Collaborative Sciences Center for ROAD SAFETY

FINAL REPORT

Machine Learning Tools for Informing Transportation Technology and Policy

11/18/2019

Mary Cummings (P.I.)
Songpo Li
Duke University

THE UNIVERSITY OF TENNESSEE KNOXVILLE
Berkeley UNIVERSITY OF CALIFORNIA
FAU FLORIDA ATLANTIC UNIVERSITY
Duke UNIVERSITY
THE UNIVERSITY of NORTH CAROLINA at CHAPEL HILL

www.roadsafety.unc.edu

U.S. DOT Disclaimer

Acknowledgement of Sponsorship

This project was supported by the Collaborative Sciences Center for Road Safety, www.roadsafety.unc.edu, a U.S. Department of Transportation National University Transportation Center promoting safety.

TECHNICAL REPORT DOCUMENTATION PAGE

1. Report No. CSCRS-R10	**2. Government Accession No.**	**3. Recipient's Catalog No.**
4. Title and Subtitle: **Machine Learning Tools for Informing Transportation Technology and Policy**		**5. Report Date** November 18th 2019
		6. Performing Organization Code
7. Author(s) Mary Cummings and Songpo Li		**8. Performing Organization Report No.**
9. Performing Organization Name and Address Humans and Autonomy Laboratory 130 North Building Duke University Durham, NC 27708		**10. Work Unit No.**
		11. Contract or Grant No. Collaborative Sciences Center for Road Safety (Grant #: 69A3551747113)
12. Sponsoring Agency Name and Address Collaborative Sciences Center for Road Safety 730 Martin Luther King Jr. Blvd., Suite 300 Chapel Hill, NC 27599		**13. Type of Report and Period Covered** Final Report (March 2018-November 2019)
		14. Sponsoring Agency Code

15. Supplementary Notes
Conducted in cooperation with the U.S. Department of Transportation, Federal Highway Administration.

16. Abstract
With rapid advances in data analytic tools, machine learning techniques are increasingly being used to study large transportation crash data sets. The objective of such approaches is to reveal underlying and potentially unknown patterns of influence between driver and pedestrian characteristics, environment factors, vehicle attributes and crash fatalities. However, machine learning results can be greatly affected by the subjectivity of the machine learning practitioner, where the practitioner subjectively selects the machine learning algorithm and algorithm parameters for a specific data set, and either this person or perhaps other people then interpret the results. Little work has been conducted to study this practitioner-induced subjectivity problem in order to understand its causes, influences, and methods for avoidance, particularly in transportation settings. To help fill this gap, two transportation datasets examining driver and pedestrian accident fatalities were analyzed with two different machine learning techniques of low and high complexity (logistic regression and neural networks). The results demonstrate that both the type of model and feature interpretation method produce different results in terms of model performance and assessment of feature importance. These outcomes that highlight more than ten opportunities for subjective decisions suggest that more work is needed in looking at how such subjective modeling and interpretation choices affect the use of machine learning models in support of policy decision making.

17. Key Words Explainable artificial intelligence, Machine Learning (ML), Logistic Regression (LR), Neural Network (NN), feature selection	**18. Distribution Statement**		
19. Security Classif. (of this report) Unclassified	**20. Security Classif. (of this page)** Unclassified	**21. No. of Pages** 34	**22. Price**

Form DOT F 1700.7 (8-72) Reproduction of completed page authorized

Contents

Machine Learning Tools for ____ 1
Informing Transportation Technology and Policy ____ 1
U.S. DOT Disclaimer ... 2
Acknowledgement of Sponsorship ... 2
Introduction ____ 5
Transportation Data Set #1: The Role of Road Infrastructure in Accidents ____ 7
Methodology ... 7
Logistic Regression for HSIS ... 8
Model Accuracy ... 8
Feature Weights ... 9
HSIS Neural Network Analysis ... 11
Model Accuracy ... 12
Method 1: Feature analysis using “Leave one out” ... 13
Method 2: Feature analysis using first layer weights ... 14
Method 3: Feature analysis using weights & standard deviation ... 15
Method 4: Feature analysis using weights, standard deviation & the NN accuracy drop ... 15
Comparison of Results ... 16
Conclusions from Transportation Case Study #1 ... 17
Transportation Data Set #2: A Not-So-Big Data Approach to Pedestrian Safety ____ 18
Methodology ... 18
NASS Logistic Regression ... 19
Model Accuracy ... 19
Feature Weights ... 20
Neural Network for NASS ... 21
Model Accuracy ... 21
Method 1: Feature analysis using “Leave one out” ... 22
Method 2: Feature analysis using first layer weights ... 22
Method 3: Feature analysis using weights & standard deviation ... 23
Method 4: Feature analysis using weights, standard deviation & the NN accuracy drop ... 23
Comparison of Results ... 24
Conclusion ____ 25
Acknowledgments ____ 25
References ____ 26
Appendix A: Variables in HSIS dataset ____ 28
Appendix B: Comparison of HSIS Original and Cleaned Data ____ 29
Appendix C: Histograms of HSIS variables before and after data imputation. ____ 30

Introduction

With rapid advances in data analytic tools, machine learning techniques are increasingly being used to study large transportation crash data sets. The objective of such approaches is to reveal underlying and potentially unknown patterns of influence between driver and pedestrian characteristics, environment factors, vehicle attributes, and crash fatalities. However, machine learning results can be greatly affected by the subjectivity of the machine learning practitioner, where the practitioner subjectively selects the machine learning algorithm and algorithm parameters for a specific data set, and either this person or perhaps other people then interpret the results. As depicted in Figure 1, when combined with biases that can be inadvertently introduced due to underlying sample selection bias (e.g., Samimi, Mohammadian et al. 2010, Gianfrancesco, Tamang et al. 2018), these practitioner-induced subjective biases introduced into a machine learning modeling process can make any resulting conclusions questionable.

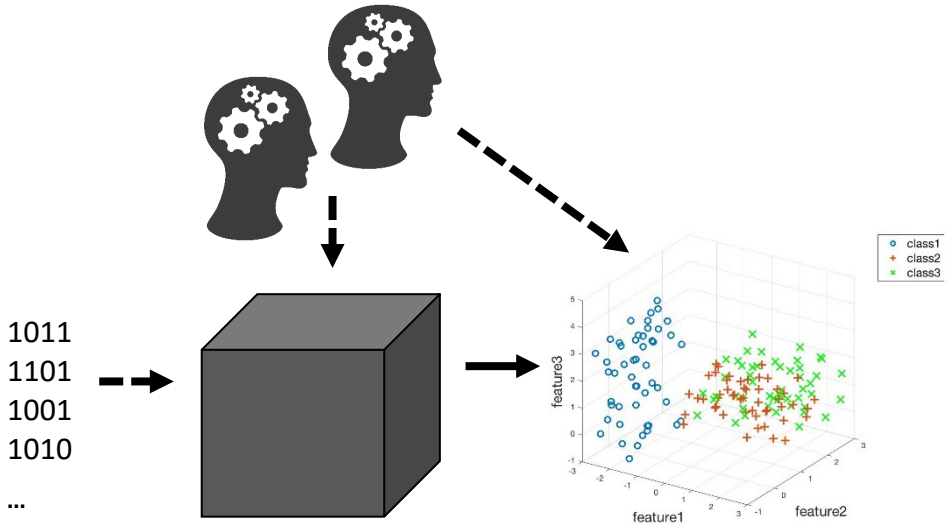


Figure 1: Depicted by a dashed line, sources of human bias in any machine learning modeling process including human subjectivity in model and parameter selection as well as sample selection bias coming from the data.

Little work has been conducted to study the practitioner-induced subjectivity problem in order to understand its causes, influences, and methods for avoidance, particularly in transportation settings. To help fill this gap, two transportation datasets examining car and pedestrian accident fatalities were analyzed with two different machine learning techniques of low and high complexity (logistic regression and neural networks respectively). While there are many other types of machine learning models that could be used (e.g., see Figure 2 and Cummings and Stimpson (2019)) these models were selected since they represent different model complexities and are both commonly used.

Figure 2. Complexity map of common machine learning models

Using both a large and small transportation data set that use various features to predict driver and pedestrian injuries and fatalities, the results of these models are compared with one another, as well as with different possible interpretations of the neural net features. The results demonstrate that both the type of model and feature interpretation method produce different results in terms of model performance and assessment of feature importance. In addition, examples of practitioner interpretations are included that span novice to expert which also exemplify how experience can modify one's interpretation of the results. These outcomes, which highlight the more than ten opportunities for subjective decisions, suggest that more work is needed in looking at how such subjective modeling and interpretation choices affect the use of machine learning models in support of decision making.

Transportation Data Set #1: The Role of Road Infrastructure in Accidents

The first data set analyzed used the Highway Safety Information System[1] (HSIS), a roadway-based data repository that provides data on a large number of accidents that include a number of roadway and traffic variables. It was developed by the Federal Highway Administration (FHWA) to help highway engineers to make better decisions about roadway design. For the purposes of this effort, the target variable (aka, the predicted or dependent variable) is whether a driver was seriously injured or killed vs. no serious injury. In machine learning modeling, once the target variable is known, the next step is to select the features, also known as predictor or independent variables.

Methodology

The selection of features in any machine learning model development effort represents a significant subjective decision point. For large data sets, there could be up to hundreds of possible features. The winnowing of such choices should be guided by the overarching research question and associated hypotheses as well as a thorough literature review and previous research and analyses using the dataset of interest.

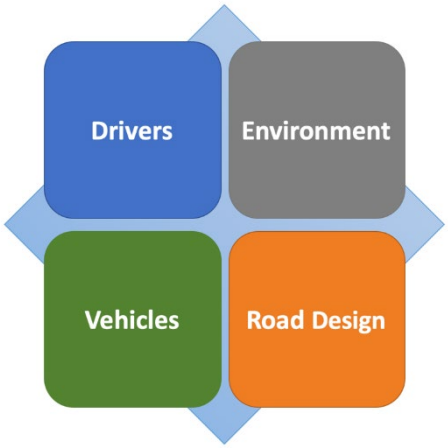


Figure 3: Classes of variables that potentially influence severity if vehicle accidents

For the HSIS data set, there were 248 features that could be selected, but using all would result in data overfitting and a loss of generalizability. We chose a set of variables that would help answer a representative transportation-related question that such a big data set could help answer which is "What road design elements most influence serious road accidents?" Given that previous research has shown that other elements like driver age, poor weather, and the type of vehicles drivers are in are also significant predictors (Peden, Scurfiled et al. 2004, Hermans, Brijs et al. 2006), we included variables from four classes as depicted in Figure 3.

We down-selected to 16 HSIS variables, as depicted in Table 1, and used data from two states, California and Washington, since they have similar data collection methods. Appendix A provides the exact categories and units for each of these variables. The final 16 features selected represent the four classes in Figure 3, but include a higher number of road safety variables since these were the primary focus. The target variable was the severity of injury in the accident, with 1 including both fatal and severe injuries because fatalities account for a very small percentage of crashes and 0 including all other injuries that were not severe.

The data had to be collected from the HSIS site in 3 separate files, which include accident, road, and vehicle files, and the files were linked using road keys, mileposts and accident keys. There were 968,371 accidents in the original data, but 53,481 records were dropped due to invalid data including outcomes, and more than three predictor variables missing. For those accident records with missing variables (N=660,675), a k-nearest-neighbor method was used to impute the missing values. K was initialized at 400 for records missing one, two, and three variables respectively. After selection and imputation, the final data set resulted in a total of 914,890 accidents.

Such complexity is typical of such large data sets but also represents potential sources of error. Two-sample Kolmogorov-Smirnov tests were performed on the original and cleaned data to examine whether the cleaned data still followed the distribution of the original data. Appendix B and C provide more details about the original dataset and cleaned dataset. Moreover, the HSIS dataset is imbalanced in that there were 23,949

[1] https://www.hsisinfo.org/

fatal / severe observations and 890,941 non-severe observations. Thus there are only 2.62% positive observations which is also typical for such large data sets describing extreme events like deaths in healthy populations. This imbalance in the data is important when evaluating such models, and this issue will be revisited in the results section.

Table 1. Variables selected for HSIS dataset

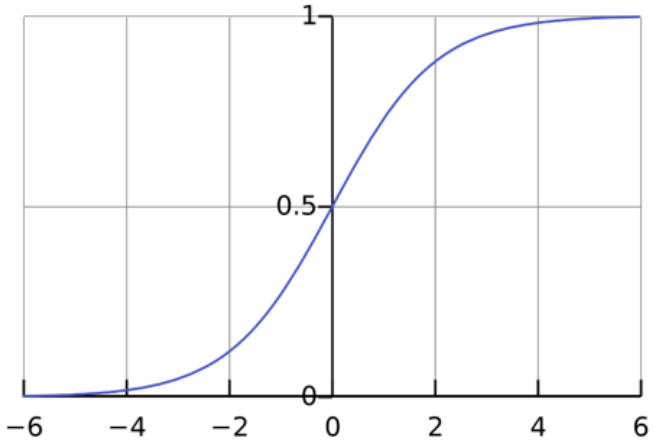
1	2	3	4	5	6	7	8
Speed limit	Average annual daily traffic	Access control type	Left shoulder width	Right shoulder width	Number of lanes	Median width	Section length
9	**10**	**11**	**12**	**13**	**14**	**15**	**16**
Light	Weather	Max driver age involved	Min driver age involved	Vehicle type	Sobriety	Urban / Rural	Lane width

Drivers Environment Vehicles Road Design

As mentioned previously, Logistic Regression (LR) and Neural Network (NN) analyses were both used to model the relationship between the injury severity and selected variables in Table 1. The following sections demonstrate how the two models performed and what insights were gained.

Logistic Regression for HSIS

Logistic Regression (LR) is a classification modeling approach that predicts a categorical variable from a set of predictor variables, also known as features. In binary LR, the variables/features attempt to predict a classification of 1 or 0 using a sigmoid function as depicted in Figure 4.



$$p(x) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x)}}$$

x = feature
β_0 = value of the criterion when features = 0
β_1 = regression coefficient
p(x) = probability of success

Figure 4. Logistic regression function

Model Accuracy

An accurate model is one that has a high success rate for predicting both fatalities and non-fatalities. However, in the case of this HSIS data set, if a model predicted every observation to be negative (non-fatal),

the overall accuracy would be 97.38% since there is only a 2.62% occurrence of fatalities. Thus, to develop an accurate model we must first select the statistical threshold to determine what constitutes a 0 or 1, without skewing the predictive results to just one class.

To determine the threshold, we iterated the threshold from 0 to 1 with step 0.01 to see how different thresholds would affect model performance, assuming the logistic regression model contains all 16 variables in Table 1. This threshold is in the region where the true positive rate (TPR) and overall accuracy cross beyond which the TPR falls exponentially. Thus, the threshold of 0.0234 balances model accuracy and TPR (Figure 5). At this value the model accuracy is 72.63% and the TPR is 71.14%.

For imbalanced data sets with binary outcomes, which is the case with this HSIS data set, model accuracy is often not a good indicator of model performance (Menardi and Torelli 2014). Another potential method that can be used to assess model performance, especially for imbalanced data, is examining the area under the Precision-Recall curve (Saito and Rehmsmeier 2015). Similar to Receiver Operator Characteristic (ROC) curves that plot true positive against false positive rates, precision-recall plots incorporate additional information. Precision is defined as the ratio of true positives to the sum of true and false positives. Recall is defined as the ratio of true positives to the sum of true positives and false negatives. Figure 6 demonstrates that the area under the curve is 0.1250 which suggests the model is correctly classifying samples it predicts as fatalities but may miss many classifications.

Feature Weights

Given that the model is acceptable, although not a particularly strong model, we then need to understand how the different 16 features contributed to the overall model. Table 2 lists the weights, exp(weights) and p-values. Those variables with a p value less than .0031 were considered significant. Because LR models produce regression coefficients for each feature that are log odds as shown in Figure 5, taking the exponential of the coefficient weights estimates the expected change in the log odds of the target variable per unit increase in the corresponding predictor variable holding the other predictor variables constant. Take, for example, variable 8 in Table 2 which is the section length. A one-unit increase in this variable increases the odds of a fatality by 6.2564 (6.2564:1). The weights less than zero decrease the odds of a fatality by 1/exp(W), so Table 2 details all the odd ratios accounting for those features with positive and negative weights.

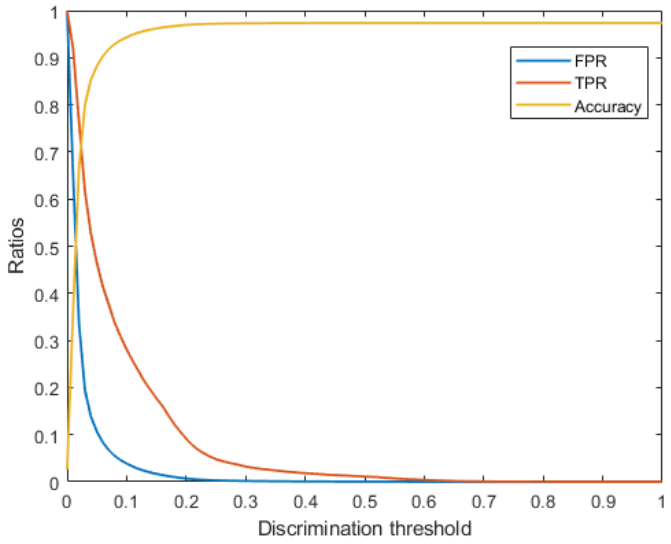


Figure 5. Different accuracies with different thresholds using LR. FPR = False Positive Rate, TPR = True Positive Rate

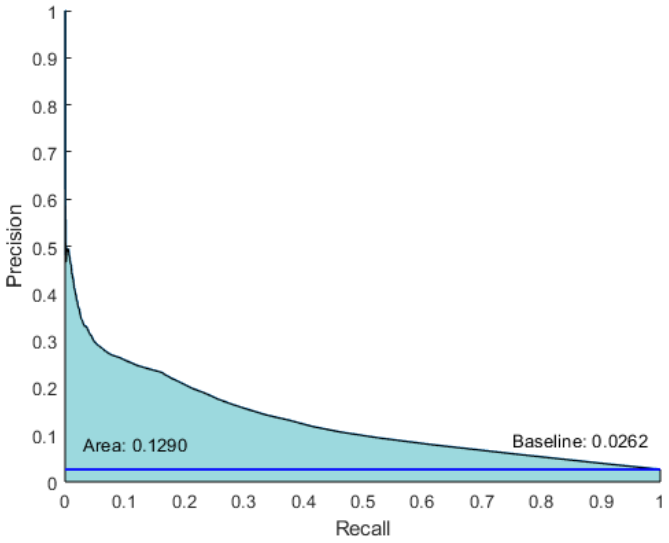


Figure 6. Precision-recall curve of the LR model. The baseline is the percentage of fatalities for the global dataset.

Table 2. Feature weight results of LR for HSIS, * indicates significance at p < 0.0031, which uses a family-wise error correction rate of α/16, where α= 0.05[2]

Feature	1	2	3	4	5	6	7	8
Weight	0.0280	-3.8286	0.0158	0.8704	-0.0052	-0.3087	-1.4184	1.8336
Exp(W)	1.0284	0.0217	1.016	2.3879	0.9948	0.7344	0.2421	6.2564
p-value	0.5522	0*	0.6124	0*	0.9610	0.0277	0*	0*
Sig. Odds Ratio		45.9979		2.3879			4.1304	6.2564
Feature	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**
Weight	0.7460	-0.1598	0.5811	0.4164	2.0926	1.6240	-0.7059	-0.1879
Exp(W)	2.1085	0.8523	1.7879	1.5164	8.1064	5.0733	0.4936	0.8287
p-Value	0*	0*	0*	0*	0*	0*	0*	0.0171
Sig. Odds Ratio	2.1085	1.1733	1.7879	1.5164	8.1064	5.0733	2.0257	

As depicted in Figure 1, determining which features are the most important is a subjective decision with different decision criteria producing different results. One common interpretation is all those features that are statistically significant should be in the model and in this case, 11 different variables mattered the most (Tables 1-3). However, another interpretation could be that only those odds ratios greater than 2 should be considered since every integer over 1 represents a 100% increase in likelihood, and such a rule would capture the features with the largest contributions to the model. As depicted in Table 3, if this threshold was selected, then the odds ratios suggest the important features would drop from 11 to 8, with Average Annual Daily Traffic (2), Vehicle type (13), Section length (8), Sobriety (14), median width (7), Left shoulder width (4), Light (9), and Urban/Rural (15) mattered the most, in this order.

[2] The selection of the .05 criteria for a p value is a subjective decision, and was selected for this study because it is a conventional norm.

In answer to the research question, “What road design variables are the most important?” the answer would be left shoulder width (feature 4), median width (feature7, more is better (Stamatiadis and Pigman 2009), and section length (feature 8, roads with longer consistent sections have more accidents) which has been seen in previous research (Hadi, Aruldhas et al. 1995). While previous research has shown that increasing the shoulder width generally leads to less fatalities (Neudorff, Jenior et al. 2016), this HSIS data set was initially slightly biased in the opposite direction (Appendices B and C). The slight difference in means of shoulder width (5.30 ft for people with serious or fatal injuries vs. 5.28 ft for those not seriously injured) may be statistically significant but is not practically significant.

Table 3. LR variable importance by decision criteria

Selection Criteria	Important Features										
Statistically significant	2	4	7	8	9	10	11	12	13	14	15
Odds Ratios > 2	2	4	7	8	9	13	14	15			

So, if a transportation engineer wanted to know what road design features mattered the most in preventing fatalities, it appears that with this model, section length is the most important (with the highest odds ratio).

Ultimately, the choice of what variables are the most important has to be made by the practitioner, and budgetary and complexity constraints could drive threshold selection. Understanding real dollars are at stake when making such decisions, and given that the LR model was not particularly strong due to the lower model accuracy, developing another model based on the very popular machine learning approach of neural networks can be investigated for comparison. This approach is detailed in the next section.

HSIS Neural Network Analysis

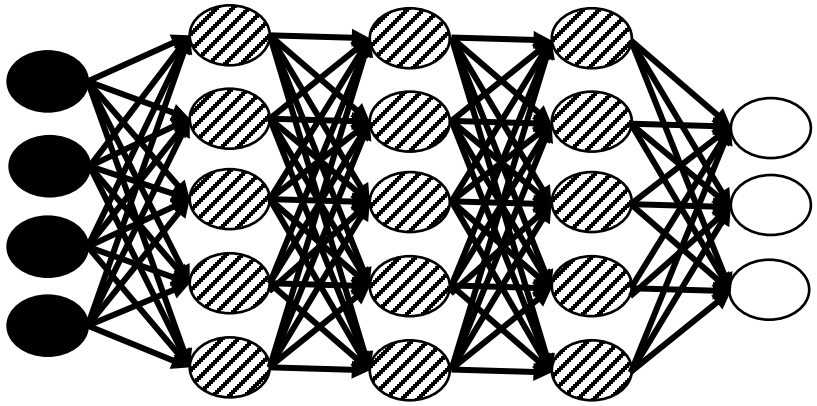


Figure 7. Demonstration of a Neural Network

Figure 7 is a representation of a neural network (NN), where black shadow and white circles represent input neurons (model inputs), hidden neurons, and output neurons (model outputs), respectively. For each link, there is a weight parameter. Looking left to right is the forward propagation (predicting process), while right to left is back propagation (the learning process). Such models must first be trained on a subset of the overall data set. In the training process, all the parameters in the network must first be initialized to make the first forward propagation. Then the cost (deviation between output and true result) is calculated. After that, parameters will be adjusted to minimize the deviation between model predictions and the desired outputs. This process is repeated until overall model accuracy cannot be improved.

Due to their multi-layered complicated structure, NNs can be very powerful and can represent non-linear high-order relationships. However, interpreting NN models can be very difficult. Unlike LR models, there are no individual weights associated with the features. Since the weights are distributed across the network, ultimately partial feature weights combine in an unknown non-linear manner to contribute to the overall model (Ibrahim 2013). Thus, interpreting features important in a NN can carry significantly more subjectivity, as depicted in Figure 1, than with LR. This will be illustrated in the results section.

In order to develop a NN[3] for the same HSIS data set described earlier, we first initialized parameters of the neural network, including:

- Input: The input layer included 16 variables which included categorical, ordinal, and continuous data (Appendix A).
- NN structure: For the hidden layer size (number of neurons in each layer), we selected 10 but examined up to 100 which did not seem to make a difference in model performance.
- Output: The output layer consisted of a single node to predict 0 (non-fatal accident) or 1 (Serious injury or fatal accident).
- Ratios: When training a NN, the data must be divided into three sets such that the first set, the training set, is iteratively used for backpropagation, the second set is for validation used to avoid overfitting, and the third set is the final testing set, used only once. The final ratios used in this effort were 68% for training, 12% for validation, and 20% for testing.

During the training process, network weights were randomly initialized. In each backpropagation iteration, the cross-entropy loss is calculated, and the model generated a gradient towards the direction the parameters were adjusted. After the NN was trained, this NN model was then used for testing. Similar to LR, the outputs from the NN model need to be discriminated by a threshold to determine the positive and negative predictions.

Model Accuracy

Often when NNs are generated, the typical primary performance metric is the overall accuracy of such a model. In the case of the imbalanced HSIS data, as explained earlier in the LR section, we need both overall accuracy and true positive rate to evaluate the model. We took the intersection of the TPR curve and the accuracy curve as the threshold, as shown in Figure 8. The model performance is then given in Table 4, with a threshold equal to 0.0259, the intersection of the accuracy curve and TPR curve. Table 4 also shows how these accuracies compare to the LR model, which are similar.

As with the LR model, given the imbalanced data set, the area under the precision-recall curve was calculated. Table 4 indicates that the NN model has a higher area under the curve as 0.1604 which suggests it may be the better performing model.

While this analysis is useful in determining whether the NN model is accurate *enough* to be useful, the above results do not lend to any clarity in terms of how features contributed to crash fatality predictions. Thus, we explored various ways to determine the influence of the relative weighting of individual features, similar to that of LR. We selected different approaches for NN feature interpretation given their prevalence on various machine learning discussion boards (Computer Science 2013, Cross Validated 2017) which suggests these are commonly used methods in practice. They are explained in detail below.

[3] The neural nets in this paper were all developed using MATLAB R2018b and the Pattern Recognition toolbox.

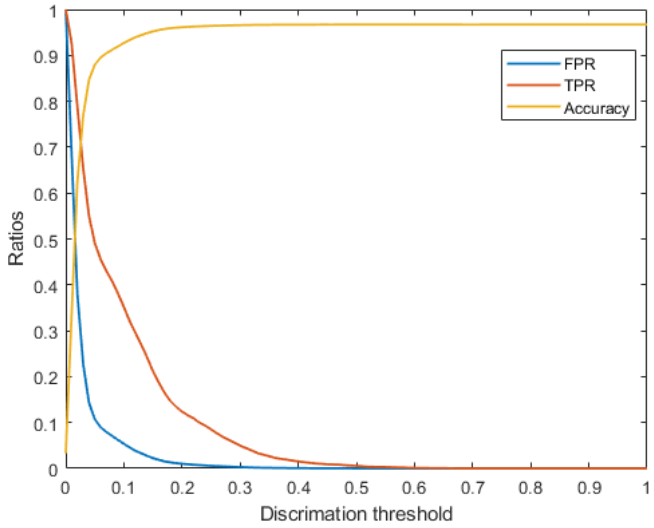


Figure 8. Different model performances with different thresholds for HSIS using NN

Table 4. Model performance for HSIS NN and LR

	NN	LR
Accuracy	71.84%	72.63%
Area under the curve (Precision-Recall)	0.1604	0.1290

Method 1: Feature analysis using "Leave one out"

One widely-used method in selecting important features in machine learning models is called "leave one out," e.g., (Xing, Jordan et al. 2001, Sung and Mukkamala 2003). In this form of a sensitivity analysis, a single feature was removed, and the NN model was re-developed with the remained features. In theory, if the removal of a feature leads to decreased model accuracy, this feature can be considered important, and vice versa. For model stability, ten NN models were separately developed with the original 16 features, and the average accuracy of the ten models became the threshold to evaluate the performance change of removing a feature.

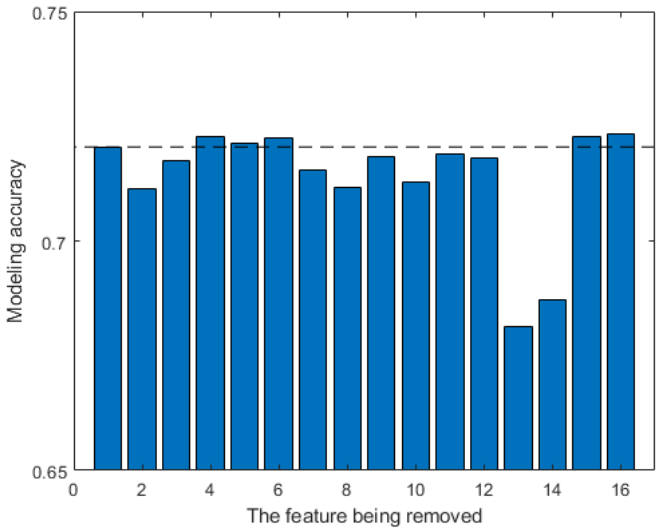


Figure 9. NN model accuracy after applying "Leave one out" calculations for the HSIS data. The dashed line represents average model accuracy.

Using this method on the model, Figure 9 demonstrates that features 13 (vehicle type) and 14 (sobriety) are the most important. An interesting observation is that after individually removing features 4 (left shoulder width), 5 (right shoulder width), 6 (number of lanes), 15 (urban/rural), or 16 (lane width) the performance of the new model increased. This increase suggests that under this model, those individual features may introduce noise in modeling the fatality relationship. As noted in the LR model, left shoulder width was a significant variable so there is a clear discrepancy between the two approaches.

Method 2: Feature analysis using first layer weights

Another approach in interpreting feature weights is looking at the weights of the first layer in a shallow neural network which is a NN of 1-2 layers (Intratora and Intratorb 2001, Guha, Stanton et al. 2005). We trained ten shallow NNs with no hidden layers, and Figure 10 shows the average weight and standard deviations of each feature across the 10 different models. Using the mean first layer feature weights in Figure 10, Table 5 summarizes those features with weight magnitudes higher than a varied weight threshold from 0 to 1.

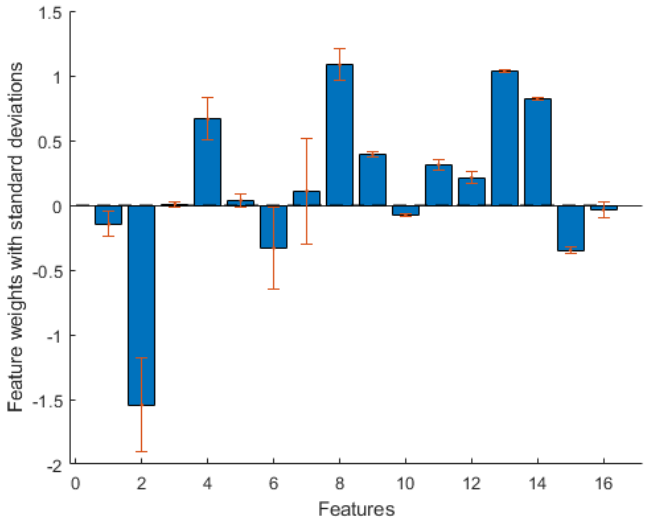


Figure 10. Feature weights generated from 10 shallow NNs for the HSIS data

Table 5 illustrates another issue with subjectivity which is where to draw the line of criticality of feature weights, understanding that the combination of features underpins a NN model. Weight thresholds are difficult to conceptualize for NNs, but given that this shallow NN is effectively the same as a LR model, it is possible to take an odds ratio approach just as for the LR model. If we subjectively decide that 2 is the correct odds ratio number ($e^{.7}$), then variables 2 (AADT), 8 (section length), 13 (vehicle type) and 14 (sobriety) are the most important for this interpretation of the NN. These results generally align with that of the LR model.

Table 5. Variable selection for weights using a shallow NN

Weight Threshold	Features that have mean higher weights than the weight threshold															
0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
0.1	1	2	4	6	7	8	9	11	12	13	14	15				
0.2	2	4	6	8	9	11	12	13	14	15						
0.3	2	4	6	8	9	11	13	14	15							
0.4	2	4	8	13	14											
0.5	2	4	8	13	14											
0.6	2	4	8	13	14											

0.7	2	8	13	14												
0.8	2	8	13	14												
0.9	2	8	13													
1.0	2	8	13													

Method 3: Feature analysis using weights & standard deviation

Using only weights to assess feature importance ignores how much the weights vary through multiple modeling iterations, as seen by the error bars in Figure 10. To address this problem with feature stability, we also investigated a weighted mean metric using the ten shallow NN models which used mean feature weight / standard deviation (SD) and is illustrated in Figure 11.

One limitation in this approach is the loss clear mapping of the feature weights, since odds ratios cannot be computed. Figure 11 illustrates that features 13 (vehicle type), 14 (sobriety), 9 (light), and possibly 15 (urban/rural) could be seen as the most important, but this is clearly a subjective judgment. Using this interpretation, no road design variable would be seen as important.

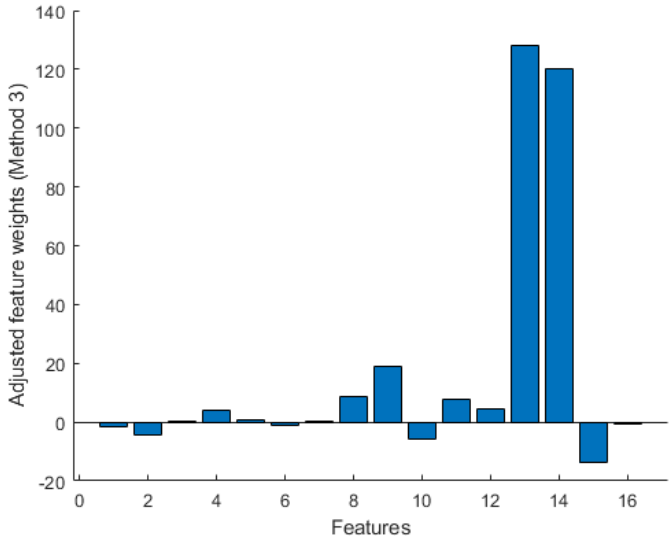


Figure 11. Feature weights/SD generated by 10 shallow NNs for the HSIS data

Method 4: Feature analysis using weights, standard deviation & the NN accuracy drop

While normalizing the feature weights by their variance helps to address the instability of some features, such a method does not examine how each feature ultimately influences overall model accuracy. To account for this, each feature weights/standard deviation was then multiplied by the overall model accuracy drop if that variable were removed (Figure 12). Thus, we combined methods 1 and 3 for a new evaluation method 4. This method, similar to other feature permutation approaches (Breiman 2001, Fisher, Rudin et al. 2018), suggests feature 13 (vehicle type) and 14 (sobriety) were dominant in model with negligible contribution by other features.

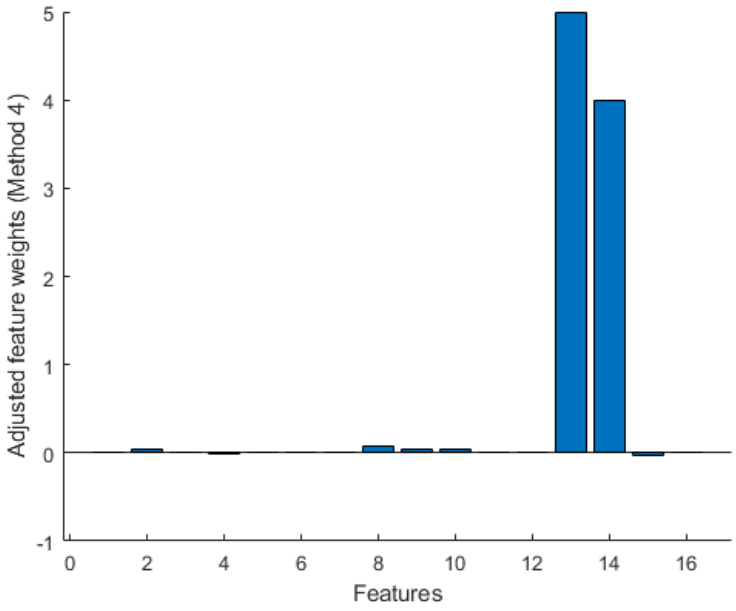


Figure 12. Feature weight/SD * accuracy drop for the HSIS data

Comparison of Results

Given the 4 different methods for NN feature interpretation, we wanted to compare the outputs with the LR's outcome. In order to make meaningful comparisons, we ranked the variables in order of importance for each method, understanding that the distance between these rankings is not directly comparable, particularly between LR and NN models. The results are listed in Table 7 with the top 5 features shaded in each set given that there were 5 models. The exception is the fourth NN interpretation method which only demonstrated two important features.

Table 7. HSIS rank orderings by different methods

Feature	LR	Method 1	Method 2	Method 3	Method 4
1 Speed limit	14	11	11	11	-
2 AADT	1	3	1	9	-
3 Access control	15	7	16	15	-
4 Left shoulder width	6	14	5	10	-
5 Right shoulder width	16	12	15	13	-
6 Number of lanes	11	13	8	12	-
7 Median width	5	6	12	16	-
8 Section length	3	4	2	5	-
9 Light	7	9	6	3	-
10 Weather	13	5	13	7	-
11 Maximum age	9	10	9	6	-
12 Minimum age	10	8	10	8	-
13 Vehicle type	2	1	3	1	1
14 Sobriety	4	2	4	2	2
15 Urban / Rural	8	15	7	4	-
16 Lane width	12	16	14	14	-

Not surprisingly, the LR model and Method 2 were in close but not exact alignment. Method 2 uses a shallow NN which is similar to LR. Method 4 for NN interpretation is a derivative of Methods 1 and 3, so this outcome

is also not surprising, save for the fact that 14 of the 16 feature contributions were negligible. In aggregate, all 5 models agreed that sobriety and vehicle type (where people driving motorcycles were at higher risk for fatalities) were relatively strong predictors of fatal or near-fatal accidents, but there was not strong consensus across the NN models about other features.

It is important to recognize that while the model interpretations are not radically different, they are different enough to cause issues if such results were used to justify policy decisions. It is unlikely that in any practical setting, a data scientist would take the time to evaluate all 5 models, and indeed, likely would just complete one if time and resource pressure existed. So, it is clear from Table 7 that the choice of machine learning model can affect results, as can the choice of feature weight interpretation if NNs are used.

Understanding that interpretation of these results is a highly subjective process, a master's student in Electrical and Computer Engineering, a junior machine learning engineer, and the PI examined these results and provided their interpretation, as follows:

Interpretation #1 (Student Engineer):
For the LR model, method 3 and method 4 are the most reasonable models, so features are important if at least two of three metrics rank in the top 5. Features 8 (section length), 13 (vehicle type), and 14 (sobriety) are the most important.

Interpretation #2 (Junior Engineer):
By counting how many times a feature ranked in the top 5, features were divided into three groups. Group 1 has features 8 (section length), 15 (vehicle type), and 16 (sobriety) both ranked in the top 5 above four times. Group 2 has feature 2 (AADT) which ranked three times in the top 5. The remaining features belong to Group 3.

Interpretation #3 (PI):
This analysis adds more evidence that sobriety and vehicle type are significant causes of fatalities for drivers, but because the model accuracy is too low, no other conclusions should be drawn from this data.

Conclusions from Transportation Case Study #1

A large transportation data set concerning driver fatalities and injuries was examined with two different machine learning models and 5 quantitative representations of the results. The two models and 5 representations were similar in some aspects but not in alignment, illustrating that the choice of model and representation strategy can alter the results. Such differences in model prediction performance has been noted in the use of other machine learning models applied to similar data sets (Iranitalab and Khattak 2017). Moreover, three different interpretations of the results by practitioners with different levels of experience demonstrate how much variability can be introduced when drawing conclusions from such studies.

If such an approach were to be used by transportation engineers to determine whether some aspect of road design is a good candidate for investment, the choice of the model could dramatically affect the outcome and ultimately dollars spent. If these engineers do not fully understand the ramifications of their choices and assumptions in the modelling process, it is not clear that the outcomes would be in the best interest of public safety.

In the next section, this exact same approach will be replicated with a pedestrian accident data set to determine how such approaches affect results from a smaller data set.

Transportation Data Set #2: A Not-So-Big Data Approach to Pedestrian Safety

While the first data set was very large with hundreds of thousands of accidents and 248 features, not all data sets are as populated, so we wanted to examine a traffic safety data set that was not as large. With the recent rise in pedestrian deaths (National Center for Statistics and Analysis 2018), it would be useful to do a similar machine learning analysis but such complete data sets about pedestrians are difficult to find. To this end, we elected to use the 1996 National Automotive Sampling System (NASS)[4], which attempts to establish the relationship between vehicle and pedestrian contact parameters along with injury type and severity, as well as impact speeds in Buffalo, Fort Lauderdale & Hollywood FL, Dallas, Chicago, Seattle, and San Antonio.[5]

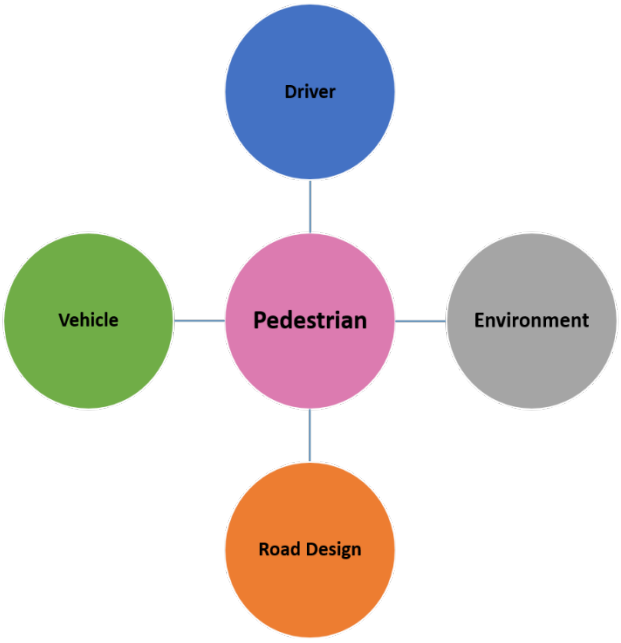


Figure 13: Model characteristics for the NASS

Methodology

The first step was to preprocess the dataset. Despite the large number of cities, there were only 549 observations of pedestrian fatalities in this dataset with 189 possible features. Such a large number of predictor variables would cause overfitting, so we needed to down select by at least an order of magnitude to preserve degrees of freedom. In addition, there were many invalid variable values. We ended with 310 observations of pedestrian fatalities with 16 features listed in Table 8, categorized by the same 4 features as in the first section with the additional category of pedestrian characteristics (Figure 13) which often account for si. Appendix D details these parameters. We elected to use 16 variables to show a comparison with the large HSIS model that also had 16 features. The target variable indicated the level of injury with 1 = fatal injury and 0 = non-fatal injury. There were no missing data in this set.

This dataset is substantially smaller than the first, which represents real world constraints but is also limiting in that machine learning algorithms perform best with much more data. In addition, this dataset is also imbalanced with 26 fatal observations and 284 non-fatal observations, so there are only 8.39% positive

[4] https://ftp.nhtsa.dot.gov/PED/96PedMan.pdf

[5] There have been recent efforts to standardize data elements between NHTSA's Fatality Analysis Reporting System (FARS) and the NASS General Estimates System (GES), more information can be found at https://www.nhtsa.gov/research-data/fatality-analysis-reporting-system-fars.

observations. While not as imbalanced as the HSIS data set, there are similar issues in determining the relevant thresholds. In this small dataset, there were no missing values.

As with the first dataset, Logistic Regression (LR) and Neural Network (NN) models were developed with similar feature weight investigations, detailed in the next sections.

Table 8. Variables selected from the NASS dataset

1	2	3	4	5	6	7	8
Month	Time	Pedestrian weight	Pedestrian age	Pedestrian gender	Pedestrian motion	Pedestrian action relative to vehicle	Pedestrian first avoidance action
9	10	11	12	13	14	15	16
Driver drinking	Speed limit	Vehicle curb weight	Driver attention to driving	Traffic way flow	Number of travel lanes	Roadway surface condition	Traffic control device functioning

Drivers Environment Vehicles Road Design Pedestrian

NASS Logistic Regression

Model Accuracy

As with the HSIS data, the NASS dataset is imbalanced, so the model accuracy must be considered alongside the definition of the threshold between 1 and 0. The LR model with all variables gives a predicted value for each observation, and the threshold between fatal and non-fatal values decides the overall accuracy and true positive rate. To examine the appropriateness of threshold values, we iterated the thresholds from 0 to 1 with step 0.01 to see how different thresholds affect the modeling performance, Figure 14. To this end, we set our threshold = 0.0988 at which the accuracy curve intersects with the TPR curve.

As with the previous case study, model performance is given in Table 9 for both model accuracy and area-under-the-precision-recall-curve approaches. The area under the precision-recall curve is 0.4254. This is much larger than the area of LR model in the HSIS data, which suggests LR functions better for this NASS data.

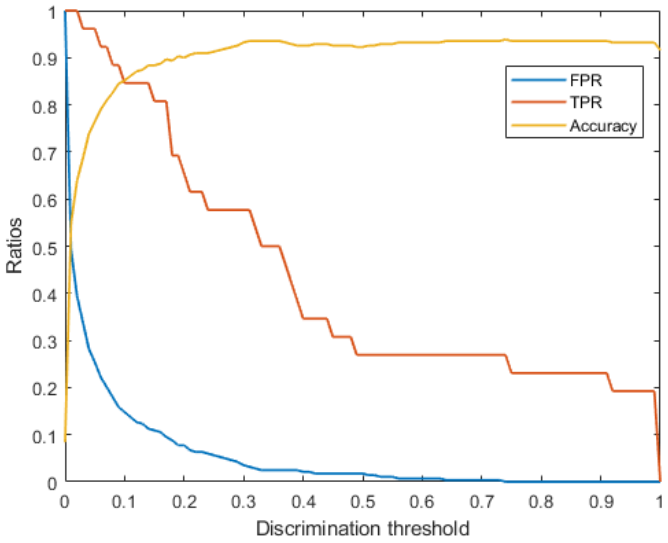


Figure 14. LR model performance with different thresholds for NASS

Table 9. LR model performance for NASS

Accuracy	84.84%
Area under the curve (Precision-Recall)	0.4254

Feature Weights

Table 10 shows weights, exp(weights) and p-values. According to this table, only feature 4 (pedestrian age) and feature 10 (speed limit) were statistically significant and the strongest predictors of fatalities for pedestrians, which has been noted in other studies (McLean, Anderson et al. 1994, Richards 2010). By these results, the odds ratio is more than 2,000:1 that a pedestrian will be killed if hit by a car at higher speed limits, the highest of which was 65 mph in the NASS dataset. This result is in line with previous research that has shown similar outcomes (Swanson, Yanagisawa et al. 2016).

Table 10. Feature weight results of NASS LR, * indicates significance at p < 0.0031, which uses a family-wise error correction rate of α/16, where α= 0.05

Feature	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
Weight	0.4455	1.9689	2.1563	4.1080	-1.3572	3.1362	0.8849	-17.4539
Exp(W)	1.5613	7.163	8.6392	60.8274	0.2574	23.0153	2.4227	0
p-value	0.6504	0.1188	0.2695	0.0024*	0.2851	0.1476	0.5990	0.1167
Sig. Odds Ratio				60.8274				
Feature	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**
Weight	24.7249	7.7414	-0.3979	-8.1344	-1.6829	-1.8427	0.7744	-1.0495
Exp(W)	5.47E+10	2301.702	0.6717	0.0003	0.1858	0.1584	2.1693	0.3501
p-Value	0.5262	0.0018*	0.8183	0.2611	0.2071	0.1618	0.4957	0.1227
Sig. Odds Ratio		2301.702						

Neural Network for NASS

To develop a NN for this dataset, the process was very similar to that described for the HSIS dataset:

- Input: The input layer included 16 variables which included categorical, ordinal, and continuous data (Appendix D).
- NN structure: For the hidden layer size (number of neurons in each layer), we selected 15.
- Output: The output layer consisted of a single node to predict 0 (non-fatal accident) or 1 (fatal accident).
- Ratios: The training, validation, testing ratios used were 0.68, 0.12, and 0.2, respectively.

Model Accuracy

Similar to the NN in the HSIS dataset, we took the intersection point of the TPR and accuracy curves as the threshold, and Figure 15 shows the TPR, FPR, and accuracy curves when iterating thresholds over 0 to 1. With a threshold of 0.1635, a single NN model achieved an accuracy of 80.65% and a true positive rate of 85.71%. However, because there is a small amount of data, the threshold could take on other values under different NN instances. When another nine NN models were separately trained using the NASS dataset, the threshold at the intersection of the TPR and accuracy curves then ranged from 0.0776-0.1924. The average model accuracy of all ten NN models was 79.84%, summarized in the Table 11.

The model accuracy varied because of the small size of the data set and represents model instability. As in the previous case study, the areas under the precision-recall curves were calculated and summarized in Table 11, which demonstrates the LR model performed better than the NN model in terms of higher accuracy and higher area under the precision-recall curve.

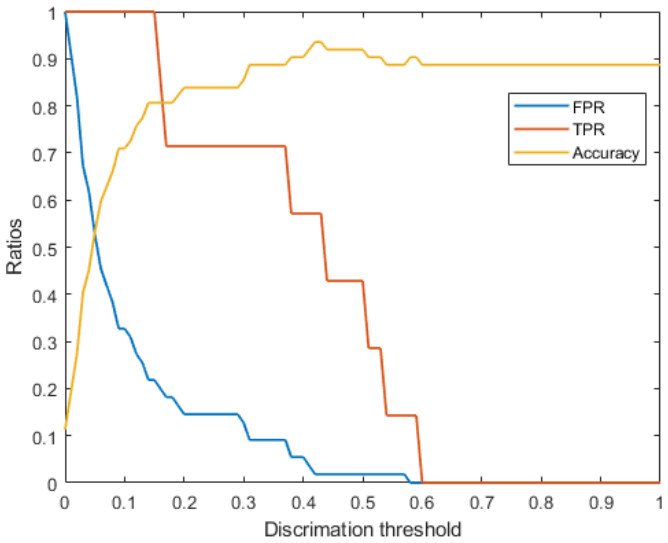


Figure 15. NN model performance with different thresholds for NASS

Table 11. Model accuracy for NASS (10 iterations for the NN)

	NN	LR
Accuracy	79.84%	84.84%
Area under the curve (Precision-Recall)	0.4231	0.4254

This model was then used to generate the feature weights in the same four ways described for the HSIS data set, detailed in the next section.

Method 1: Feature analysis using "Leave one out"

In this method, a single feature was removed with ten NN models, and the change in model accuracy was recorded, Figure 16. Using this method, Figure 16 demonstrates feature 1 (month), 4 (pedestrian age), and 6 (pedestrian motion) are the most important features as they resulted in the largest accuracy drops.

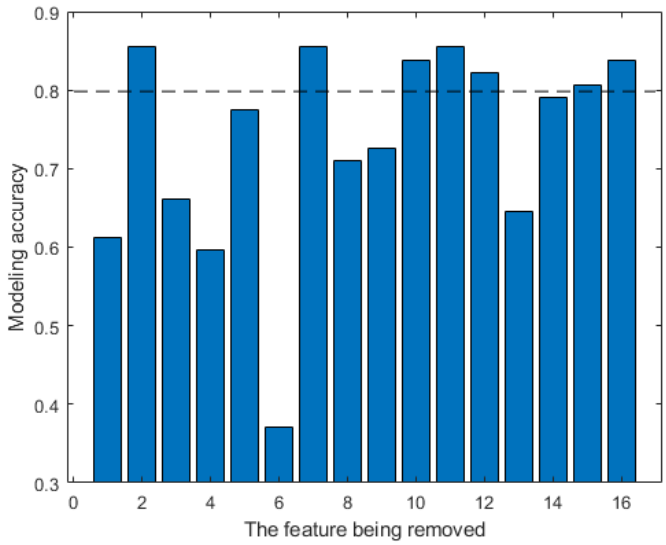


Figure 16. NN model accuracy after applying "Leave one out" accuracies for the NASS data. The dashed line represents average model accuracy.

Method 2: Feature analysis using first layer weights

Just as for the HSIS dataset, we trained ten shallow NNs for the NASS dataset, and Figure 17 shows this result. If we apply the same threshold criteria as for the HSIS data set of 0.7, features 3 (pedestrian weight), 4 (pedestrian age), 6 (pedestrian motion), 8 (first avoidance action), 9 (driver drinking), and 10 (speed limit) are the most important.

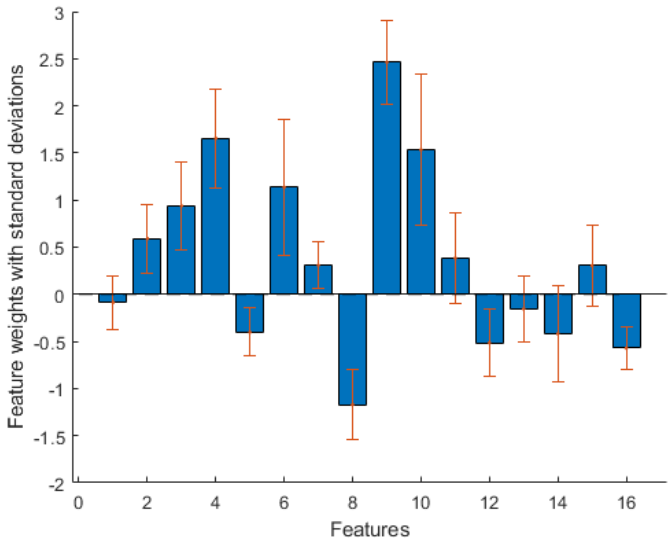


Figure 17. Feature weights generated by a shallow NN for pedestrian data

Method 3: Feature analysis using weights & standard deviation

As with the HSIS data, we also used mean weights / standard deviation (SD) from the shallow NNs to account for the stability of a particular feature. From Figure 18, feature 4 (pedestrian age), 8 (first avoidance action), 9 (driver drinking), and 16 (traffic light functioning) are the most important.

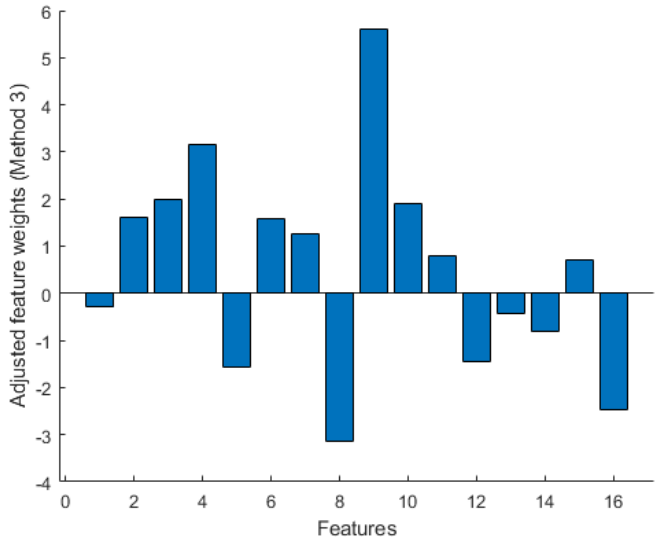


Figure 18. NN feature weights / SD for the NASS data

Method 4: Feature analysis using weights, standard deviation & the NN accuracy drop

Lastly, we examined weighting the feature weights by the NN model accuracy changes. From Figure 19, we can see that feature 3 (pedestrian weight), 4 (pedestrian age), 6 (pedestrian motion), 8 (first avoidance action), and 9 (driver drinking) are the most important in this combined method.

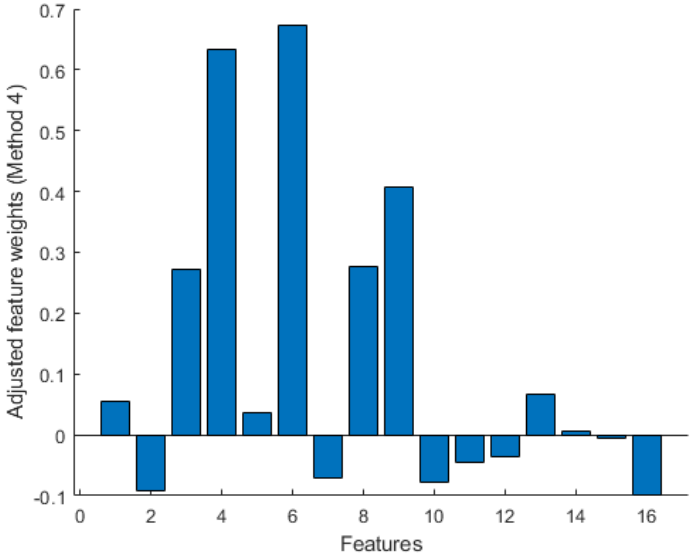


Figure 19. (Weights * accuracy drop / SD) generated by NN for NASS pedestrian data

Comparison of Results

As with the HSIS data, we ranked the features by order of importance for the different methods, with the top 5 highlighted. For the LR model, only variables that were statistically significant are listed. There is overall agreement across the methods that pedestrian age (feature 5) is a significant factor in whether a pedestrian will be killed if hit which has been seen in other research (Swanson, Yanagisawa et al. 2016). However, there was no consensus across the models for the leading factor. The speed limit and whether a driver had been drinking compete for the first and third spots, but none of the NNs agree with the LR model about speed limit.

Table 12. NASS rank orderings by the different methods

Feature	LR	Method 1	Method 2	Method 3	Method 4
1 Month	-	3	16	16	7
2 Time	-	14	7	7	-
3 Pedestrian weight	-	5	6	5	5
4 Pedestrian age	2	2	2	2	2
5 Pedestrian sex	-	8	11	9	8
6 Pedestrian motion	-	1	5	8	1
7 Action relative to vehicle	-	15	13	11	-
8 First avoidance action	-	6	4	3	4
9 Driver drinking	-	7	1	1	3
10 Speed limit	1	12	3	6	-
11 Vehicle curb weight	-	16	12	13	-
12 Driver attention	-	11	9	10	-
13 Traffic way flow	-	4	15	15	6
14 Number of travel lanes	-	9	10	12	9
15 Surface condition	-	10	14	14	-
16 Traffic light functioning	-	13	8	4	-

If the goal of a transportation engineer in analyzing this data was to determine roadway elements that could contribute to pedestrian deaths, the only variables that seemed potentially relevant were the traffic way flow, ranked 4th in method 1 and whether a traffic light was functioning, ranked 4 for method 3. However, because of the model instability due to the small sample size, the NN results should be viewed with caution.

Again, to demonstrate how much interpretation of these data can vary, three different perspectives are given regarding the outcomes.

Interpretation #1 (Student Engineer):
Features are important if at least three ranked in the top 5. Given this, feature 3 (pedestrian weight), 4 (pedestrian age), 6 (pedestrian motion), 8 (first avoidance action), and 9 (driver drinking) are the most important features.

Interpretation #2 (Junior Engineer):
By counting how many times a feature ranked in the top 5, features were divided into three groups. Group 1 has features 4 (pedestrian age) that ranked five times in the top 5. Then Group 2 has features 3 (pedestrian weight), 6 (pedestrian motion), 8 (first avoidance action), 9 (driver drinking), and 10 (speed limit) which ranked in the top 5 three times and twice. Other features belong to Group 3, the least important group.

Interpretation #3 (PI):
Given the low model accuracy of the NN model as well as its instability, I have more confidence in the LR model. Speed limit and pedestrian age appear to be strong predictors for fatalities, which is in agreement with previous studies, and so future work should target interventions that specifically address these two variables.

Conclusion

Transportation analysts are inundated with requests to apply popular machine learning modeling techniques to data sets to uncover never-before-seen relationships that could potentially revolutionize safety, congestion and mobility. To demonstrate some of the pitfalls in engaging in such analytics, which include subjectivity at several points in the modeling process, we developed Logistic Regression (LR) and Neural Network (NN) models for a driving injury/fatality and pedestrian fatality datasets. We then developed 5 different representations for each data set, one LR and one NN, with 4 different feature interpretations commonly used in the machine learning community.

This study showed that when attempting to determine if road design variables significantly influenced driver injuries and fatalities, the answer is unclear, with many possible interpretations of the results. In the pedestrian model, results indicated that speed and age mattered, but any conclusions could not be drawn about road design variables. However, interpretations could be very different depending on the model and parameters selected.

These modeling attempts highlighted several points of subjectivity:

- Picking the model to be used
- Picking which features should be included out of large data sets
- Determining whether to drop cases with missing data or to generate missing data estimates
- Picking a p value for LR significance
- Deciding numbers of neurons for hidden NN layers
- Picking the maximal training iteration for the NN training process
- Picking stopping rules for training performance factors (software dependent)
- Selecting data training and testing ratios
- Picking threshold values between fatal/non-fatal values
- Choosing thresholds between important/unimportant features
- Deciding across models with slightly different results, what the actual important features were

These models highlighted other core issues not often discussed in practical applications of these methods which include issues with imbalanced data which occurs when one class of data (non-fatalities in our data sets) significantly dominates over the other (fatalities). Unfortunately, such imbalance is a typical characteristic of transportation data sets, and if practitioners blindly apply statistical packages without understanding the underlying nature of the assumptions, then results could be negatively affected.

Another significant issue with the use of such powerful but potentially brittle analytical tools is a lack of checks and balances for result generation and interpretations. In this study, we generated 5 different interpretations for two different data sets, but it would be very unusual for practitioners to go to this level of analysis to determine how different models may differ and why. Experience (or lack thereof) ultimately guides the construction and interpretation of such models, and this reality represents a significant source of subjectivity. Thus, there is a very real possibility that decisions could be made from results generated by models that are not exactly wrong, but also are not exactly correct. Such inherent data analytic weaknesses need to be accounted for when policymakers make decision based on machine learning-generated results.

Acknowledgments

This research was funded by the US Department of Transportation's University Transportation Center grant through the University of North Carolina's Collaborative Sciences Center for Road Safety. Yaoyu Wang aided in collecting, cleaning, and processing the data.

References

Breiman, L. (2001). "Random Forests." Machine Learning Journal 45(1): 5-32.

Computer Science. (2013). "What can be learned from the weights in a neural network?" Retrieved 24 July, 2019, from https://cs.stackexchange.com/questions/10295/what-can-be-learned-from-the-weights-in-a-neural-network.

Cross Validated. (2017). "Deep learning: How do I know which variables are important?" Retrieved 24 July, 2019, from https://stats.stackexchange.com/questions/261008/deep-learning-how-do-i-know-which-variables-are-important.

Cummings, M. L. and A. Stimpson (2019). Identifying Critical Contextual Design Cues Through a Machine Learning Approach. AAAI AI Magazine Special Issue on Computational Context. W. Lawless and D. Sofge. Palo Alto, CA.

Fisher, A., C. Rudin and F. Dominici (2018). "All Models are Wrong but many are Useful: Variable Importance for Black-Box, Proprietary, or Misspecified Prediction Models, using Model Class Reliance." arXiv:1801.01489

Gianfrancesco, M. A., S. Tamang, J. Yazdany and G. Schmajuk (2018). "Potential Biases in Machine Learning Algorithms Using Electronic Health Record Data." JAMA internal medicine 178(11): 1544–1547.

Guha, R., D. T. Stanton and P. C. Jurs (2005). "Interpreting Computational Neural Network Quantitative Structure−Activity Relationship Models: A Detailed Interpretation of the Weights and Biases." Journal of Chemical Information and Modeling 45(4): 1109-1121.

Hadi, M. A., J. Aruldhas, L. Chow and J. A. Wattleworth (1995). " Estimating safety effects of cross-section design for various highway types using negative binomial regression." Transportation Research Record (1500): 169-177.

Hermans, E., T. Brijs, T. Stiers and C. Offermans (2006). The Impact of Weather Conditions on Road Safety Investigated on an Hourly Basis. 85th Transportation Research Board (TRB) Annual Meeting. Washington DC.

Ibrahim, O. M. (2013). "A comparison of methods for assessing the relative importance of input variables in artificial neural networks." Journal of Applied Sciences Research, 9(11): 5692-5700.

Intratora, O. and N. Intratorb (2001). "Interpreting neural-network results: a simulation study." Computational Statistics & Data Analysis 37(3): 373-393.

Iranitalab, A. and A. Khattak (2017). "Comparison of four statistical and machine learning methods for crash severity prediction." Accident Analysis & Prevention 108: 27-36.

McLean, A. J., R. W. G. Anderson, M. J. B. Farmer, B. H. Lee and C. G. Brooks (1994). Vehicle Speeds and the Incidence of Fatal Pedestrian Collisions. Federal Office of Road Safety. Australia, The University of Adelaide. I.

Menardi, G. and N. Torelli (2014). "Training and assessing classification rules with imbalanced data." Data Mining and Knowledge Discovery 28(1): 92–122.

National Center for Statistics and Analysis (2018). 2017 Fatal Motor Vehicle Crashes: Overview. NHTSA. Washington, DC, Department of Transportation,.

Neudorff, L., P. Jenior, R. Dowling and B. Nevers (2016). Use of Narrow Lanes and Narrow Shoulders on Freeways: A Primer on Experiences, Current Practice, and Implementation Considerations F. H. Administration. Washington DC, US Department of Transportation.

Peden, M., R. Scurfiled, D. Sleet, D. Mohan, A. A. Hyder, R. Scurfield, E. Jarawan and C. Mathers (2004). World report on road traffic injury prevention Geneva, Switzerland, World Health Organization.

Richards, D. C. (2010). Relationship between Speed and Risk of Fatal Injury: Pedestrians and Car Occupants. Transport Research Laboratory. London, Department of Transport,.

Saito, T. and M. Rehmsmeier (2015). "The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets." PLoS One 10(3): e0118432.

Samimi, A., A. Mohammadian and K. Kawamura (2010). An online freight shipment survey in US: Lessons learnt and a non-response bias analysis. 89th Annual Transportation Research Board Meeting, Washington DC, Transportation Research Board of the National Academies.

Stamatiadis, N. and J. Pigman (2009). Impact of Shoulder Width and Median Width on Safety. Washington DC, Transportation Research Board.

Sung, A. H. and S. Mukkamala (2003). Identifying important features for intrusion detection using support vector machines and neural networks IEEE Symposium on Applications and the Internet, Orlando, FL.

Swanson, E. D., M. Yanagisawa, W. Najm, F. Foderaro and P. Azeredo (2016). Crash Avoidance Needs and Countermeasure Profiles for Safety Applications Based on Light-Vehicle-to-Pedestrian Communications John A.Volpe National Transportation Systems Center. Washington DC, US Department of Transportation.

Xing, E. P., M. I. Jordan and R. M. Karp (2001). Feature Selection for High-Dimensional Genomic Microarray Data Eighteenth International Conference on Machine Learning, Williamstown, MA.

Appendix A: Variables in HSIS dataset

Table A-1: Variables, Descriptions and Categories of the HSIS Variables

	Variables	Description	Category
1	Speed limit	Miles per hour from 10 to 70	Road Design
2	AADT	Average annual daily traffic ranging from 0 to 775,446 cars	Environment
3	Access control	0: no access control; 1: expressway - partial access control; 2: freeway - full access control	Road Design
4	Left shoulder width	Left should width (in increasing direction of the roadway), feet, ranging from 0 to 80.	Road Design
5	Right shoulder width	Right shoulder width (in increasing direction of roadway), feet, ranging from 0 to 40	Road Design
6	Number of lanes	From 1 to 25	Road Design
7	Median width	Feet ranging from 0 to 800	Road Design
8	Section length	Stretch of road that is consistent in terms of certain road characteristics (e.g. shoulder widths, lane number, lane width …). Miles, ranging from .01 to 13.543.	Road Design
9	Light	0: daylight, 1: dusk/dawn, 2: dark but with street light, 3: dark without street light, 4: dark with street light not functioning	Environment
10	Weather	0: clear or cloudy, 1: weather may influence driving, including raining, snowing, wind and fog	Environment
11	Maximum age	Maximum age of driver involved in the accident, ranging from 0 to 105	Drivers
12	Minimum age	Minimum age of driver involved in the accident ranging from 0 to 104	Drivers
13	Vehicle type	0: normal, 1: heavy, 2: motorcycle	Vehicles
14	Sobriety	0: not impaired, 1: impaired	Drivers
15	Urban / Rural	0: rural, 1: urban	Environment
16	Lane width	Feet ranging from 0 to 150	Road Design

Appendix B: Comparison of HSIS Original and Cleaned Data

Two-sample Kolmogorov-Smirnov tests were performed on the original and processed data. The mean and median values are listed below. A p-value less than 0.0031 indicates the original data and processed data have different distributions, highlighted in grey.

Table B-1: HSIS Original vs. Cleaned Data

		Original		Processed		p-value from KS test	
		Fatal	Nonfatal	Fatal	Nonfatal	Fatal	Nonfatal
1 Speed limit (miles/hour)	Mean	61.8528	64.104	61.7897	63.5592	0.0003	0
	Median	70	70	65	70		
2 AADT	Mean	81,381.17	127,366.6	82,410.72	127,924	1	0.0066
	Median	37,500	123,793	39,500	125,000		
3 Access control	Mean	1.7709	1.7655	1.7713	1.7656	1	1
	Median	2	2	2	2		
4 Left shoulder width (feet)	Mean	5.3011	5.2813	5.3114	5.2831	0.9988	0.0015
	Median	5	5	5	5		
5 Right shoulder width (feet)	Mean	7.4593	7.9696	7.5126	7.9963	0.9958	0
	Median	8	10	8	10		
6 Number of lanes	Mean	5.1141	6.4949	5.1551	6.5247	1	1
	Median	4	6	4	6		
7 Median width (feet)	Mean	25.1334	26.8089	25.4172	26.9297	1	1
	Median	15	20	16	20		
8 Section length (miles)	Mean	0.7619	0.4156	0.7574	0.4119	0.9774	0
	Median	0.344	0.21	0.34	0.21		
9 Light	Mean	1.1167	0.6965	1.1208	0.6949	1	0.5021
	Median	0	0	0	0		
10 Weather	Mean	0.0649	0.0906	0.0640	0.0841	1	1
	Median	0	0	0	0		
11 Maximum age	Mean	46.2080	46.2169	46.011	46.2775	1	0.0050
	Median	47	47	46	46		
12 Minimum age	Mean	35.1059	32.1884	35.0022	32.1289	1	0
	Median	31	28	30	28		
13 Vehicle type	Mean	0.5117	0.1489	0.5067	0.1506	0.1511	1
	Median	0	0	0	0		
14 Sobriety	Mean	0.2593	0.0621	0.2527	0.0604	1	0.2292
	Median	0	0	0	0		
15 Urban/Rural	Mean	0.1305	0.2312	0.0860	0.1649	0	0
	Median	0	0	0	0		
16 Lane width (feet)	Mean	27.6578	34.2761	27.9214	34.4224	1	0.9540
	Median	14	30	15	30		

Appendix C: Histograms of HSIS variables before and after data imputation.

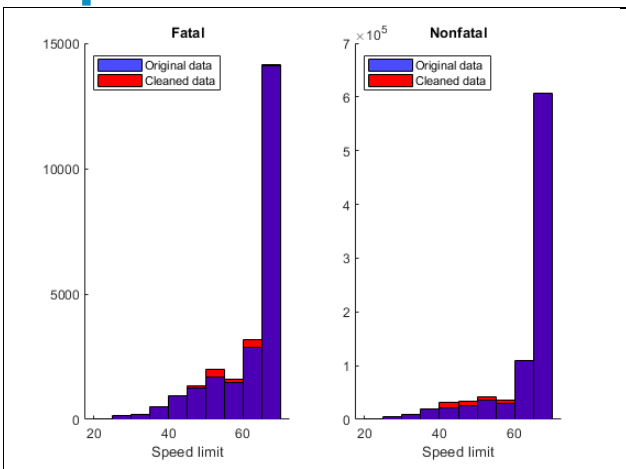


Fig. C-1 Distributions of the speed limit

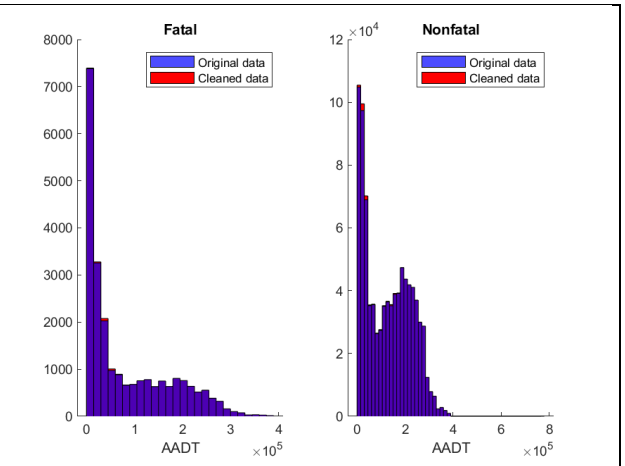


Fig. C-2 Distributions of the AADT

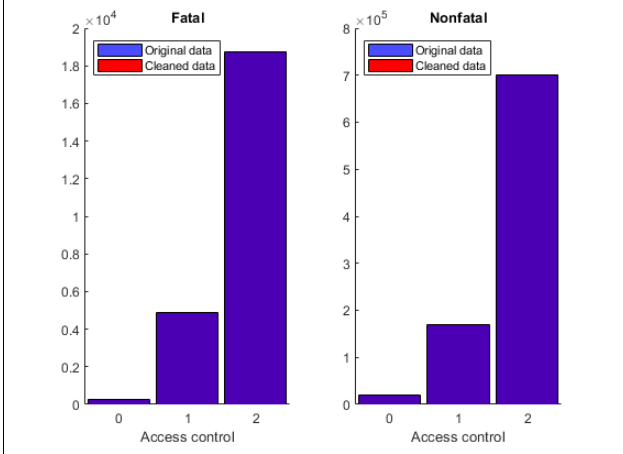


Fig. C-3 Distributions of the access control

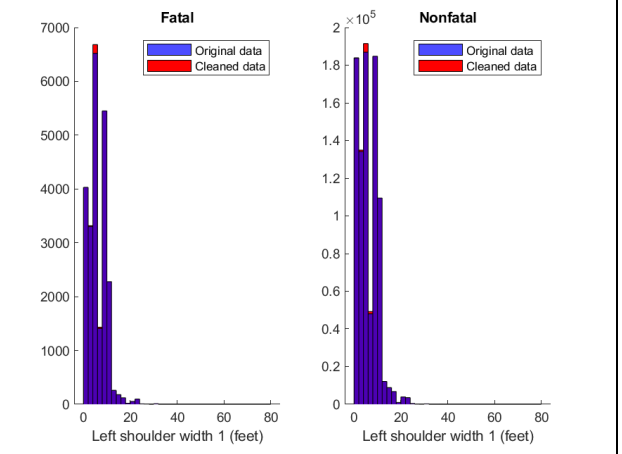


Fig. C-4 Distributions of the left shoulder width 1

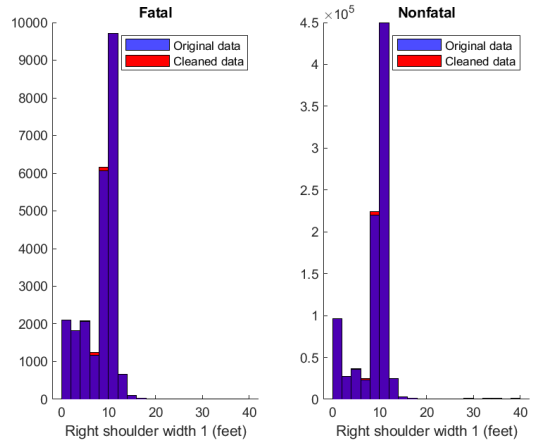


Fig. C-5 Distributions of the right shoulder width 1

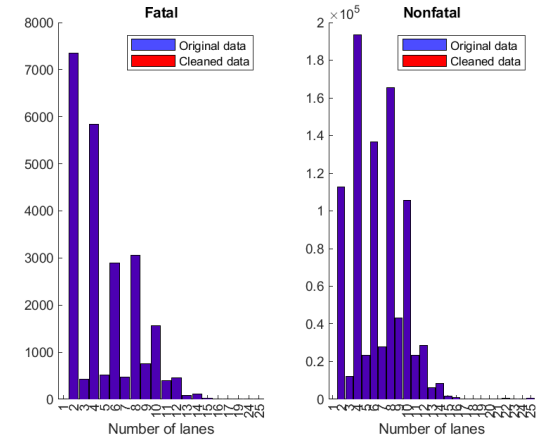


Fig. C-6 Distributions of the number of lanes

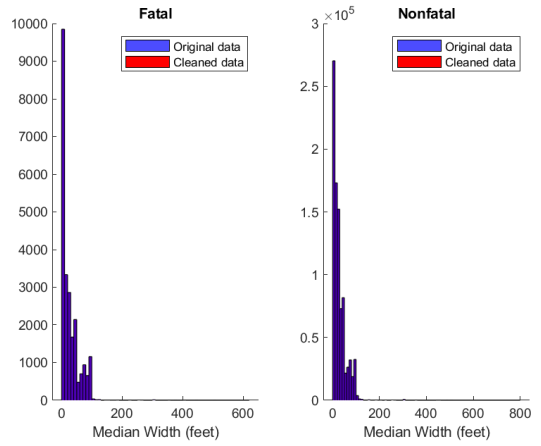


Fig. C-7 Distributions of the median width

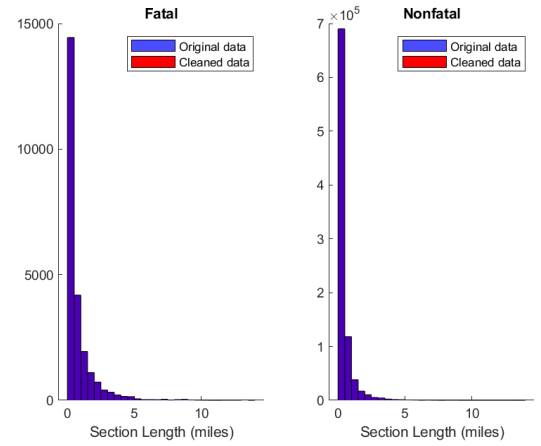


Fig. C-8 Distributions of the section length

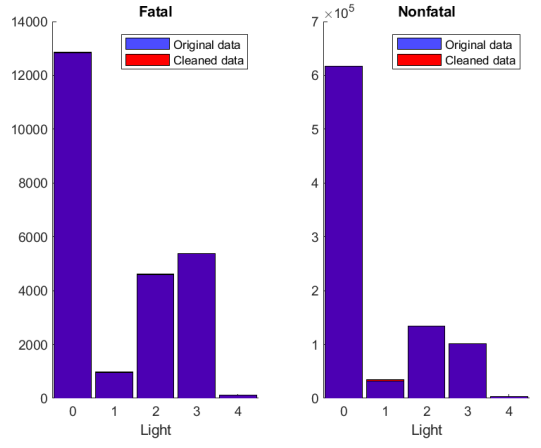


Fig. C-9 Distributions of the light

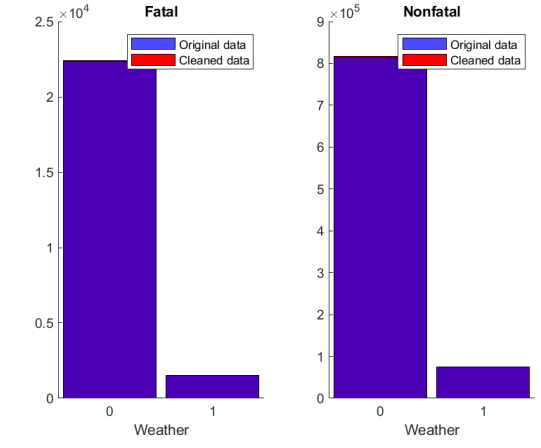


Fig. C-10 Distributions of the weather

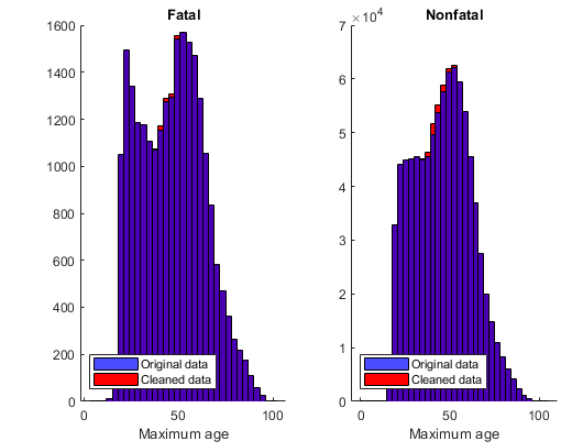


Fig. C-11 Distributions of the maximum age

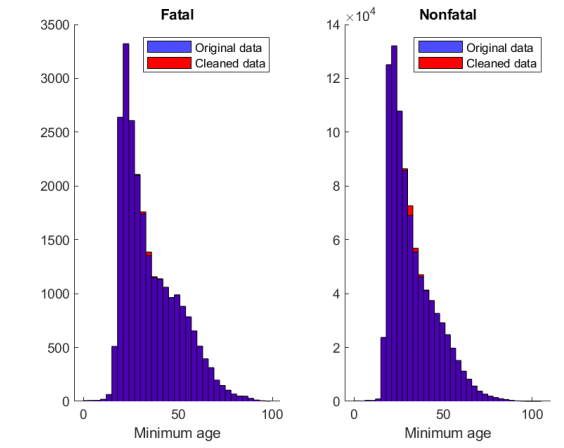


Fig. C-12 Distributions of the minimum age

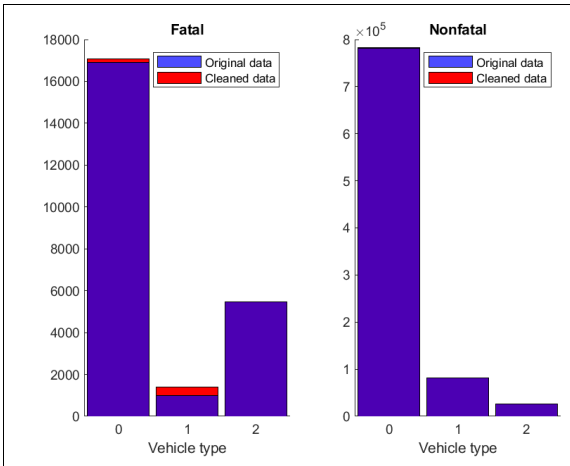


Fig. C-13 Distributions of the vehicle type

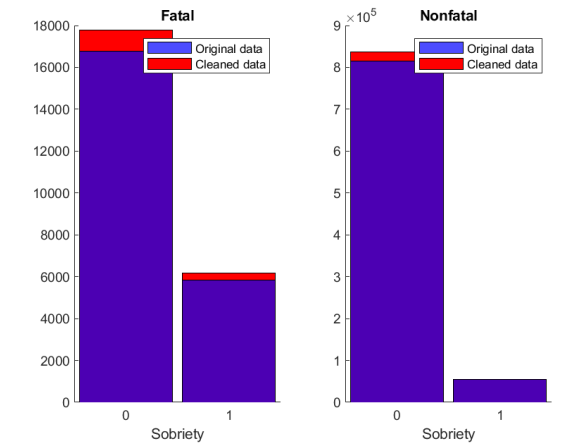


Fig. C-14 Distributions of the sobriety

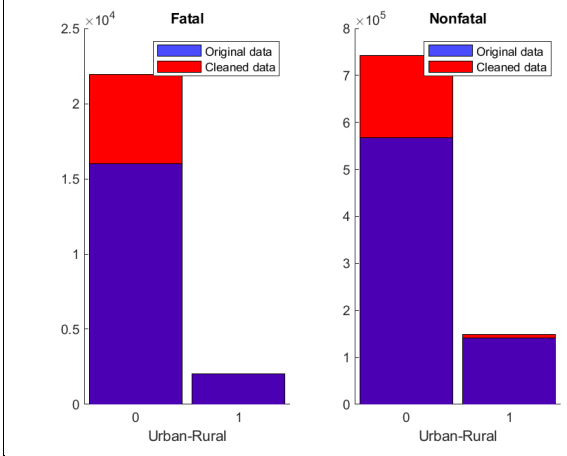


Fig. C-15 Distributions of the urban/rural

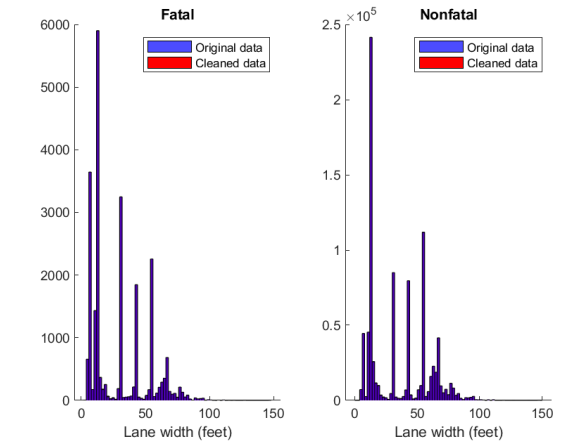


Fig. C-16 Distributions of the lane width

Appendix D: Variables in the NASS Pedestrian Dataset

Table D-1: Variables, Descriptions and Categories of the NASS Variables

	Variables	Description	Category
1	Month	1-12: the month	Environment
2	Time	3 or 4 digits representing time, e.g. 1523 means 15:23, 830 means 8:30	Environment
3	Pedestrian weight	Weight in kg ranging from 9 to 150	Pedestrian
4	Pedestrian age	Actual age ranging from 2 to 93	Pedestrian
5	Pedestrian sex	1: male, 2: female	Pedestrian
6	Pedestrian motion	0: not moving, 1: walking slowly, 2: walking rapidly, 3: running or jogging, 4: hopping, 5: skipping, 6: jumping, 7: falling/ stumbling or rising	Pedestrian
7	Action relative to vehicle	00: stopped, 01: crossing road straight, 02: crossing road diagonally, 03: moving in road with traffic, 04: moving in road against traffic, 05: off road approaching road, 06: off road going away from road, 07: off road moving parallel, 08: off road crossing driveway, 09: off road moving along driveway	Pedestrian
8	First avoidance action	00: no avoidance actions, 01: stopped, 02: accelerated pace, 03: ran away (along vehicle path), 04: jumped, 05: turned toward vehicle, 06: turned away from vehicle, 07: dove or fell away, 11: vault corner of vehicle, 12: vault onto vehicle, 13: brace against vehicle, 14: crouched and braced hands against vehicle	Pedestrian
9	Driver drinking	0: not drinking, 1: drinking	Driver
10	Speed limit	Speed limit in km/h ranging from 16 to 105	Vehicle
11	Vehicle curb weight	Actual value / 10 in kilogram ranging from 73 to 293.	Vehicle
12	Driver attention (prior to recognition of critical event)	1: full attention to driving, 2: distracted by other occupant, 3: distracted by moving object in vehicle, 4: distracted by outside person/object/event, 5: talking on cellular phone or CB radio, 6 sleeping or dozing while driving	Driver
13	Traffic way flow	1: not physically divided (two-way traffic), 2: divided trafficway - median strip without positive barrier, 3: divided trafficway - median strip with positive barrier, 4 one-way trafficway	Road Design
14	Number of travel lanes	Ranging from 1 to 7	Road Design
15	Surface condition	1: dry, 2: wet, 3: snow and slush, 4: ice, 5: sand/dirt/oil	Environment
16	Traffic light functioning	0: no traffic control, 1: not functioning, 2: functioning	Environment

730 Martin Luther King Jr. Blvd.
Suite 300
Chapel Hill, NC 27599-3430
info@roadsafety.unc.edu

www.roadsafety.unc.edu