NHTSA HW

Summary
This HW has two parts. First, we look at models that are easy to understand in terms of what predictor variables effect the head injury criterion. Some algorithms are not as useful because either they are ensemble of many tree models or the predictor variables have undergone transformation and hence it is difficult to understand the exact relationship. So only a few models are chosen for interpretation. I also highlight some of the challenges faced during the model building process. I have considered multiple options and then decided on the best option for this dataset. The second part of the paper publishes the performance for multiple models with respect to error rate. A standard error [SE] for the error rate is given which is based on the following formula

$$SE = sqrt((p*(1-p))/n)$$

Where p is average error rate and n is the sample size. Finally I conclude by highlighting what I learned about the data and the different algorythms.

Feature Engineering

From the baseline model given by Porf. Loh I removed the following variables from the data files AX1, AX2, AX3, AX4, AX5, AX6, AX7, AX8, AX9, AX10, AX11, AX12, AX13, AX14, AX15, AX16, AX17, AX18, AX19, AX20, AX21, BMPENG, SILENG, APLENG, DPD1, DPD2, DPD3, DPD4, DPD5, DPD6, LENCNT, DAMDST, CRHDST, YEAR. Most of these variables are removed since these are measurement after the test.

AX1 to AX21 were post test measurement and hence we excluded these from the model. DPD1 to DPD6 are damage profile measurement taken after the test and hence I removed these variables as well. LENCNT, DAMDST, CRHDST are also variables that characterize the damage to the vehicle after the test. YEAR, I think is not a very useful predictor in this model since MODELD would be more intuitive in identifying vehicles introduced in certain years. BMPENG, SILENG, APLENG were removed from the model since these condition such as bumper engagement and side engagement can only determined after the test.

Converting variables with more than 53 variables into dummy variables. MAKED has 71 levels and hence this variable cannot be handled in random forest, ctree, cforest or any other linear model.

Dropping Variables with more than 53 levels: Certain algorithms such as Random Forest, ctree and cforest cannot handle variables that have more than 53 levels. Also these could not be converted to dummy variables since they had more than 500 levels. E.g ModelD.

Issues with colineartiy in the data

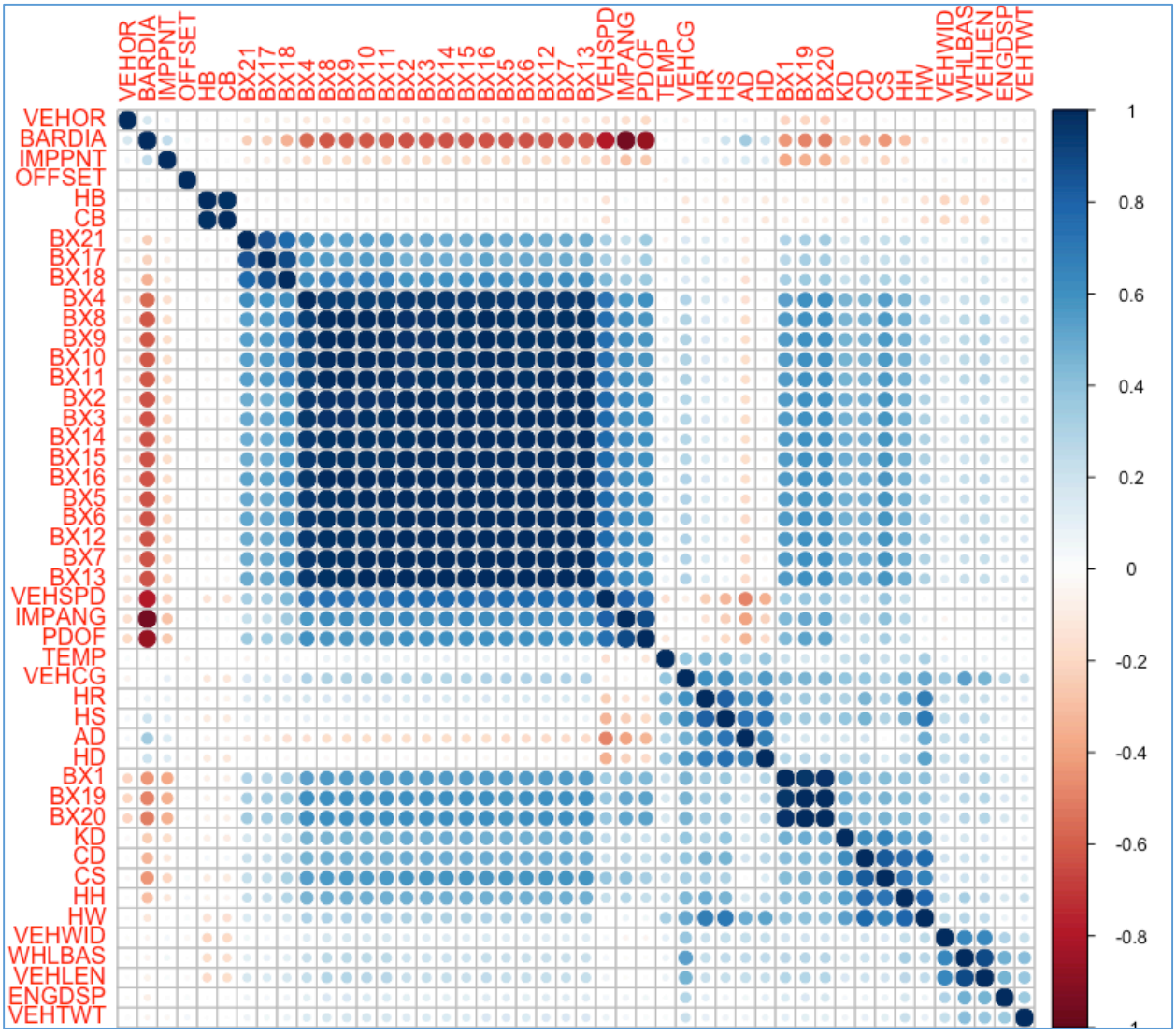
Since some of the variables measure dimensions of the vehicle or barrier, these measure are highly collinear. This can be seen in the below visualization.

I considered two options to deal with this issue

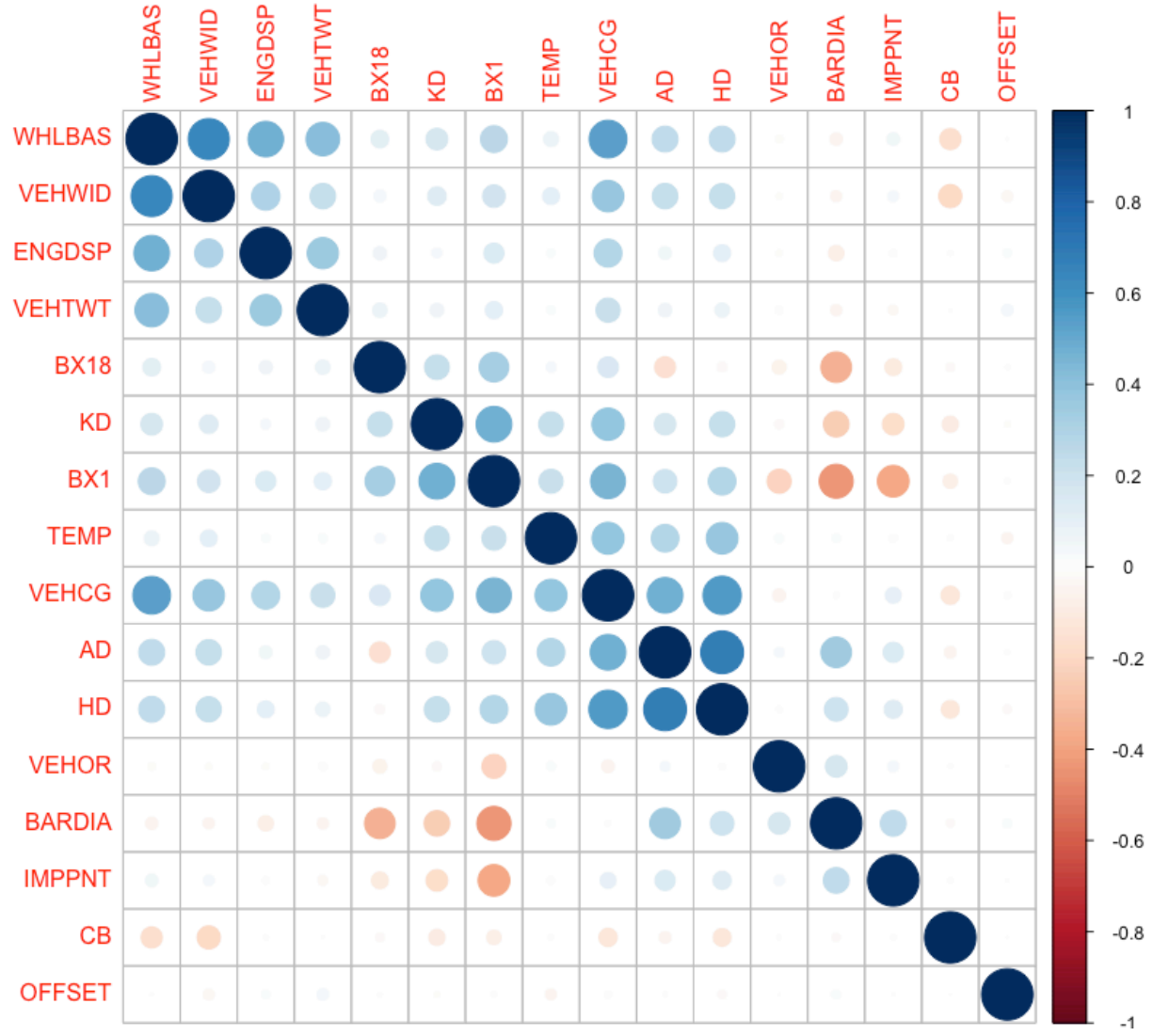
a) Dropping the highly collinear variables
b) Generation Principal Component (PCA) for numeric variables

I set a cut off of 0.7 when deciding to drop variables. The number of numeric predictor variables reduces from 54 to 16. This drastic reduction in number of variables would lead to loss in predictive power. On the other hand, using PCA would lead to making the model uninterpretable. Since I was using tree models to analyze feature importance and was trying to compare predictive performance of different algorithm I decided to go ahead with PCA.

Correlation Matrix of all variables



Correlation Matrix of Variables with correlation < 0.7



Missing data handling

Missing data was one of the biggest challenge with this dataset. Since there are approximately 400 observations with complete data out of approximately 3000 total observations. Dropping these observations from the dataset would not be advisable. Only 10 columns out of the 110 columns have complete data. Hence dropping columns with missing data would not make be advisable. Mean and mode imputation would be too naïve hence I chose to use one the more sophisticated imputation. MICE and Amelia failed due to collinearity issue in the data. Missforrest worked since it used randomforect for numeric data imputation.

The following factor variables have missing values
SEPOSN, CTRL2, TKSURF, TKCOND, ENGINE, TRANSM, STRSEP, COLMEC, MODIND

Missing values for factor variables were replace with “MISSING”. I decided against doing a mode missing data since the missingness may not be at random.

The following continuous variables have missing data
CS, AD, HD, KD, HB, NB, CB, KB, TEMP, IMPANG, OFFSET, IMPPNT, ENGDSP, VEHTWT, CURBWT, WHLBAS, VEHLEN, VEHWID, VEHCG, BX1, BX2, BX3, BX4, BX5, BX6, BX7, BX8, BX9, BX10, BX11, BX12, BX13, BX14, BX15, BX16, BX17, BX18, BX19, BX20, BX21, VEHSPD, CRBANG, PDOF, CARANG, VEHOR

Missforest was used to impute missing data and since the algorithm took too long for multiple iterations hence I restricted the max number of iteration to 1. Also Missforest was used because of colinearity issue discussed earlier. MICE failed because of colinearity issues.

Analysis of driver side

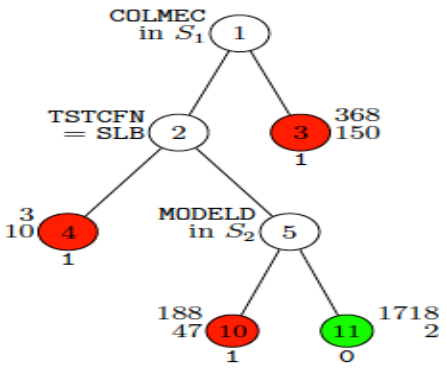
To understand which factors influence head injury criterion, I used the following algorithms

a) Guide tree
b) Ctree
c) Rpart
d) Random forest

We will also look at which factor are different when comparing driver side and passenger side.

a) Guide tree

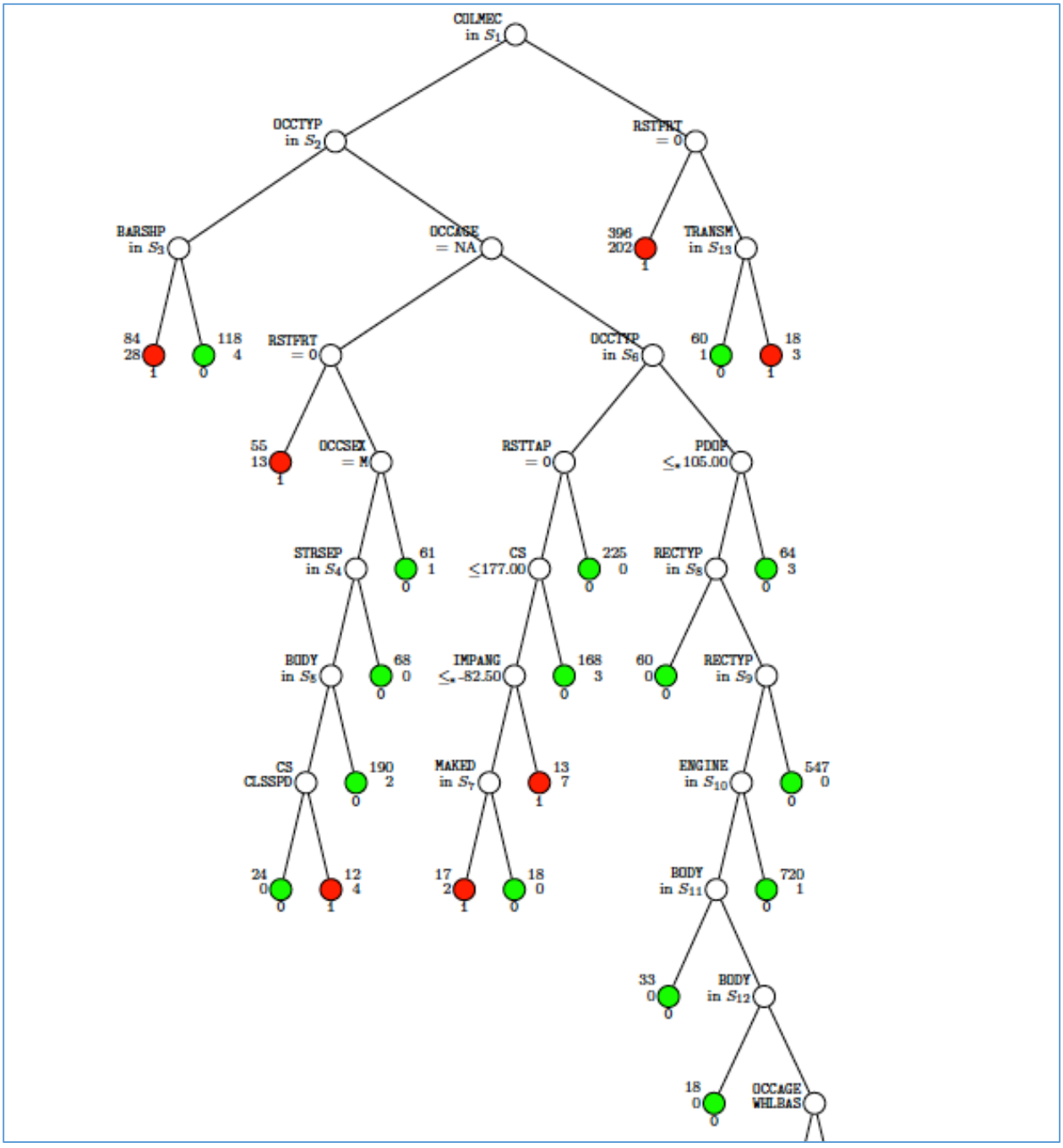
Decision Tree guide for driver



GUIDE 0.50-SE classification tree for predicting HIC2 using equal priors and unit misclassification costs. At each split, an observation goes to the left branch if and only if the condition is satisfied. Set S_1 = {BWU, NA, NAP, UNK}. Set S_2 = {ACHIEVA, AEROSTAR, ASTRO, BEAUVILLE, BLAZER, BROUGHAM, CAPRICE, CARAVAN, CENTURY, CORSICA, DE VILLE, DURANGO, ELANTRA, EX35, EXPLORER, GRAND AM, INTRIGUE, LUMINA, METRO, RAM, RANGER, RAV4, REGAL, S-10, S10 BLAZER, SEDONA, SIDEKICK, SILVERADO, SL2, SPACECAB, SPORTAGE, SPORTVAN, TACOMA, TRAILBLAZER, TRANS SPORT, TROOPER II, VIGOR, VUE, XT}. Predicted classes (based on estimated misclassification cost) printed below terminal nodes; sample sizes for HIC2 = 0 and 1, respectively, beside nodes. Second best split variable at root node is RSTDPL.

Colapse Mechanism [Colmec] is the most important variable in predicting HIC. Specifically, Behind Wheels Unit [BWU], MISSING[NA], Not applicable [NAP], UNK [Unknown] split to a subtree that has higher number of HIC =0. It seems that BWU is better at protecting against head injury for the driver. The next split on test setup [TSTCFN] separates based on whether an entire car is used in the crash test. This doesn't seem to be very good predictor for real life crashes. Hence I removed TSTCFN for my next tree. I also changed the SE 0.00 to get a bigger tree.

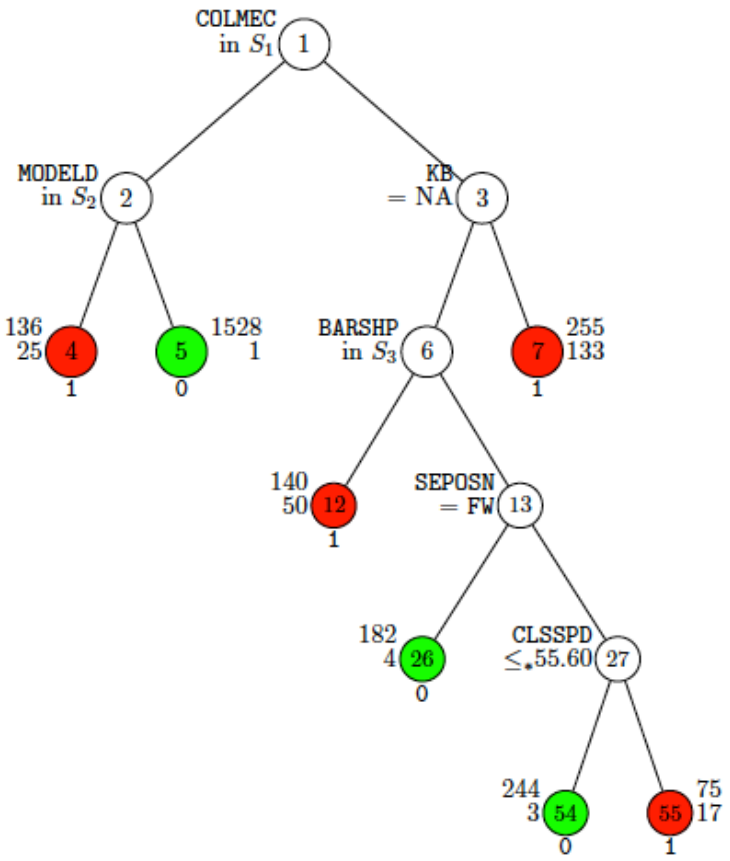
Decision guide tree with SE 0.00 driver



GUIDE 0.50-SE classification tree for predicting HIC2 using equal priors and unit misclassification costs. At each split, an observation goes to the left branch if and only if the condition is satisfied. The symbol '$\leq_*$' stands for '$\leq$ or missing'. For splits on categorical variables, values not present in the training sample go to the right. Set S_1 = {BWU, NA, NAP, UNK}. Set S_2 = {E2, OT, P5, S3, WS}. Set S_3 = {LCB, POL}. Set S_4 = {SP, UN}. Set S_5 = {PU, VN}. Set S_6 = {HT, S2}. Set S_7 = {BMW, BUICK, CHEVROLET, CODA, JEEP, MAZDA, TOYOTA}. Set S_8 = {NA, UNK}. Set S_9 = {FMT, OTH}. Set S_{10} = {4CIF, S6IF}. Set S_{11} = {3H, 4S, VN}. Set S_{12} = {2C, UV}. Set S_{13} = {AF, AR}. Predicted classes (based on estimated misclassification cost) printed below terminal nodes; sample sizes for HIC2 = 0 and 1, respectively, beside nodes. Second best split variable at root node is OCCTYP.

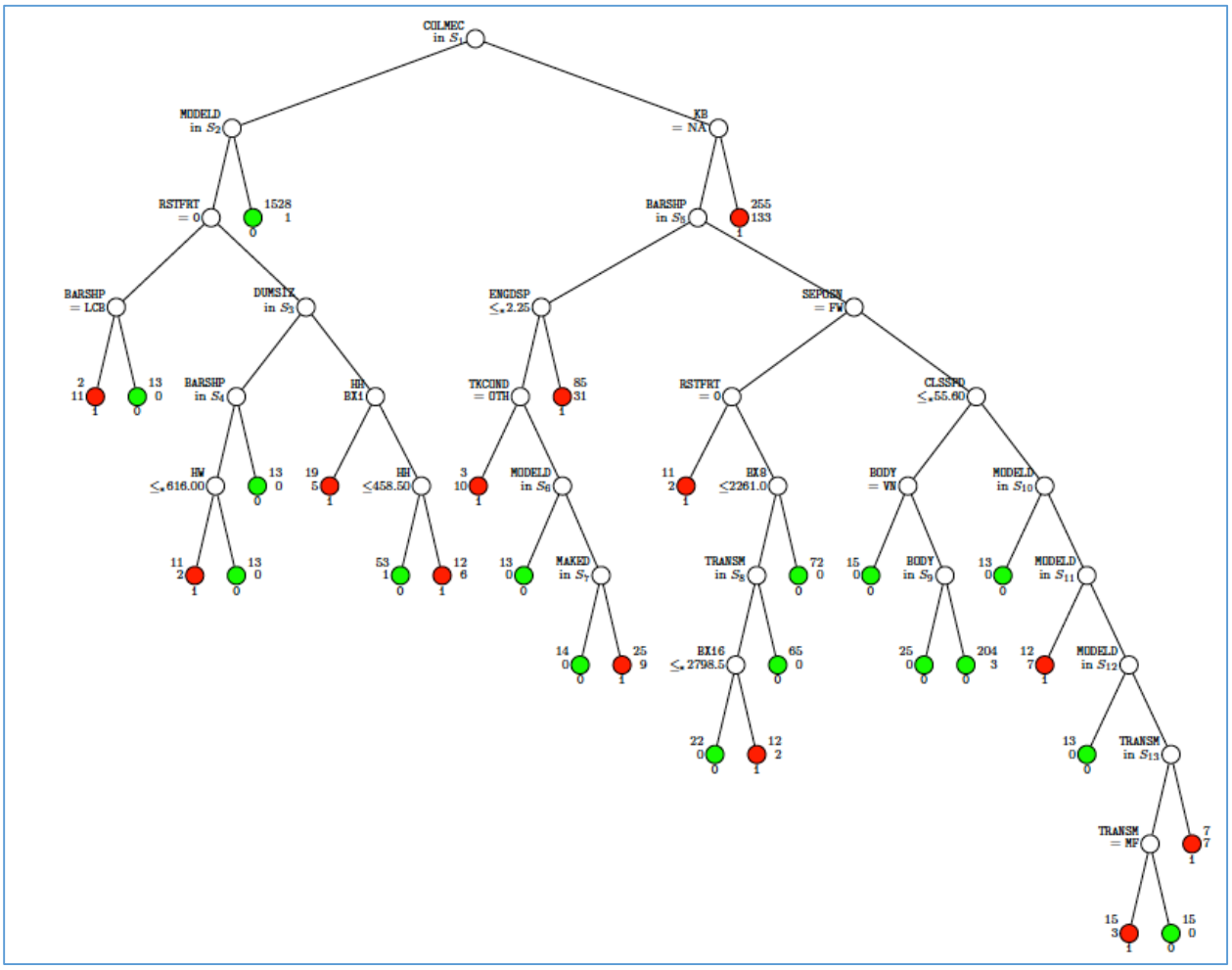
Even in this tree colmec is best variable for split. But here the second split on RSTFRT[front air bag] is important. Impact angle < -82.50 split makes sense since the impact will be toward the driver side of the vehicle. Van and pick up body types are more likely to have greater head injuries for driver.

Decision guide tree passenger



GUIDE 0.50-SE classification tree for predicting HIC2 using equal priors and unit misclassification costs. At each split, an observation goes to the left branch if and only if the condition is satisfied. The symbol '$\leq_*$' stands for '$\leq$ or missing'. Set S_1 = {BWU, NA, UNK}. Set S_2 = {ASTRO, BLAZER, CAPRICE, CARAVAN, CENTURY, DE VILLE, ELANTRA, EX35, EXPLORER, GRAND AM, INTRIGUE, LEGACY, RAM, RANGER, RAV4, S-10, S10 BLAZER, SILVERADO, SPORTAGE, TACOMA, TRAILBLAZER, VIGOR, VUE}. Set S_3 = {LCB, UNK}. Predicted classes (based on estimated misclassification cost) printed below terminal nodes; sample sizes for HIC2 = 0 and 1, respectively, beside nodes. Second best split variable at root node is RSTDPL.

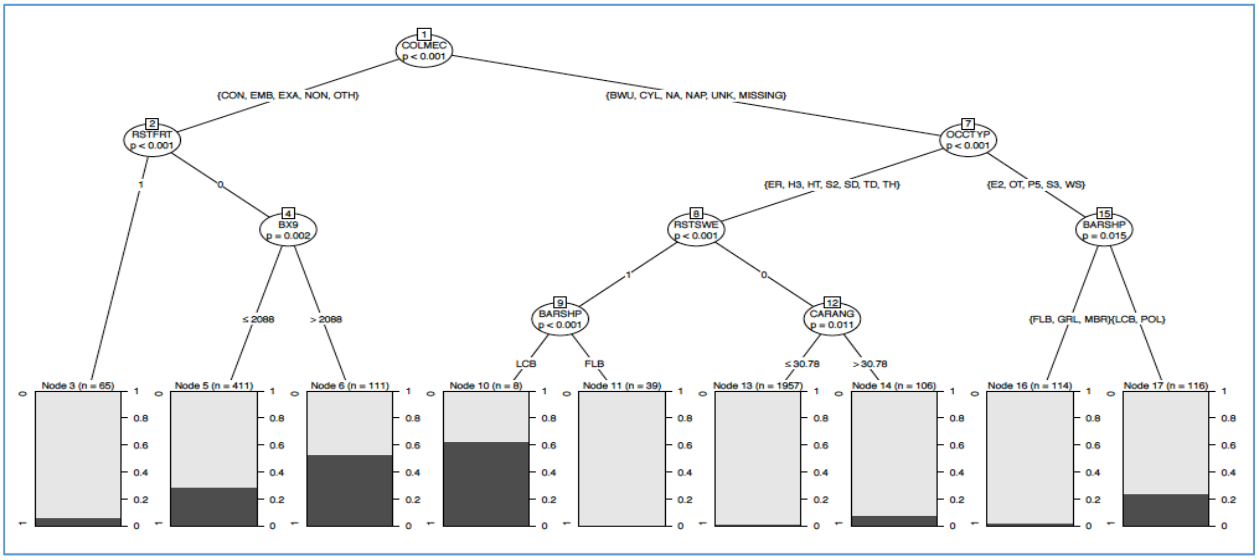
Decision guide tree with SE 0.00 for passenger



When we compare guide tree models for driver and passenger collapse mechanism and frontal airbags are important in both the models. For the passenger side the seat position that is safest is forward of center position. Also closing speed of less than 56 MPH has a drastically less likelihood pf head injury.

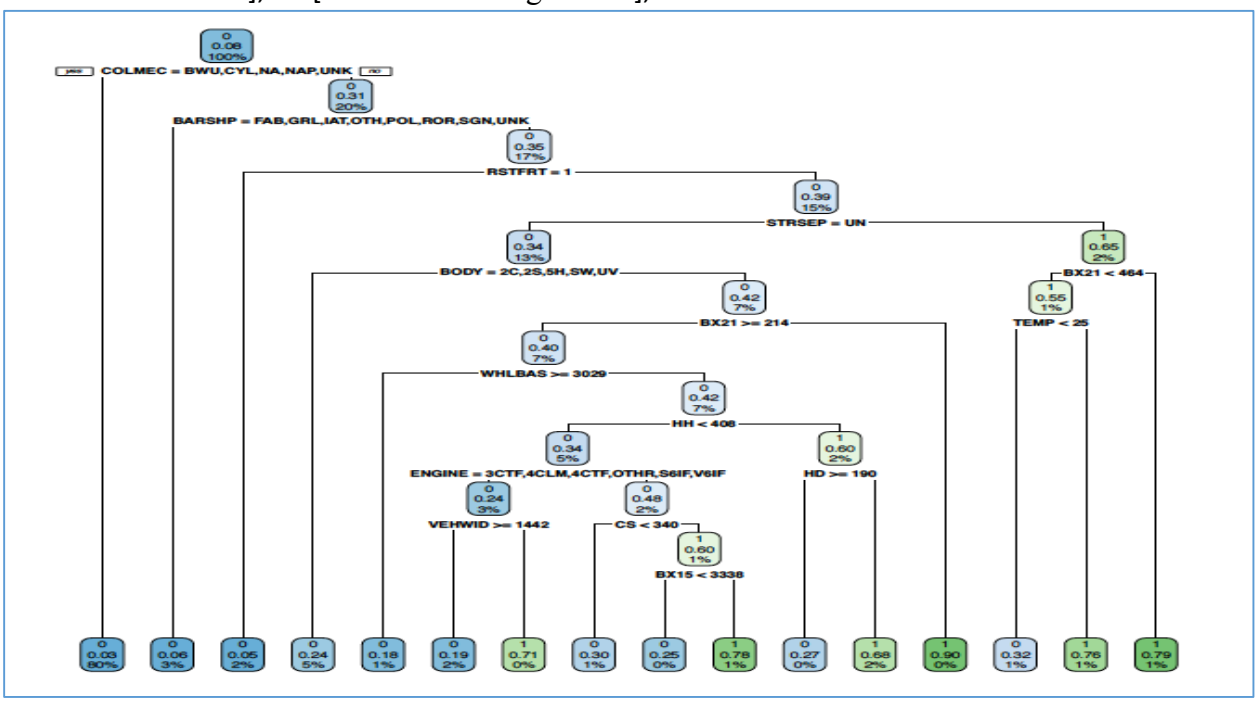
b) Ctree

The Ctree model is similar to guide model with colmec and RSTFRT being important factors effecting head injury criterion. Ctree also points to RSTSWE[STR. WHEEL - EA ENERGY ABSORBING] is important factor that helps predict head injury. Having a energy absorbing steering reduces the risk of head injury. Fixed hard barrier shapes of LCB [LOAD CELL BARRIER] and POL [POLE] have adverse effect on the head injury.



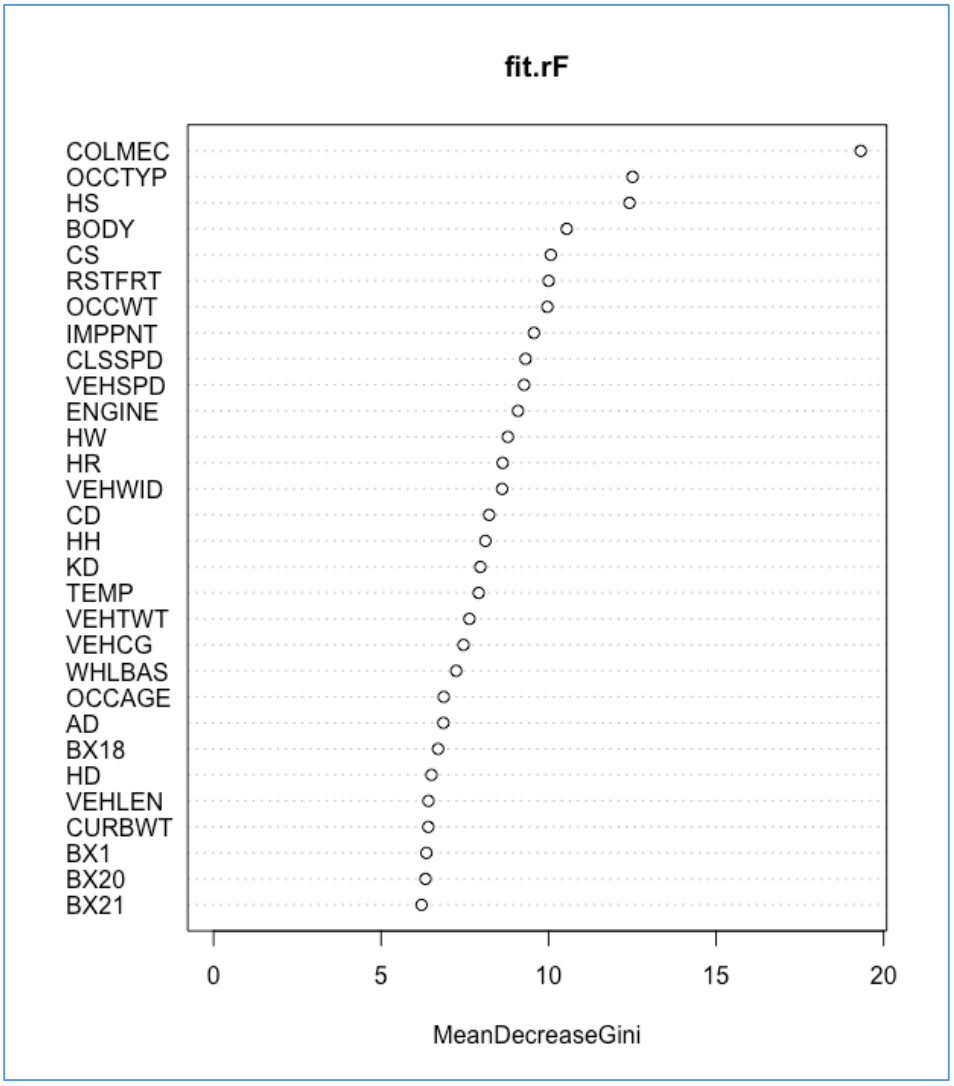
c) rpart

rpart model is similar to guide and ctree models. Apart from the main variables identified in the guide and ctree model, the rpart model points at other variables such as HH [Head to Windshield Header], CS [Chest to Steering Wheel],



d) Random Forest

Random forest also have Colmec, RSTPRT as one its top predictors.



Algorithm Prediction Performance

	# of Error passenger	Proportion Error	SE	# of Error driver	Proportion	SE
svm	26.22	0.09	0.02	31.00	0.09	0.02
rpart	27.11	0.10	0.02	33.00	0.10	0.02
random forest	**25.44**	**0.09**	**0.02**	**30.78**	**0.09**	**0.02**
qda	119.11	0.43	0.03	122.78	0.37	0.03
lda	29.11	0.10	0.02	33.56	0.10	0.02
lm con	27.78	0.10	0.02	32.22	0.10	0.02
lm	27.11	0.10	0.02	32.00	0.10	0.02
guide tree	25.89	0.09	0.02	30.78	0.09	0.02
guide forest	**24.56**	**0.09**	**0.02**	**29.00**	**0.09**	**0.02**
cforest	runs too long			runs too long		
glmnet	27.67	0.10	0.02	29.11	0.09	0.02

	passenger time (sec)	driver time (sec)
svm	1.34	1.19
rpart	1.14	1.85
random forest	**6.36**	**5.10**
qda	0.09	0.08
lda	0.09	0.11
lm con	0.05	0.04
lm	0.04	0.04
guide tree	43.84	170.67
guide forest	**161.73**	**152.72**
cforest	runs too long	runs too long
glmnet	65.74	74.79

Conclusion

Rpart, Ctree and guide tree have been very helpful in explain relationship between predictors and the head injury criterion. From prediction accuracy standpoint guide forest outperformed all algorithm but it was slowest of all algorithm. Certain algorithms have drawback such as not being able to handle missing values, not being able handle factor variables and not being able handle variables with more than 53 levels. Since guide forest did face these issues, it was able to take full advantage of the data. Cforest was taking very long to complete and hence I decided not to run that algorithm.