Lab 3 –Tables and Answers

2. Building the classifiers:

Build a classifier (‘Test’) identifying your injury of interest using either diagnosis or body part as your ‘Truth’. \*Other variables may be used to define your case population. The classifier must use at least three terms from the narrative field to start. Structured data may also be used.

**Classifier 1:**

**I started with terms that appeared in the random sampling most often and that I thought would appear for basketball ankle injuries. The three terms in the first classifier are ‘SPRAIN’, ‘STRAIN’, and ‘STRAINED’. I also restricted my data set to high school age kids between 14-19.**

data basketball; set injuries;

where prod1=1205 AND (age > 14 AND age < 19);

if index(narrative,'SPRAIN') > 0 then narr\_sprain=1;

else narr\_sprain=0;

if index(narrative,'STRAIN') > 0 then narr\_basketball=1;

else narr\_basketball=0;

if index(narrative,'STRAINED')>0 then narr\_strained=1;

else narr\_strained=0;

if bdpt=37 then baketball\_ankles=1;

else baketball\_ankles=0;

run;

**Classifier 2:**

**Building on the first classifier, I added another term that appeared often for ankle injuries, but less than the first three by adding the term ‘TWISTED’.**

data basketball; set injuries;

where prod1=1205 AND (age > 14 AND age < 19);

if index(narrative,'SPRAIN') > 0 then narr\_sprain=1;

else narr\_sprain=0;

if index(narrative,'STRAIN') > 0 then narr\_basketball=1;

else narr\_basketball=0;

if index(narrative,'STRAINED')>0 then narr\_strained=1;

else narr\_strained=0;

if index(narrative,'TWISTED')>0 then narr\_twisted=1;

else narr\_twisted=0;

if bdpt=37 then baketball\_ankles=1;

else baketball\_ankles=0;

run;

**Classifier 3:**

**I changed the "TWISTED" term to "FX" (fracture) to see the difference on AUC each narrative term has independently on the classifier.**

data basketball; set injuries;

where prod1=1205 AND (age > 14 AND age < 19);

if index(narrative,'SPRAIN') > 0 then narr\_sprain=1;

else narr\_sprain=0;

if index(narrative,'STRAIN') > 0 then narr\_basketball=1;

else narr\_basketball=0;

if index(narrative,'STRAINED')>0 then narr\_strained=1;

else narr\_strained=0;

if index(narrative,'FX')>0 then narr\_fx=1;

else narr\_fx=0;

if bdpt=37 then baketball\_ankles=1;

else baketball\_ankles=0;

run;

**Classifier 4:**

**I added both 'FX' and 'TWISTED' from Classifier 2 and 3 to see is there was a compound effect and there was.**

**With 'FX' the AUC was 0.8363 while 'TWISTED' had an AUC of 0.8630, together they slightly improve the Classifier 4 AUC to 0.8653.**

data basketball; set injuries;

where prod1=1205 AND (age > 14 AND age < 19);

if index(narrative,'SPRAIN') > 0 then narr\_sprain=1;

else narr\_sprain=0;

if index(narrative,'STRAIN') > 0 then narr\_basketball=1;

else narr\_basketball=0;

if index(narrative,'STRAINED')>0 then narr\_strained=1;

else narr\_strained=0;

if index(narrative,'FX')>0 then narr\_fx=1;

else narr\_fx=0;

if index(narrative,'TWISTED')>0 then narr\_twisted=1;

else narr\_twisted=0;

if bdpt=37 then baketball\_ankles=1;

else baketball\_ankles=0;

run;

**Classifier 5:**

**This Classifier (Classifier 5) included all the narrative terms from Classifier 4 and included the term ‘FRACTURE’ as well and was the most complex Classifier I built.**

**This is also the best classifier as it has the largest AUC and smallest ASE**

data basketball; set injuries;

where prod1=1205 AND (age > 14 AND age < 19);

if index(narrative,'SPRAIN') > 0 then narr\_sprain=1;

else narr\_sprain=0;

if index(narrative,'STRAIN') > 0 then narr\_basketball=1;

else narr\_basketball=0;

if index(narrative,'STRAINED')>0 then narr\_strained=1;

else narr\_strained=0;

if index(narrative,'FX')>0 then narr\_fx=1;

else narr\_fx=0;

if index(narrative,'TWISTED')>0 then narr\_twisted=1;

else narr\_twisted=0;

if index(narrative,'FRACTURE')>0 then narr\_fracture=1;

else narr\_fracture=0;

if bdpt=37 then baketball\_ankles=1;

else baketball\_ankles=0;

run;

4. Using 50/50 test/train, 5-fold cross validation, 10-fold cross validation, and at least 100 bootstrap sets, calculate the prevalence, sensitivity, specificity, positive predictive value, and negative predictive value each time.

A screenshot of a cell phone

Description automatically generatedThe following are the results for the 5 CV Fold (top) and 10 CV Fold (bottom). For the 5 CV the sensitivity is 0.6840 and the specifity is 0.8522. For the 10 CV is the sensitivity is 0.6769 and the specificity is 0.8542.

A screenshot of a cell phone

Description automatically generated

A screenshot of a cell phone

Description automatically generatedThe following is my bootstrap which produced my PPV and NPV. The PPV properly classified 2464 cases as 1’s (people with injured ankles) and misclassified 863 people. The NPV properly classified 8296 cases as 0’s (people who do not have injured ankles) and misclassified 1400 people.

This table shows the prevalence of ankle injuries from all basketball injuries in my dataset. Of the 9696 basketball injuries, 3327 of them are for the ankle, translating to 25.55% of data being ankle injuries.

A screenshot of a cell phone

Description automatically generated

5. To rank my models, I am going to use the ACU from the 10 CV Fold chart for each of the 5 classifiers. From best to worst, according to the AUC’s, the Classifiers rank as 5 (0.8674), 4 (0.8653), 2 (0.8630), 3 (0.8363), 1 (0.8350).

A screenshot of a cell phone

Description automatically generatedA screenshot of a cell phone

Description automatically generatedA screenshot of a cell phone

Description automatically generatedA screenshot of a cell phone

Description automatically generated

A screenshot of a cell phone

Description automatically generated