2020

Prediction of Fatality Crashes with Multilayer Perceptron of Crash Record Information System Datasets

Thanh Hung Duong
University of Houston

Fengxiang Qiao
Texas Southern University

Jyh-Haw Yeh
Boise State University

Yunpeng Zhang
University of Houston

 https://doi.org/10.1109/ICCICC50026.2020.9450248

Prediction of Fatality Crashes with Multilayer Perceptrons on Crash Records Information System Dataset

***Abstract*—This**

Keywords—vehicle crash, data analysis, multilayer perceptrons,

I. INTRODUCTION

According to the Texas Department of Transportation, nearly every minute there is one reportable crash in Texas in 2018. [1] The number of deaths was 3639 and the economic loss of all crashes in Texas has been estimated at approximately 38.4 billion US dollars per year. As the population and number of annually manufactured vehicles increase gradually, vehicle crash prediction has become even more extremely important task and many scientists have tried to solve this challenging problem. Generally, they have carried out that task via three types of vehicle crash predictions. Firstly, the crash frequency or crash probability in a geographic space in a time period was chosen as a response variable, while the explanatory variables include information of traffic, road, driver, etc.[2] In order to model the relationship between independents and dependent variable, the Poisson regression or its modifications are applied since most of the crash data is count data. [3] The second approach is to predict the crash risk in real-time by using traffic flow characteristics obtained from traffic surveillance systems. [4], [5] Normally, statistical models like the Bayesian model or logistic regression have been applied to obtain the relationship between crash risk and real-times factors. [6], [7]. And finally, the crash severity analysis has attracted a significant number of researchers since it could access risk factors. These factors might be characteristics of the crash, driver, vehicle. For example, Keall et al. studied the effect of alcohol, age, and the number of passengers on the injury fatality of drivers in the night-time. [8] A logistic regression model was fitted to the crash data in New Zealand to expose those relationships. The result showed that teenage or drunk drivers are prone to fatal accidents. Each type of vehicle crash predictions plays different roles in the efforts of improving highway safety and this paper would concentrate on the last one – the prediction crash fatality.

In the crash fatality analyses, the dependent variables are either crash fatality or injury severity. While crash fatality has only two unique values: yes or no, the injury severity has different numbers of levels, which depends on the classification of each country or state. For example, the U.S. National Safety Council (NSC) developed the "KABCO" injury scale to classify injuries and they are K-Fatal, A-Incapacitating injury, B – Non-incapacitating injury, C – Possible injury, and O – No injury. [9] Since those target variables are categorical variables, statistical tools such as logit models, multivariate models or Bayesian model have been applied intensively in the reported analyses. [10]For instance, the logit model has been used since 1994 in the research of Shibata et al., in which they estimated the influence of driver characteristics such as license, alcohol, driving speed. [11] Later, in 2011, Zhu et al., used this model to conduct a comprehensive analysis on the injury severity of large-truck crashes. [12] Those statistical models have been proven to be useful in many cases but still, they have two primary drawbacks, including the linear function characteristic and valid-assumption requirement. [13], [14] Most of these models assumed certain statistical conditions of the data and applied linear functions form between target and explanatory variables. However, in many cases, the relations between crash severity and risk factors are non-linearity and when the model assumptions cannot be satisfied, the estimates of risk factors may be biased.

Recently, the artificial neural network (ANN) has drawn significant attention due to its success in many areas. Among ANNs, Multilayer Perceptron (MLP), a feed-forward ANN, has been applied frequently in vehicle crash analysis since it is a simple but powerful tool for linear/non-linear regression, and it does not need any assumptions. Among a large number of reported applications of ANN in crash prediction, the number of researches that solely concentrated on crash severity is inadequate. [4], [15]–[17] Delen et al. developed 8 different ANN models to capture the non-linear relationships between injury severity and 5 main groups of crash factors, including person, vehicle, environment, accident, and other information. [18] Moreover, a sensitivity analysis was conducted to determine the importance of crash-related factors. However, although their models were for binary classification, the evaluation metric was only the accuracy, the other important metrics such as precision, recall were not mentioned. Later, Zeng and Huang implemented the convex combination algorithm into the neural network model to reduce the load of computation and to prevent over-fitting. [13] Despite having fewer nodes than the normal neural network, the optimized neural network attained NN, achieved remarkable results in terms of both prediction accuracy and computation time. Nonetheless, they had to combine the fatalities with incapacitating injuries to reduce the imbalance of the dataset as the model might be biased. Similarly, in the most recent reported crash severity analysis, Banerjee and Khadem converted 5 injury severity level models into 3-class and binary models, though they did a highly detailed analysis. Moreover, their study focused intensively on the alcohol-related crashes.

In this study, we propose a novel simple model of MLP that can provide a high accuracy prediction of crash fatality within a short time and it can handle large and heavily imbalanced datasets. The dataset, which is obtained from the Highway Safety Information System, consists of all reported crashes in Texas from June to July 2019. The result was compared with the traditional logit model in terms of performance and computing time. Furthermore, via our NN model, a sensitivity analysis was conducted to estimate the roles of risk factors.

The structure of the rest of this paper is that section II is a description of the dataset. Next, section III shows our methodology in data preparation, the traditional logistic model, and our novel MLP model. Then the results and discussion were represented in the Section IV.

II. DATASET

In Texas, the crash data is collected by the law enforcement officers in Texas Department of Transportation (TxDOT). Normally, public can obtain the crash data via a query tool but if the needed amount of data is large, they must make requests directly to the TxDOT's Crash Records Information System (CRIS). The original crash dataset we retrieved contains collected information from 324,117 crashes from June to July 2019 in Texas. It has 172 difference features and is in csv format. After filtering out the ones whose number of missing values is larger than 20% of the total cases, 19 features were extracted, and 2 more calculated features are added, as showed in the Table 1. As all cases that contains missing values were removed, the number of crashed in the final dataset is 287,847.

Table 1 List of all variables

Target variable	
Crash Fatality	Indicates that the crash involved one or more fatalities
Independent variables	
Time factors	
Crash Date	Date on which crash occurred
Crash Time	Time crash occurred
Day of Week	The day of the week that the crash occurred
Road factors	
Roadway System	type of roadway on which crash occurred
Roadway Part	part of roadway on which crash occurred
Speed limit	Speed limit
Toll_Road	Toll Road/Toll Lane
Road_Constr_Zone	Indicates whether the crash occurred in or was related to a construction, maintenance, or utility work zone
Road_Constr_Zone_Wrkr	Indicates whether workers were present in the road construction zone at the time of the crash
Entr_Road	Entering Roads
Roadway Alignment	The geometric characteristics of the roadway at the crash site
Surface condition	The surface condition (wet, dry, etc.) present at the time and place of the crash
At intersection	Indicates if the crash occurred at an intersection.
Other factors	
Longitude	GPS coordinate
Latitude	GPS coordinate
Light condition	The type and level of light that existed at the time of the crash
Weather condition	The prevailing atmospheric condition reported by the officer at the time of the crash
Traffic Control	Type of traffic control at the scene of the crash
SCV	Indicates whether this is a single-vehicle crash
Number of Units	Number of crash-involved vehicles

III. METHODOLOGY

A. Data preprocessing and wrangling

Since the dataset contains different types of variables, while machine learning models only accept the numerical input, it needs to be prepared. For example, the crash date and time were merged and converted to a Python DateTime object. As all crashes occurred in the same year 2019 and more than 90% of the crashes happened in June, the crash data is filtered out observations in another month. Moreover, we already selected Day of Week variable, only time variable was kept. Next, the range of crash time is more important than the exact crash time so crash time was categorized as dawn (2AM-6:59AM), morning (7AM-11:59AM), afternoon (12PM-4:59PM), evening (5PM-8:59PM) and night (9PM-1:59AM). Then the crash time was converted into dummy/indicator variables. Similarly, this kind of processing cycle was applied to all other categorical variables to not only convert them into numerical variables but also increase their impact on models. Besides, duplicated rows or rows that contain missing values were removed. The final prepared dataset contains 287,847 cases with 101 numerical explanatory variables. The target variable is the crash fatality, which has 2 unique values 0 and 1. Since the number of fatal crashes is 1640 cases, which is 0.57% of total crashes, the data set is heavily imbalanced.

For evaluating machine learning models, the dataset is divided into training, validation, and test sets, which have sizes of 184,221, 46,056, and 57,570 cases, respectively. Then all those sets were standardized to make sure that they are internally consistent.

B. Traditional statistical regression model

Because the dependent variable indicates whether the crash is fatal, it can be predicted via binary classification models like logistic regression. In logistic regression, instead of predicting the class of the target variable directly, the model focuses on predict the probability of the event that the target variable is true. [19]The probability function, or logistic function, can be expressed as the following equation:

$$p(X) = \frac{e^{\beta_0+\beta_1 X}}{1+e^{\beta_0+\beta_1 X}} \quad (1)$$

In this study, the logistic model in the package statsmodels.api in Python was applied.

C. MLP model

The term neural network involved in a large type of deep learning model. [20]In this paper, we used the most common neural network, which can be called single hidden layer back-propagation network or single layer perception since it is simple

but powerful. Typically, our MLP model consists of an input layer, one hidden layer, a dropout layer and an output layer, as illustrated in the Figure 1. In our case, the number of nodes in input and output layers are 100 and 1, respectively as the input data have 100 features and the output is the probability that a crash is fatal. The number of nodes in the hidden layer was tuned between 10 and 50 to optimize the model. To handle the highly imbalance of the data set, two techniques, which are putting weight on the minority class and resampling, were applied.

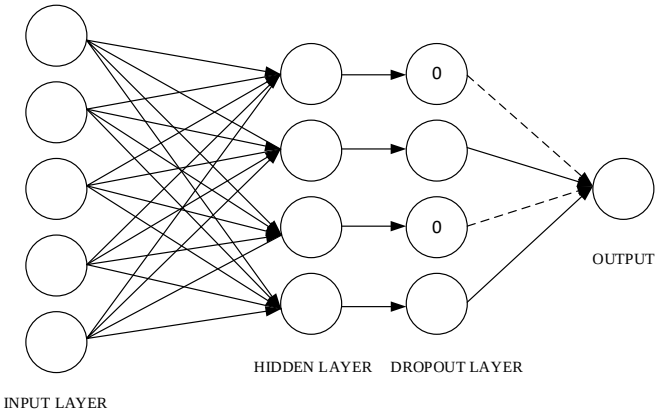

Figure 1 The Multilayer Perceptrons Structure used in the paper.

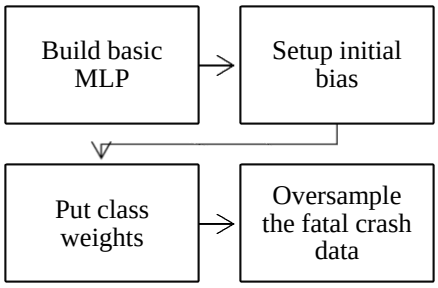

Figure 2 The flowchart of model developing

The Figure 2 shows the steps were conducted to develop the neural network model. Firstly, a simple MLP model was built based on the architecture in Figure 1 with 10 nodes in the hidden layer. The activations for hidden layer and output layers were ReLU and Sigmoid functions, respectively. The optimizer was adam with learning rate at 0.001, which can handle very noise and sparse gradients.[21] The training will stop early if there is no improvement of the AUC for the test set. As the neural network models are extremely sensitive to the initial weights, by setting the correct initial bias, a significant amount of computation can be reduced. [22], [23] The initial bias was computed by using the following formula:

$$initial\_bias = log(\frac{number\ of\ fatal\ crashes}{total\ number\ of\ crashes}) \quad (2)$$

Next, in order to get the attention of model on the minority class, class weights were calculated and applied. The weight for each class was a ratio of the total number of crashes to the number of cases in that class. After that, we tried to reduce the imbalanced in the data via random oversampling approach for the fatal class.

D. Evaluation metrics

In binary classification, the importance evaluation metrics contains not only the accuracy but also precision, recall, and the AUC. The accuracy is the ratio between the total number of correct prediction and the total cases. However, in cases of imbalanced dataset, if all observations are predicted as non-fatal crashes, we will obtain a very high accuracy. Therefore, other evaluation metrics must be included to estimate the model's performance correctly.

IV. RESULTS & DISCUSSION

A. Traditional statistical models

The logit() function from statsmodels.api package was implemented to fit the training set with 'bfgs' method. However, as the dataset is imbalanced, model was not converged, as a result, the pseudo R-squared was 0.107. Even though the accuracy was 99.4 %, no fatal crashes were detected. Hence, both recall and precision were zero.

B. MLP models

Table 2 show the evaluation metrics after each step. At the beginning, the initial model provided similar result of logit() function, which was high accuracy but was unable to detect any fatal crashes. After putting the weights for each class, the accuracy dropped to 88.09%, but 214 fatal crashes were detected correctly. Hence the improvement in the AUC was significant, from 78.80% to 81.12%. Similarly, the oversampling technique helped increase the area under the curve moderately. The outstanding performance of the final model was illustrated in the ROC plots in the Figure 3. However, the trade-off was a significant drop in accuracy, which is from 99.45% to 74.71%. In details, percentage of correctly predicted fatal crashes were 76.82%, while that of correctly non-fatal predictions are 74.70%. Those accuracies are slightly higher the reported accuracies in the study of Banerjee et al.[24]

Table 2 Evaluation metrics for test set at each step of model developing

Step	Accuracy	Precision	Recall	AUC
Basic model	99.45	0	0	78.8
Class weights	88.09	1.85	67.94	81.12
Oversampling	74.71	1.63	76.82	82.79

Next, in order to check the feature importance, the function PermutationImportance() from eli5 package was applied. The main idea of this function is that when a feature is not available, the amount of decreases of model scores will reflect the importance of that feature. This sensitivity analysis on the input were carry out on the oversampled dataset and top ten highest weights are shown in the Table 3. The result indicates that the factor that whether a crash is a single-vehicle crash has the biggest weight on the output. However, since among 10 highest weight, 4 features belong to the road alignment factor, which sum up to 0.0262, it can be said that road alignment has the highest impact on the probability of fatal crashes. Moreover,

other factors can affect the fatal possibility are dark condition, interstate highway, cloudy weather, signal light.

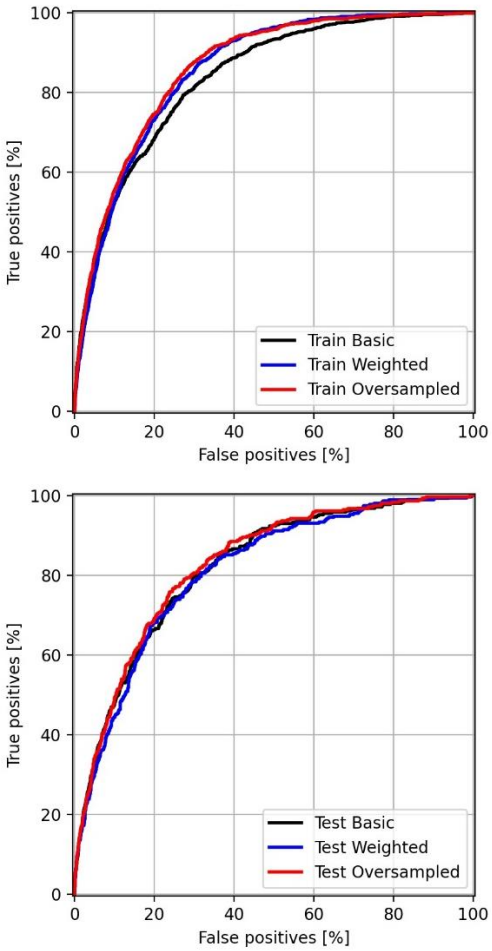

Figure 3 The ROC plots for training set and test set

Table 3 Top 10 highest weights in the model

Rank	Weight	Feature
1	0.0190 ± 0.0009	SVC
2	0.0151 ± 0.0019	Road_Algn_ID_STRAIGHT, LEVEL
3	0.0109 ± 0.0009	Light_Cond_ID_DARK, NOT LIGHTED
4	0.0083 ± 0.0002	Rpt_Rdwy_Sys_ID_INTERSTATE
5	0.0069 ± 0.0006	Road_Algn_ID_STRAIGHT, GRADE
6	0.0034 ± 0.0006	Wthr_Cond_ID_CLOUDY
7	0.0033 ± 0.0007	Traffic_Cntl_ID_SIGNAL LIGHT
8	0.0029 ± 0.0012	Traffic_Cntl_ID_NONE
9	0.0021 ± 0.0004	Road_Algn_ID_CURVE, LEVEL
10	0.0021 ± 0.0001	Road_Algn_ID_STRAIGHT, HILLCREST

V. CONCLUSION

In this paper, a simple binary artificial neural network model was proposed to predict the fatal crashes in a highly imbalanced dataset. The class weight and oversampling techniques were applied to improve the performance of the model. As a result, the achieved AUC was 82.79% and the overall accuracy were 74.71%, which is slightly better than previous reported study. Furthermore, sensitivity analysis was conducted based on the model to estimate the importance of features. The result showed that single-vehicle crash factor, geometric characteristics of the roadway, and light conditions affect significantly to the probability of fatal crash. For future work, more approach for handling imbalanced dataset such as Synthetic Minority Oversampling Technique (SMOTE), Tomek links, should be applied to improve the model. And another approach is to add more features from crash data to the model input.

REFERENCES

[1] "Texas Motor Vehicle Traffic Crash Facts Calendar Year 2018," 2019.

[2] Q. Zeng, H. Huang, X. Pei, S. C. Wong, and M. Gao, "Rule extraction from an optimized neural network for traffic crash frequency modeling," *Accid. Anal. Prev.*, vol. 97, pp. 87–95, Dec. 2016.

[3] D. Lord, S. P. Washington, and J. N. Ivan, "Poisson, poisson-gamma and zero-inflated regression models of motor vehicle crashes: Balancing statistical fit and theory," *Accid. Anal. Prev.*, vol. 37, no. 1, pp. 35–46, 2005.

[4] C. Lee, B. Hellinga, and F. Saccomanno, "Real-Time Crash Prediction Model for Application to Crash Prevention in Freeway Traffic," *Transp. Res. Rec. J. Transp. Res. Board*, vol. 1840, no. 1, pp. 67–77, Jan. 2003.

[5] W. Hu, X. Xiao, D. Xie, T. Tan, and S. Maybank, "Traffic accident prediction using 3-D model-based vehicle tracking," *IEEE Trans. Veh. Technol.*, vol. 53, no. 3, pp. 677–694, May 2004.

[6] M. Abdel-Aty, N. Uddin, A. Pande, M. F. Abdalla, and L. Hsia, "Predicting Freeway Crashes from Loop Detector Data by Matched Case-Control Logistic Regression," *Transp. Res. Rec. J. Transp. Res. Board*, vol. 1897, no. 1, pp. 88–95, Jan. 2004.

[7] C. Oh, J. S. Oh, and S. G. Ritchie, "Real-time hazardous traffic condition warning system: Framework and evaluation," *IEEE Trans. Intell. Transp. Syst.*, vol. 6, no. 3, pp. 265–272, 2005.

[8] M. D. Keall, W. J. Frith, and T. L. Patterson, "The influence of alcohol, age and number of passengers on the night-time risk of driver fatal injury in New Zealand," *Accid. Anal. Prev.*, vol. 36, no. 1, pp. 49–61, Jan. 2004.

[9] "Highway Safety Improvement Program Manual - Safety | Federal Highway Administration." [Online].

Available: https://safety.fhwa.dot.gov/hsip/resources/fhwasa09029/sec4.cfm. [Accessed: 21-Apr-2020].

[10] P. T. Savolainen, F. L. Mannering, D. Lord, and M. A. Quddus, "The statistical analysis of highway crash-injury severities: A review and assessment of methodological alternatives," *Accid. Anal. Prev.*, vol. 43, no. 5, pp. 1666–1676, Sep. 2011.

[11] A. Shibata and K. Fukuda, "Risk factors of fatality in motor vehicle traffic accidents," *Accid. Anal. Prev.*, vol. 26, no. 3, pp. 391–397, Jun. 1994.

[12] X. Zhu and S. Srinivasan, "A comprehensive analysis of factors influencing the injury severity of large-truck crashes," *Accid. Anal. Prev.*, vol. 43, no. 1, pp. 49–57, Jan. 2011.

[13] Q. Zeng and H. Huang, "A stable and optimized neural network model for crash injury severity prediction," *Accid. Anal. Prev.*, vol. 73, pp. 351–358, Dec. 2014.

[14] Q. Zeng, H. Huang, P. Xu, and M. Ma, "Developing an Optimized Artificial Neural Network to Predict Traffic Crash Injury Severity," in *CICTP 2014*, 2014, pp. 2396–2407.

[15] R. Abduljabbar, H. Dia, S. Liyanage, and S. A. Bagloee, "Applications of artificial intelligence in transport: An overview," *Sustainability (Switzerland)*, vol. 11, no. 1. MDPI AG, 02-Jan-2019.

[16] N. Fallah, H. Gu, K. Mohammad, S. A. Seyyedsalehi, K. Nourijelyani, and M. R. Eshraghian, "Nonlinear Poisson regression using neural networks: A simulation study," *Neural Comput. Appl.*, vol. 18, no. 8, pp. 939–943, 2009.

[17] A. Abdulhafedh, "Crash Frequency Analysis," *J. Transp. Technol.*, vol. 06, no. 04, pp. 169–180, 2016.

[18] D. Delen, R. Sharda, and M. Bessonov, "Identifying significant predictors of injury severity in traffic accidents using a series of artificial neural networks," *Accid. Anal. Prev.*, vol. 38, pp. 434–444, 2006.

[19] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An introduction to Statistical Learning*, vol. 7, no. 10. 2000.

[20] J. Hastie, Trevor, Tibshirani, Robert, Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. 2009.

[21] D. P. Kingma and J. L. Ba, "Adam: A method for stochastic optimization," in *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*, 2015.

[22] J. F. Kolen, J. F. Kolen, J. F. Kolen, J. B. Pollack, and J. B. Pollack, "Back Propagation is Sensitive to Initial Conditions," *COMPLEX Syst.*, vol. 4, pp. 860--867, 1990.

[23] G. Thimm and E. Fiesler, "Neural network initialization," Springer, Berlin, Heidelberg, 1995, pp. 535–542.

[24] B. Snehanshu Banerjee and N. K. Khadem, "Factors Influencing Injury Severity in Alcohol-Related Crashes: A Neural Network Approach Using HSIS Crash Data," *ITE J.*, vol. 89, no. 3, 2019.