Homework Set 6 - COMPAS ⚖

Purpose: In this homework you will explore "soft" metadata (e.g. provenance). This component is critical for placing the data in context. However it can be more challenging to work with because of the non-rigorous nature. In particular you will explore how concepts like 'fairness' are dependent on the viewpoint of the observer. (N.B. The review section is made up of items from previous homework's review sections, this is meant to help prepare you for the mid-term exam).

Instructions: Put your solutions for all problems in a google colaboratory notebook. Create a level-1 header for each problem and then present the solution in subsequent cells. The header must contain the problem number followed by "confidence x" where x is an integer from 1 to 10 with 10 representing high confidence. For full credit you must show your work. The submission is made via a link on UVACanvas.

Domain Extras: This assignment looks into the criminal justice system and the concept of 'fairness'. Both of these are complicated and when combined can be rather tricky to understand. In particular when newer ideas and technologies come up against the established tradition of the criminal justice system.

Data

This assignment explores the data and analysis behind the infamous Pro Publica article "Machine Bias". The GitHub site can be found here:
https://github.com/propublica/compas-analysis/tree/master

6.1 Pipeline: Preparation (10ψ)

- Load the COMPAS data and identify the variables used in the analysis (charge severity, priors, demographics, age, sex, COMPAS scores, two-year recidivism).
- Apply ProPublica's inclusion criteria: charge date within 30 days of arrest; valid recidivism flag; exclude traffic-only offenses; restrict to people who either recidivated within two years or had at least two years of follow-up; exclude rows with missing COMPAS score.
- Report sample size before and after filtering and the percentage of the original sample retained.
- Summarize the analysis sample: counts or proportions by age category, race, and sex; distribution of COMPAS score (Low / Medium / High).

6.2 Pipeline: Analysis (10ψ)

- Test whether COMPAS scores differ by race and age after controlling for other factors: fit a logistic regression with outcome Medium/High vs Low and predictors including gender, age category, race, priors, charge type, and two-year recidivism.
- Interpret key coefficients (e.g., odds of higher score for Black vs white, young vs middle-aged) and the reference category.
- Assess predictive accuracy of COMPAS for recidivism (e.g., Cox proportional hazards model and concordance).
- Quantify directions of bias: overprediction (high-score defendants who did not reoffend) and underprediction (low-score defendants who did reoffend) by race.

6.3 Pipeline: Visualization (10ψ)

- Summarize the filtered sample with bar charts: demographics by age category, race, and sex; distribution of COMPAS score levels.
- Visualize score distribution by race (e.g., decile scores and Low/Medium/High counts by race).
- Visualize directions of bias: e.g., proportion of high-score defendants who did not reoffend, by race; proportion of low-score defendants who did reoffend, by race.

6.4 Pipeline: Design Focus (10ψ)

Different stakeholders use different definitions of fairness when evaluating risk scores like COMPAS. ProPublica emphasized disparities in false positive rates (e.g., Black defendants labeled higher risk but not reoffending at higher rates than white defendants). Northpointe (COMPAS's creator) rebutted by saying *predictive parity* and *accuracy equity* are what matter. In response to the discussion Kleinberg, Mullainathan, and Raghavan showed that these definitions of fairness conditions are *incompatible*. Furthermore, Rudin et al. argue that focusing on competing fairness definitions is misplaced without *transparency*; they partially reconstruct COMPAS and question ProPublica's assumptions (e.g., dependence on age).

Task: Chose one of the articles below and contrast how it handles the concept of 'fairness' compared to Pro Publica.

- Northpointe rebuttal: `https://go.volarisgroup.com/rs/430-MBX-989/images/ProPublica_Commentary_Final_070616.pdf`
- Kleinberg, Mullainathan, Raghavan, "Inherent Trade-Offs in the Fair Determination of Risk Scores," `https://arxiv.org/abs/1609.05807`

- Rudin, Wang, Coker, "The Age of Secrecy and Unfairness in Recidivism Prediction," Harvard Data Science Review: `https://hdsr.mitpress.mit.edu/pub/7z10o269/release/7`

Explain the difference, and why there is a conflict. Conclude with your view: which consideration(s) should matter most when deciding whether to use a tool like COMPAS?

6.5 Pipeline: Press Release (10ψ)

Role-play: You are a member of the Albemarle County Board of Supervisors. The board must take a public position on the use of algorithmic risk-assessment tools (such as COMPAS) in *pretrial* decision making. Your task is to draft a press release that states and defends the board's position.

Choose one of the following positions (and write to it):

1. **ProPublica-style fairness:** The board is concerned that risk tools produce racially disparate outcomes—e.g., Black defendants are more likely than white defendants to be labeled higher risk but not reoffend. The board supports scrutiny of algorithms on these grounds.

2. **Northpointe-style defense:** The board finds that when properly evaluated (e.g., predictive parity, accuracy equity), risk tools perform similarly across groups and can be used in pretrial decisions.

3. **Rudin et al.–style transparency:** The board prioritizes transparency over any single fairness definition; it opposes proprietary tools and supports interpretable, auditable alternatives.

4. **Algorithms are not fair:** The board opposes the use of algorithmic risk scores in pretrial decisions on the grounds that true fairness cannot be achieved (citing Kleinberg, Mullainathan, and Raghavan).

Your press release must include:

- A **headline** that clearly states the board's position.
- One or two **paragraphs** that explain the position and the board's argument.
- A **chart** (figure) that supports the argument—e.g., a comparison of rates by group, a summary of accuracy or fairness metrics, or another relevant visualization from your analysis or from the course materials. The chart must be referenced in the text and labeled (e.g., "Figure 1").

Write in the voice of an official board communication: clear, concise, and directed at the public and the press.

6.6 Apply in new context (10ψ)

The purpose of this problem is to apply the principles you have learned about soft metadata (provenance, licensing, etc.) to a new context.

- Find a new article of data journalism
- Identify and explain how it communicates the provenance, licensing, and ethical considerations of the data.
- Assess how well the article uses GitHub (or equivalent) to report and share the data and code necessary to reproduce the analysis. Share this assessment in one paragraph.

6.7 Review: Definitions (20ψ)

Provide definitions, or describe the distinction, for the following (one or two sentences each):

6.7.1 Accuracy vs. Precision

6.7.2 Seed

6.7.3 Stratified Sampling

6.7.4 Schema

6.7.5 Primary Key vs. Foreign Key

6.7.6 Relational Algebra

6.7.7 Metadata

6.7.8 Provenance vs. DMP

6.7.9 Representative Sample

6.7.10 FAIR vs. CARE

6.8 Review: Multiple Choice (20ψ)

Instructions: Select the best answer for each question. Do not show work.

6.8.1 What is the fundamental principle behind Monte Carlo simulation?

(A) Using deterministic algorithms to find exact solutions

(B) Using random sampling to estimate numerical results

(C) Minimizing computational cost by reducing sample size

(D) Fitting analytical functions to experimental data

6.8.2 Many-to-many (M:N) relationships require a junction table because:

(A) SQL databases cannot store arrays in a single cell

(B) It improves query performance

(C) A foreign key in one table can only reference one row in another table

(D) It is required by the ERD standard

6.8.3 In a double-blind clinical trial, which groups are unaware of treatment assignments?

(A) Only the participants

(B) Only the researchers administering the treatment

(C) Both the participants and the researchers

(D) Only the data analysts

6.8.4 What is the primary purpose of randomization in a randomized controlled trial (RCT)?

(A) To ensure the study has enough participants

(B) To minimize selection bias and balance confounding variables between groups

(C) To make the study easier to conduct

(D) To reduce the cost of the study

6.8.5 In the ANSI/SPARC three-level architecture, the level that describes how data is actually stored on disk is the:

(A) Conceptual level

(B) External level

(C) Logical level

(D) Physical level

6.8.6 The cardinality constraint "one-to-many" in an ER diagram means:

(A) Each entity instance participates in exactly one relationship instance

(B) One instance of entity A can be associated with many instances of entity B, and vice versa

(C) One instance of entity A can be associated with many instances of entity B, but each B is associated with at most one A

(D) The relationship is optional on both sides

6.8.7 Which of the following best defines metadata?

(A) Data that describes other data

(B) The largest or most important values in a dataset.

(C) A backup copy of a dataset stored in a different location.

(D) The statistical summary (e.g., mean, median) of a numeric variable.

6.8.8 Provenance in the context of metadata refers to:

(A) The cost of storing and processing the data.

(B) The origin and history of the data

(C) The number of rows and columns in a table.

(D) The visual layout of a report or dashboard.

6.8.9 A data dictionary is used to:

(A) Document what each field means

(B) Store the actual data values in a compressed form.

(C) Define security rules for who can edit the data.

(D) Automatically correct errors in the data.

6.8.10 This homework set is:

(A) Too short and Too easy

(B) Too short and Too hard

(C) Too long and Too easy

(D) Too long and Too hard