Homework 8: COMPAS and Fairness

36-313, Fall 2022

Due at 6 pm on Thursday, 3 November 2022

Our data set this week comes from the analysis, performed by the news organization ProPublica, of the "COMPAS" risk prediction scores for Broward County, Florida[1]. ProPublica compiled a data set on everyone arrested in Broward County over a certain time span, for whom the police or the jails had calculated a COMPAS score, and follow-up information about whether they had been re-arrested. (The course homepage provides further reading about this controversy. You don't *have* to read them for this assignment, but they can't hurt.) Specifically, our data file, `compas_violence.csv`, tracks the following information (in order):

- The age of each arrestee;
- Their age, binned into categories;
- Their sex;
- Their race;
- Their COMPAS score[2] for risk of violence (1–10, 1 being low and 10 high);
- Whether they were charged with a felony[3] (F) or misdemeanor (M);
- Count of priors[4]
- Whether they had a subsequent conviction for violence within two years. This is called "recidivism".

Notation: In this problem set, Y is the recidivism variable, 1 if the arrestee was re-arrested for violence within 2 years, and 0 otherwise. $\hat{Y}$ is the prediction of Y. The "positive" class will be recidivism, $Y = 1$, so a false positive means $Y = 0$ but $\hat{Y} = 1$, and a false negative means $Y = 1$ but $\hat{Y} = 0$.

Before using the data, filter it to remove all the arrestees who aren't either black or white[5]. When these questions refer to "everyone", it means "all blacks and all whites".

1. **Features and race**

    a. (3) Using histograms or other suitable plots, show the distribution of (i) age, (ii) number of priors and (iii) COMPAS scores for (α) everyone, (β) blacks and (γ) whites. (You should have 3 plots, each with 3 curves, but a 3×3 array of plots will get partial credit.)

    b. (5) You should see that black arrestees (as a population) are younger than white arrestees. Consider a rule which guesses that an arrestee is black if their age $< t$, and white if they are $\geq t$. For each t

[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. We are only using the score for violence in this assignment.

[3] American law distinguishes between two kinds of crimes. Felonies are more serious crimes, punishable by (in most states) a year or more of imprisonment, or, in some situations, death. Misdemeanors are punishable by shorter terms of imprisonment (typically in city or county jails rather than state or federal prisons) and/or fines. Most crimes of violence are felonies, but not all felonies are crimes of violence: fraud, drug dealing, and tax evasion, for instance, are all felonies.

[4] This appears to be the count of prior *convictions* for crimes (not just arrests).

[5] This is partly so that we don't have to worry about more than two-way comparisons, and partly because some of the other racial categories have only a small number of members in the data set, which would complicate your coding for some questions without teaching you much.

from 18 to the maximum age in the data set, plot the fraction of arrestees whose race would be correctly classified by this rule.

c. (5) Similarly, black arrestees (as a population) have more priors than white arrestees. Plot the accuracy of a rule which guesses arrestees with $\geq t$ priors are black and others are white, as a function of t. (You'll need to work out the range of suitable thresholds.)

d. (5) Similarly, black arrestees (as a population) have higher COMPAS scores. Consider a rule which guesses an arrestee is black if their COMPAS score $\geq t$, and white otherwise. Plot the fraction of arrestees who would be correctly classified for all thresholds t from 1 to 11.

e. (3) How reliably can an arrestee's race be inferred from their age? From their priors? From their COMPAS score? Explain in words, referring to the plots you drew in Q1a–1d.

f. (5) Consider the claim that "Predicting recidivism from age is just a disguised way of predicting recidivism from race". Give one reason in favor of this statement, and one against, based on Q1a–1e.

g. (5) Do the same for claim that "Predicting recidivism from the number of priors is just a disguised way of predicting recidivism from race".

h. (5) Do the same for the claim that "Predicting recidivism form COMPAS scores is just a disguised way of predicting recidivism from race".

2. **Accuracy and Error Rates of COMPAS** Suppose we predict recidivism for everyone whose COMPAS score reaches some threshold t, so $\hat{Y} = 1$ if $COMPAS \geq t$ and $\hat{Y} = 0$ otherwise. Since the scores are integers from 1 to 10, $t = 1$ would predict recidivism for everyone, and $t = 11$ would predict recidivism for no one.

   a. (5) **Accuracy** Plot the classification accuracy of the COMPAS score as a function of the threshold t. Include a horizontal line showing the baseline accuracy which we could achieve by predicting the same label for everyone. For what thresholds (if any) does COMPAS improve on this baseline? (There should be 11 points on this plot.)

   b. (5) **FNR vs. FPR** Plot the false negative rate (on the vertical axis) against the false positive rate (on the horizontal axis). (Again, there should be 11 points on the plot.) If instead of using the COMPAS score or any other information, we *randomly* labeled a fraction p of arrestees as violent, we'd have an FPR of p and an FPR of $1 - p$. Include a line showing the performance of this baseline randomized classifier. Describe the trade-off COMPAS makes between the two types of error. Is there evidence here that COMPAS is better than random labeling?

3. **Calibration of COMPAS**

   a. (4) For each level (1–10) of the COMPAS score, find the actual frequency of recidivism, i.e., what fraction of arrestees with that score were, in fact, violent recidivists. Do this for (i) blacks, (ii) whites and (iii) both together. Plot the results. (Ideally, you should have one plot with three curves, but three plots with one curve each will get partial credit.)

   b. (3) Repeat you plot from Q3a, but now add suitable error bars of ± 2 standard errors to all your estimated proportions. *Hints*: (i) If n trials each have success probability p, successes are independent across trials, and we observe x total successes, we can estimate $\hat{p} = x/n$, with approximate standard error $\sqrt{\hat{p}(1-\hat{p})/n}$. (What's "success" here? What's n?) (ii) `segments()` may be helpful for drawing.

   c. (4) Does the COMPAS score appear to be calibrated, or equally calibrated for both blacks and whites? Justify your answer by referring to what you found in Q3a and Q3b.

4. **Disparity in COMPAS**

   a. (5) Predictions/decisions have **demographic parity** when the fraction of positive predictions is the same across groups. For races, this would mean that $P(\hat{Y} = 1|\text{Race})$ is the same across races. Plot the fraction of arrestees with $\hat{Y} = 1$ as a function of threshold for (i) blacks alone,

(ii) whites alone, and (iii) everyone. At what thresholds does COMPAS come closest to (or reach) demographic parity?

b. (5) Predictions have **parity of predictive accuracy** when they are equally accurate for different groups in the population. Re-do your plot of accuracy against threshold from Q2a, showing separate curves for whites, for blacks, and for the whole population. At what thresholds does COMPAS come closest to (or achieve) parity of predictive accuracy?

c. (5) Predictions have **parity of error rates** when error rates are equal across different groups in the population. Make a plot of false positive rates against threshold, showing separate curves for whites, for blacks, and for the whole population. At what thresholds does COMPAS come closest to (or achieve) parity of false positives?

d. (5) Define the **FPR difference** as the false positive rate for blacks *minus* the false positive rate for whites. Make a plot showing the FPR difference against the accuracy as the threshold varies. Describe the trade-off, if any, between parity and accuracy.

e. (5) We have assumed that if we use the COMPAS score, we need to apply the *same* threshold t to both whites and blacks. If we allowed there to be different thresholds for the two groups, could we achieve parity of false positive rates? If not, explain why not. If so, what would the common false positive rate be, what would the false negative rates be, and what would the accuracies be?

5. (7) **Advising Riverdale** Suppose that Riverdale County is considering adopting COMPAS, and that you have been hired by a member of the county council to advise them about this decision. (You can assume that Riverdale County, while fictional, is otherwise very similar to Broward County, where the data come from.) Summarize what you have learned from this analysis about the ways in which COMPAS is or is not accurate and fair. Give an argument in favor of using COMPAS, an argument for using a different model instead of COMPAS, and an argument against using statistical model at all.

6. (1) **Timing** How long, roughly, did you spend on this assignment?

Presentation 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. 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 plots and tables are generated by code included in the R Markdown file. 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. All parts of all problems are answered with actual coherent sentences, and raw computer code or output are only shown when explicitly asked for. Text from the homework assignment, including this rubric, is included only when relevant, not blindly copied.

Extra Credit

A (10)

The main problems have asked you to look at whether COMPAS is fair across races. We can also ask about whether it is fair across sexes. Re-do the parts of Q1, Q3 and Q4 which called for racial comparisons to look at the disparity between the sexes.

References