INFO370 Problem Set: Is the model fair?

March 5, 2023

Introduction

This problem set has the following goals:

1. Use confusion matrices to understand a recent controversy around racial equality and criminal justice system.
2. Use your logistic regression skills to develop and validate a model, analogous to the proprietary COMPAS model that caused the above-mentioned controversy.
3. Give you some hands-on experience with typical machine learning workflow, in particular model selection with cross-validation.
4. Encourage you to think over the concept of fairness, and the role of statistical tools in the policymaking process.

Background

Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) algorithm is a commercial risk assessment tool that attempts to estimate a criminal defendents recidivism (when a criminal reoffends, i.e. commits another crime). COMPAS is reportedly one of the most widely used tools of its kind in the U.S. It is often used in the US criminal justice system to inform sentencing guidelines by judges, although specific rules and regulations vary.

In 2016, ProPublica published an investigative report arguing that racial bias was evident in the COMPAS algorithm. ProPublica had constructed a dataset from Florida public records, and used logistic regression and confusion matrix in its analysis. COMPAS's owners disputed this analysis, and other academics noted that for people with the same COMPAS score, but different races, the recidivism rates are effectively the same. Even more, as Kleinberg *et al.* (2016) show, these two fairness concepts (individual and group fairness) are not compatible. There are also some discussion included in the lecture notes, ch 12.2.3 (admittedly, the text is rather raw).

The COMPAS algorithm is proprietary and not public. We know it includes 137 features, and deliberately excludes race. However, another study showed that a logistic regression with only 7 of those features was equally accurate!

Note: Links are optional but very helpful readings for this problem set!

Dataset

The dataset you will be working with, *compas-score-data*, is based off ProPublicas dataset, compiled from public records in Florida. However, it has been cleaned up for simplicity. You will only use a subset of the variables in the dataset for this exercise:

age Age in years

c_charge_degree Classifier for an individuals crime–*F* for felony, *M* for misdemeanor

race Classifier for the recorded race of each individual in this dataset. We will only consider *Caucasian*, and *African-American* here.

age_cat Classifies individuals as under 25, between 25 and 45, and older than 45

sex "Male" or "Female".

priors_count Numeric, the number of previous crimes the individual has committed.

decile_score COMPAS classification of each individuals risk of recidivism (1 = low . . . 10 = high). This is the score computed by the proprietary model.

two_year_recid Binary variable, 1 if the individual recidivated within 2 years, 0 otherwise. This is the central outcome variable for our purpose.

Note that we limit the analysis with the time period of two years since the first crime—we do not consider re-offenses after two years.

1 Is COMPAS fair? (50pt)

The first task is to analyze fairness of the COMPAS algorithm. As the algorithm is proprietary, you cannot use this to make your own predictions. But you do not need to predict anything anyway–the COMPAS predictions are already done and included as *decile_score* variable!

1.1 Load and check (2pt)

Your first tasks are the following:

1. (1pt) Load the COMPAS data, and perform the basic checks.
2. (1pt) Filter the data to keep only Caucasian and African-Americans.

   All the tasks below are about these two races, there are just too few other offenders.

1.2 Aggregate analysis (20pt)

COMPAS categorizes offenders into 10 different categories, starting from 1 (least likely to recidivate) till 10 (most likely). But for simplicity, we scale this down to two categories (low risk/high risk) only.

1. (2pt) Create a new dummy variable based off of COMPAS risk score (*decile_score*), which indicates if an individual was classified as low risk (score 1-4) or high risk (score 5-10).

   Hint: you can do it in different ways but for technical reasons related the tasks below, the best way to do it is to create a variable "high score", that takes values 1 (decile score 5 and above) and 0 (decile score 1-4).

2. (4pt) Now analyze the offenders across this new risk category:

   (a) What is the recidivism rate (percentage of offenders who re-commit the crime) for low-risk and high-risk individuals?

   (b) What are the recidivism rates for African-Americans and Caucasians?

   Hint: 39% for Caucasians.

3. (7 pt) Now create a confusion matrix comparing COMPAS predictions for recidivism (low risk/high risk you created above) and the actual two-year recidivism and interpret the results. In order to be on the same page, let's call recidivists "positives".

   Note: you do not have to predict anything here. COMPAS has made the prediction for you, this is the *high risk* variable you created in 1. See the referred articles about the controversy around COMPAS methodology.

   Note 2: Do not just output a confusion matrix with accompanying text like "accuracy = x%, precision = y%". Interpret your results such as "z% of recidivists were falsly classified as low-risk, COMPAS accurately classified k% of individuals, etc."

4. (7pt) Find the accuracy of the COMPAS classification, and also how its errors (false negatives and false positives) are distributed–compute precision, recall, false positive rate and false negative rate.

   We did not talk about *FPR* and *FNR* in class, but you can consult Lecture Notes, section 6.1.1 *Confusion matrix and related concepts*.

   Would you feel comfortable having a judge to use COMPAS to inform sentencing guidelines? What do you think, how well can judges perform the same task without COMPAS's help? At what point would the error/misclassification risk be acceptable for you? Do you think the acceptable error rate should be the same for human judges and for algorithms?

   Remember: human judges are not perfect either!

1.3 Analysis by race (28pt)

1. (2pt) Compute the recidivism rate separately for high-risk and low risk African-Americans and Caucasians.

   Hint: High risk AA = 65%.

2. (6pt) Comment the results in the previous point. How similar are the rates for the the two race groups for low-risk and high-risk individuals? Do you see a racial disparity here? If yes, which group is it favoring? Based on these figures, do you think COMPAS is fair?

3. (6pt) Now repeat your confusion matrix calculation and analysis from 3. But this time do it separately for African-Americans and for Caucasians:

   (a) How accurate is the COMPAS classification for African-Americans and for Caucasians?

   (b) What are the false positive rates (false recidivism rates) *FPR*?

   (c) The false negative rates (false no-recidivism rates) *FNR*?

   Hint: FPR for Caucasians is 0.22, FNR for African-Americans is 0.28.

4. (6pt) If you have done this correctly, you will find that COMPAS's percentage of correctly categorized individuals (accuracy) is fairly similar for African-Americans and Caucasians, but that false positive rates and false negative rates are different. In your opinion, is the COMPAS algorithm "fair"? Justify your answer.

5. (8pt) Does your answer in 4 align with your answer in 2? Explain!

   Hint: This is not a trick question. If you read the first two recommended readings, you will find that people disagree how you define fairness. Your answer will not be graded on which side you take, but on your justification.

---

2 Can you beat COMPAS? (50pt)

COMPAS model has created quite a bit controversy. One issue frequently brought up is that it is "closed source", i.e. its inner workings are not available neither for public nor for the judges who are actually making the decisions. But is it a big problem? Maybe you can devise as good a model as COMPAS to predict recidivism? Maybe you can do even better? Let's try!

2.1 Create the model (30pt)

Create such a model. We want to avoid explicit race and gender bias, hence you do *not* want to include gender and race in order to avoid explicit race/gender bias. Finally, let's analyze the performance of the model by cross-validation.

More detailed tasks are here:

1. (6pt) Before we start: what do you think, what is an appropriate model performance measure here? A, P, R, F or something else? Maybe you want to report multiple measures? Explain!

2. (6pt) you *should not* use variable `decile score` *that originates from COMPAS model*. Why?

3. (8pt) Now it is time to do the modeling. Create a logistic regression model that contains all explanatory variables you have in data into the model. (Some of these you have to convert to dummies). Do *not* include the variables discussed above, *do not* include race and gender in this model either to avoid explicit gender/racial bias.

   Use 10-fold CV to compute its relevant performance measure(s) you discussed above.

4. (10pt) Experiment with different models to find the best model according to your performance indicator. Try trees and k-NN, you may also include other types of models. Include/exclude different variables. You may also do feature engineering, e.g. create a different set of age groups, include variables like age^2, age^2, interaction effects, etc. But do not include race and gender.

   Report what did you try (no need to report the full results of all of your unsuccessful attempts), and your best model's performance. Did you got better results or worse results than COMPAS?

2.2 Is your model more fair? (20pt)

Finally, is your model any better (or worse) than COMPAS in terms of fairness? Let's use your model to predict recidivism for everyone (i.e. all data, ignore training-testing split), and see if you managed to FPR and FNR for African-Americans and Caucasians are now similar.

1. (6pt) Now use your model to compute the two-year recidivism rates by race and your risk prediction (replicate 1.3-1). Is your model more or less fair than COMPAS?
2. (6pt) Compute FPR and FNR by race (replicate 1.3-3 the FNR/FPR question). Is your model more or less fair than COMPAS?
3. (8pt) Explain what do you get and why do you get it.

Finally tell us how many hours did you spend on this PS.

References

Kleinberg, J., Mullainathan, S. and Raghavan, M. (2016) Inherent trade-offs in the fair determination of risk scores, Tech. rep., arXiv.