Homework 7: COMPAS

36-402, Spring 2025

Due at 6 pm on Thursday, **13 March 2025**

AGENDA: Practice with logistic regression (again), with classification trees, and with evaluating deliberately-obscure black-box classifiers.
READING: Chapter 13 of the textbook, on tree models.
TIMING: This looks long but it's pretty repetitive.

Our data set this week comes from the analysis, performed by the news organization ProPublica, of the "COMPAS" risk prediction scores as used in Broward County[1], Florida. COMPAS is a a complicated and propietary algorithm, developed and sold by a company called NorthPointe to local governments across America, which is used to assess the likelihood that people who have been arrested will commit violent crimes if released before their trial[2]. The company does not say exactly how COMPAS works, just that it's a statistical model based on over 100 features.

Specifically, our data file tracks the following information, for (supposedly) everyone who was arrested in Broward County and given a COMPAS score in the years 2013 and 2014:

- Their age;
- Their sex;
- Their race;
- Their COMPAS score for risk of violence (1–10);
- Whether they were charged with a felony (F) or misdemeanor (M)[3]
- Count of prior convictions (not just arrests), or "priors";
- Whether they had a subsequent arrest (not necessarily conviction) for violence in Florida within two years, or "recidivism".

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

[2] COMPAS actually calculates separate scores for risk of "failure to appear" at trial, risk of committing any type of crime if released, and risk of violence if released. This data set only contains the score for risk of violence.

[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 cases, death. "Misdemeanors" are punishable by shorter terms of imprisonment (typically in city or county jails rather than state 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.

You will construct your own models for predicting recidivism. Some questions will ask you about the models' predictions about the following (hypothetical) individuals, arrested after a brawl at the Riverdale Diner

- Archie, a 19 year old white male with one prior, charged with a felony.
- Betty, a 22 year old white female with two priors, charged with a misdemeanor.
- Chuck, a 34 year old black male with no priors, charged with a misdemeanor.
- Veronica, a 42 year old Hispanic female with 12 priors, charged with a felony.

Notation: Y stands for actual recidivism, $\hat{Y}$ for predicted recidivism based on some model or score.

1. *Data preparation* The file `testing_set` contains the row numbers of a randomly-chosen 20% of the data points, which you will use as (what else?) a testing set. The other 80% of the rows will be your training set. In all subsequent problems, fit all models on the training set. If you need to do any cross-validation, do it *within* the training set. Whenever you evaluate prediction accuracy or quality, use the testing set.
   (a) (2) In your own words, why is it important to evaluate predictions *only* on the testing set?
   (b) (2) In your own words, why is it important to *randomly* divide the data?
   (c) (1) What proportion of people in the testing set are re-arrested for violence within two years?
2. *Logistic regression, estimated* Fit a logistic regression to predict recidivism from age and priors. (Remember to only use the training set from Q1 here.)
   (a) (3) Report the coefficients, including 95% confidence intervals, in a table. For full credit, the confidence interval should be calculated using resampling of rows; if you cannot get that to work, R's default calculation will get you partial credit.
   (b) (3) Give an interpretation, in words, of the coefficients. (Imagine describing them to a judge who does not know a lot of statistics.)
   (c) (4) What predicted probabilities of recidivism does the model give for each of the 4 arrestees? What are 95% confidence intervals for those probabilities? (Again, use bootstrap resampling for full credit, R's defaults for partial credit.)
3. *Logistic regression, evaluated* Evaluate the model from Q2 on the testing set from Q1. Classify people in the testing as a predicted recidivists, $\hat{Y} = 1$, if their probability of recidivism is $\geq t$, for a range of thresholds $t = 0, 0.01, 0.02 \ldots 1.0$ (that is, 101 thresholds in total).

(a) (3) Plot the accuracy, defined to be $\mathbb{P}\left(Y = \hat{Y}\right)$, as a function of the threshold t. (There should be 101 points in your plot.) At what threshold is accuracy maximized? Add a horizontal line showing the accuracy of a model which always returns the *same* value of $\hat{Y}$. (Which value should it return, 0 or 1?) Include this horizontal line in all later plots of accuracy vs. threshold for other models.

(b) (2) Plot the false positive rate or FPR, $\equiv \mathbb{P}\left(\hat{Y} = 1 | Y = 0\right)$, as a function of t. Where is FPR minimized?

(c) (2) Plot the false negative rate or FNR, $\equiv \mathbb{P}\left(\hat{Y} = 0 | Y = 1\right)$, as a function of t. Where is FNR minimized?

(d) (4) Plot FNR (vertical axis) versus FPR (horizontal). (Again, there should be 101 points in this plot.) Explain why, if someone *randomly* set $\hat{Y} = 1$ with probability p, they would have an FPR of p and an FNR of $1 - p$, and add the corresponding diagonal line to your plot, and to all later plots of FNR vs. FPR. Does your logistic regression add any predictive value over the random procedure? Explain, by referring to your plot.

Hint: You'll find it helpful to write one function which takes in t and returns accuracy, FPR and FNR, and then apply it to a vector of thresholds.

4. *GAM, estimated* Fit a generalized additive model for the same variables as in Q2.

   (a) (1) Plot the partial response functions. (If you've done this right, they should be obviously nonlinear.) Bootstrapping is not required here.

   (b) (3) Give an interpretation, in words, of the partial response functions.

   (c) (3) What predicted probabilities of recidivism does the model give for each of the 4 arrestees? What are 95% confidence intervals for those probabilities? (Again, use bootstrap resampling for full credit, R's defaults for partial credit.)

5. *GAM, evaluated* Repeat (most of) Q3 for this model.

   (a) (1) Plot accuracy versus threshold t. Where is accuracy of the GAM maximized?

   (b) (2) Plot FNR vs FPR.

   (c) (4) Is the GAM a better classifier than the logistic regression? Explain, by referring to the plots in this question and Q3.

   Hint: Write (or revise) the function in Q3 so it works with a vector of predicted probabilities that could come from any model.

6. *Tree, fitted* Fit a classification tree to predict recidivism, using all the variables *except* the COMPAS score. Use the 'tree' package with its default settings. (Make sure you fit a classification and not a regression tree.)

(a) (2) Give a plot of the tree, labeled to show the splits at each interior node and the predictions at each leaf node.

(b) (2) Which variables does the tree actually use to make its predictions?

(c) (4) Describe, in words, how the tree you have fit works. (If you have done this properly, the tree is quite small and easy to summarize in natural language.)

(d) (4) What predicted probabilities of recidivism does the tree give to each of the four people arrested at the diner? For full credit, include 95% confidence intervals for those probabilities from bootstrap; if you cannot figure out how to get that to work, you can get partial credit using the standard formula for a confidence interval for a proportion.

Hint: The sampling distributions here can be very multi-modal, so it's safer to use the "quantile" bootstrap than the basic, "pivotal" one we usually use. (See chapter 6 for a discussion.) The function `bootstrap.ci-revised.R` on the class website will let you make the switch rather painlessly.

7. *Tree, evaluated* Repeat (most of) Q3 and Q5 for this model.

(a) (1) Plot accuracy versus threshold t. (This plot should have 101 points.)

(b) (2) Plot FNR vs FPR. (This plot should have many fewer than 101 points — why?)

(c) (3) Is the tree a better classifier than the logistic regression? Than the GAM? Refer to the plots in Q3, Q5 and Q7 to back up your answer.

8. *COMPAS, evaluated* The COMPAS score is a number from 1 to 10; it is not supposed to be a probability estimate, but increasing scores are supposed to indicate increasing risk of violence.

(a) (2) Set $\hat{Y} = 1$ if the COMPAS score is $\geq t$. Using thresholds $t = 1, 2, \ldots 11$, plot accuracy versus t. Why does this one need to go to eleven?

(b) (1) Plot FNR versus FPR. (Again, this plot should have 11 points.)

(c) (3) Does COMPAS work better as a classifier than any of the models you have built? Refer to your earlier results to justify your answer.

9. *Calibration*

(a) (3) Following Chapter 11, and the solutions to Homework 6, make calibration plots for the logistic regression, the GAM, and the tree. (Be sure you are evaluating calibration on the testing set.) — Either three separate plots, or one combined plot, is OK here, whichever you are able to do more clearly.

(b) (2) Are any of these three models notably better, or worse, calibrated than the others?

(c) (1) For each level of the COMPAS score, calculate the frequency of recidivism, and plot that frequency versus the score. For full credit, add ± 2 standard error bars, and explain your calculation of the standard errors.

(d) (2) Explain why the plot you have just made for COMPAS is not a calibration plot in the same sense as previous ones.

(e) (2) Explain why the plot for COMPAS should be monotonically increasing, if the COMPAS score is working as it is supposed to. Is it?

10. *Residuals*

(a) (2) Consider any binary, 0-1 valued random variable Y, and any other variable X, with $\mu(x) = \mathbb{E}[Y|X = x]$ as usual. Define $R = \frac{Y-\mu(X)}{\sqrt{\mu(X)(1-\mu(X))}}$. Show that $\mathbb{E}[R|X = x] = 0$ and $\mathrm{Var}[R|X = x] = 1$ for all x.

(b) (2) Suppose you have a model which makes a probability prediction for a binary Y. Explain how to calculate a residual which should always have mean 0 and variance 1, if the model is correct.

(c) (2) Calculate residuals for your logistic regression, your GAM, and your tree. Plot them against the predicted probabilities, and use a smoothing spline to add an estimate of the conditional mean residual function. (Be sure to do all this using the testing set.) For which models, if any, are the residuals (nearly) flat around 0? (Produce one plot with three sets of points, or three plots with one set of points each, whichever you can make clearer.)

(d) (2) Plot the squared residuals for your three models against the predicted probabilities, and use a smoothing spline to add an estimate of the conditional mean squared residual function. For which, if any, are the squared residuals (nearly) flat around 1? (Again, 3 plots or 1 plot, which ever you can make clearer.)

(e) (1) Explain why you cannot make a residual plot of this sort for COMPAS.

Hint: In-class exercises.

11. *Recommendations* (6) The government of Riverdale County wants to adopt a statistical model to screen arrestees for pre-trial release. If you had to pick one model from among the three models you estimated, which would you recommend to the county government, and why? If it was a choice between one of these three models and COMPAS, what would you recommend, and why? You can assume that Riverdale County is very similar to Broward County, where the data were gathered. (For full credit, refer to specific findings in earlier questions.)

12. (1) 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.

(In Gradescope, assign *all* pages to this rubric.)

EXTRA CREDIT: DISPARITIES It is often important, if controversial, to know whether statistical methods or algorithms work equally well for different social categories. Differences in outcomes or performance across categories are often called "disparities".

1. *Disparities by race* Consider a binary division of race, into blacks/African Americans and everyone else[4].

   (a) (5) For each of your three models, and for COMPAS, make plots of FNR versus FPR, calculating error rates for (i) blacks alone, and (ii) non-blacks alone. Does any model, at any threshold, achieve equal error rates across racial groups? If not, how close (or far) do they come?

   (b) (5) Repeat your calibration plots for your three models, and the not-quite calibration plot for COMPAS, but now calculate frequency separately for (i) blacks and (ii) non-blacks. Which models, if any, are equally calibrated across racial groups?

2. *Disparities by sex* Now do the same analysis but contrast female versus male, instead of black versus non-black.

   (a) (5) As in EC1a.

   (b) (5) As in EC1b.

[4]The COMPAS data, relying on race as assessed by the police, classifies arrestees as "African American", "Asian", "Caucasian" (i.e., white), "Hispanic", "Native American" or "Other". There are so few Asian and Native American arrestees that doing separate analyses for them doesn't tell us much, while the distributions for "Hispanic" and "Other" are actually quite close to those for "Caucasian", and all three are obviously different from the distributions for "African American".