Homework 12: COMPAS

36-462/662, Fall 2019

Due at 10 pm on Wednesday, 20 November 2019

Our data set this week comes from the analysis of the COMPAS risk prediction scores performed by ProPublica for Broward County[1], Florida. (See the readings for Lecture 23, and `https://github.com/propublica/compas-analysis/`.) Specifically, our data file tracks the following information:

- The age of each suspect;
- Their age, binned into categories;
- Their sex;
- Their race;
- Their COMPAS score[2] for risk of violence (1–10);
- Their COMPAS score, binned into categories of "Low" risk (1–4), "Medium" (5–7) or "High" (8–10);
- Their COMPAS score, binned into categories of "Low" (1–4) and "Medium or High" (5–10);
- Whether they were charged with a felony (F) or misdemeanor (M);
- Count of priors[3].
- Whether they had a subsequent conviction for violence within two years.

1. (10) *Online questions* are online, and orient you towards next week's homework.
2. Load the data.

[1] Mostly: Fort Lauderdale, in the greater Miami metropolitan area.

[2] COMPAS calculates separate scores for risk of "failure to appear" at trial, risk of committing any type of crime, and risk of violence. This is specifically the score for risk of violence.

[3] I cannot quite tell from the documentation whether this is convictions or arrests (I think it's convictions), or whether, for this data on *violent* recidivism, only priors for violence are counted (I think not).

(a) (5) Narrow down the data set so it contains only blacks and whites. How many rows are you left with? Can you check that everyone remaining in the data is either black or white, and that all blacks and all whites in the original data are in this reduced data set?

(b) (5) Using histograms or other density estimates, show the distribution of (i) ages, (ii) number of priors, and (iii) COMPAS scores for (α) everyone, (β) whites only and (γ) blacks only. Ideally, you should have either a 3×3 matrix of plots, or three plots with three curves each.

(c) (5) How easy would it be to predict whether an arrestee was white or black from their COMPAS score? From their age?

3. Suppose we predict whether someone is at risk of committing a violent crime by thresholding their COMPAS score at 4, i.e., we predict no violence for those with a "Low" score, and violence for those with a high or medium score. For all of the following problems, do all calculations for (i) everyone, (ii) blacks alone, and (iii) whites alone. For full credit, have R do all the calculations using (elements of) named variables — there should be no hard-coded numbers you would have to adjust if the data-set changed.

   (a) (2) What is the confusion matrix? What is the over-all inaccuracy?

   (b) (2) What are the rates of false positives and false negatives?

   (c) (2) What are the positive and negative predictive values?

   (d) (4) Does COMPAS show (i) demographic parity, (ii) parity of error rates and/or (iii) parity of predictive accuracy?

4. *Calibration*

   (a) (2) Plot the actual frequency of recidivism for (i) all low-risk people, (ii) low-risk blacks and (iii) low-risk whites, and again for those categorized as being high or medium risk.

   (b) (3) Plot the actual frequency of recidivism against the numerical, 1–10 risk score for (i) everyone, (ii) blacks and (iii) whites.

   (c) (5) Using methods from homework 8, how much information does race give about recidivism, conditional on the risk score?

   (d) (5) How much information does the number of priors add about recidivism, conditional on the risk score?

   (e) (5) How much information does sex add about recidivism, conditional on the risk score?

   (f) (5) Does the COMPAS score appear to be calibrated, or equally calibrated for both blacks and whites? Explain your answer by referring to the earlier parts of this problem.

5. *Equalizing error rates*

(a) (7) Consider applying different thresholds, in the range 1–10, to the risk score. For each threshold, calculate the over-all FPR, the FPR for blacks, and the FPR for whites. Plot these as a function of the threshold. (You should have one graph with three curves.) Is there any threshold at which the FPR is equalized across races, that is, can you achieve FPR parity? If so, what is the over-all accuracy at this threshold, and is the accuracy the same for both blacks and whites?

(b) (8) Let's define the "violation of parity" as the absolute value of the difference in FPRs between the races. Calculate the over-all error rate and the violation of parity for each threshold, and plot the results. Describe, in words, what this curve tells us about the trade-off between efficiency (in the sense of having a low over-all error rate) and fairness (in the sense of parity of false positives).

(c) (5) Consider applying a *different* threshold for each race. Is there a pair of thresholds which will achieve parity? If so, what is the over-all error rate with those thresholds, and the error rate for each race separately?

6. *Summarizing* (10) Imagine you are hired by member of the council of a county which is considering adopting COMPAS. Summarize what you have learned from this analysis about the ways in which COMPAS is or is not accurate and fair. Based on this, how would you recommend that the county use COMPAS, if at all? (You may assume that the county is also in Florida, and generally similar to Broward County.) Would it make a difference to your recommendation whether the council member was black, white, or something else?

Rubric (10): The text is laid out cleanly, with clear divisions between problems and sub-problems. The writing itself is well-organized, free of grammatical and other mechanical errors, and easy to follow. Questions which ask for a plot or table are answered with both the figure itself and the command (or commands) use to make the plot. Plots are carefully labeled, with informative and legible titles, axis labels, and (if called for) sub-titles and legends; they are placed near the text of the corresponding problem. All quantitative and mathematical claims are supported by appropriate derivations, included in the text, or calculations in code. Numerical results are reported to appropriate precision. Code is properly integrated with a tool like R Markdown or knitr, and both the knitted file and the source file are submitted. The code is indented, commented, and uses meaningful names. All code is relevant to the text; there are no dangling or useless commands. All parts of all problems are answered with actual coherent sentences, and never with raw computer code or its output.

Extra credit (5 points): The original analysis of this data by ProPublica included using a logistic regression to predict whether blacks were more likely than whites to be assigned to the high risk category, *conditional on* the other features, including whether or not they proved to be recidivists. (They were.)

What notion of fairness is this testing? Can you think of an innocent, fair explanation for this apparent disparity?