Bayesian and Neural Network Analysis Utilizing the NHTSA Crash Safety Database

Gabrielle Mack

CS 634 Final Project

Dr. Jason Wang

SPRING 2017

Table of Contents

Abstract: *Head injury criterion (HIC) is a measurement used to detect the amount of damage the head can sustain in the event of a collision. The metric is used widely in safety equipment manufacturing for sports as well as vehicle crash safety tests. The National Highway Traffic and Safety Administration (NHTSA) maintains a database of crash safety results for all vehicles manufactured from 1969-2017. The intent of this paper is to utilize Bayesian Networks to illustrate causality between factors that can affect an HIC was well as attempt to predict HIC with the use of a Neural Network.*

History of Safety Mechanisms

From the dawn of the first production vehicle built by Karl Benz in 1886, automotive safety has been a major concern for the population. Although seatbelts were patented in 1885, their use was not truly advocated until the 1930s. It would be another 40 years before Australia would enact the first seatbelt law in 1970. A decade would pass before New York enacted the initial seatbelt law for the United States in 1984. Likewise, it would take 10 more years for the remaining states to follow suit. Airbags, which were patented in 1951, would not be offered as an option in production vehicles until 1974. One more decade would pass before they were installed as standard equipment in fleets, a result the Federal Motor Vehicle Safety Standard passed in 1984. Crumple zones were patented in 1937 but were not used until 1959. As time has passed, these mandatory features have grown to include Electronic Stability Control, Anti-lock Brakes, Child Safety and Booster Seats, Safety Glass, and much more. Each of these mechanisms provides a vital component to keeping drivers and passengers safely operating motor vehicles nationwide. This requires the oversight from a regulatory body in the United States known as the NHTSA.

National Highway Traffic and Safety Administration (NHTSA)

Mounting pressure from the American public regarding automobile legislation led to the creation of the Department of Transportation (DOT) in 1966. Subsequent departments within the DOT would be created to manage both highway and vehicle safety standards. In 1970, these departments merged to create the NHTSA. The NHTSA is responsible for establishing and maintaining federal motor vehicle safety standards, emissions standards, and vehicle import legislation. The organization also maintains multiple data repositories used for research:

- ***Fatality Analysis Reporting System*** – Nationwide census of fatality data gathered from vehicular crashes across the United States. (Link)
- ***Vehicle Crash Test Database –*** Engineering data from various crash safety tests used to assess safety ratings in vehicles. (Link)
- ***Biomechanics Test Database*** – Experimental data used for developing Anthropomorphic Test Devices (crash dummies) and the injuries that are sustained by them. (Link)
- ***Component Test Database*** – Engineering data gathered from various instrumentation during crash tests. (Link)
- ***Crash Safety Models*** – Simulation models that are available per vehicle to support research. (Link)

As a regulatory body, the NHTSA provides this data to help ensure the safety and wellbeing of America's motorists while also lending its influence to engineering practices and scientific advancement.

Head Injury Criterion

Head Injury Criterion measures damage to the head caused by rapid deceleration. Multiple crash safety mechanisms have been developed to manage rapid deceleration including seatbelts, airbags, and crumple zones. These serve as a means to minimize the amount of g force exerted on an occupant of a vehicle. HIC scores are assessed to vehicles that have undergone crash safety testing by the NHSTA. Originally proposed in 1999, the acceptable HIC score has decreased from 1000 to 700. The formula for calculating Head Injury Criterion is as follows:

$$HIC = \max(t_1 \text{ or } t_2)\left\{(t_2 - t_1)\left[\frac{1}{t_2 - t_1}\int_{t_1}^{t_2} a(t)\,dt\right]^{2.5}\right\}$$

Where t1 and t2 represent the time period for which complete deceleration occurs. Vehicles without airbags will have a much higher head injury criterion.

.Table ES.1: Proposed Head Injury Criterion for Various Dummy Sizes

Dummy Type	Large § Male	Mid-Sized Male	Small Female	6 Year Old Child	3 Year Old Child	1 Year Old Infant
Existing HIC_{36} Limit	NA	1000	N/A	N/A	N/A	N/A
Proposed HIC_{15} Limit	700	700	700	700	570	390

§ The Large Male (95th percentile Hybrid III) is not currently proposed for inclusion in the SNPRM, but the performance limits are listed here for completeness.

Source: Development of Improved Injury Criteria for the Assessment of Advanced Automotive Restraint Systems II – Eppinger Et Al, NHTSA

Dataset

The dataset used in this study was provided by the National Highway Traffic and Safety Administration.(NHTSA). It contains 16,454 records of crash safety data from vehicles manufactured from 1965-2017. This particular database is valuable as it illustrates over time the advent and efficacy of improved safety mechanisms. The data was retrieved from the occupancy table of the database containing over 14,000 rows of crash dummy data, the intent of which is to analyze head injury criterion. Subsequent rows created for test barriers were removed, as well as rows for commercial vehicles such as limousines and buses. For this analysis, the focus is strictly on non-commercial, personal vehicles manufactured from 1965-2017.

Preprocessing

The dataset is made available online via multiple pipe separated files which were imported into a MySQL database for manipulation. The NHTSA performs multiple types of testing using barriers, walls, and other vehicles. For the purposes of this study, the focus is specifically on the occupancy table. This table lists results and measurements for tests that contained crash test dummies. This table was then joined with the vehicle table using TESTNO primary key. This particular query resulted in a total of 18,460 rows to account for test barriers. Rows were subsequently removed for the following conditions:

- HIC was less than 9999, as this is a default value used for tests in which HIC was not recorded.
- MAKE was equal to OTHER or NHTSA, as these rows were specific to barriers used in the test
- BX1-BX21 are null or have default 9999 value. These fields are pre-test measurements that are necessary for establishing a potential causal pattern in the Bayesian network.
- AX1-AX21 are null or have default 9999 values. These fields are post-test measurements that are necessary for establishing a potential causal pattern in the Bayesian network.

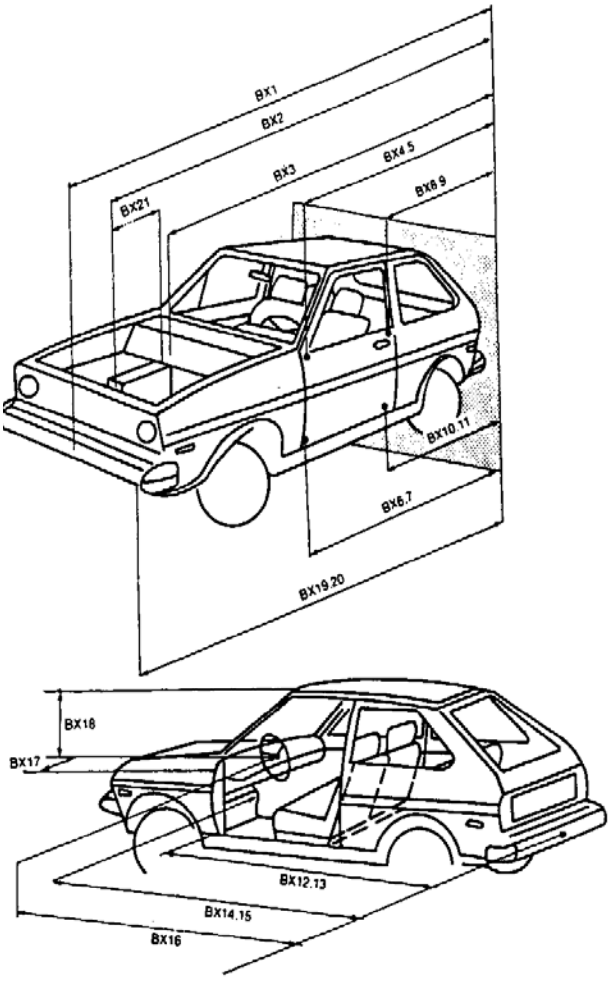
The resulting subset of data used in the study amounted to 726 rows. The vehicle table also required modification for the BODY variable, as both networks required numerical data for each model. As such, the preprocessed data was altered according to the following tables:

Label	Field Value	Description
1	2C	Two Door Coupe
2	2S	Two Door Sedan
3	3H	Three Door Hatchback
4	4P	Four Door Pickup
5	4S	Four Door Sedan
6	5H	Five Door Hatchback
7	CV	Convertible
8	EX	Extended Cab Pickup
9	MV	Minivan
10	PU	Pickup Truck
1	SW	Station Wagon
12	UV	Utility Vehicle
13	VN	Van

Occupant Location ID	Occupant Location
1	Left Front Seat (Driver Seat)
2	Right Front Seat (Front Passenger Seat)
3	Right Rear Seat
4	Left Rear Seat
5	Center Front Seat (Older Vehicles)
6	Center Rear Seat
7	Left Third Seat
8	Center Third Seat

The database also contained 21 unique measurement fields of data collected before (BX) and after (AX) each crash to assess vehicle damage during testing. Initially, all fields were used for analysis, as is the case for the Neural Network, but the fields were truncated to only include BX/AX 4,5,8, and 9 for the Bayesian Network. These particular measurements are crucial points that would affect head injury for side impact collisions.

BX1 - Total Length of Vehicle at Centerline
BX2 - Rear Surface of Vehicle to Front of Engine
BX3 - Rear Surface of Vehicle to Firewall
BX4 - Rear Surface of Vehicle to Upper Leading Edge of Right Door
BX5 - Rear Surface of Vehicle to Upper Leading Edge of Left Door
BX6 - Rear Surface of Vehicle to Lower Leading Edge of Right Door
BX7 - Rear Surface of Vehicle to Lower Leading Edge of Left Door
BX8 - Rear Surface of Vehicle to Upper Trailing Edge of Right Door
BX9 - Rear Surface of Vehicle to Upper Trailing Edge of Left Door
BX10 - Rear Surface of Vehicle to Lower Trailing Edge of Right Door
BX11 - Rear Surface of Vehicle to Lower Trailing Edge of Left Door
BX12 - Rear Surface of Vehicle to Bottom of A Post of Right Side
BX13 - Rear Surface of Vehicle to Bottom of A Post of Left Side
BX14 - Rear Surface of Vehicle to Firewall, Right Side
BX15 - Rear Surface of Vehicle to Firewall, Left Side
BX16 - Rear Surface of Vehicle to Steering Column
BX17 - Center of Steering Column to A Post
BX18 - Center of Steering Column to Headliner
BX19 - Rear Surface of Vehicle to Right Side of Front Bumper
BX20 - Rear Surface of Vehicle to Left Side of Front Bumper
BX21 - Length of Engine Block



Pretest Measurement Data (BX1 - BX21)

Hypotheses

Based on preliminary analysis of the data, certain assumptions were made. For example, it is expected that there will be a marked increase in HIC as the years decrease. This is because safety mechanisms such as airbags were not required in vehicles until 1983. As such, there should be a decrease in HIC as the years increase due to the enactment of the Federal Motor Vehicle Safety Standard. There should be a an increase in HIC for vehicles with higher speed ratings or vehicles with higher measurement differences post test, as this should be indicative of a greater amount of force against the vehicle thereby resulting in a higher HIC. From a database standpoint, we should also observe a relationship between vehicle body type and HIC, as there is a disproportionate amount of rows for four door sedans compared to other vehicles. Lastly, there should be some type of relationship between HIC and occupant location (ie driver, passenger seat) within the vehicle. This is to demonstrate the difference between not having access to front impact airbags in the backset.

Bayesian Networks and Causal Analysis

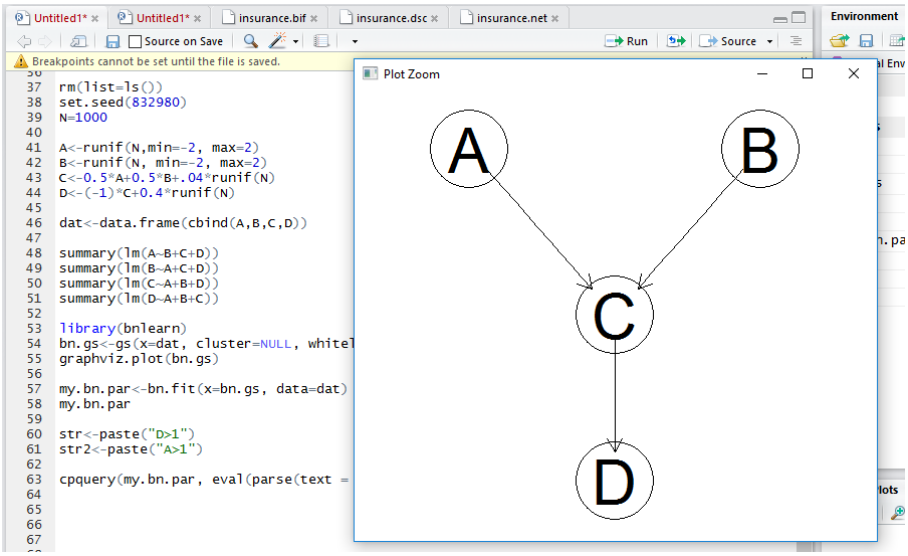
Bayesian networks utilize joint conditional probability distributions to perform causal analysis. They are also referred to as “belief networks” or “probabilistic networks”. They consist of a directed acyclic graph (DAG) representing variable relationships and associated conditional probability tables. Changes within the Bayesian network propagate throughout the network with network training. Nodes within the network represent variables in a child/descendent formality between each of the random variables, essentially mapping probabilistic relationships through a series of arcs. Bayesian networks utilize Bayes’ Rule:

$$\underbrace{P(A|B)}_{\text{Posterior}} = \frac{\overbrace{P(A|B)}^{\text{Likelihood}} * \overbrace{P(A)}^{\text{Prior}}}{\underbrace{P(B)}_{\text{Marginal Likelihood}}}$$

Arcs between random variables are known as edges and represent conditional dependencies. Nodes that are not connected via edges are considered to be conditionally independent of one another. Bayesian network graphs are usually presented with a set of conditional probability tables (CPTs) along with the acyclic graph.

BNLEARN Package and Bayesian Network Example

For the implementation of the Bayesian Network, R version 3.3 was installed alongside R Studio version 1.0.136. BNLEARN version 4.1.1 was also installed along with Bioconductor for graphviz (chart plotting). The BNLEARN package provides constraint, score, and hybrid based algorithms for Bayesian Learning. The example below illustrates the creation of a rudimentary Bayesian Network via the manual input of individual node values. Each node is directed, that is, the edges dictate which nodes is conditionally dependent of another.



Example: Manually coded Bayesian Network, Lecture from Prof. Justin Esarey, Rice University (link)

After the network is learned (in this case manually), the bnlearn function `cpquery` can be called to retrieve the Conditional Probability Tables for the network.

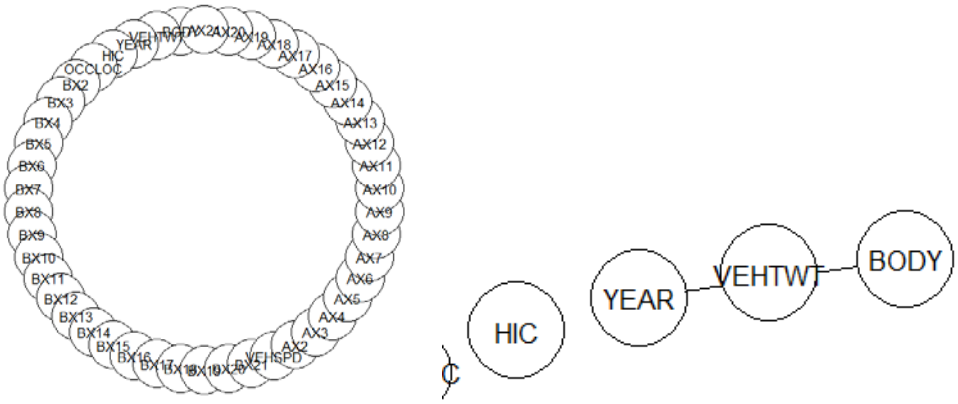
```
Console F:/E_DOCS/Documents/Education/CS/CS 634/Project Files/Final/bnlearn/nhtsa/
= NULL, debug = FALSE, optimized = TRUE, strict = FALSE, undirecte
> graphviz.plot(bn.gs)
> my.bn.par<-bn.fit(x=bn.gs, data=dat)
> my.bn.par

  Bayesian network parameters

  Parameters of node A (Gaussian distribution)

Conditional density: A
Coefficients:
(Intercept)
 0.02422979
Standard deviation of the residuals: 1.120489

  Parameters of node B (Gaussian distribution)

Conditional density: B
Coefficients:
(Intercept)
-0.05694167
Standard deviation of the residuals: 1.153254

  Parameters of node C (Gaussian distribution)

Conditional density: C | A + B
Coefficients:
(Intercept)            A            B
 0.01990602   0.49982336   0.50003527
Standard deviation of the residuals: 0.0115489

  Parameters of node D (Gaussian distribution)

Conditional density: D | C
Coefficients:
(Intercept)            C
  0.1977800   -0.9925974
Standard deviation of the residuals: 0.1155686
```

Example: CPTs from manually coded Bayesian Network, Lecture from Prof. Justin Esarey, Rice University (link)

Results: NHTSA Bayesian Network

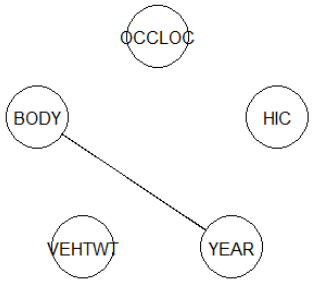
Rather than inputting the network manually, the approach is to use the Grow/Shrink (gs) algorithm to learn the network associations automatically. Preprocessed data was generated in a comma separated (csv) file. Systemic testing commenced through many different iterations of data modification.

The first preprocessed file contained a total of 47 original columns, none of which showed any type of relationship except for between year, vehicle weight, and body.



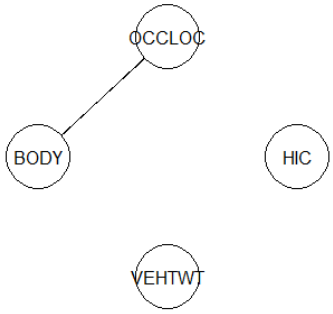
Attempt 1: Original 47 columns showing only a relationship between YEAR, VEHWT, and BODY

Focusing on the year, weight, body relationship, the majority of columns were truncated, to which only VEHWT, YEAR, HIC, and Occupant Location were remaining. This time, the network revealed only a relationship with body and year.



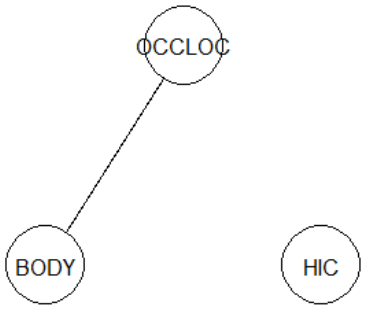
Attempt 2: Deletion of independent rows yielded only a relationship between BODY and YEAR.

Removing year resulted in a relationship between body and Occupant location:



Attempt 3: Deletion of independent rows yielded only a relationship between BODY and OCCLOC (Occupant Location).

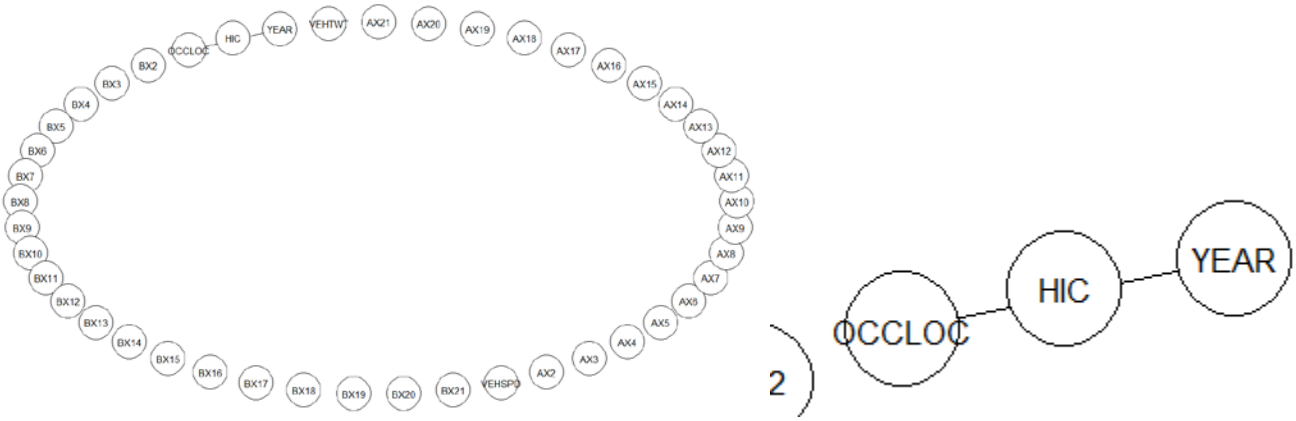
Removing vehicle weight revealed the same relationship.



Attempt 4: Deletion of VEHTWT (Vehicle Weight) yielded only a relationship between BODY and OCCLOC (Occupant Location)

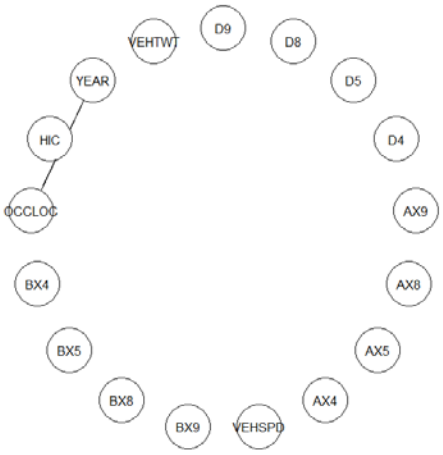
Finally, looking for a statistical relationship between occupant location and HIC revealed no statistical relationship.

A different approach was taken in which only records of the same vehicle type were selected. In this case, vehicles that were only 4 four door sedans were analyzed. This resulted in a relationship for HIC between the occupant location and year. The dataset was further truncated to focus specifically on the driver seat of this subset, but this did not reveal any discernable changes.



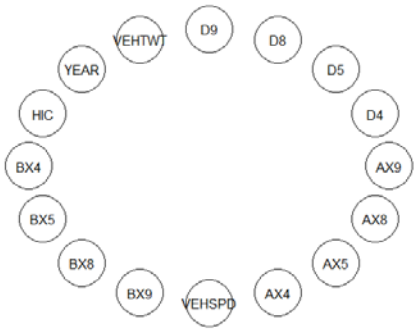
Attempt 5, 6: Filtering the preprocessed data set utilizing only vehicles of the same type (4 door sedan) and Occupant Location (Driver Seat) yielded the same results.

Another approach was to take the differences in measurements provided by each of the crash tests and compare them against each value for HIC. Measurements were values AX4, AX5, AX8, and AX9, which pertain to the driver's side A and B pillar doorway for a side impact crash. This resulted in an arc between simply occupant location and year, bypassing a value for HIC.



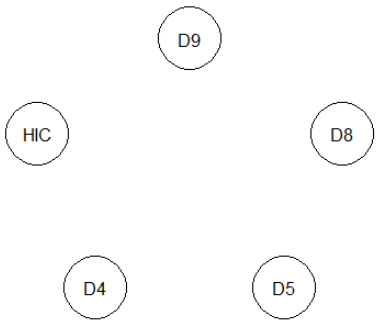
Attempt 7: Taking the difference in pre and post test AX/BX measurements yielded only a result between Occupant Location and Year.

Next, using the same file, records were filtered to only include one occupant location for the driver, but this resulted in no relationships:



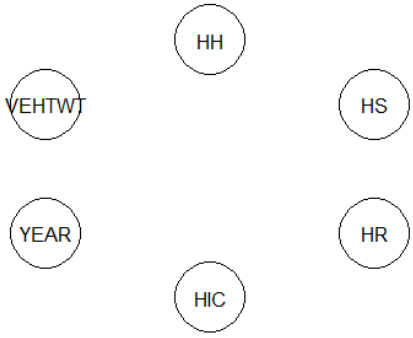
Attempt 8: Filtering pre and post test measurements for driver side impacts yielded no relationship.

Running an analysis strictly between HIC and the measurement distances did not yield any viable arcs:



Attempt 9: Analysis of HIC and driver side impact measurement criteria yielded no results.

Taking a different approach, additional data was retrieved from the occupants table to include measurements between the head of the crash test dummy and the interior of the vehicle. This data was filtered to include only driver side data. These new fields included HH-Head to Windshield Header, HW – Head to Windshield, HR – Head to Side Header, and HS – Head to Side Window. This data did not yield any interesting patterns for HIC:



Attempt 10: Additional head measurements from the Occupants table is imported for causal analysis

Artificial Neural Networks

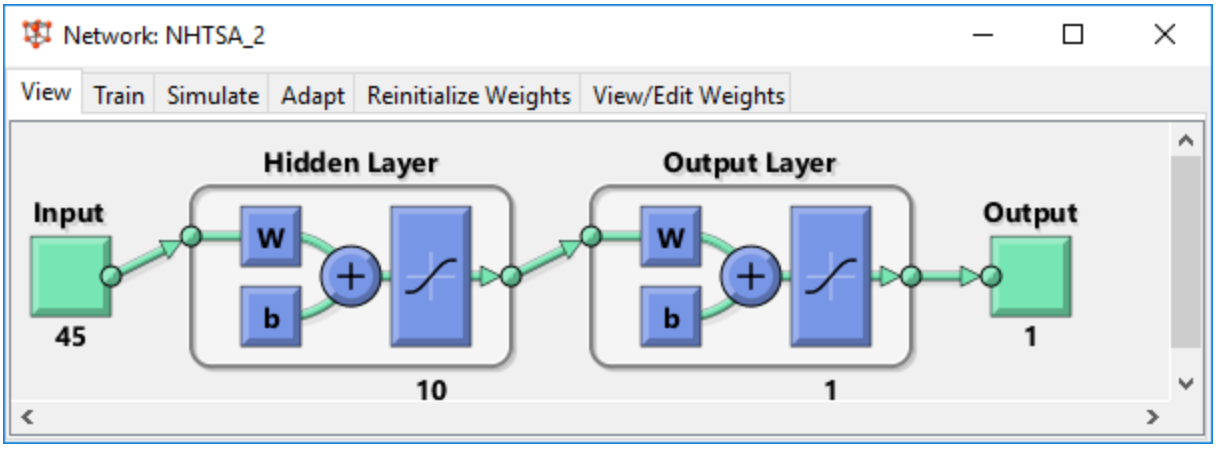
Whereas Bayesian networks detect causality within a dataset, Artificial Neural Networks (ANNs) are used for predictive results. Much like the human brain, ANNs are composed of neurons that when trained, can produce predictive results. This approach uses a set of mathematical weights to map input values to output values thereby establishing or extracting a clear pattern of machine learning. A network is able to learn or establish these patterns through repeated training intervals in which data is supplied to it. The weights supplied to the network can be modified to obtain the preferred result.

Testing an ANN model requires that the dataset be divided into two partitions. Data used to train the model is known as the training set. Data used to test the model simulation is known as the testing set. Algorithms in this step are generated and modified to prove the hypothesis for the business case. This iterative process continues until the goal for the model is reached.

MATLAB and Neural Network Example

The goal of this Neural Network was to attempt to predict HIC values given the existing dataset. MATLAB version R2016b (9.1.0) using the Neural Network (nntool) toolbox was installed and utilized. The data was partitioned twice for training and testing different versions of the Neural Network.

Of the 685 rows available from the preprocessed data 500 were used for training the first network and the remaining 185 were used for testing the simulation. This resulted in the original 45 columns being used for training input, with the output variable being the predicted HIC.

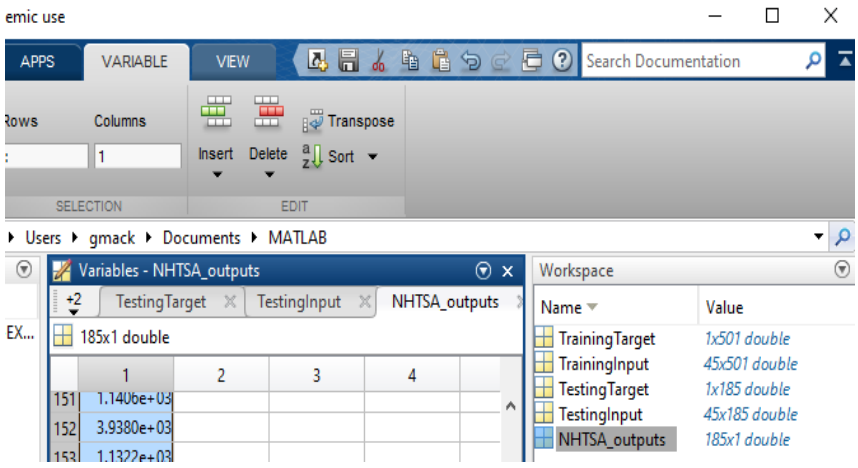


Attempt 1: Neural Network Diagram

The network in all instances contained 10 hidden layers. Weights were not modified from the weights created by `nntool`. The columns listed below were used to predict a score:

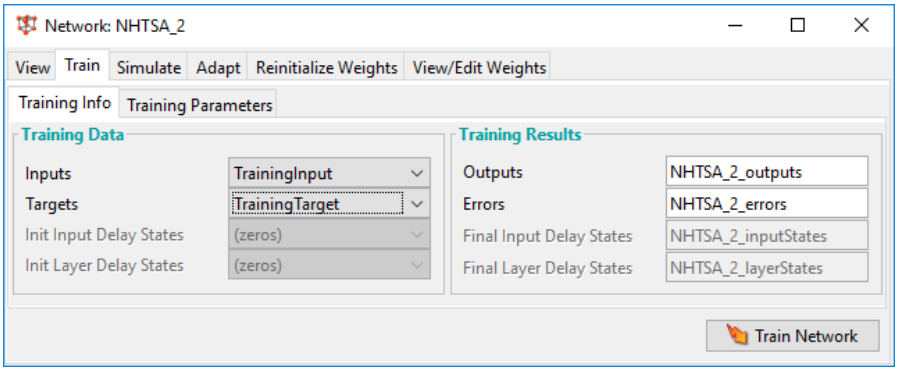
Data Columns		
BODY	BX15	AX7
VEHTWT	BX16	AX8
YEAR	BX17	AX9
OCCLOC	BX18	AX10
BX2	BX19	AX11
BX3	BX20	AX12
BX4	BX21	AX13
BX5	VEHSPD	AX14
BX6	AX2	AX15
BX7	AX3	AX16
BX8	AX4	AX17
BX9	AX5	AX18
BX10	AX6	AX19
BX11		AX20
BX12		AX21
BX13		HIC
BX14		

The columns were then separated and transposed for use in the network.



Attempt 1: Training and Testing input as 501 and 185 rows respectively

The network was then trained several times using the training data set of 501 rows:



Attempt 1: Training using the 501 rows in the training data set

Training is an iterative process by which the network learns the behavior of the data through a series of algorithms. Preliminary training yielded the following results:

Network 1 (501 training, 185 testing)				
Attempt	**Training**	**Validation**	**Test**	**All**
1st	.45	.40	.31	.40
7th	.70	.46	.52	.64
10th	.77	.71	.19	.65
25th	.84	.77	.20	.72
40th	.84	.85	.92	.85
50th	.87	.84	.77	.85

The scores for the network continued to increase after every iteration of testing. However, simulation results did not indicate a proper predictive score for HIC when compared against the target values. At the 25th training attempt original HIC values were totaled along with the simulated values to produce a 73% difference in data scoring.

Original HIC Total Value	**Simulated HIC Total Values**	**Percentage Difference**
278830	482661	73.10%

Likewise, at the 50th attempt of training the network, values had increased by almost 100%.

Original HIC Total Value	**Simulated HIC Total Values**	**Percentage Difference**
278830	556023	99%

A second network using the same dataset was created and yielded the following results, but simulated HIC values still did not align with the actual HIC values.

Network 2 (501 training, 185 testing)				
Attempt	**Training**	**Validation**	**Test**	**All**
1st	.04	.00	.38	.03
10th	.53	.37	.61	.51
20th	.39	.53	.80	.51
30th	.58	.69	.81	.63
40th	.61	.61	.73	.63

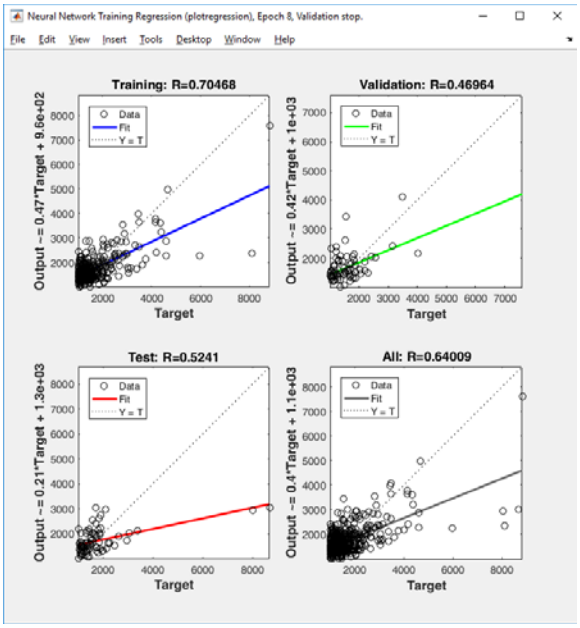
Finally, a 3rd attempt where the training and test datasets were modified to increase the training dataset and decreasing the testing dataset yielded the following results.

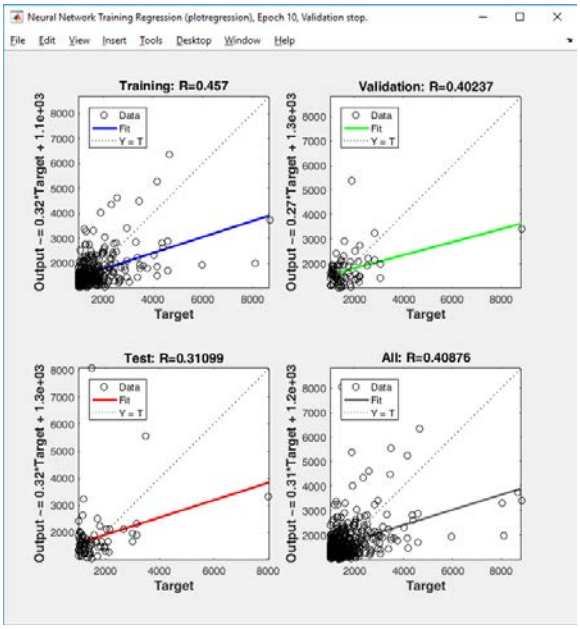
Network 3 (585 training, 101 testing)				
Attempt	**Training**	**Validation**	**Test**	**All**
1st	.01	0	.08	.004
10th	.01	0	0	.013
20th	.02	.122	.07	.013
30th	.04	.09	.03	.01
40th	.03	0	.07	.013

Original HIC Total Value	**Simulated HIC Total Values**	**Percentage Difference**
278830	116747	-58%

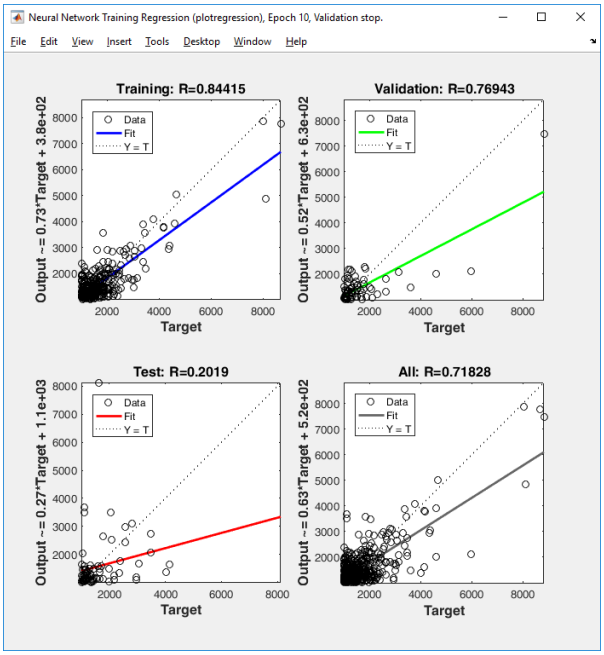
2202	2435	11%	1001	-55%
1597	2402	50%	1001	-37%
1093	1824	67%	1001	-8%
1184	7298	516%	1001	-15%
1200	1973	64%	1001	-17%
1328	3797	186%	1001	-25%
1568	2334	49%	1001	-36%
1071	2434	127%	1001	-7%
1051	4574	335%	1001	-5%
278830	556023	99%	116747	-58%

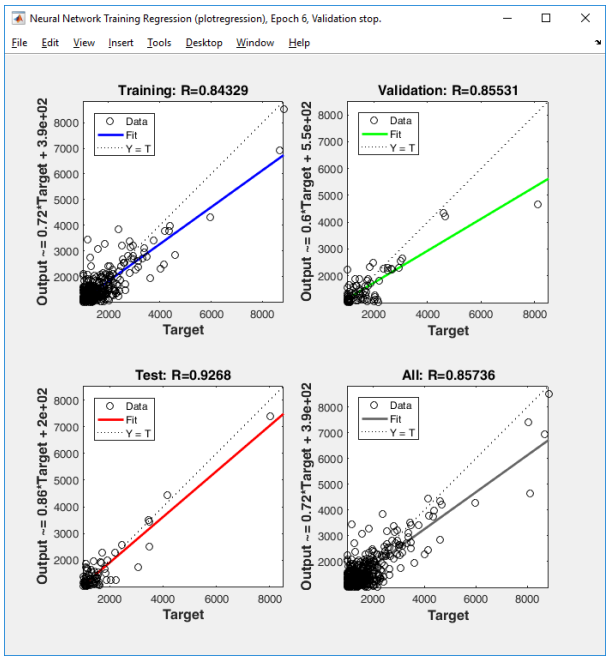
This testing application yielded a 58% decrease. In addition, it is interesting to note that the majority of the dataset had a value of 1001 as the calculated HIC value.





Attempt 1: 7th and 1st Iterations of training respectively





Attempt 1: 25th and 40th iterations of training respectively

Observations and Conclusion

Both network instances did not yield the expected results or confirm any of the preliminary hypotheses. Observations can be made that the nature of the data may not be consistent in its current state with predicting HIC values. In hindsight, the data itself is highly fragmented and may not be sufficient for these particular methods of machine learning. For example, although the dataset does contain tests from many different types of vehicles, actual metrics pertaining to the safety features of the vehicle are not included in the dataset. Different body types of vehicles may have skewed the way the

Neural Network could learn about effect of vehicle weight and speed on HIC. Different crash test dummy models are also used in each test, and the metrics for those are also lacking, as well as logistical information regarding the actual barriers used in the tests. Although barrier information was omitted from the dataset, the metrics regarding weight, speed, and force could be used to attempt to predict HIC. Each vehicle tested is unique in the dataset with respect to test type. The same model is not administered the same type of test more than once. Even the use of pre and post test measurements is not an accurate method for which to depend on, as each vehicle has a completely different construction from the next. This results in a rather sporadic dataset of different vehicles for which pattern recognition and machine learning might not be able to create a distinguishable pattern.

References

Websites:

Federal motor vehicle safety standards and regulations: https://one.nhtsa.gov/cars/rules/import/FMVSS/index.html

NHTSA Crash Safety Database: https://www-nrd.nhtsa.dot.gov/database/veh/veh.htm

NHTSA Research Division: https://one.nhtsa.gov/Research

R Version Documentation: https://stat.ethz.ch/R-manual/R-devel/library/base/html/Version.html

Head Injury Criterion: http://www.intmath.com/applications-integration/hic-head-injury-criterion.php

Books and Papers:

R. Nagarajan, M. Scutari and S. Lèbre (2013). Use R!, Vol. 48, Springer (US). ISBN-10: 1461464455

Lawrence J. Mazlack, "Considering Causality in Data Mining", International Conference on Software Engineering – 2008.

John Binder, "Adaptive Probabilitstic Networks with Hidden Variables", Machine Learning, 1997