Subjectivity in the Creation of Machine Learning Models

MARY L. CUMMINGS*

Electrical and Computer Engineering, Duke University

SONGPO LI

Electrical and Computer Engineering, Duke University,

Transportation analysts are inundated with requests to apply popular machine learning modelling techniques to data sets to uncover never-before-seen relationships that could potentially revolutionize safety, congestion and mobility. However, the results from such models can be influenced not just by biases in underlying data, but also through practitioner-induced biases. To demonstrate the significant number of subjective judgments made in the development and interpretation of machine learning models, we developed Logistic Regression and Neural Network models for transportation-focused data sets including those looking at driving injury/fatalities and pedestrian fatalities. We then developed five different representations of feature importance for each data set, including different feature interpretations commonly used in the machine learning community. Twelve distinct judgments were highlighted in the development and interpretation of these models, which produced inconsistent results. Such inconsistencies can lead to very different interpretations of the results, which can lead to errors of commission and omission, with significant cost and safety implications if policies are erroneously adapted from such outcomes.

CCS CONCEPTS • Supervised learning by classification • Neural networks • Modeling methodologies

Additional Keywords and Phrases: Bias, interpretable machine learning, transportation, logistic regression, subjectivity

ACM Reference Format:

Mary L. Cummings and Songpo Li. 2020. Subjectivity in the Creation of Machine Learning Models. J. Data and Information Quality XX, X, Article X (Month 2020), XX pages.

1 INTRODUCTION

With rapid advances in data analytic tools, machine learning techniques are increasingly being used to study large data sets in order to reveal underlying and potentially unknown patterns that would not likely otherwise be discovered. However, such techniques are coming under increasing scrutiny with widely publicized flaws in machine learning (ML) applications like racial bias in criminal recidivism predictions [1] and IBM's failure to effectively use Watson in healthcare diagnostics [2]. These and other similar events have led to increased calls

* Authors' addresses: M. Cummings and S Li, Duke University, Durham, NC, 27708, USA; emails: {m.cummings, songpo.li}@duke.edu

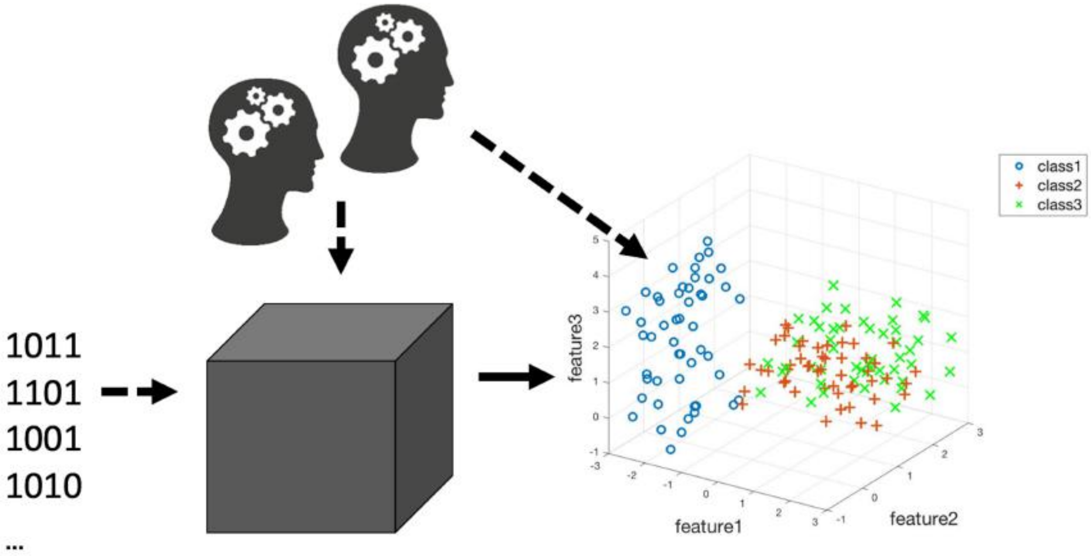


Figure 1: Depicted by a dashed line, sources of human bias in any machine learning modeling process including human subjectivity in model and parameter selection as well as sample selection bias coming from the data.

for artificial intelligence regulation [3] and warnings from major companies that AI presents significant risks to company products and reputations [4].

Research efforts are increasing to identify and correct bias in ML datasets [5], as well as making outputs more explainable [6] or interpretable through sensitivity analyses likes counterfactual explanations [7]. However, while such efforts are important, there has been substantially less focus on examining those subjectivities and biases that can be introduced into ML models through the many seemingly small choices that modelers make when developing predictive models. Machine learning results can be greatly affected by the subjectivity of the ML practitioner, where the practitioner subjectively selects the ML algorithm and is associated parameters for a specific data set. 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., [8, 9]), these practitioner-induced subjective biases introduced into a machine learning modeling process can make resulting conclusions vary widely.

Little work has studied this practitioner-induced subjectivity problem in order to understand its causes, influences, and methods for avoidance. This paper contributes to the growing body of literature on ML bias and subjectivity through a case study where two transportation data sets are analyzed with two different machine learning techniques of low and high complexity (logistic regression and neural networks). While there are many other types of machine learning models that could be used (e.g., see Figure 2 and [10]) these models were selected since they represent increasing model complexity and are both commonly used.

Using both a large and small transportation data set that include 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. The results demonstrate that both the type of model and feature interpretation method can 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 exemplify how experience can modify one's interpretation of the results. These outcomes, which highlight the more than ten opportunities for subjective decisions in model construction, suggest that more work is needed in looking at how

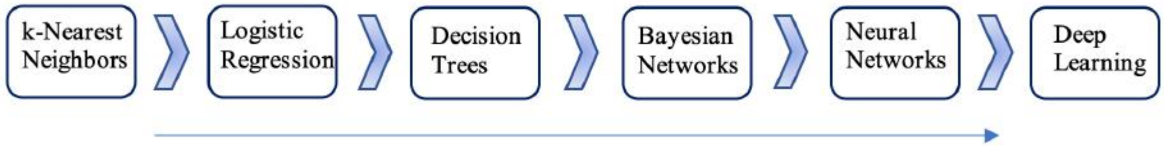


Figure 2: Complexity map of common machine learning models.

such subjective modeling and interpretation choices affect the suitability of machine learning models in support of decision making.

This paper is organized as follows: Section 2 illustrates sources of subjectivity by applying a logistic regression and then a neural network analysis to a large transportation data set, including analyses of model performance and different possible interpretations of feature weights. Section 3 repeats the logistic regression and neural network analyses on a small data set, include feature weight interpretations based on different levels of experience. Section 4 concludes with a comparison of the results between the two dataset analyses and a discussion of the number and nature of subjective choices made in the development and interpretation of models, which can have very real negative implications in practice.

2 A LARGE TRANSPORTATION DATA SET ANALYSIS

The first data set analyzed used the Highway Safety Information System[1] (HSIS), a roadway-based data repository that provides crash and fatality data that include a large number of roadway, infrastructure and traffic variables. It was developed by the Federal Highway Administration (FHWA) to help highway engineers to make better decisions on roadway and infrastructure design. For the purposes of this effort, the classification target variable was whether a driver was seriously injured or killed (1) vs. no serious injury (0).

The feature selection process for the HSIS data set represents a significant subjective decision point. There were 248 features that could be selected in HSIS but using them all would likely result in data overfitting and a loss of generalizability. We first choose a set of variables that typical transportation engineers would want to answer with a large data set, which is "What road and infrastructure design elements most influence serious road accidents?" We also based our feature selection on previous research that has shown additional elements like driver age, poor weather, and the vehicle type are also significant predictors of fatalities [11, 12]. This resulted in 16 HSIS variables, as depicted in Table 1. Appendix A1 provides the exact categories and units for each of these variables.

The data was collected 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 more than three predictor variables missing as well as a missing target variable. 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.

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

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. Moreover, the resulting data set was imbalanced, in that there were 23,949 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 a later section.

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.

Table 1: Variables selected for the HSIS data set, including the associated number. Gray shading indicates a potential road design variable.

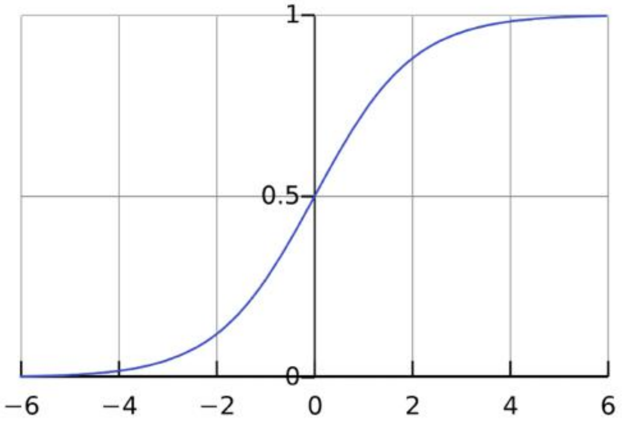
Speed Limit (1)	Avg. Annual Daily Traffic (2)	Access Control Type (3)	Left Shoulder Width (4)	Right Shoulder Width (5)	Number of Lanes (6)	Median Width (7)	Section Length (8)
Light (9)	Weather (10)	Driver Max Age (11)	Driver Min Age (12)	Vehicle Type (13)	Sobriety (14)	Urban vs. Rural (15)	Lane Width (16)

2.1 Logistic Regression Model

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 3. The features vector is notated as x, and $\boldsymbol{\beta}$ is a vector of regression coefficients corresponding to each feature. Scaler β_0 is a bias term that shifts the sigmoid function left or right. An accurate LR 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 determined where the true positive rate (TPR) and overall accuracy curves cross, beyond which the TPR falls exponentially. This led to a threshold of 0.0234, resulting in a model accuracy of 72.63%, with a TPR of 71.14%. However, for imbalanced data sets with binary outcomes like this data set, model accuracy is often not a good indicator of model performance [13]. Another potential method that can be used to assess model performance, especially for imbalanced data, is examining the area under the Precision-Recall curve [14].

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. This area under the curve equalled 0.1290, which suggests that the model is correctly classifying samples it predicts as fatalities, but may miss many classifications (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 3: Logistic Regression Function

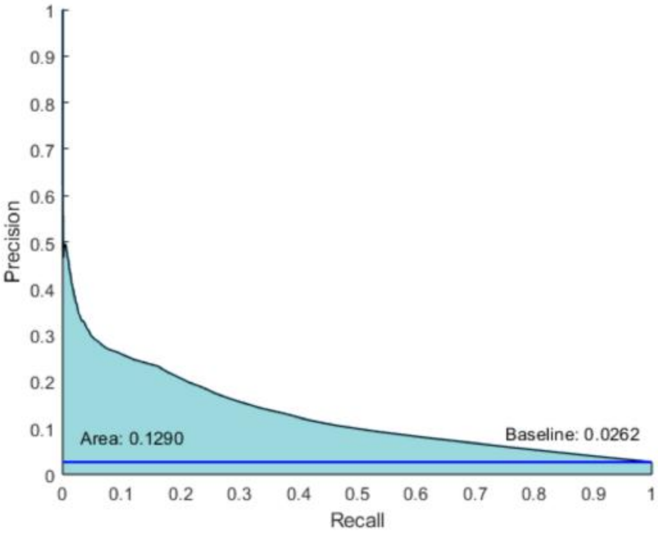


Figure 4: Precision-recall curve of the LR model. The baseline is the percentage of fatalities for the global dataset.

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

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	

Given that the LR model is acceptable, although not particularly strong, we then needed to understand how the different 16 features in Table 1 contributed to the overall model. Table 2 lists the weights, exp(weights) and p-values. Those variables with a p value less than 0.0031 were considered significant. Because LR models produce regression coefficients for each feature that are log odds, 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.

Determining which features are the most important is a subjective decision, with different decision criteria producing different results. One common interpretation is that all those features that are statistically significant should be in the mode, and in this case, 11 different variables mattered the most (Table 2). 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 (more is better [15], feature 7), and section length (feature 8, roads with longer consistent sections have more accidents), which has been seen in previous research [16]. While previous research has shown that increasing the shoulder width generally leads to less fatalities [17], this data set was slightly biased in the opposite direction.

So, if a transportation engineer wanted to know what road and infrastructure 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 that 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.

2.2 Neural Network Model

Neural networks (NNs) are often used to classify outcomes, similar to LR, but instead of a direct input and output computation, NNs contain layers where an input layer feeds hidden layers to develop a system of weighted connections that produce a classification at the output layer (as in Figure 5). Each unit in the hidden layer and output layer contains an activation function $f(\sum w \cdot i + b)$, where i is a vector of input variables, w is a corresponding weight vector, and b is the bias. The activation function could be a sigmoid, hyperbolic tangent, or rectified linear unit function. Assembly of the activation functions allow for representation of high-order nonlinear relations. Such models must first be trained on a subset of the overall data set and in the training process, all parameters in the network must first be initialized to make the first forward propagation. Then, the cost (deviation between output and a true result) is calculated. After that, parameters are adjusted to minimize the deviation between model predictions and the desired outputs. This process is repeated until overall model accuracy cannot be improved.

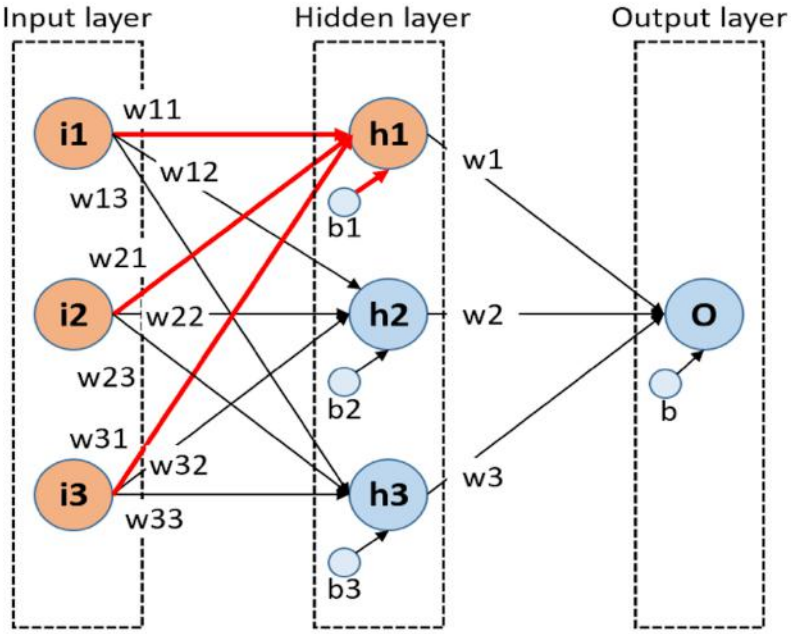


Figure 5: Illustration of a Neural Network

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 [18]. 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 a later section.

In order to develop a NN[2] for the same data set described earlier, we first initialized parameters of the neural network, including:

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

- Input: The input layer included 16 variables which included categorical, ordinal, and continuous data (Appendix A.1).
- NN structure: For the hidden layer size (number of neurons in each layer), we selected 10 but examined up to 100, which did not 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).
- Tests 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 was calculated, and the model generated a gradient towards the direction the parameters were adjusted. Similar to LR, the outputs from the NN model needed a threshold to determine the positive and negative predictions which was 0.0259, resulting in a model accuracy of 71.84%. As with the LR model, the area under the precision-recall curve was calculated and equaled 0.1604 as shown in Figure 6. So, while the overall model accuracy was slightly worse than the LR model, the area under the precision-recall curve was slightly better.

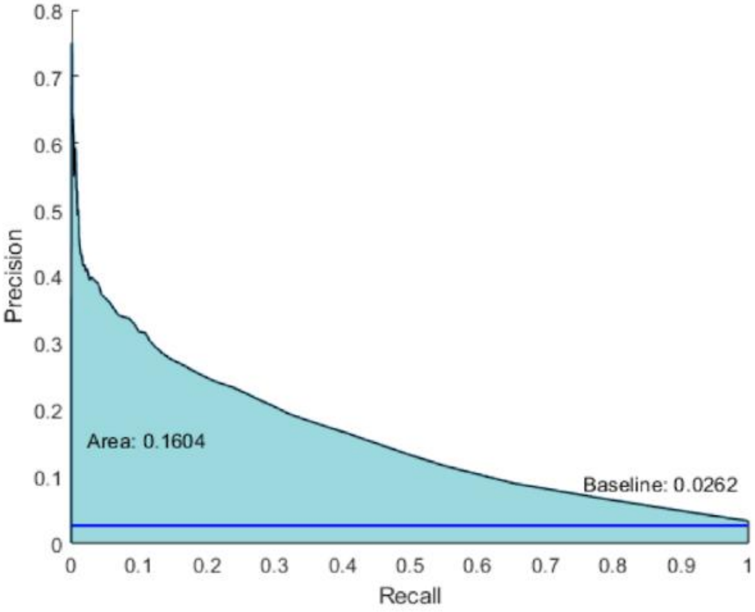


Figure 6: Precision-recall curve of the NN model. The baseline is the percentage of fatalities for the global dataset.

2.2.1 *Interpreting feature weights*

If an analyst decided that such a model was acceptable, the next step would be interpreting the feature weights. NN results do not provide any clarity in terms of how individual features contribute to predictions, so 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 [19, 20], which suggests these are commonly used methods in practice.

The *first method* we explored is a widely-used method in selecting important features in machine learning models is called “leave one out,” e.g., [21, 22]. In this form of a sensitivity analysis, a single feature is removed, and the NN model is re-developed with the remaining features. If the removal of a feature leads to decreased model accuracy, this feature can be considered important, and vice versa. For this application, 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.

Using this method on the model, Figure 7a 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 the model.

The *second method* used in interpreting feature weights looks at the weights of the first layer in a shallow neural network, which is a NN of 1-2 layers [23, 24]. We trained ten shallow NNs with no hidden layers just for the purpose of feature selection, and Figure 7b shows the average weight and standard deviations of each feature across the 10 different models. Figure 7b also illustrates another issue with subjectivity, which is where to draw the line of criticality for feature weights, which is up to the individual practitioner.

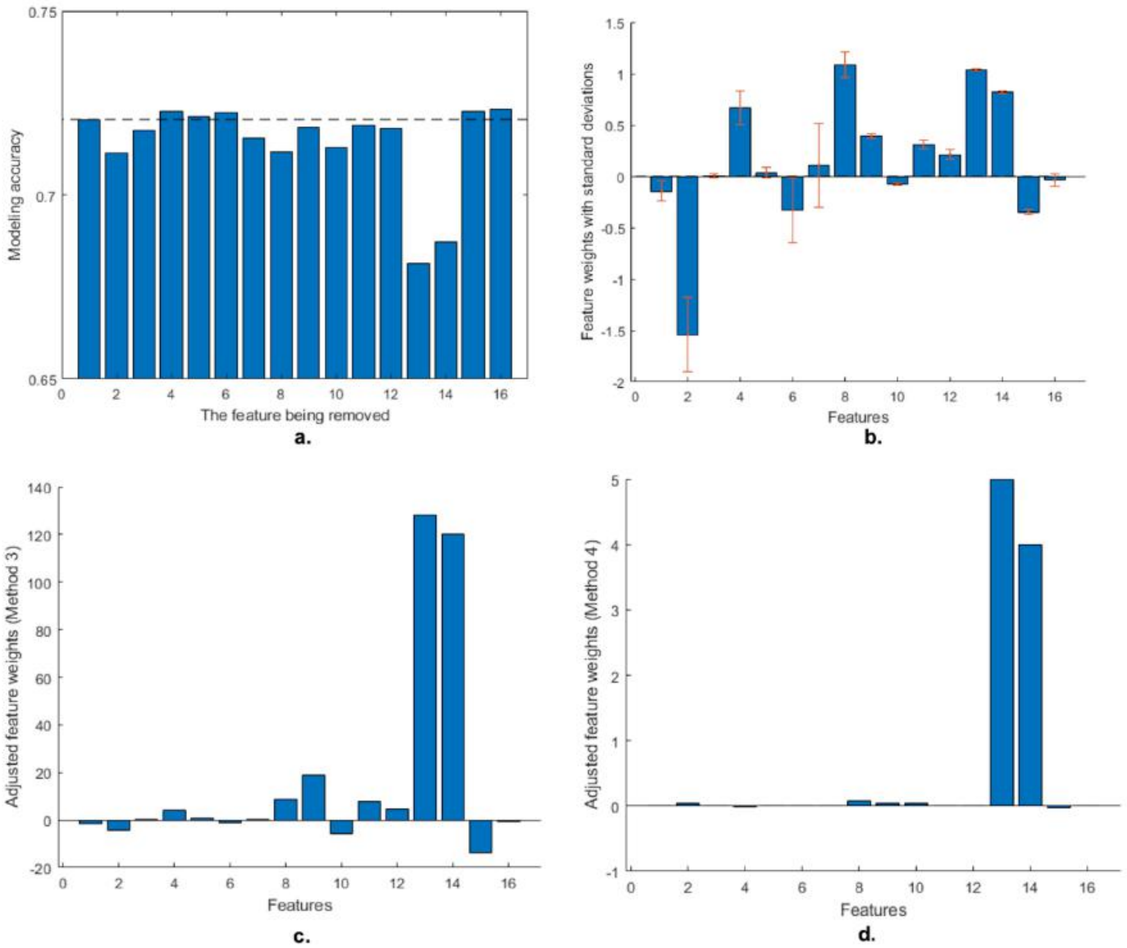


Figure 7: Four different methods of feature interpretation, (a) Leave one out, (b) Shallow neural networks (c) Feature weighted by standard deviation (d) Feature weighted by standard deviation and drop in model accuracy.

Because the use of weights only to assess feature importance ignores how much the weights vary through multiple model iterations, as seen by the error bars in Figure 7b, the *third method* we applied used a weighted mean metric of mean feature weight/standard deviation (SD), as seen in Figure 7c. While 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 or infrastructure design variable would be seen as important.

Lastly, 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, the *fourth method* of feature interpretation weighted each feature weights/standard deviation by the overall model accuracy drop if that variable was removed (Figure 7d). Thus, we combined methods 1 and 3 for a new evaluation method 4. This method, similar to other feature permutation approaches [25, 26], suggests feature 13 (vehicle type) and 14 (sobriety) were dominant in the model, with negligible contribution by other features.

2.3 Model Comparison

Given the four different methods for NN feature interpretation, we wanted to compare these outputs with the LR model 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 3 with the top five features shaded in each set, given that there were five models. The exception is the fourth NN interpretation method which only demonstrated two important features.

Table 3: HSIS rank orderings by different feature interpretation 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 clear from Table 3 that the choice of machine learning model can affect results, as can the choice of feature weight interpretation if NNs are used. Moreover, understanding that interpretation of these results is a highly subjective process, and experience can significantly affect such outcomes. For example, when three different engineers of various levels of experience were asked to interpret these results, the following was reported:

- **Interpretation #1 (Student Engineer):**
  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. So, 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), all ranked in the top 5 at least 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 (Senior Engineer):**
  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.

If such an approach was used by transportation engineers to determine whether some aspect of road or infrastructure 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 summary, for this first data set of factors influencing driver fatalities and injuries, two different machine learning models and five quantitative representations of the results were generated. The two models and five 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 [27]. 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.

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

3 A SMALL TRANSPORTATION DATA SET ANALYSIS.

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 [28], we elected to use the 1996 National Automotive Sampling System (NASS)[3], 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 and Hollywood FL, Dallas, Chicago, Seattle, and San Antonio. As with the driver fatalities, transportation engineers would be interested in determining if any relationships between roadway and infrastructure elements and pedestrian fatalities could be elucidated.

Despite the large number of cities, there were only 549 observations of pedestrian fatalities in this data set, with 189 possible features. Such a large number of predictor variables would cause overfitting so we down-selected 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 4, categorized similarly as in the first section with the additional category of pedestrian characteristics. Appendix A.2 details these parameters. We elected to use 16 variables to show a comparison with the large HSIS model that also had 16 independent variables. 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.

Table 4: Variables selected for the NASS data set, including the associated number. Gray shading indicates a potential road or infrastructure design variable.

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

This data set 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 data set is also imbalanced with 26 fatal observations and 284 non-fatal observations, so there are only 8.39% positive observations. While not as imbalanced as the HSIS data set, there are similar issues in determining the relevant thresholds.

As with the first data set, Logistic Regression (LR) and Neural Network (NN) models were developed as in the previous data set exploration, with similar feature weight investigations. The overall accuracy of the LR model was 84.84% with an area under the prevision-recall curve of 0.4254. The NN model accuracy was 76.77% with an average area under the prevision-recall curve of 0.3612 (standard deviation of .1329). While LR is clearly the better choice for such a model, given the small data set size, it is very common mistake for practitioners to use NN models on smaller data sets, so we include it here to show how such a choice could influence overall interpretation.

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

3.1 Model Comparison

As with the HSIS data, we ranked the features by order of importance for the different methods, with the top 5 highlighted in Table 5. For the LR model, only variables that were statistically significant are listed (alpha = 0.05). 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 [29]. 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.

If the goal of a transportation engineer in analyzing this data was to determine roadway and infrastructure 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 4th for method 3. However, there was no strong model consensus and whether tax payer dollars should be spent due to recommendations from such models is questionable. Again, to demonstrate how much interpretation of these results can vary, three different perspectives are given regarding the outcomes.

Table 5: NASS rank orderings by the different models and feature interpretation 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 gender	-	8	11	9	8
6. Pedestrian motion	-	1	5	8	1
7. Pedestrian action relative to vehicle	-	15	13	11	-
8. Pedestrian first avoidance action	-	6	4	3	4
9. Sobriety	-	7	1	1	3
10. Speed limit	1	12	3	6	-
11. Vehicle curb weight	-	16	12	13	-
12. Driver distraction	-	11	9	10	-
13. Traffic way flow	-	4	15	15	6
14. Number of lanes	-	9	10	12	9
15. Roadway surface condition	-	10	14	14	-

16. Traffic control device function	-	13	8	4	-

- **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 only 1 feature, 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 (Senior Engineer):**
  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 agrees with previous studies, and so future work should target interventions that specifically address these two variables.

These three interpretations have some overlap but there is disagreement about a key aspect, which is the posted speed limit. The most experienced person felt this was a critical variable that should be acted upon, i.e., that DOTs should potentially lower speed limits in places where pedestrians were at high risk. However, the student engineer did not list this variable as important and the junior engineer felt speed limit might be important but ranked it 6th in order of importance. If the student or junior engineer's results were deemed correct, speed limits may not be deemed critical and such an error of omission could cause tangible safety concerns. Speed has an exponential relationship to pedestrian injury [30], so if data scientists who build ML models are not familiar with domain-specific literature, they could miss an important clue that suggests predictive models that do not include speed for pedestrian injury are likely wrong.

4 DISCUSSION

Transportation analysts are inundated with requests to apply popular machine learning modelling techniques to data sets to uncover never-before-seen relationships that could potentially revolutionize safety, congestion and mobility. To demonstrate some of the subjectivity and bias pitfalls in engaging in such analytics, two case studies were developed to answer the notional question of which road and infrastructure design variables in a dataset mattered for possibly reducing deaths in driving and pedestrian accidents.

To demonstrate how different choices in modelling technique could lead to different outcomes, Logistic Regression (LR) and Neural Network (NN) machine learning models were constructed on a large and a small transportation data set. 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. The models were then interpreted to determine which variables mattered in terms of possible interventions.

We chose two different data sets, one with 968,371 observations and the other at just 549 because with the rising popularity of ML techniques, there are increasing efforts to demonstrate the effectiveness of using of neural networks on small data sets. While some researchers claim that deep learning algorithms can achieve superior classification accuracy on data sets of only a few hundred points [31], others disagree and argue that ML-based methods are inherently brittle, even with large data sets [32, 33]. Thus, we wanted to explore just how much results could vary if a modeler chose to use a NN on a small data set instead of a more traditional and appropriate model like LR

For both data sets, results varied significantly depending on the model selected. Results also varied for the large and small data sets when four different approaches were used for NN feature weighting interpretation. Our results across both data sets and modelling approaches 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. No set of outcomes was dominant, even for the large data set, demonstrating that for real-world data sets, the choice of modelling approach and interpretation can dramatically influence possible results, and therefore any accompanying recommendations.

To illustrate why such varying model interpretations are so important, suppose the Junior Engineer's Group 1 variables in the first case study were decided to be the ones that received attention. Such people are the ones typically assigned these analyses, and increasingly organizations have few senior people to cross-check such results [34]. Recall that the goal was to determine where a Department of Transportation should spend money on infrastructure to reduce fatalities, for the large data set, the section length was the only road variable that the Junior Engineer ranked high in reducing fatalities. The practical implications of this potential error of commission are that some roads would need to be redesigned to reduce fatalities, since section length refers to the lengths of straightaways, curves, etc. (longer section lengths are better). On average, it would cost $4M/mile to expand and reconfigure such a road [35], so a wrong decision can have significant fiscal consequences.

The pedestrian model suffered from similar problems of extreme model variability. In this case, the neural network models were considered to be *good enough* models by the student and junior engineers, which led to them not adequately considering the role of speed limit. In this case study, speed limit is likely a critical variable that could have been missed. This error of omission could lead to inaction in reducing speed limits, which ultimately could save pedestrian lives if its importance is understood.

When looking at both the construction and the interpretation of the models through the lenses of these two case studies, we propose that that beyond ensuring that the underlying data is free from bias, there are a number of other opportunities for the introduction of subjectivity and bias which include:

- Picking the modelling approach 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 statistical 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 binary unbalanced outcomes,

- Choosing thresholds between important/unimportant features,
- Determining whether model accuracy is *good enough,*
- Deciding what the actual important features for a model.

This effort also highlights other core issues not often discussed in practical applications of these methods which include issues with imbalanced data, i.e., 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 as well as in other domains like healthcare with infection rates, and also in the financial world for events like fraudulent transactions. If practitioners blindly apply statistical packages without understanding the underlying nature of the assumptions, then results could be negatively affected.

It is possible that the underlying transportation data sets in the public repositories we accessed carry more uncertainty due to human construction, which led to the widely varying outcomes. However, data quality is a problem that is rampant across most domains where ML models are constructed. Moreover, we only demonstrated the use of two machine learning modelling approaches in the two transportation case studies, so these results cannot generalize to all forms of ML models (Fig. 2). However, there will be similar sources of subjectivity and bias in any application of a probabilistic analysis that requires making judgments about model parameters and thresholds, as well as the inherent subjectivities in interpretation.

5 CONCLUSION

While often heralded as transformative decision support tools, machine learning models can be significantly biased and subjective, and these weaknesses are a real problem for the commercial sector today. Because of such problems, recently Alphabet warned investors that AI can present serious risk to companies, stating *"…new products and services, including those that incorporate or utilize artificial intelligence and machine learning, can raise new or exacerbate existing ethical, technological, legal, and other challenges, which may negatively affect our brands and demand for our products and services and adversely affect our revenues and operating results. [4]"*

While it has been well established that underlying data can bias machine learning outcomes, how other subjective decisions about model construction and interpretation can influence such outcomes is less well known. To this end, this paper demonstrated that for both large and small data sets, two different machine learning approaches had twelve distinct points of subjectivity. The resulting models had similar accuracies, but very different outcomes in terms of determining which variables in the model should be considered critical actionable variables. Experience with such models produced very different interpretations of the outcomes, which could lead to errors of commission and omission, i.e., when some variables are thought to be more important than the really are or when critical variables are overlooked because of their low ranking in feature importance interpretation. Such errors have real world cost and safety consequences.

A significant issue with the use of such powerful but potentially brittle analytical tools is a lack of checks and balances for result generation and interpretation. In this study, we generated five 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 decisions based on machine learning-generated results.

Given the potential high number of sources of subjectivity and bias in both data selection and model development and interpretation, the data science community needs to be more transparent about these points of weakness. Moreover, courses should be developed that combine probabilistic modeling tools with decision science courses that look at biases in how humans perceive data relationships and how modelers' own experiences influence their perceptions of outcomes. Other future work includes the need to investigate data quality and stability of resulting model outcomes. For example, it may be that transportation data is especially susceptible to unstable models due to significant human interpretation in accident reports which is not seen in financial data sets. Another rich area of inquiry is how to optimize model feature selection given previous findings and sensitivity analysis results.

Understanding then that subjectivity and bias in ML models can come from more than just the underlying data, how should practitioners develop models that are robust, and more importantly, do no harm? First, this case study suggests it is important to both have experience in a data modeling team as well as subject matter expertise. Applying models of any kind to a domain that is not well understood raises the risk of irrelevant results, at best, but potentially serious consequences at worst. Taking the time to deeply understand a domain takes time and money, but it is not clear companies are willing to commit the resources for assumption and limitation checking, as well as in-depth sensitivity analyses. In addition, marketing departments are often disconnected from the actual engineers so performance claims are sometimes made to win a contract that engineers may struggle to meet. Experienced management that understands realistic timelines, the importance of gathering critical background information and conducting comprehensive sensitivity analyses is the key to ensuring models can be deployed with confidence.

ACKNOWLEDGMENTS

This work was sponsored by the US Department of Transportation University Transportation Center in partnership with the University of North Carolina Collaborative Science Center for Road Safety. Yaoyu Wang aided in collecting, cleaning, and processing the data. The reviewers were very helpful in shaping this work.

REFERENCES

[1] K. Hao. (2019) AI is sending people to jail—and getting it wrong. *Technology Review*.

[2] E. Strickland. (2019) IBM Watson, Heal Thyself. *IEEE Spectrum*.

[3] M. MacCarthy, "AI needs more regulation, not less," Brookings Institution, Washington DC, 2020.

[4] [J. Vincent. "Google and Microsoft warn investors that bad AI could harm their brand." The Verge. https://www.theverge.com/2019/2/11/18220050/google-microsoft-ai-brand-damage-investors-10-k-filing (accessed 26 June, 2020).

[5] N. T. Lee, P. Resnick, and G. Barton, "Algorithmic bias detection and mitigation: Best practices and policies to reduce consumer harms," Brookings Institution, Washington DC, 2019.

[6] S. Chandler. "How Explainable AI Is Helping Algorithms Avoid Bias." Forbes. https://www.forbes.com/sites/simonchandler/2020/02/18/how-explainable-ai-is-helping-algorithms-avoid-bias/#4c16d79e5ed3 (accessed 26 June, 2020).

[7] C. Fernández-Loría, F. Provost, and X. Han, "Explaining Data-Driven Decisions made by AI Systems: The Counterfactual Approach," *arXiv,* 2020.

[8] M. A. Gianfrancesco, S. Tamang, J. Yazdany, and G. Schmajuk, "Potential Biases in Machine Learning Algorithms Using Electronic Health Record Data," *JAMA internal medicine,* vol. 178, no. 11, pp. 1544–1547, 2018, doi: 10.1001/jamainternmed.2018.3763.

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

[10] M. L. Cummings and A. Stimpson, "Identifying Critical Contextual Design Cues Through a Machine Learning Approach," in *AAAI AI Magazine Special Issue on Computational Context*, W. Lawless and D. Sofge Eds. Palo Alto, CA, 2019.

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

[12] M. Peden *et al.*, "World report on road traffic injury prevention " World Health Organization, Geneva, Switzerland, 2004.

[13] G. Menardi and N. Torelli, "Training and assessing classification rules with imbalanced data," *Data Mining and Knowledge Discovery,* vol. 28, no. 1, pp. 92–122, 2014.

[14] T. Saito and M. Rehmsmeier, "The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets," *PLoS One,* vol. 10, no. 3, p. e0118432, 2015, doi: 10.1371/journal.pone.0118432.

[15] N. Stamatiadis and J. Pigman, "Impact of Shoulder Width and Median Width on Safety," Transportation Research Board, Washington DC, 2009.

[16] M. A. Hadi, J. Aruldhas, L. Chow, and J. A. Wattleworth, "Estimating safety effects of cross-section design for various highway types using negative binomial regression," *Transportation Research Record* no. 1500, pp. 169-177, 1995.

[17] L. Neudorff, P. Jenior, R. Dowling, B. Nevers, "Use of Narrow Lanes and Narrow Shoulders on Freeways: A Primer on Experiences, Current Practice, and Implementation Considerations", US Department of Transportation, Washington DC, 2016.

[18] O. M. Ibrahim, "A comparison of methods for assessing the relative importance of input variables in artificial neural networks," *Journal of Applied Sciences Research,,* vol. 9, no. 11, pp. 5692-5700, 2013.

[19] Cross Validated. "Deep learning : How do I know which variables are important?" Stack Exchange. https://stats.stackexchange.com/questions/261008/deep-learning-how-do-i-know-which-variables-are-important (accessed 24 July, 2019).

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

[21] E. P. Xing, M. I. Jordan, and R. M. Karp, "Feature Selection for High-Dimensional Genomic Microarray Data " in *Eighteenth International Conference on Machine Learning*, Williamstown, MA, 2001, pp. 601-608.

[22] A. H. Sung and S. Mukkamala, "Identifying important features for intrusion detection using support vector machines and neural networks " in *IEEE Symposium on Applications and the Internet*, Orlando, FL, 2003.

[23] O. Intratora and N. Intratorb, "Interpreting neural-network results: a simulation study," *Computational Statistics & Data Analysis,* vol. 37, no. 3, pp. 373-393, 2001.

[24] R. Guha, D. T. Stanton, and P. C. Jurs, "Interpreting Computational Neural Network Quantitative Structure−Activity Relationship Models: A Detailed Interpretation of the Weights and Biases," *Journal of Chemical Information and Modeling,* vol. 45, no. 4, pp. 1109-1121, 2005.

[25] A. Fisher, C. Rudin, and F. Dominici, "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* 2018.

[26] L. Breiman, "Random Forests," *Machine Learning Journal,* vol. 45, no. 1, pp. 5-32, 2001.

[27] A. Iranitalab and A. Khattak, "Comparison of four statistical and machine learning methods for crash severity prediction," *Accident Analysis & Prevention,* vol. 108, pp. 27-36, 2017, doi: 10.1016/j.aap.2017.08.008.

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

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

[30] B. C. Tefft, "Impact Speed and a Pedestrian's Risk of Severe Injury or Death," AAA Foundation for Traffic Safety, Washington DC, 2011.

[31] M. Olson, A. J. Wyner, and R. Berk, "Modern Neural Networks Generalize on Small DataSets," presented at the 32nd Conference on Neural Information Processing Systems, Montréal, Canada, 2018.

[32] D. Heaven, "Deep Trouble for Deep Learning," *Nature,* vol. 574, pp. 163-166, 2019.

[33] [33] M. A. Alcorn *et al.*, "Strike (with) a Pose: Neural Networks Are Easily Fooled by Strange Poses of Familiar Objects," *arXiv: 1811.11553,* 2018.

[34] R. Zwetsloot, R. Heston, and Z. Arnold, "Strengthening the U.S. AI Workforce," Center for Security and Emerging Technology, Washington DC, 2019.

[35] F. Elswick, " How Much Does It Cost to Build a Mile of Road?," vol. 2020, ed: Midwest, 2016.

A APPENDICES

Variable descriptions, ranges and categories are detailed in the following sections for the two data sets used for the two case studies.

A.1 HSIS Data Parameters

	Variables	Description
1	Speed limit	Miles per hour from 10 to 70.
2	AADT	Average annual daily traffic ranging from 0 to 775,446 cars.
3	Access control	0: no access control; 1: expressway - partial access control; 2: freeway - full access control.
4	Left shoulder width	Left should width (in increasing direction of the roadway), feet, ranging from 0 to 80.
5	Right shoulder width	Right shoulder width (in increasing direction of roadway), feet, ranging from 0 to 40.
6	Number of lanes	From 1 to 25.
7	Median width	Feet ranging from 0 to 800.
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.
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.
10	Weather	0: clear or cloudy, 1: weather may influence driving, including raining, snowing, wind and fog.
11	Maximum age	Maximum age of driver involved in the accident, ranging from 0 to 105.
12	Minimum age	Minimum age of driver involved in the accident ranging from 0 to 104.
13	Vehicle type	0: normal, 1: heavy, 2: motorcycle.
14	Sobriety	0: not impaired, 1: impaired.
15	Urban / Rural	0: rural, 1: urban.
16	Lane width	Feet ranging from 0 to 150.

A.2 NASS Data Parameters

	Variables	Description
1	Month	1-12: the month.
2	Time	3 or 4 digits representing time, e.g. 1523 means 15:23, 830 means 8:30.
3	Pedestrian weight	Weight in kg ranging from 9 to 150.
4	Pedestrian age	Actual age ranging from 2 to 93.
5	Pedestrian gender	1: male, 2: female.
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.
7	Pedestrian 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.
8	Pedestrian 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.
9	Sobriety	0: not drinking, 1: drinking.
10	Speed limit	Speed limit in km/h ranging from 16 to 105.
11	Vehicle curb weight	Actual value / 10 in kilogram ranging from 73 to 293.
12	Driver distraction	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.
13	Traffic way flow	1: not physically divided (two-way traffic), 2: divided traffic way - median strip without positive barrier, 3: divided traffic way - median strip with positive barrier, 4 one-way traffic way.
14	Number of lanes	Ranging from 1 to 7.
15	Roadway surface condition	1: dry, 2: wet, 3: snow and slush, 4: ice, 5: sand/dirt/oil.
16	Traffic control device function	0: no traffic control, 1: not functioning, 2: functioning.