**\*\*\* The purpose of this file is to provide my SAS code in a format it can be coded on, the “comments” for the code have been edited for readability. \*\*\***

libname neisshw '/home/u36241478/my\_courses/nicholassoulakis0/lib/neiss';

proc format

cntlin=neisshw.neiss\_fmt;

run;

title 'Frequency of the NEISS formats';

title2 'Note: PROD N=1,115';

proc freq data=neisshw.neiss\_fmt;

tables fmtname;

run;

**/\* Adding day of week format to see if injury is more likely on a particular day \*/**

proc format;

value dowf

1 = 'Sunday'

2 = 'Monday'

3 = 'Tuesday'

4 = 'Wednesday'

5 = 'Thursday'

6 = 'Friday'

7 = 'Saturday';

run;

**/\* Combine three years of NEISS data \*/**

**/\* Apply formats for all revelant variables \*/**

**/\* Combine NARR1 and NARR2 fields into one with CAT function \*/**

**/\* Create date variable with WEEKDAY,WEEK,MONTH,YEAR functions \*/**

**/\* Note: Sunday=1, Saturday=7 in the WEEKDAY function \*/**

**/\* Create a random number to pick a selection of injuries. \*/**

data injuries; set neisshw.neiss2015 neisshw.neiss2016 neisshw.neiss2017;

drop fmv race raceoth diagoth;

format sex gender.;

format race race.;

format bdpt bdypt.;

format diag diag.;

format disp disp.;

format loc loc.;

format dow dowf.;

format prod1 prod.;

narrative=cat(narr1,narr2);

dow=weekday(trmt\_date);

week=week(trmt\_date);

month=month(trmt\_date);

year=year(trmt\_date);

if dow in (1,7) then we=1; else we=0;

if dow in (2,3,4,5,6) then wd=1; else wd=0;

call streaminit(4567);

randno=rand('uniform');

run;

**I used a random sampling query to see what terms come up most and used that to guide what narrative terms I used to build my classifiers**

title 'Random Selection of Basketball ankle Injuries Narratives';

proc sql outobs=25;

select randno, diag label='Diagnosis', prod1, narr1, narr2

from injuries

where bdpt=37 and prod1=1205

order by randno;

quit;

title 'Distribution of Head injuries by Sampling Stratum';

title2 'Note: National Estimate is the Sum of the wt Variable';

proc sql outobs=25;

select stratum label='Sampling Stratum',

count(\*) as freq label='Frequency',

sum(wt) as sum\_wt label='National Estimate'

from injuries

where bdpt=37 and prod1=1205 /\* bdpt=37 diag=64\*/

group by stratum

order by freq desc;

quit;

**CREATING 5 CLASSIFIERS**

**The area under the curves (AUC) for 5 and 10 Fold CV's of the model with the same variables were the same, however other metrics such as the average square error (ASE) would be different when comparing the two. The ASE for the 10-fold CV was lower.**

**I included the AUC's for Model 5 so you can compare different AUC's for each model**

**How I chose which model was the best: the model with the largest AUC and secondly, the lowest ASE.**

**Classifier 1:**

**Model 1 AUC = 0.7936**

**Model 2 AUC = 0.8042**

**Model 3 AUC = 0.8171**

**Model 4 AUC = 0.8185**

**Model 5 AUC = 0.8350**

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:**

**Model 1 AUC = 0.8220**

**Model 2 AUC = 0.8323**

**Model 3 AUC = 0.8421**

**Model 4 AUC = 0.8437**

**Model 5 AUC = 0.8630**

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:**

**Model 1 AUC = 0.7988**

**Model 2 AUC = 0.8081**

**Model 3 AUC = 0.8207**

**Model 4 AUC = 0.8224**

**Model 5 AUC = 0.8363**

**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:**

**Model 1 AUC = 0.8258**

**Model 2 AUC = 0.8360**

**Model 3 AUC = 0.8459**

**Model 4 AUC = 0.8477**

**Model 5 AUC = 0.8653**

**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:**

**Model 1 AUC = 0.8279**

**Model 2 AUC = 0.8384**

**Model 3 AUC = 0.8484**

**Model 4 AUC = 0.8500**

**Model 5 AUC = 0.8674**

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

**This is also the best classifer 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;

**This code is used to calculate the prevalence of my injury from my dataset. In the result we get that 25.55% of the basketball injuries in the dataset are specifically ankle injuries.**

title 'Frequency of Basketball Ankle Injuries among high school-age population';

proc freq data=basketball;

tables baketball\_ankles;

run;

**/\* 5 CV VALIDATION \*/**

**I started with only the narr\_ variable to build my model.**

title '5 Fold VC Model 1';

title2 'Variables Defining Model 1 Case Population: narr\_';

proc hpsplit data=basketball cvcc cvmodelfit

assignmissing=similar

cvmethod=random (5);

prune reducederror;

id NEK psu wt stratum;

weight wt;

class baketball\_ankles narr\_:;

model baketball\_ankles(event='1') =narr\_:;

output out=bball\_cv05\_model01;

run;

**I added location as I thought it would be a good predictor of injuries (gym, school, etc).**

title '5 Fold VC Model 2';

title2 'Variables Defining Model 2 Case Population: narr\_, loc';

proc hpsplit data=basketball cvcc

cvmodelfit

assignmissing=similar

cvmethod=random (5);

prune reducederror;

id NEK psu wt stratum;

weight wt;

class baketball\_ankles narr\_: loc;

model baketball\_ankles(event='1') =narr\_: loc;

output out=bball\_cv05\_model02;

run;

**I thought males were more likely to injure their ankles than females because in the random sampling there were more YOM’s (year old males) than YOF’s (year old females).**

title '5 Fold VC Model 3';

title2 'Variables Defining Model 3 Case Population: narr\_, loc, sex';

proc hpsplit data=basketball cvcc

cvmodelfit

assignmissing=similar

cvmethod=random (5);

prune reducederror;

id NEK psu wt stratum;

weight wt;

class baketball\_ankles narr\_: loc sex;

model baketball\_ankles(event='1') =narr\_: loc sex;

output out=bball\_cv05\_model03;

run;

**Some people’s disposition might give them a higher propensity for injuries, so I added this variable as well.**

title '5 Fold VC Model 4';

title2 'Variables Defining Model 4 Case Population: narr\_, loc, sex, disp';

proc hpsplit data=basketball cvcc

cvmodelfit

assignmissing=similar

cvmethod=random (5);

prune reducederror;

id NEK psu wt stratum;

weight wt;

class baketball\_ankles narr\_: loc sex disp;

model baketball\_ankles(event='1') =narr\_: loc sex disp;

output out=bball\_cv05\_model04;

run;

**I thought the day of the week might be important as high school games tend to be on certain days of the week so I added the dow day of week variable.**

title '5 Fold VC Model 5';

title2 'Variables Defining Model 5 Case Population: narr\_, loc, sex, disp, dow';

proc hpsplit data=basketball cvcc

cvmodelfit

assignmissing=similar

cvmethod=random (5);

prune reducederror;

id NEK psu wt stratum;

weight wt;

class baketball\_ankles narr\_: loc sex disp dow;

model baketball\_ankles(event='1') =narr\_: loc sex disp dow;

output out=bball\_cv05\_model05;

run;

**/\* 10 CV VALIDATION \*/**

**The same variables were used in the models for 5** **CV and 10 CV.**

title '10 Fold VC Model 1';

title2 'Variables Defining Model 1 Case Population: narr\_';

proc hpsplit data=basketball cvcc cvmodelfit

assignmissing=similar

cvmethod=random (10);

prune reducederror;

id NEK psu wt stratum;

weight wt;

class baketball\_ankles narr\_:;

model baketball\_ankles(event='1') =narr\_:;

output out=bball\_cv10\_model01;

run;

title '10 Fold VC Model 2';

title2 'Variables Defining Model 2 Case Population: narr\_, loc';

proc hpsplit data=basketball cvcc

cvmodelfit

assignmissing=similar

cvmethod=random (10);

prune reducederror;

id NEK psu wt stratum;

weight wt;

class baketball\_ankles narr\_: loc;

model baketball\_ankles(event='1') =narr\_: loc;

output out=bball\_cv10\_model02;

run;

title '10 Fold VC Model 3';

title2 'Variables Defining Model 3 Case Population: narr\_, loc, sex';

proc hpsplit data=basketball cvcc

cvmodelfit

assignmissing=similar

cvmethod=random (10);

prune reducederror;

id NEK psu wt stratum;

weight wt;

class baketball\_ankles narr\_: loc sex;

model baketball\_ankles(event='1') =narr\_: loc sex;

output out=bball\_cv10\_model03;

run;

title '10 Fold CV Model 4';

title2 'Variables Defining Model 4 Case Population: narr\_, loc, sex, disp';

proc hpsplit data=basketball cvcc

cvmodelfit

assignmissing=similar

cvmethod=random (10);

prune reducederror;

id NEK psu wt stratum;

weight wt;

class baketball\_ankles narr\_: loc sex disp;

model baketball\_ankles(event='1') =narr\_: loc sex disp;

output out=bball\_cv10\_model04;

run;

title '10 Fold CV Model 5';

title2 'Variables Defining Model 5 Case Population: narr\_, loc, sex, disp, dow';

proc hpsplit data=basketball cvcc

cvmodelfit

assignmissing=similar

cvmethod=random (10);

prune reducederror;

id NEK psu wt stratum;

weight wt;

class baketball\_ankles narr\_: loc sex disp dow;

model baketball\_ankles(event='1') =narr\_: loc sex disp dow;

output out=bball\_cv10\_model05;

run;

**/\* Coding Probability for 10 CV Fold model \*/**

data basketball\_code; set bball\_cv10\_model05;

if P\_baketball\_ankles1 > 0.50 then test=1;

else test=0;

run;

**/\* BOOTSTRAP - 1000 REPS \*/**

title 'Bootstrap Analysis - 1000 Reps';

proc surveyfreq data=basketball\_code varmethod=bootstrap (reps=1000);

tables baketball\_ankles\*test / row column cl alpha=0.05 plots=all;

weight wt;

strata stratum;

cluster psu;

run;