INFO370 Problem Set 8: Applied Modeling

December 4, 2020

Introduction

This problem set has three 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. Encourage you to think over the role of statistical tools and AI in our policymaking process.

1 Is COMPAS fair? (60pt)

Background

Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) algorithm is a commercial risk assessment tool that attempts to estimate a criminal defendent's 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 US. 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.

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 is based off ProPublica's 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:

c_charge_degree Classifier for an individual's crime–F for felony, M for misdemeanor

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

age Age in years

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

sex Classifier for the recorded sex of each individual in this dataset. Male or female.

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

decile_score COMPAS classification of each individual's risk of recidivism (1 = low . . . 10 = high).

score_text COMPAS classification of each individual's risk of recidivism (text: Low, Medium, or High)

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

1. (2pt) Load the COMPAS data, and perform a basic sanity checks.

2. (2pt) Filter the data to keep only only Caucasian and African-Americans.

3. (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).

4. (6pt) Now analyze the offenders across this new risk category:

   (a) What is the recidivism rate for low-risk and high-risk individuals?

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

5. (10 pt) Now create a confusion matrix comparing COMPAS predictions for recidivism (low risk/high risk) and the actual two-year recidivism and interpret the results.

   Note: you do not have to predict anything here. COMPAS has made the prediction for you, this is the variable you created in 3 based on `decile_score`. 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 N% of individuals, etc."

6. (12pt) Note the accuracy of the COMPAS classification, and also how its errors were distributed. Would you feel comfortable having a judge to use COMPAS to inform sentencing guidelines? At what point would the error/misclassification risk be acceptable for you?

   Remember: human judges are also not perfect!

7. (14pt) Now, you repeat your confusion matrix calculation and analysis from 5. But this time do it separately for African-Americans and for Caucasians:

   (a) How accurate is the COMPAS classification for African-American individuals? For Caucasians?

   (b) What are the false positive rates $\mathsf{FPR} = \mathsf{FP}/\mathsf{N} = \mathsf{FP}/(\mathsf{FP} + \mathsf{TN})$?

   (c) The false negative rates $\mathsf{FNR} = \mathsf{FN}/\mathsf{P} = \mathsf{FN}/(\mathsf{FN} + \mathsf{TP})$?

8. (12pt) If you have done this correctly, you will find that COMPAS's true negative and true positive percentages are similar for African-American and Caucasian individuals, but that false positive rates and false negative rates are different. Look again at the overal recidivism rates in the dataset for Black and White individuals. In your opinion, is the COMPAS algorithm 'fair'? Justify your answer.

   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? (40pt)

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!

We proceed as follows:

- Note that you *should not* use variables *that originate from COMPAS model* (i.e. `score_text` and `decile_score`). You should also not use the outcome variables (`two_year_recid` and `is_recid`). Do you see why?
- First we devise a model that explicitly does *not* include gender and race. Your task is to use cross-validation to develop the best model you can develop based on the available variables.
- Thereafter we add gender and see if gender improves the model performance.
- And finally we also add race and see if race has an additional explanatory effect, i.e. does race help to improve the performance of the model.

1. (8pt) 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) Now it is time to do the modeling. Create a logistic regression model that contains all the (permissible) 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.

   We also *do not* include race and gender in this model to avoid explicit gender/racial bias.

   Use 10-fold CV to compute its relevant performance measure(s) you discussed above.
3. (6pt) Experiment with different models to find the best model according to your preformance indicator. (Include/exclude different variables, you may also do feature engineering, and try other model categories). Still do not include race and gender.

   Report your best model's performance. Is it better or worse than for the COMPAS model?
4. (4pt) Now add *sex* to the model. Does it help to improve the performance?
5. (4pt) And finally add *race*. Does the model improve?
6. (12pt) Discuss the results. Did you manage to beat COMPAS? Do gender and race help to improve your predictions? What should judges do when having access to such models? Should they use such models?

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