3. Nodes Analysis:

(You may include summary graphics or tables, but no coding is required. Excel-level analysis is sufficient.)

Head Injury Classification		my_def (True condition)	
		1	**0**
my_test (Predicted condition)	**1**	**TP**	**FP**
	0	**FN**	**TN**
/	/	**Sensitivity = TP/(TP+FN)**	**Specificity = TN/(TN+FP)**

I will use above table for the nodes analysis.

The five nodes I choose to analyze are node 36, node 32, node 28, node 8, and node 7.

1. Node 36

Node 36 is one of the topmost nodes in the decision tree since it has 17 leaves under this node. 17 terminal-node leaves mean there are some next-level splits follow under the node 36. It includes the "LOC to be 0, SCALP to be 0, HEADS to be 0, HD to be 0, HEADACHE to be 0, COLLISION to be 0, CHI to be 0, CONCUSSION to be 0, and HEAD to be 0". In node 36, there are 51 rows of data (0.68%) are cases, and 7,477 (99.32%) rows are as non-cases. However, node 36 classifies all the data into 0, which are non-cases (no head injury).

Node 36 uses "AND" to connect 9 conditions. When all of the above keywords are missing simultaneously, the prediction result for head injuries will be "0" (non-case, no head injury predicted). Otherwise, the prediction will be a "one" (cases, head injury predicted). All the result predictions under node 36 are classified as 1 at the "test" column by looking at the "INVESTIGATE_NODES" dataset. This is a very high-level and general large condition because there are very fewer narratives that will contain all of the above keywords exactly. Once a case data satisfied the conditions (all the keywords included), it must not be a head injury. To improve the identity case, I can apply "proc sql" with clause WHERE my_def = 1 and my_test = 0 to show all the true cases that have not been identified, which are the false negatives. In this case, 99.31 percent are actually no head injury involved, which means 0.68% of data are misclassified. For this node, I should pay more attention to the false negatives, since there are approximately 0.68% of cases involved in a head injury but we identified these were not head injuries. I should try to find some common keywords in these data and classify them correctly in the further tree splits to improve the sensitivity (true positive rate) and specificity (true negative rate).

Evidences and examples:

NEK	myd	bdpt	narr1	narr2	_Node	test
170250073	1	75 - HEAD	10YO M AT SCHOOL WHEN HIT ON FOREHEAD WITH FOOT	DENT. DX:HEAD TRAU	36	0
170347292	1	75 - HEAD	*13YOF,HIT BY FOOTBALL TO HEAD,NAUSE,DIZZY,DX:CONCUSSION		36	0
170554941	1	75 - HEAD	11YOM SLIPPED AND FELL HITTING BACK OF HEA PLAYIN	SSION	36	0
170613910	1	75 - HEAD	16YOM PLAYING FOOTBALL AND TACKLED HARD, WEARI	N	36	0
170700841	1	75 - HEAD	14YOM DEVELOPED A HEADACHE,NUMBNESS TO HANDS,	DEELING WHILE AT FC	36	0
170826572	1	75 - HEAD	A 17YOM FELL TO GROUND, SUSTAINED SKULL FX WHILE PLAYING FLAG FOOTBA		36	0
170842220	1	75 - HEAD	12 YOM AT FOOTBALL PRACTICE COLLIDED HELMETS W/A	SION	36	0
170848272	1	75 - HEAD	15 YOM INJURED HEAD,THIGH,PLAYING FOOTBALL. DX-C	THIGH	36	0
170863990	1	75 - HEAD	15 YOM INJURED HEAD,PLAYING FOOTBALL. DX-ACUTE NO	NINTRACTABLE HA	36	0
170907461	1	75 - HEAD	*16YOM,ABD/N/V,P;LAYS FOOTBALL,PRACTICE HIT MULT	HELMET,DX:CONCUSS	36	0
170910209	1	75 - HEAD	10YM PLAYING FOOTBALL, IN FULL GEARS, GOT TACKLED	H/A>>CONCUSSION	36	0
170912779	1	75 - HEAD	12YM PLAYING FOOTBALL, TACKLED&TOOK A SHOULDER	AD "GLASSY EYES">>C	36	0
170921097	1	75 - HEAD	15 YOM PLAYING FOOTBALL WITHOUT HELMET, GOT HIT	N	36	0
170922261	1	75 - HEAD	16YOM-? CONCUSSION-POSSIBLY HAPPENED IN FOOTBALL-@ SCHOOL		36	0
170922968	1	75 - HEAD	11 YOM PLAYINIG FOOTBALL AND COLLIDED WITH ANOTI	ON	36	0
170923693	1	75 - HEAD	12YM WAS PLAYING FOOTBALL, TOOK HARD HIT TO HEAD	D >>CONCUSSION	36	0
170939167	1	75 - HEAD	*16YOM,HEADACHE,PLAYING FOOTBALL AFTER DEVELOP	T BY ANOTHER PLAYEF	36	0
170942097	1	75 - HEAD	9 YOM PLAYING FOOTBALL AND GOT TACKLED, HELMET C	N.	36	0
170947033	1	75 - HEAD	13YM PLAYING FOOTBALL IN THE GYM CLASS&RAN INTO	C>>CHI/LAC	36	0
170958186	1	75 - HEAD	11 YOM PLAYING FOOTBALL WHEN ANOTHER PLAYER RAI	ONCUSSION	36	0
170958834	1	75 - HEAD	10 YOM INJURED PLAYING FOOTBALL. DX-CONCUSSION		36	0
170960619	1	75 - HEAD	13YM YTD PLAYING FOOTBALL GAME, TACKLED&STRUCK	T TDY C SEVERE H/A&	36	0
170962209	1	75 - HEAD	15YOM INJURED PLAYING FOOTBALL	DX: POSTCONCUSSIVI	36	0
170965016	1	75 - HEAD	15YOM HIT HELMET TO H ELMET PLAYING FOOTBALL GAM	N	36	0
170968321	1	75 - HEAD	12YOM HIT DURING FOOTBALL GAME C/O EHADACHES AI	N@	36	0

2. Node 32

Node 32 is a very similar level node as node 36. It is also one of the topmost nodes in our decision tree. There are 15 leaves follow under this node. This node includes the keywords occurrence condition: NECK to be 0, LAC to be 0, FX is 0, CONT is 0, COLLISION to be 1, CHI to be 0, CONCUSSION to be 0, and HEAD is 0. In node 32, there are 48 data rows (87.27% of data) are cases, and 7 rows of data (12.73% of data) are non-cases. However, node 32 classifies all the data to be 1 (cases, head injuries occurred).

Node 32 uses “AND” to connect 8 conditions. As all the above conditions are satisfied at the same time, the data will be classified as 1, and otherwise, the data will be classified as 0 for the “test” column. Node 32 is a high-level classification since it includes many conditions, and very few data could satisfy all of them concurrently. This classification does delivery some information. Since all the keywords occurrences will be 0 but COLLISION is 1, “COLLISION” is a unique word to extract from the narrative. This means when all other keywords are missing but COLLISION is used in the description, it will very likely be a head injury. Node 32 splits all of the data to be 1 for a head injury, and a few data to be 0. This split successfully increases the number of true positives, but I should recheck the miss-classification for the false positive rate. I could use “proc sql” with clause WHERE my_def = 0 and my_test = 1 to show the cases are wrongly classified as a head injury but actually not the head. I could avoid some of the keywords in the next split to improve the specificity if the false positive rate is reduced.

Evidences and examples:

NEK	mydef	bdpt	narr1	narr2	_Node	test
170610388	0	30 - SHOULDER	11 YOM C/O LEFT SHOULDER PAIN AFTER COLLISION WITI	FLAG FOOTBALL, DX:	32	1
170907915	0	31 - UPPER TRUNK	A 13YOM WAS IN COLLISION WHILE PLAYING FOOTBALL, RIB PAIN		32	1
170924422	0	35 - KNEE	15YOM INJ R KNEE PLAY' FOOTBALL AFTER COLLISION: KNEE PN R		32	1
170955442	0	76 - FACE	17 YOM HELMET TO HELMET COLLISION IN FOOTBALL PR/	ANDIBLE.	32	1
171029296	0	32 - ELBOW	8YOM ELBOW PAIN- COLLISION PLAYING FOOTBALL AT HOME		32	1
171059250	0	31 - UPPER TRUNK	11YOM HAS CHEST PAIN AFTER COLLISION PLAYING FOOT	AGO; CHEST PAIN	32	1
171070055	0	83 - FOOT	15YM BEGAN C/O FOOT PAIN P TACKLE FOOTBALL COLLIS	>>CONTS	32	1

3. Node 28

Node 28 is a low-level node in our decision tree. There are 12 terminal-node leaves follow under node 28. This means that there has already been some successful classification based on the previous nodes. This node includes the keywords occurrence condition: R is 0, HD is 0, HEADACHE is 0, COLLISION is 0, CHI is 0, CONCUSSION is 0, and HEAD is 0. When all the conditions are satisfied, the prediction for my_test will be classified as 1. Otherwise, the head injury prediction will be classified as 0. Node 28 classifies all the data into 1, which are cases (head injury). However, 12 cases (92.31%) are truly head injuries, and only 1 case (7.69%) is actually non-cases (no head injury).

Node 28 uses “AND” to connect 7 conditions. As all the conditions are met, the head injury prediction test will be 1, and otherwise, the prediction test will be 0. This is still a long condition that very seldom narrative data will meet all the conditions simultaneously. Node 28 is only responsible for distinguishing 13 data. Since 92.41% percent of the data are classified as head injury, it is very possible that the false positives exist. Some of the non-head injury cases are tested to be cases. The specificity is affected by this node. I should check the data using “proc sql” with clause WHERE my_def = 0 and my_test = 1 to show the misclassification rows. The only false negative prediction should not be influenced. I need to pull out the false positive data to check the keywords extraction for the next level split to improve the specificity by reducing the false positive rate.

Evidence and example:

NEK	mydef	bdpt	narr1	narr2	_Node	test
171076645	0	76 - FACE	17YOM EVAL FACIAL INJURY X 1 DAY, STS WAS PLAYING F	HIT HIS NOSE AGAINS	28	1

4. Node 8

Node 8 is a very low-level node in the decision tree model. There are only 2 leaves follow this node 8. This node includes keywords occurrence condition:
CONCUSSION is 0, STRAIN is 0, and HEAD is 0. If the three conditions are met at the same time, the prediction result will be a 0 (No head injury). Otherwise, the case will be classified as 1 (head injury occurred). Node 8 classifies all the data to be non-cases (no head injury occurred). However, 9 cases (34.62%) are actually cases (there are head injuries involved) and the other 17 cases (65.38%) are truly non-cases.

Node 8 only uses “AND” to connect 3 conditions. Unlike the previous three nodes I talked about, this node classifies for the data that are more even than the previous nodes. Almost 35% of the cases data are head injuries and 65% of cases data are not head injury. If all three words, “CONCUSSION, STRAIN, and HEAD”, are missing, it will be unlikely to be a head injury. False negatives are affected by node8 classification, and the specificity may be reduced. The false-negative rate could be high here due to the easy-reach condition. I should post the data by using "proc sql" with clause WHERE my_def = 1 and my_test = 0. I should check the keyword extractions and add some conditions into the next level tree split to increase the sensitivity by reducing the false-negative rate.

Evidences and examples:

NEK	myd	bdpt	narr1	narr2	_Node	test
170554893	1	75 - HEAD	11YOM PLAYING FOOTBALL AND RAN BACKWARDS, TIPPI	HEAD INJURY, NECK S	8	0
170926331	1	75 - HEAD	15 YOM NECK PAIN HIT ON RT SIDE BASE OF NECK DURINC	ATHLETIC FIELD DX CL	8	0
170947929	1	75 - HEAD	14 YOM WAS INJURED WHILE PLAYING FOOTBALL W/ BLU	HEAD INJURY DX: CL	8	0
171013244	1	75 - HEAD	CHI, CERVICAL STRAIN. 8 YOM WAS PLAYING FOOTBALL	WHEN HE HIT HIS HEAD	8	0
171026467	1	75 - HEAD	8 YOF INJURED HEAD, NECK,PLAYING FOOTBALL. DX-CLO	NECK STRAIN	8	0
171028977	1	75 - HEAD	13YOM-PT WAS IN FOOTBALL GAME WAS THE QUARTERB	LEFT SIDE FALLING ON	8	0
171034634	1	75 - HEAD	CHI, CERVICAL STRAIN. 9 YOM WAS PLAYING FOOTBALL V	HEAD WITH ANOTHEI	8	0
171042585	1	75 - HEAD	5YOM HIT HEAD PLAYING FOOTBALL WITH BROTHER, SIST	IT ON SOMEONE'S ELB	8	0
171108289	1	75 - HEAD	12YOM TACKLED DURING FOOTBALL GAME. C/O NECK PA	AL STRAIN	8	0

5. Node 7

Node 7 is one of the last internal nodes in the decision tree split for head injury prediction. There is only one leaf following this node. All 5 cases 100% of data under this node are classified as cases and no non-cases predicted. Node 7 has conditions:
CONCUSSION is 1, STRAIN is 1, and HEAD is 1. The prediction result is 1, which is the head injury predicted. Otherwise, the tree will classify as no head injury occurred. The cases under Node 7 are truly cases, which are head injuries, and no datum is actually non-case. Therefore node 7 classifies all the data correctly and purely.

Node 7 uses “AND” to connect 3 conditions. The conditions for Node 7 are almost the same as Node 8, but they may under different split branches from disparate mother nodes (the previous level nodes). Node 8 says when three conditions are 0, the prediction result is 0, and node 7 says when three conditions are 1, the prediction result is 1. Node 7 is only responsible to classify 6 data, and all of them satisfy the three conditions. All three narratives include keywords of “CONCUSSION”, “STRAIN”, and “HEAD”. When all three keywords that are appeared in the narrative, the case must be a head injury. Since node 7 classifies all the data into the true positives and is the last level internal node, and the conditions are the same as node 8’s conditions under different branches. I would presume the keywords classification is relatively accurate. However, I should still use "proc sql" to extract data by using WHERE my_def = 0 and my_test = 1 to check the false positives to see if any miss classification occurred. It will be unlikely to continue to improve the sensitivity or specificity, but this is a very good example node to visualize the split result for leaves of the decision tree model. After I pulled out the data, all the cases were correctly classified.

Evidence and examples for correct predictions:

NEK	mydef	bdpt	narr1	narr2	_Node	test
170831866	1	75 - HEAD	14 YOM C/O HEAD INJURY S/P HITTING HELMET TO HELM	AT A FOOTBALL GAME	7	1
170933130	1	75 - HEAD	13YOM-PT WAS IN FOOTBALL PRACTICE HEAD TO HEAD T/	PADS. DX- CONCUSSI(	7	1
170949944	1	75 - HEAD	11 YOM WAS GOING HEAD TO HEAD DOING DRILLS IN FO	DX: CLOSED HEAD IN.	7	1
171031053	1	75 - HEAD	14YOM HEAD TO HEAD WITH ANOTHER PLAYER IN FOOTB	HEAD AND NECK PAIN	7	1
171109298	1	75 - HEAD	16 YOM INJ TO HEAD PLAYING FOOTBALL TACKLED HIT HE	TIC FIELD DX CONCUS	7	1