression
  - ◆ K-Nearest Neighbors
  - ◆ Random Forests

- ➔ **Measuring Unfairness:** Kendall Tau correlation between Predicted *Y* and *S*.
- ➔ **Measuring Accuracy:** Percent of correctly classified defendants

COMPAS Dataset (dsld)

- **DsldFairML** wrappers incorporate **unfairness parameter** [0,1] to reduce predictive power of *race* at some cost in model accuracy.
  - Fair Ridge Regression, Fair Generalized Ridge Regression (Scutari et. al, Komiyama et. al)
  - Zafar's Linear Regression, Zafar's Logistic Regression (Zafar. et al)
- We set unfairness parameter for race at **0.01** and measure fairness vs. utility trade-offs.

- **DsldEDFFair** (Matloff and Zhang)**:** We omit *race* entirely, and also account for the effect of proxies using the **deWeightPars parameter** to increase fairness at cost of model accuracy.
  - Fair Ridge Linear/Logistic Regression
  - Fair K-Nearest Neighbors
  - Fair Random Forests
- Using **dsldOHunting**, we can identify possible proxies as **age** and **number of prior arrests**.
- Set deWeightPars to **0.01** to reduce both of their predictive power.

Results Table

Algorithm	S-Corr (Black)	S-Corr (White)	S-Corr (Hispanic)	Accuracy
Logistic Regression	0.210	-0.156	-0.106	0.734
K-Nearest Neighbors	0.224	-0.138	-0.162	0.731
Random Forests	0.175	-0.123	-0.100	0.777
dsldFgrrm	0.012	0.00039	-0.0228	0.731
dsldZlm	-0.0372	0.059	-0.036	0.633
dsldQeFairRidgeLog	0.197	-0.147	-0.097	0.735
dsldQeFairKNN	0.167	-0.107	-0.114	0.747
dsldQeFairRF	0.135	-0.078	- 0.106	0.780

Discussion

- ➔ Fairness in Machine Learning is an increasingly growing and important topic, especially with the application of extremely complex AI algorithms throughout different sectors.

- ➔ DSLD provides several statistical and graphical tools for detecting and measuring discrimination and bias – racial, gender, age, etc.

- ➔ Students are encouraged to try out further examples. Our current paper and the Quarto Book extends the examples and analysis that were highlighted in today's presentation

- ➔ Other potential use cases:
  - ◆ Quantitative analysis in instruction and research in the social sciences.
  - ◆ Corporate HR analysis and research.
  - ◆ Litigation involving discrimination and related issues.
  - ◆ Concerned citizenry.

Thank you!

Questions?