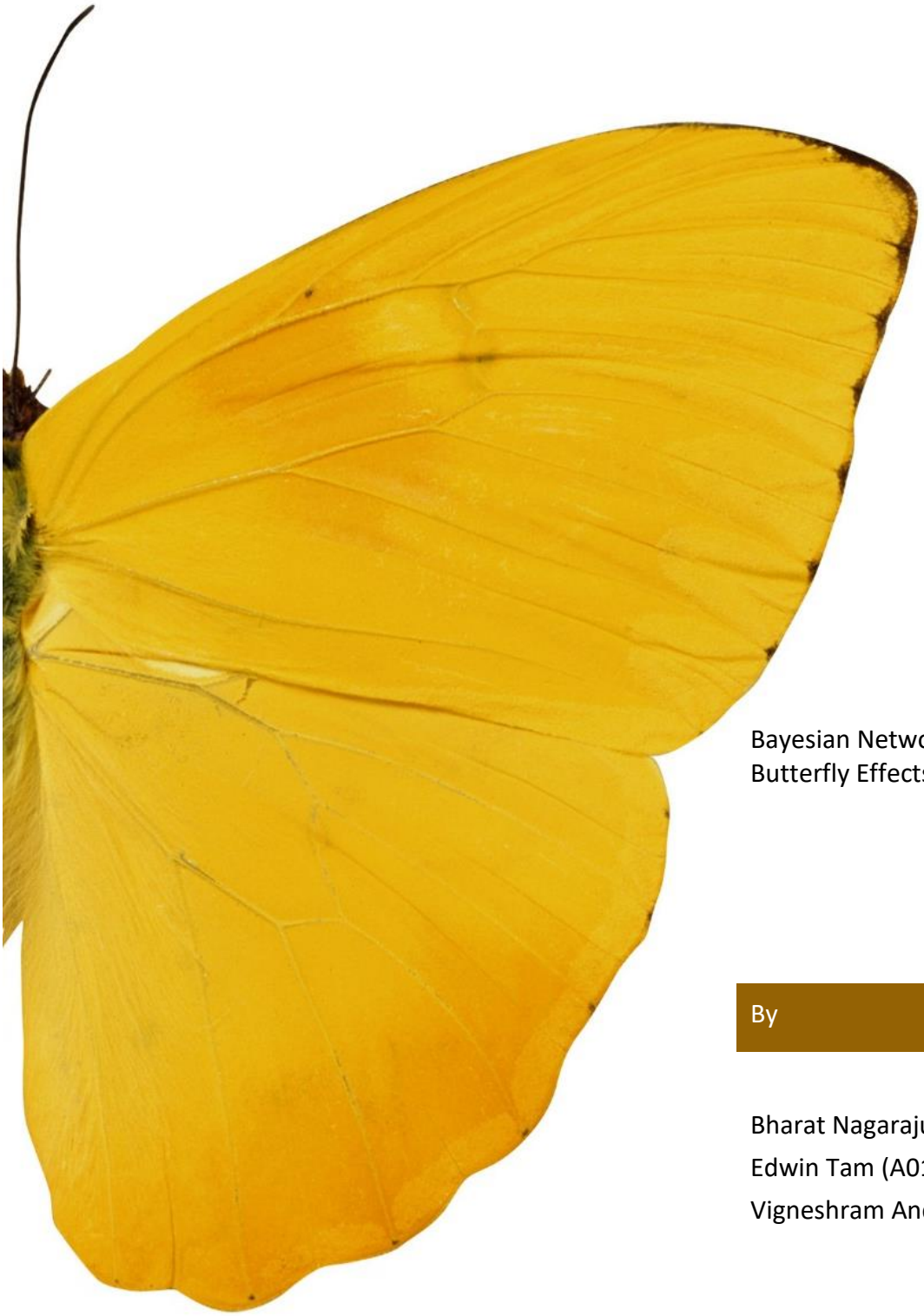


Bayesian Networks
Butterfly Effects

By

Bharat Nagaraju (A0178258N)
Edwin Tam (A0178396J)
Vigneshram Andiappan Selvaraj (A0178215A)

CONTENTS

DOMAIN & APPROACHES

JUST WHAT ARE THE FACTORS IMPACT VEHICLE SAFETY?

We are given a car crash dataset from Bayesia Website to clean, review, and infer from using 2 Bayesian Networks.

The dataset contains 19 usable features, of which OA_MAIS is the target variable (it measures injury severity). Before we describe the preparation, network generation, and results etc. – let's have a looksee into the dataset. We'll use a mixture of Genie & Excel to get the work done.

Dataset Shape: 21 columns X 20,247 observations

Continuous Variables

Variable	Mean	Variance	StdDev	Min	Max	Count	Missing %	Remarks
GV_CURBWGT	1617.3	154900.5	393.6	670	4310	20204	0.2%	
GV_DVLAT	0.0	169.5	13.0	-114	118	14049	30.6%	
GV_DVLONG	-14.8	311.8	17.7	-145	84	14049	30.6%	
GV_ENERGY	505.2	416986.1	645.7	4	9852	14049	30.6%	
GV_LANES	3.3	1.8	1.4	1	7	20244	0.0%	
GV_MODELYR	2003.6	7.6	2.8	2000	2012	20247	0.0%	
GV_OTVEHWGT	1630.2	169212.7	411.4	640	4540	18147	10.4%	
GV_SPLIMIT	40.7	126.4	11.2	0	75	20016	1.1%	
OA_AGE	40.2	301.8	17.4	0	97	20190	0.3%	
OA_HEIGHT	170.8	115.5	10.7	59	216	17508	13.5%	
OA_MAIS	0.9	1.1	1.0	0	6	19203	5.2%	
OA_MANUSE	0.9	0.1	0.3	0	1	19774	2.3%	
OA_WEIGHT	78.7	385.9	19.6	28	150	17599	13.1%	

VE_ORIGAVTW	154.7	58.6	7.7	105	185	20014	1.2%	Not Used
VE_WHEELBAS	281.0	824.7	28.7	141	481	20238	0.0%	Not Used
VE_PDOF_TR	152.6	4557.2	67.5	5	355	18298	9.6%	
GV_FOOTPRINT	4.4	0.4	0.6	2.4684	7.7952	20010	1.2%	

Categorical Variables

GV_WGTCDTR	Count	Percentage		OA_BAGDEPLY	Count	Percentage
Passenger Car	12474	62%		Deployed	9593	47%
Truck (<=10000 lbs.)	2542	13%		Not Deployed	10654	53%
Truck (<=6000 lbs.)	5231	26%				

OA_SEX	Count	Percentage		VE_GAD1	Count	Percentage
(missing)	234	1%		(missing)	789	4%
Female	9938	49%		Front	11793	58%
Male	10075	50%		Left	3145	16%
				Rear	1741	9%
				Right	2779	14%

You'll see these Files

1. vehicle_safety_final_binned.csv: Dataset that's actually used by the networks
2. DataPreprocessing.R: R Script to do data cleaning and imputation
3. vehicle_safety_final_impute.csv: Dataset after imputation is done.
4. vehicle_safety_final_Naive_Bayes: Naïve Bayesian Network (Open in GeNie)
5. vehicle_safety_final_TAN: Tree Augmented Bayesian Network (Open in GeNie)

PREPARING THE DATA

The Dataset provided by NASS (National Automotive Sampling Systems) contains attributes which has various parameters to understand the nature of accidents and the fatality rates.

Initially there were more than 400 variables, later the selection was limited to 21 variables which were relevant to the variables in EPA/NHSTA research.

The rest of this section covers Data Cleaning, Imputation, and Binning.

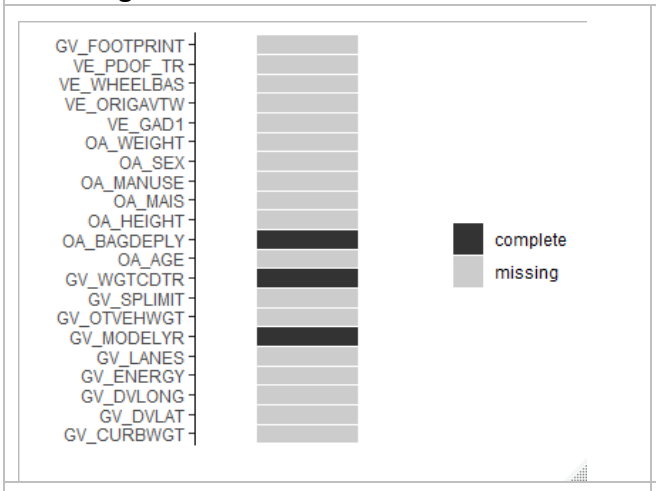
Data Cleaning

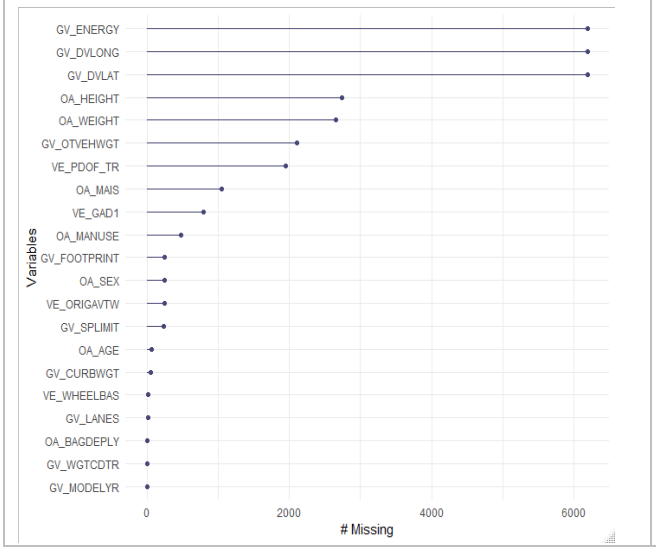
The data set contained 20,247 records. OA-MAIS which measures the injury is our target variable.

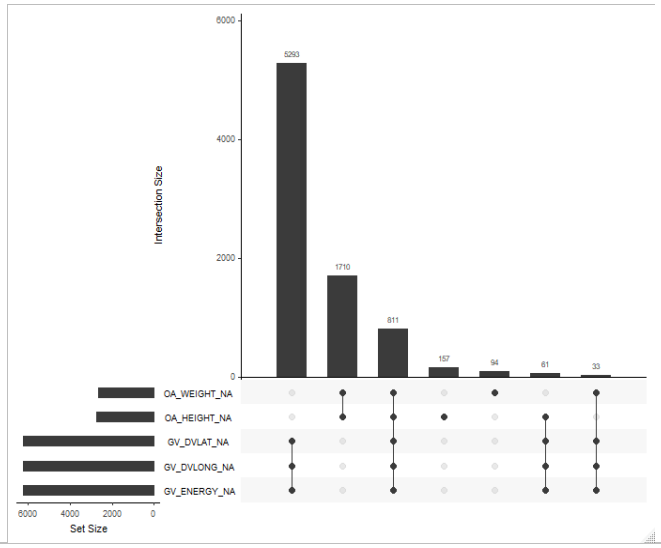
It is observed form the dataset that there are 1,044 records against the target variable which either have ‘NA’ or empty values.

These records constitute only 5% of the data set and hence removed it by deleting.

#Missing Values for Each of the columns before cleaning





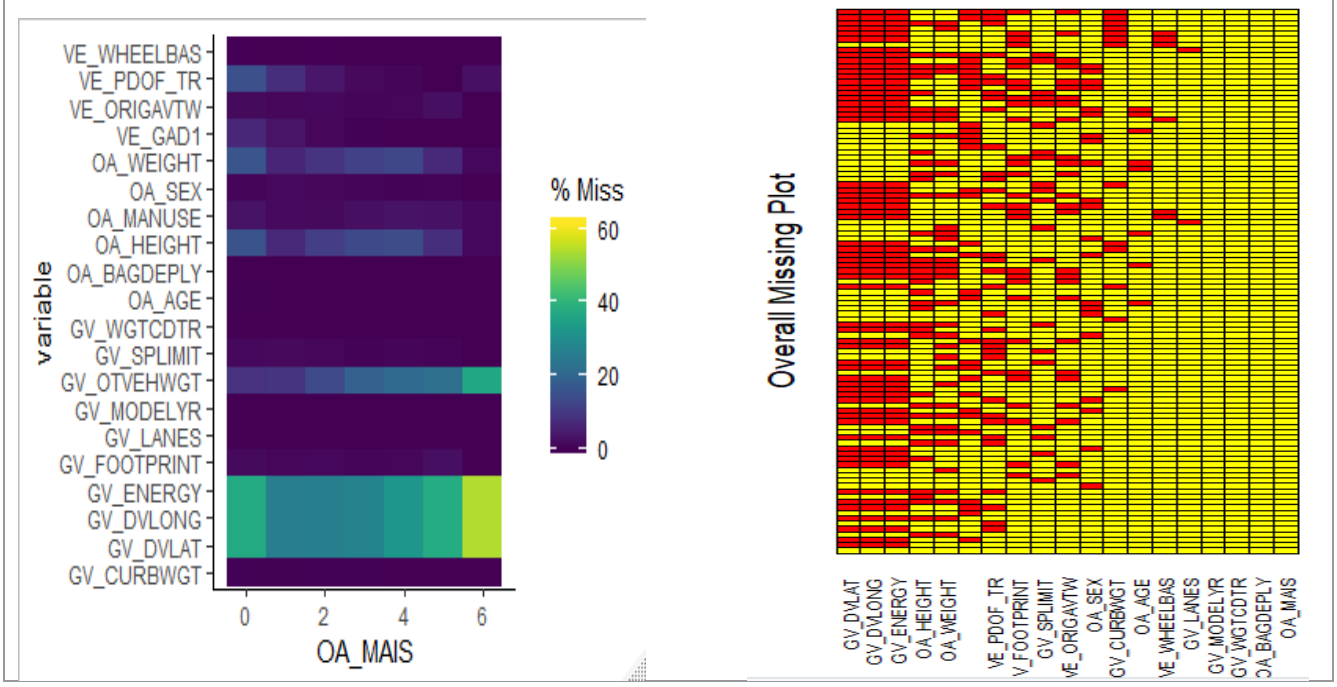


From the above charts, GV_DVLAT, GV_DVLONG and GV_ENERGY have a maximum amount of missing data. We will impute to balance the missing data

Data Imputation

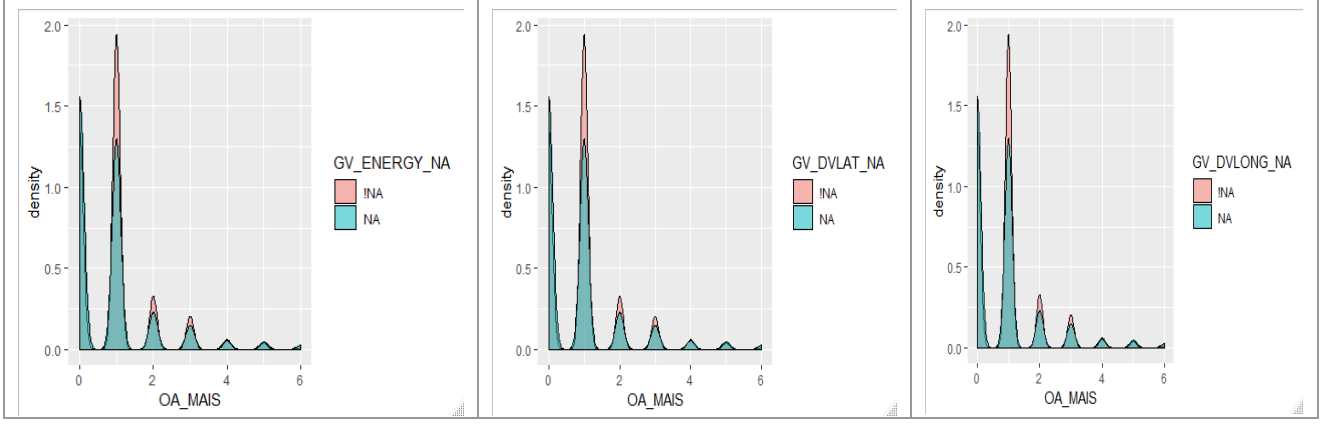
We analyzed all continuous variables' missing data and its pattern in the data set. Below is the plot depicting the same. Assuming the data is missing at Random, we decide to impute the data using PMM using 'MICE' package in R.

Breadth of Missing data



Top 3 Variables – Where is the missing data?

Overlay the values of predictor when the values are missing and when they are not.



IMPUTATION USING PMM (PREDICTIVE MEAN MATCHING)

Using MICE package, we create multiple imputations instead of single imputation techniques such as mean. This takes care of the uncertainty of the missing data. The missing data are assumed to be missing at random and it imputes the data one variable at a time by following PMM (Predictive Mean Matching) model.

DATA BINNING: INPUT VARIABLES

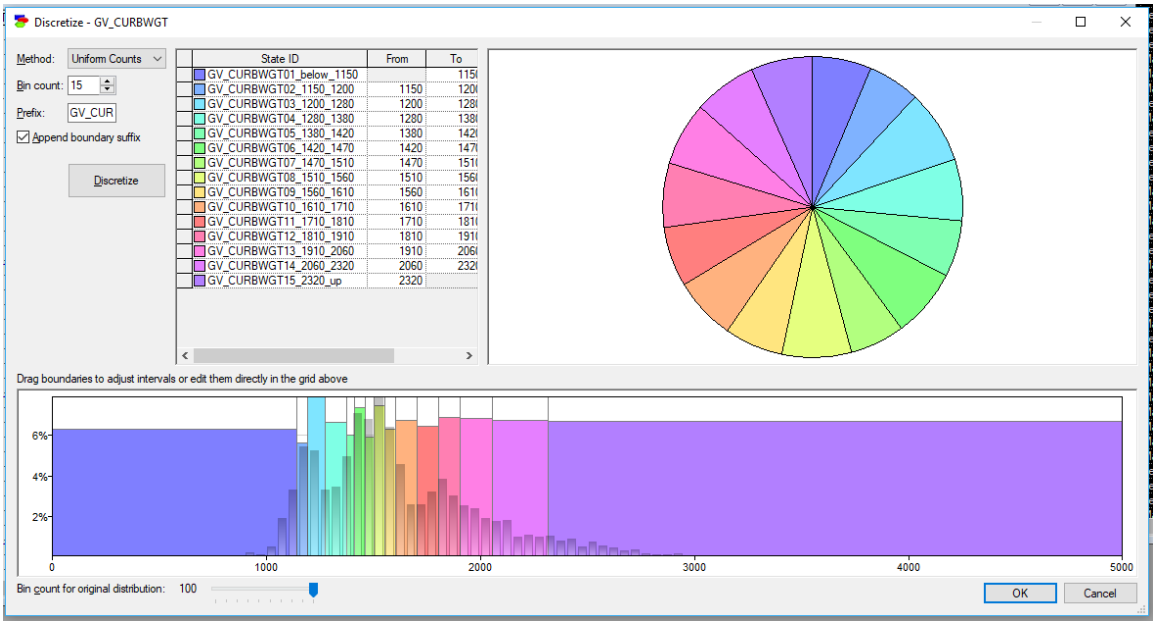
The naive and tree augmented Bayes network that we are building does not accept continuous variables as input, therefore we must group the continuous variables present in our dataset into distinct ranges called bins. We have thirteen continuous variables that we must discretize to bins. We can choose the bin sizes based on either equal number of observations in each bin, equal range for all bins (quartile, percentile, etc.) or in a hierarchical way or we can also use custom bin sizes.

Binning helps us in figuring the outliers and anomalous data in our data-set and to reduce the noise and non-linearity in our data. We have used a mixture of hierarchical, uniform width and based on range based on the distribution of the variable.

We then assign the label prefix to them to make the ranges more meaningful. In the example below, we have binned the passenger age as described below.

GV_CURBWGT is prefixed for all the bins created

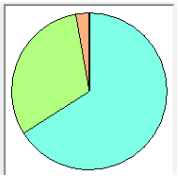
After binning all the continuous input variables our data-set attains the following structure as showed in the image below. All the variables have been labelled accordingly.



Input Bins are mostly in the range of 5-10 because the predictor Y-Variable has almost about 5 classes or bins.

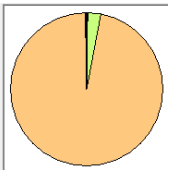
BINNING for all Variables

State ID	From	To	Count
GV_CURBWGT...		885	15
GV_CURBWGT...	885	1695	12653
GV_CURBWGT...	1695	2555	6000
GV_CURBWGT...	2555	3715	525
GV_CURBWGT...	3715		10



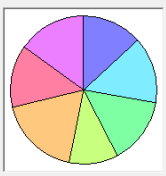
GV_CURBWGT

State ID	From	To	Count
GV_DVLAT1_below...		-103	2
GV_DVLAT2_n103...	-103	-70.5	11
GV_DVLAT3_n70_...	-70.5	-53	28
GV_DVLAT4_n53_...	-53	-24.5	527
GV_DVLAT5_n24_43	-24.5	43.5	18578
GV_DVLAT6_43_99	43.5	99.5	56
GV_DVLAT7_99_up	99.5		1



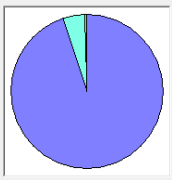
GC_DVLAT

State ID	From	To	Count
GV_DVLONG1_...		-29	2537
GV_DVLONG2_...	-29	-21	2799
GV_DVLONG3_...	-21	-16	2807
GV_DVLONG4_...	-16	-13	2069
GV_DVLONG5_...	-13	-8	3462
GV_DVLONG6_...	-8	-2	2636
GV_DVLONG7_...	-2		2893



GV_DVLONG

State ID	From	To	Count
GV_ENERGY1_...		1607.5	18218
GV_ENERGY2_...	1607.5	4282.5	888
GV_ENERGY3_...	4282.5	6970.5	83
GV_ENERGY4_...	6970.5	8231.5	3
GV_ENERGY5_...	8231.5		11

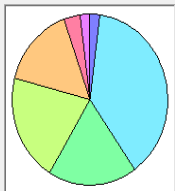


GV_ENERGY

State ID	From	To	Count
GV_MODELYR1...		2001.5	5247
GV_MODELYR2...	2001.5	2003.5	5007
GV_MODELYR3...	2003.5	2005.5	3944
GV_MODELYR4...	2005.5	2009.5	4501
GV_MODELYR5...	2009.5		504

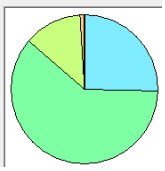
GV_MODELYR

State ID	From	To	Count
GV_Lanes1_bel...		1.5	389
GV_Lanes2_1_2	1.5	2.5	7349
GV_Lanes3_2_3	2.5	3.5	3520
GV_Lanes4_3_4	3.5	4.5	3922
GV_Lanes5_4_5	4.5	5.5	2982
GV_Lanes6_5_6	5.5	6.5	651
GV_Lanes7_6_up	6.5		390



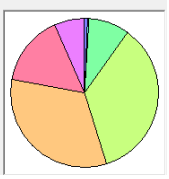
GV_LANES

State ID	From	To	Count
GV_OTVEHWG...		760	15
GV_OTVEHWG...	760	1355	4858
GV_OTVEHWG...	1355	2075	11697
GV_OTVEHWG...	2075	2835	2443
GV_OTVEHWG...	2835	3730	167
GV_OTVEHWG...	3730	4320	13
GV_OTVEHWG...	4320		10



GV_OTVEHWGT

State ID	From	To	Count
GV_SPLIMIT1_b...		7.5	125
GV_SPLIMIT2_7...	7.5	17.5	61
GV_SPLIMIT3_1...	17.5	27.5	1711
GV_SPLIMIT4_2...	27.5	37.5	6783
GV_SPLIMIT5_3...	37.5	47.5	6278
GV_SPLIMIT6_4...	47.5	57.5	2973
GV_SPLIMIT7_5...	57.5		1272

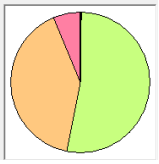


GV_SPLIMIT

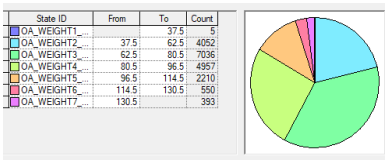
State ID	From	To	Count
OA_AGE1_belo...		11	2
OA_AGE2_11_28	11	28.5	6248
OA_AGE3_28_36	28.5	36.5	2999
OA_AGE4_36_48	36.5	48.5	4140
OA_AGE5_48_64	48.5	64.5	3760
OA_AGE6_64_76	64.5	76.5	1298
OA_AGE7_76_up	76.5		756

OA_AGE

State ID	From	To	Count
OA_HEIGHT1_b...		97.5	21
OA_HEIGHT2_9...	97.5	124.5	5
OA_HEIGHT3_1...	124.5	138.5	15
OA_HEIGHT4_1...	138.5	171.5	10158
OA_HEIGHT5_1...	171.5	186	7798
OA_HEIGHT6_1...	186	208.5	1203
OA_HEIGHT7_2...	208.5		3

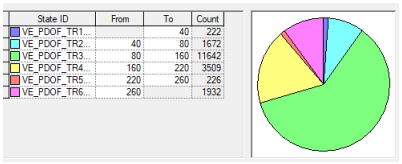


OA_HEIGHT



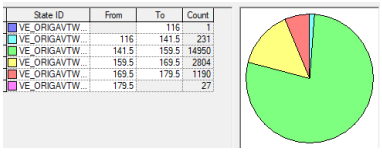
State ID	From	To	Count
OA_WEIGHT1_...		37.5	5
OA_WEIGHT2_...	37.5	62.5	4052
OA_WEIGHT3_...	62.5	80.5	7036
OA_WEIGHT4_...	80.5	96.5	4957
OA_WEIGHT5_...	96.5	114.5	2210
OA_WEIGHT6_...	114.5	130.5	550
OA_WEIGHT7_...	130.5		393

OA_WEIGHT



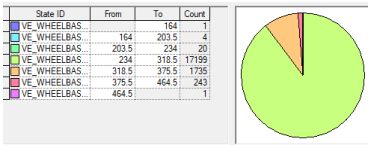
State ID	From	To	Count
VE_PDOF_TR1...		40	222
VE_PDOF_TR2...	40	80	1672
VE_PDOF_TR3...	80	160	11642
VE_PDOF_TR4...	160	220	3509
VE_PDOF_TR5...	220	260	226
VE_PDOF_TR6...	260		1932

VE_PDOF_TR



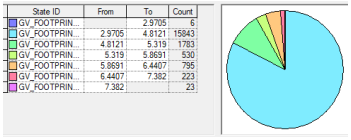
State ID	From	To	Count
VE_ORIGAVTW...		116	1
VE_ORIGAVTW...	116	141.5	231
VE_ORIGAVTW...	141.5	159.5	14950
VE_ORIGAVTW...	159.5	169.5	2804
VE_ORIGAVTW...	169.5	179.5	1190
VE_ORIGAVTW...	179.5		27

VE_ORIGAVTW



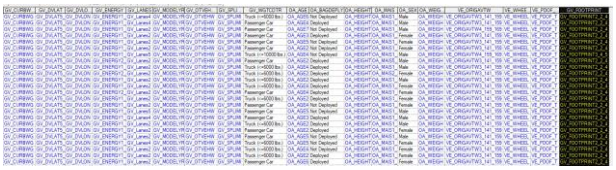
State ID	From	To	Count
VE_WHEELBAS...		164	1
VE_WHEELBAS...	164	203.5	4
VE_WHEELBAS...	203.5	234	20
VE_WHEELBAS...	234	318.5	17199
VE_WHEELBAS...	318.5	375.5	1735
VE_WHEELBAS...	375.5	464.5	243
VE_WHEELBAS...	464.5		1

VE_WHEELBAS



State ID	From	To	Count
GV_FOOTPRIN...		2.9705	6
GV_FOOTPRIN...	2.9705	4.8121	15843
GV_FOOTPRIN...	4.8121	5.319	1783
GV_FOOTPRIN...	5.319	5.8691	530
GV_FOOTPRIN...	5.8691	6.4407	795
GV_FOOTPRIN...	6.4407	7.382	223
GV_FOOTPRIN...	7.382		23

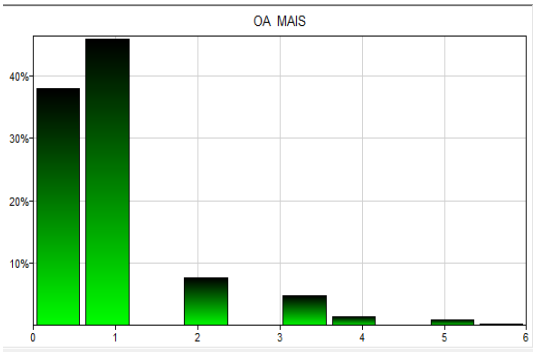
GV_FOOTPRINT



Data Set Snapshot

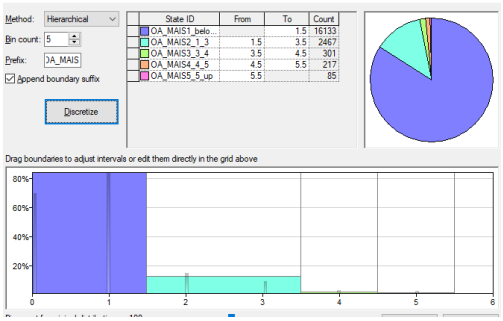
DATA BINNING: OUTPUT VARIABLE

Distribution of injury levels



The target variable OA_MAIS, rates injury by body zone and according to relative importance. Scored on a 6-point scale, lowest score of 0 indicates a no injury and max score of 6 indicates a severe/fatal injury. Each injury level has a certain probability of death to it as follows:

No Injury (0) - NIL
Minor Injury (1) - 0%
Moderate (2) - 2%
Serious Injury (3) - 10%
Severe/Injury (4) - 50%
Critical/Injury (5) - 50%
Maximum/Injury (6) - 100%



State ID	From	To	Count
OA_MAIS1_belo...		1.5	16133
OA_MAIS2_1_3	1.5	3.5	2467
OA_MAIS3_3_4	3.5	4.5	301
OA_MAIS4_4_5	4.5	5.5	217
OA_MAIS5_5_up	5.5		85

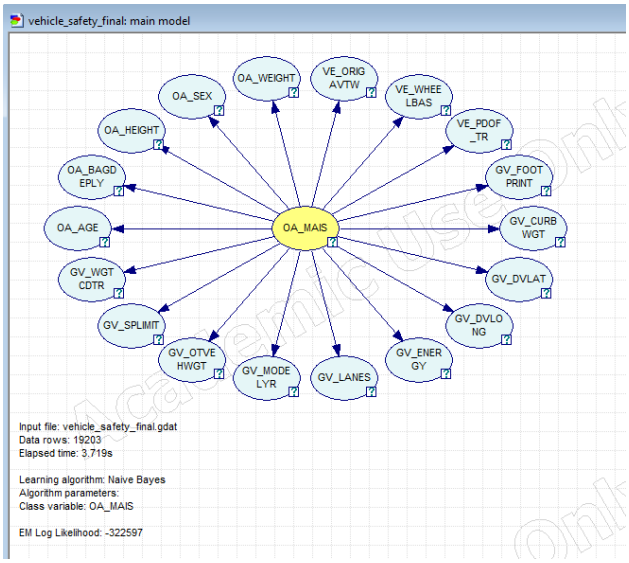
Based on the distribution, we have categorized the bins into 5 levels as shown in the fig

OUR NETWORK MODELS

We used GeNIe to build two types of Bayesian network models.
1) Naïve Bayes Network
2) Tree Augmented Naïve Bayes (TAN) Network

NAIVE BAYES

NETWORK DIAGRAM



ACCURACY

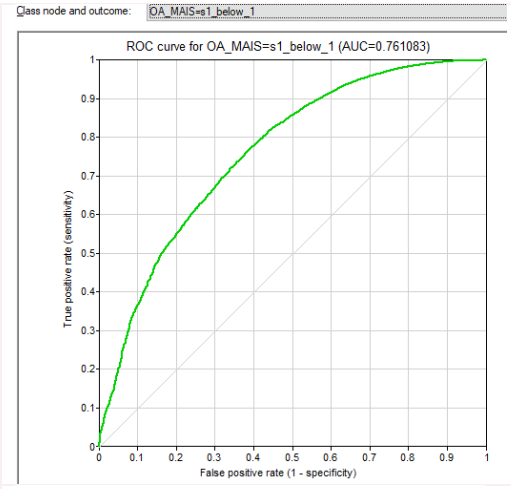
Confusion Matrix	Accuracy

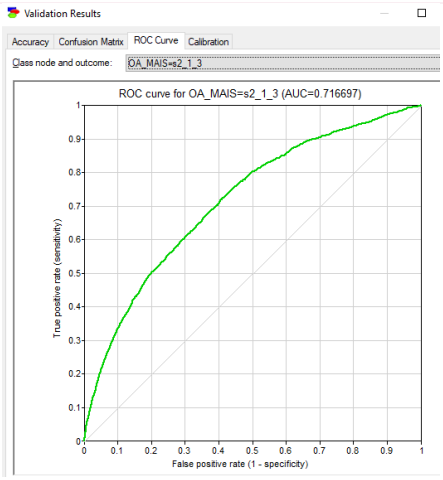
Class node: OA_MAIS

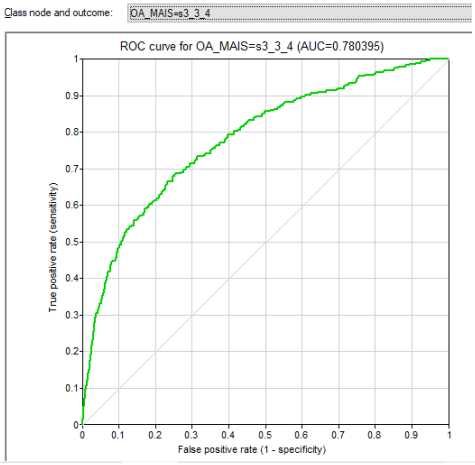
		Predicted				
		s1_below_1	s2_1_3	s3_3_4	s4_4_5	s5_5_up
Actual	s1_below_1	**15669**	431	13	14	6
	s2_1_3	1983	**409**	17	31	27
	s3_3_4	198	73	**6**	14	10
	s4_4_5	111	59	11	**21**	15
	s5_5_up	24	36	4	11	**10**

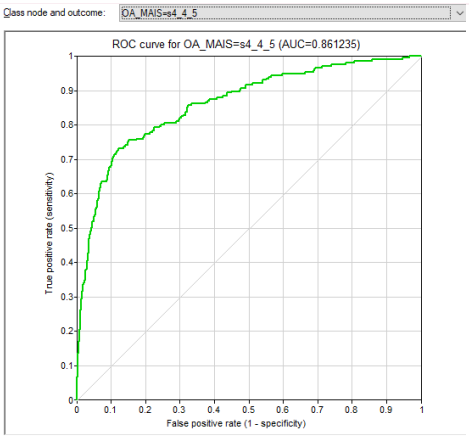
OA_MAIS = 0.839192 (16115/19203)
- s1_below_1 = 0.971239 (15669/16133)
- s2_1_3 = 0.165788 (409/2467)
- s3_3_4 = 0.0199336 (6/301)
- s4_4_5 = 0.0967742 (21/217)
- s5_5_up = 0.117647 (10/85)

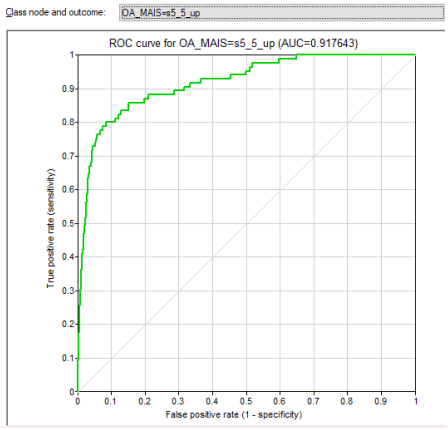
Area under the Curve (AUC)





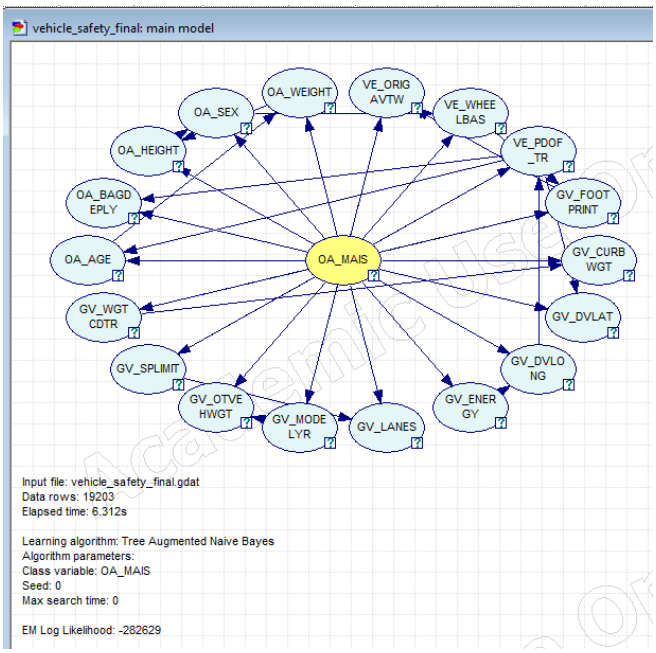






TREE AUGMENTED NAÏVE BAYES

NETWORK DIAGRAM



ACCURACY

Confusion Matrix

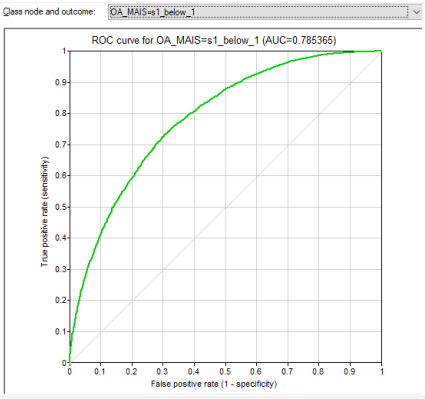
Class node: OA_MAIS

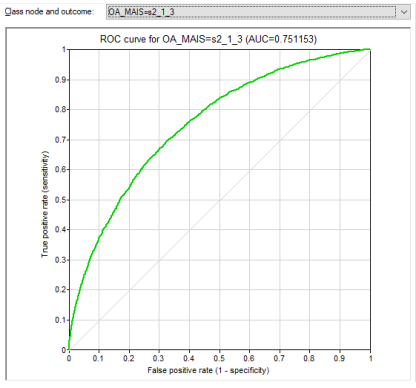
		Predicted				
		s1_below_1	s2_1_3	s3_3_4	s4_4_5	s5_5_up
Actual	s1_below_1	**15777**	313	21	20	2
	s2_1_3	2011	**396**	16	32	12
	s3_3_4	183	76	**35**	4	3
	s4_4_5	102	62	3	**46**	4
	s5_5_up	24	28	3	7	**23**

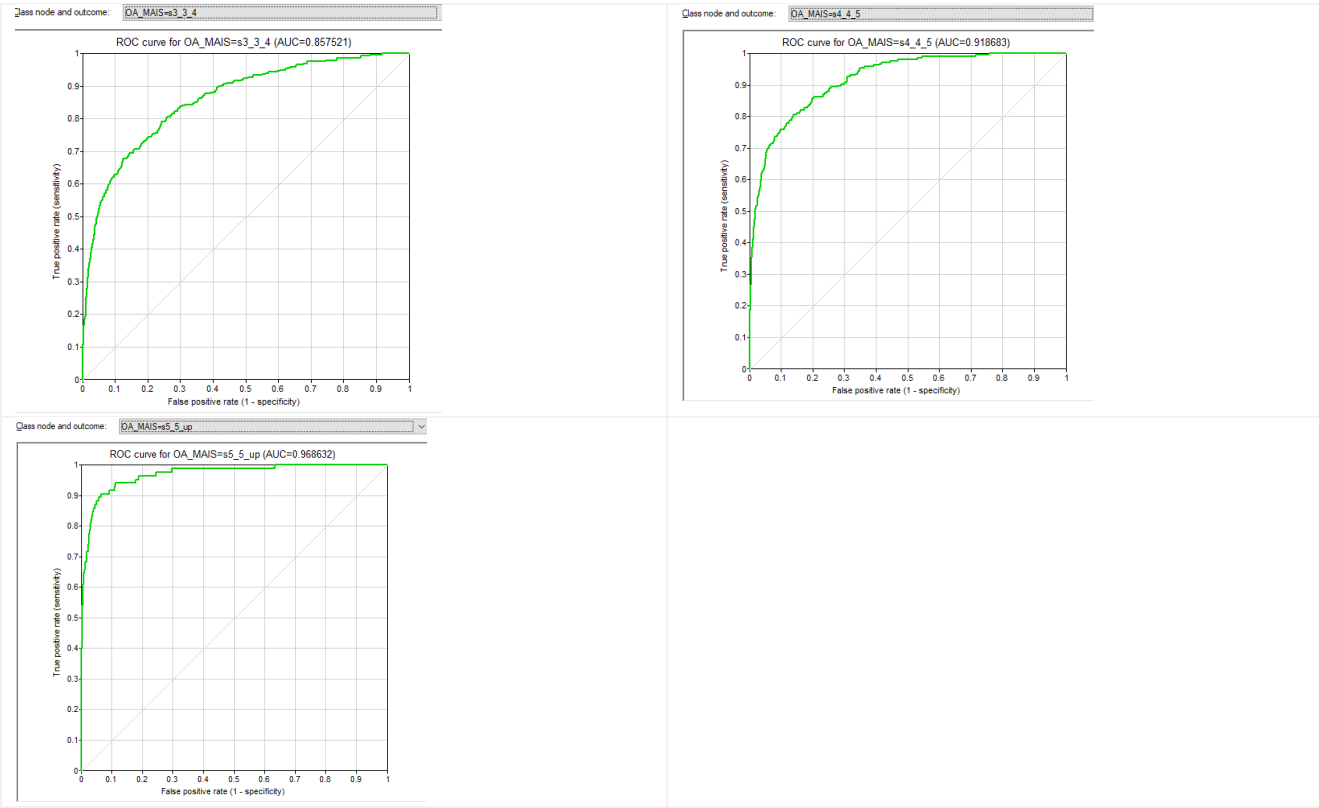
Accuracy

OA_MAIS = 0.847628 (16277/19203)
- s1_below_1 = 0.977933 (15777/16133)
- s2_1_3 = 0.160519 (396/2467)
- s3_3_4 = 0.116279 (35/301)
- s4_4_5 = 0.211982 (46/217)
- s5_5_up = 0.270588 (23/85)

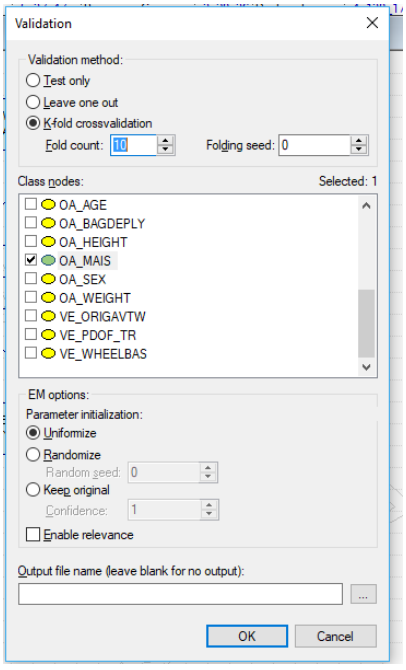
Area under the Curve (AUC)







VALIDATION



We used 10-fold cross validation technique while learning the above two network models. It helps in evaluating our predictive models by partitioning the binned data set into training set and test set.

The data set will be randomly sampled into 10 equal sized subsamples. Out of these single subsamples is retained as test data to evaluate the model and the remaining subsamples are used to train the model. This process is repeated 10 times and the results from this are averaged to give out a single estimation. The advantage of this method is that all the data in the data set are being considered while training as well as testing the model.

HOW GOOD ARE OUR NETWORK MODELS?

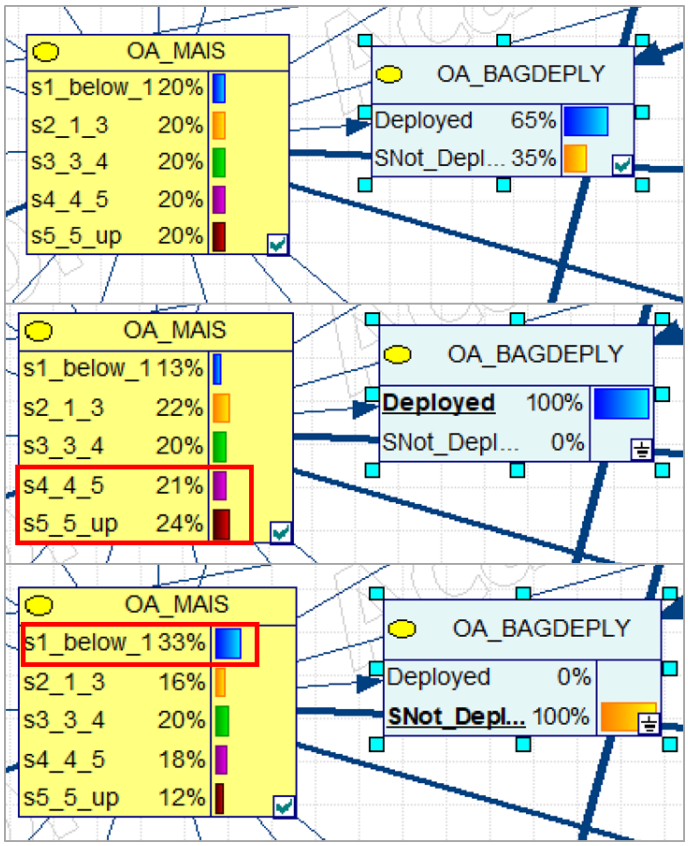
Both networks had similar accuracy rates (about 0.8). They do very well in predicting class 0; however, they do shockingly worse for the other classes. Only TAN had a slight edge in accuracy in other classes.

We suspect it is the dataset distribution – most of the observations were skewed towards class 1. For the rest of this section, we will use the TAN Network.

Intuitively, there are a number of questions that we use to interrogate the network.

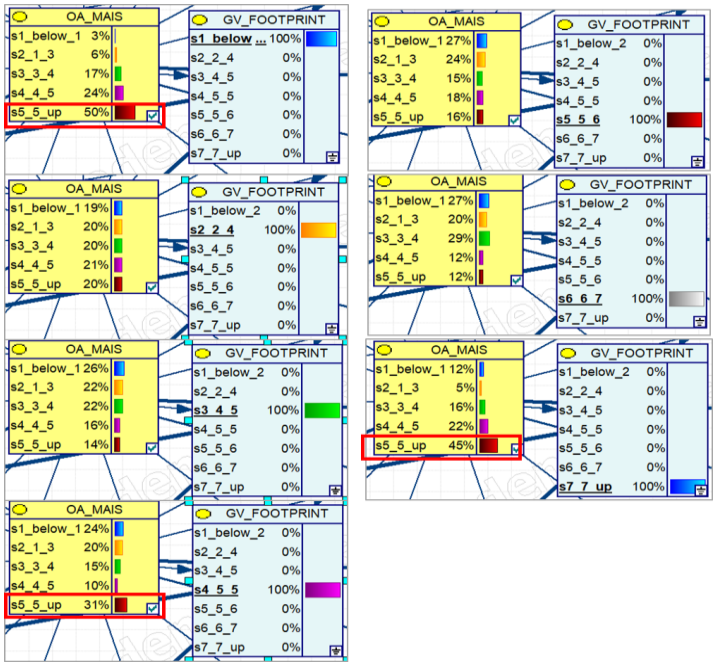
1. Would deployed airbags result in lesser injury?
2. Would greater vehicle footprint result in greater injury?
3. Which gender is likely to face greater injury?
4. For each injury state, what is likely to cause it?

1. Would deployed airbags result in lesser injury?



Oddly enough it doesn't. In fact, it increases the likelihood of severe injuries (See red boxes). Drivers seem to be better off without air bags.

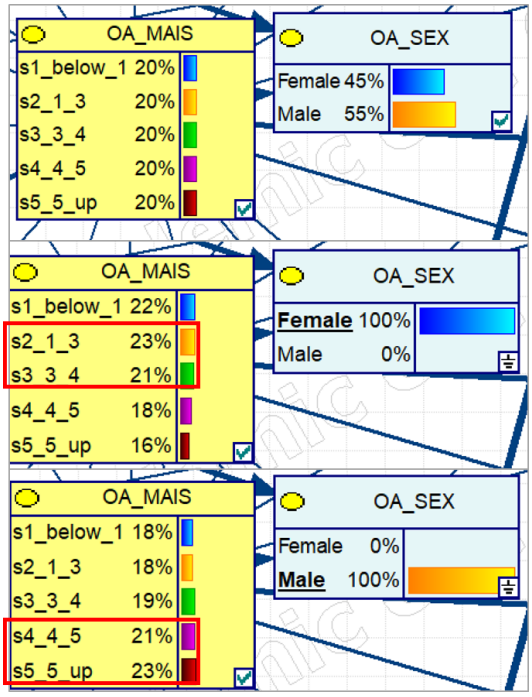
2. Would greater vehicle footprint result in greater injury?



Oddly enough, the smallest and largest vehicles have the highest rates of serious injuries.

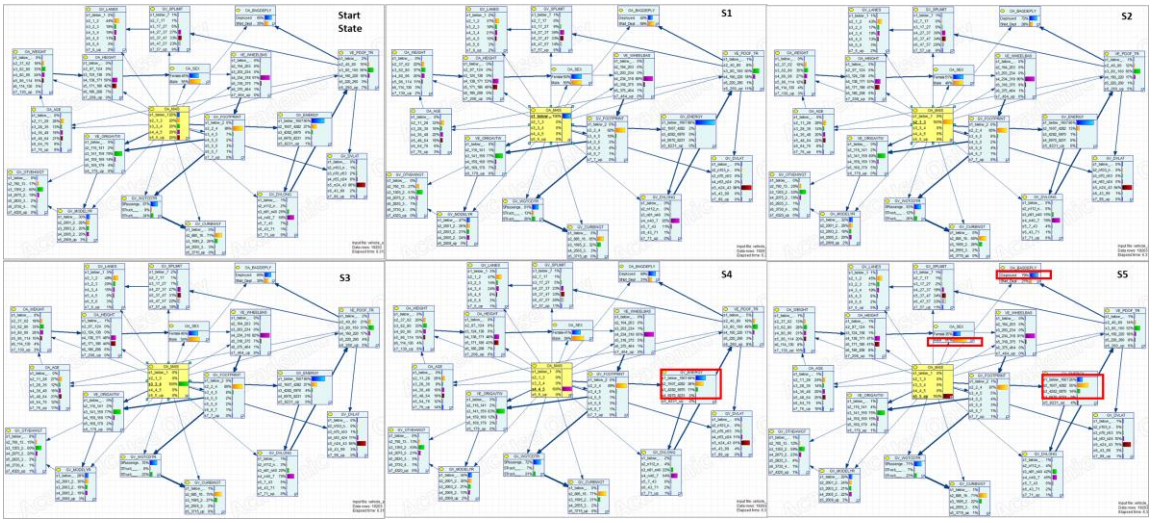
It could simply be due to the protective coverage and amount of energy transfer respectively.

3. Which gender is likely to face greater injury?

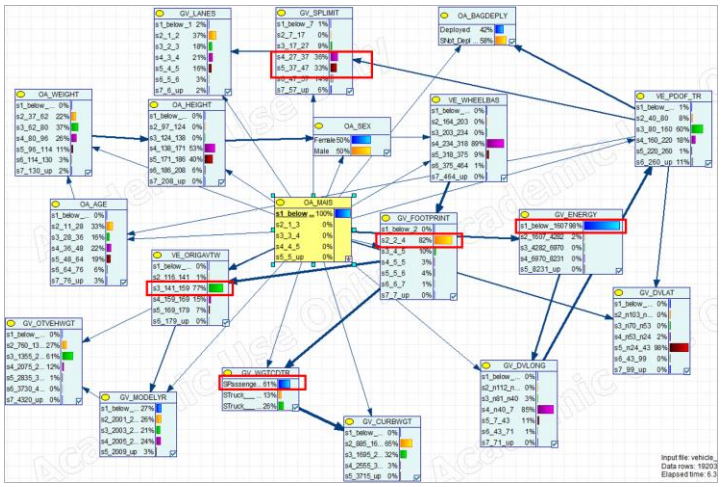


There is a bit of difference but not by much. It's likely due to weight and height difference of the gender.

4. For each injury state, what is likely to cause it?



The differentiation is quite minute until you get to more serious injury states. In this case, we believe that the accident is more likely to have males, with deployed air bags, and have energy levels of s2_1607_4282.



For accidents without injury, we can say that the vehicle was traveling between 27 to 47 MPH, likely to have a small vehicle size, likely to be a passenger vehicle, and somehow the energy was below 1607J. It is a toss up between male and female drivers.

Conclusion

We have compared two kinds of Bayesian networks – Naïve Bayes & Tree Augmented Network (TAN). The Naïve Bayes model assumes independence between attributes given the class, whereas TAN (belonging to the augmented network type) considers dependence between its class and other attributes. As seen from the network and its analysis there are some features with mutual influence and this information captured by TAN is reflected in its slight performance improvement. We also posed some questions and using Exploratory Network Analysis, found some insights on the different factors influencing vehicle safety.

REFERENCES

1. Bayesian Lab
2. Naniar
3. Bind_Shadow
4. Binning
5. PMM Imputation
6. C. Gower. A general coefficient of similarity and some of its properties. Biometrics, 27:857--874, 1971.