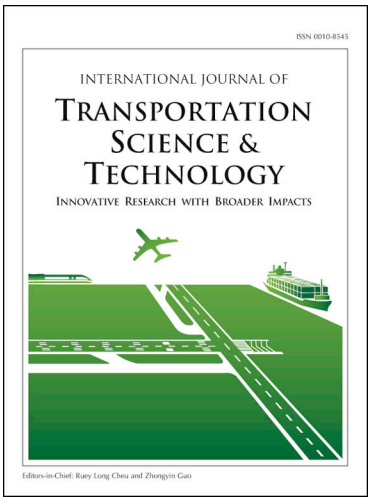
Journal Pre-proofs



Application of Machine Learning Models and SHAP to Examine Crashes Involving Young Drivers in New Jersey

Ahmed Sajid Hasan, Mohammad Jalayer, Subasish Das, Md. Asif Bin Kabir

PII: S2046-0430(23)00034-5
DOI: https://doi.org/10.1016/j.ijtst.2023.04.005
Reference: IJTST 329

To appear in: *International Journal of Transportation Science and Technology*

Received Date: 7 January 2023
Revised Date: 28 March 2023
Accepted Date: 17 April 2023

Please cite this article as: A. Sajid Hasan, M. Jalayer, S. Das, Md. Asif Bin Kabir, Application of Machine Learning Models and SHAP to Examine Crashes Involving Young Drivers in New Jersey, *International Journal of Transportation Science and Technology* (2023), doi: https://doi.org/10.1016/j.ijtst.2023.04.005

Application of Machine Learning Models and SHAP to Examine Crashes Involving Young Drivers in New Jersey

ABSTRACT

Motor vehicle crashes are the leading cause of the death of teenagers in the United States. Young drivers have shown their higher propensity to get involved in crashes due to using a cellphone while driving, breaking the speed limit, and reckless driving. This study analyzed motor vehicle crashes involving young drivers using New Jersey crash data. Specifically, four years of crash data (2016-2019) was gathered and analyzed. Different machine learning (ML) methods, such as Random Forest, Light GBM, Catboost, and XGBoost, were used to predict the injury severity. The performance of the models was evaluated using accuracy, precision, and recall scores. In addition, interpretable ML techniques like sensitivity analysis and Shapley values were conducted to assess the most influential factors' impact on young driver-related crashes. The results revealed that XGBoost performed better than Random Forest, CatBoost, and LightGBM models in crash severity prediction. Results from the sensitivity analysis showed that multi-vehicle crashes, angular crashes, crashes at intersections, and dark-not-lit conditions had increased crash severity. A partial dependence plot of SHAP values revealed that speeding while in clear weather had a higher likelihood of injury crashes, and multi-vehicle crashes at the intersection had more injury crashes. We expect that the results obtained from this study will help policymakers and practitioners to take appropriate countermeasures to improve the safety of young drivers in New Jersey.

Keywords: Random Forest, Boosting Method, Young Driver Involved Crash, Sensitivity Analysis, Shapley Values, New Jersey

1. INTRODUCTION

Young drivers are susceptible to fatal crashes around the globe. Usually, their involvement in fatal crashes is disproportionately high compared to their numbers on the road (Racioppi et al., 2004). According to the Centers for Disease Control and Prevention and the National Highway Traffic Safety Administration, motor vehicle crashes are the leading cause of death for young people in the United States (Center of Disease Control, 2017; National Highway Traffic Safety Administration, 2019). In the United States, 2,738 teenagers died in motor vehicle crashes in 2020, a 14% increase from 2019 (IIHS, 2022). The state of New Jersey averaged 60 traffic fatalities resulting from young (16-20 years old) driver-involved crashes from 2013-2017. In fact, over half (55%) of the victims of these crashes are people from other age groups (NJDHTS, 2018). For young drivers, many factors such as inexperience, risk, and sensation-seeking behavior, impairment, cell phone use while driving, other in-vehicle distractions, and speeding can contribute to crash occurrence (Groeger, 2006; C. Lee & Abdel-Aty, 2008; Simons-Morton et al., 2005).

To improve young driver safety, it is important to identify factors related to crash frequency and severity. This study contributes to state of the art by conducting a comprehensive literature review of the existing works and summarizing the important contributing factors and the methods used to interpret them. To thoroughly understand crashes involving young drivers, site and state-specific factors must be investigated. This study addresses it by examining the factors that contribute to the severity of young driver-involved crashes in New Jersey (2016–2019). To predict the severity of young driver crashes, this study further employed several ML models. A feature

importance model was utilized to determine the top variables contributing to crash severity. To investigate the influence of the top contributing factors on crash severity, sensitivity analysis, and Shapley values were evaluated. A pairwise comparison of top contributing factors is further performed to interpret the combined influence of the variables on the crash severity. The findings of this study can help engineers, practitioners, and legislators control the severity of young driver-involved crashes in New Jersey.

This paper is organized as follows. The introduction section describes the importance of studying young driver involved crashes. The introduction is followed by a comprehensive literature review section summarizing past works on young driver-involved crashes. The methods section describes the data, the study design, and the description of the models used for crash severity prediction. This section is followed by the results and discussions section, where the results of the study are explained and interpreted. Finally, the conclusion of the study is presented.

2. LITERATURE REVIEW

Previous researchers from various perspectives analyzed young drivers' behavior. First, some studies investigated the reasons behind young drivers' high crash risk compared to other drivers. These factors were inexperience, in-vehicle distractions (handheld cell phones and tuning radios), underestimation of risk, alcohol use, and their intention to reach the destination faster (Groeger, 2006; C. Lee & Abdel-Aty, 2008; Simons-Morton et al., 2005).

Some previous studies investigated behavioral factors of young drivers associated with crash involvement. These studies demonstrated that driver anger, impulsiveness, and sensation-seeking are the noteworthy personality factors contributing to young drivers involved in crashes (Dahlen et al., 2005; Deffenbacher et al., 2003). In the existing literature, various factors such as driver characteristics, crash attributes, roadway features, environmental conditions, and vehicle characteristics have been discovered to impact the severity of young driver-involved crashes. In the existing literature, various factors (including driver characteristics, crash attributes, roadway features, environmental conditions, and vehicle characteristics) have been discovered to impact the severity of young driver-involved crashes. Speeding (Ferguson, 2013; (Rolison and Moutari, 2020); distracted driving (Gershon et al., 2017, 2019); driver age and gender (Keating & Halpernfelsher, 2008; Rhodes & Pivik, 2011, Rahman et al., 2021); risky behavior (Simons-Morton et al., 2005); impaired driving (Simons-Morton et al., 2005); failure to yield (Rahman et al., 2021); and peer passengers (Bingham et al., 2016; Micucci et al., 2019) have been identified as having the most significant impact on the severity of young driver-involved crashes. Previous research has also discovered that vehicular characteristics, such as the number of vehicles involved and the type of vehicles, are frequently influenced by crash severity (Bates et al., 2014). Several studies have also discovered that environmental factors such as lighting (Lin et al., 2020), road surface conditions (Rolison and Moutari, 2020), and weather conditions (Lin et al., 2020; Rolison and Moutari, 2020) have a significant impact on crash severity in young driver-involved crashes. Crash characteristics, such as crash type and the presence of a curve, have also been shown to significantly contribute to the severity of young driver-involved crashes (Lin et al., 2020). As traffic flow varies with the time of day, temporal features like the time of the day and the day of the week (Williams, 2003) significantly impact the severity of young driver-involved crashes. Roadway geometric features like the number of lanes (Lin et al., 2020), functional classification of the road (H. Y. Chen et al., 2009), traffic control devices (Lin et al., 2020), and speed limit (Andrey et al., 2013), are also found to contribute to the crash severity involving young drivers.

Previous researchers have developed two types of prediction models for crash severity: the parametric and the non-parametric models. The crash severity is used as the dependent variable in these analyses. Crash severity is discrete and has multiple categories corresponding to various crash severity levels defined by the KABCO scale (K: Killed, A: Incapacitating injury, B: Non-incapacitating injury, C: Possible injury, O: Not injured). Recently, researchers used a variety of statistical methods in their crash severity analyses. The most used models are regression models, discrete choice models (logit and probit), multinomial logit models, mixed logit models, hierarchical modeling, rule-based association approach, and classification tree-based models (Baireddi et al., 2018; Das, Dutta, et al., 2019; Dong et al., 2017; Hasan, Orvin, et al., 2022; Islam et al., 2022; Jalayer et al., 2021; Jalayer et al., 2017; Jalayer & Zhou, 2016; Jalayer, Pour-Rouholamin, et al., 2018; Jalayer, Shabanpour, et al., 2018; Penmetsa & Pulugurtha, 2018; Rahman et al., 2021; Roque et al., 2021; Taylor et al., 2018; Q. Wu et al., 2014). In crash severity analysis, traditional statistical models (e.g., logistic regression, Bayesian logistic regression) have also been used (Abdel-Aty et al., 2007; Ahmed et al., 2012; Das, Bibeka, et al., 2019). Most of these studies examined the impact of contributing variables (roadway features, environmental variables, and driver behavior) on crash severity. Because parametric models are based on assumptions and relationships between dependent and independent variables, their use is limited when those assumptions are incorrect (Wang & Kim, 2019). However, non-parametric or ML models are not based on predefined relationships or assumptions, which makes them learn the outcome of dependent variables from a large dataset with diverse explanatory variables. This learning method makes ML models useful for crash severity analysis. Some previous studies have demonstrated the advantages of using ML over traditional statistical models (Iranitalab & Khattak, 2017; Wang & Kim, 2019). Machine Learning Models like Random Forest (D. Li et al., 2017; Mafi et al., 2018), SVM (Support Vector Machine) (C. Chen et al., 2016; Hasan, Kabir, et al., 2022; Z. Li et al., 2012; Mokhtarimousavi et al., 2019), and Neural Networks (Sameen & Pradhan, 2017; Zeng et al., 2016) have been used extensively in the field of transportation safety. In some recent studies, boosting methods like XGBoost (Hasan et al., 2021; D. Lee et al., 2019) have also been used for crash severity prediction. However, few studies on young driver-involved crashes emphasized comparison amongst traditional ML models (e.g., SVM, RF) with more recent methods (boosting methods like CatBoost, XGBoost, Light GMB).

Researchers have used explainable AI to explain the impacts of the important variables in the crash severity analysis (Das et al., 2020). Previous studies have used sensitivity analysis, permutation feature importance, and partial dependence plot (Hasan et al., 2022; Christoph Molnar, 2019; Hasan et al., 2021; Williamson et al., 2015). Shapley Additive Values are also used in recent studies (Parsa et al., 2020). However, using both sensitivity analysis and SHAP values could help better comprehend the impact of the most influential variables. To the best of the authors' knowledge, none of the previous studies on young driver-involved crashes have utilized more than one interpretable ML, which we addressed and compared in this study (sensitivity analysis and SHAP interpretation).

This study contributes to traffic safety analysis of young drivers by summarizing the existing practices and gaps in predicting young driver-involved crashes. A comparison of the prediction performance of four ML approaches is further performed. The top contributing factors are identified using three different feature selection techniques. Then, this study further analyzed the sensitivity of the critical categories of the top contributing factors. Afterward, a SHAP dependence plot is used to interpret the combined impact of important variables on the injury severity of young driver crashes. The explainable AI findings would be especially useful in

assisting transportation agencies in identifying the causes and patterns of crash severity. These findings would also help develop effective safety countermeasures to reduce the number of crashes involving young drivers in New Jersey. Additionally, the methodology of this study would be useful in future studies to analyze the crash severity for other types of crashes (e.g., heavy vehicle-involved crashes, older driver-involved crashes, and crashes due to distracted driving). Moreover, the outcomes of the ML models will help future researchers choose an appropriate technique for crash severity analysis of their dataset.

3. METHODS

3.1 Description of Database

We used four years (2016–2019) of young driver-involved crash data in New Jersey, which included a total of 170,343 crashes (8,151 injury, 34,451 possible injury, and 127,741 no injury). 'Injury' type of crashes were obtained after merging fatal crashes with major and minor injury crashes. 'No Injury' crashes are crashes that result in Property Damage Only (PDO). All incomplete and erroneous data records from the collected raw data were discarded. For instance, variables unrelated to crash severity, such as "Crash Number" and "Behavioral Countermeasures," as well as the variables containing a high number of missing values, were eliminated. Based on the literature review and engineering judgments, 22 independent variables are selected for further analysis, divided into six categories: temporal features (i.e., season, day of the week, and time of day), driver characteristics (i.e., alcohol or drugged driver involved, distracted driving involved, and cell phone in use), roadway features (i.e., intersection, highway type, median type, temporary traffic control zone, area, total pedestrian and/or bicyclist involved, unrestrained occupant involved, and run-off road involved), environmental conditions (i.e., environmental conditions, light conditions, and surface conditions), vehicle characteristics (i.e., total vehicles involved and unsafe speed involved), and the crash attributes (i.e., crash type, and curve-related). Table 1 provides a summary of the chosen variables associated with crash severity. According to the table, 77.3% of the crashes occurred during the day, and 10.4% of the late-night crashes involved injury. Also, 66.7% of the crashes were found to not be within intersection boundaries, and 39.6% occurred where the median was not present. Distracted driving accounted for about 55% of the total crashes. Some features, like total pedestrian and/or bicyclist involvement, unrestrained occupant involvement, run-off road involved, unsafe speed involved, and alcohol or drugged driver, involved showed skewness towards the crashes, which cause injury.

Table 1. Distribution of Key Features

Explanatory Variables	Injury	%	Possible Injury	%	No Injury	%	Total	Frequency (%)
Total Crashes	8151	5	34451	20	127741	75	170343	100
Temporal Variables								
Season								
Fall	1994	4.4	9529	21.0	33762	74.6	45285	26.6
Spring	2004	5.1	8018	20.2	29646	74.7	39668	23.3
Summer	2628	5.6	9540	20.5	34462	73.9	46630	27.4
Winter	1525	3.9	7364	19.0	29871	77.1	38760	22.8
Day of Week								
Weekday	5778	4.4	26338	20.3	97877	75.3	129993	76.3

Explanatory Variables	Injury	%	Possible Injury	%	No Injury	%	Total	Frequency (%)
Weekend	2373	5.9	8113	20.1	29864	74.0	40350	23.7
Time of Day								
Day (6 AM to 6 PM)	5499	4.2	26390	20.0	99773	75.8	131662	77.3
Evening (6 PM to 12 AM)	1874	6.0	6530	20.9	22822	73.1	31226	18.3
Late Night (12 AM to 6 AM)	778	10.4	1531	20.5	5146	69.0	7455	4.4
Roadway Features								
Intersection								
Near rail crossing	3	2.5	25	20.8	92	76.7	120	0.1
Not Within Intersection Boundaries	4826	4.2	20401	18.0	88417	77.8	113644	66.7
Within Intersection Boundaries	3322	5.9	14025	24.8	39232	69.3	56579	33.2
Temporary Traffic Control Zone								
Construction	125	4.6	568	20.9	2031	74.6	2724	1.6
Maintenance	5	4.1	34	27.6	84	68.3	123	0.1
Other	8018	4.8	33829	20.2	125563	75.0	167410	98.3
Utility	3	3.5	20	23.3	63	73.3	86	0.1
Highway Type								
Divided	1496	4.4	7066	20.7	25584	74.9	34146	20.0
Dual/Dual	113	4.9	427	18.5	1765	76.6	2305	1.4
Undivided	3798	5.6	15921	23.5	48085	70.9	67804	39.8
Unknown	2744	4.2	11037	16.7	52307	79.1	66088	38.8
Median Type								
Curbed	279	4.5	1249	20.4	4608	75.1	6136	3.6
None	3764	5.6	15807	23.4	47840	71.0	67411	39.6
Painted	38	5.5	161	23.5	487	71.0	686	0.4
Positive	902	4.2	4555	21.0	16230	74.8	21687	12.7
Unknown	2743	4.2	11038	16.7	52304	79.1	66085	38.8
Unprotected	425	5.1	1641	19.7	6272	75.2	8338	4.9
Total Pedestrian and/or Bicyclist Involved								
No	7575	4.5	33775	20.0	127403	75.5	168753	99.1
Yes	576	36.2	676	42.5	338	21.3	1590	0.9
Unrestrained Occupant Involved								
No	7406	4.4	33614	20.0	126676	75.5	167696	98.4
Yes	745	28.1	837	31.6	1065	40.2	2647	1.6
Area								
Rural	638	10.0	1327	20.9	4384	69.1	6349	3.7
Unknown	1943	3.7	8320	15.9	42017	80.4	52280	30.7
Urban	5570	5.0	24804	22.2	81340	72.8	111714	65.6
Run-off Road Involved								
No	5812	3.8	30367	19.9	116593	76.3	152772	89.7
Yes	2339	13.3	4084	23.2	11148	63.4	17571	10.3
Environmental Conditions								
Environmental Conditions								
Adverse	1640	4.3	7606	20.1	28569	75.5	37815	22.2
Clear	6511	4.9	26845	20.3	99172	74.8	132528	77.8
Light Conditions								
Dark-Lit	2061	5.7	7764	21.5	26279	72.8	36104	21.2
Dark-Not Lit	614	7.8	1442	18.2	5847	74.0	7903	4.6
Dawn/Dusk	335	5.6	1214	20.4	4395	73.9	5944	3.5
Daylight	5134	4.3	24013	20.0	91148	75.8	120295	70.6
Unknown	7	7.2	18	18.6	72	74.2	97	0.1
Surface Condition								
Dry	6367	4.9	26337	20.3	97018	74.8	129722	76.2

Explanatory Variables	Injury	%	Possible Injury	%	No Injury	%	Total	Frequency (%)
Wet	1784	4.4	8114	20.0	30723	75.6	40621	23.8
Vehicle Characteristics								
Total Vehicles involved								
1	2780	11.1	4493	17.9	17771	71.0	25044	14.7
2	4269	3.3	24311	18.7	101772	78.1	130352	76.5
3 or more	1102	7.4	5647	37.8	8198	54.8	14947	8.8
Unsafe Speed Involved								
No	6552	4.2	31095	20.1	117367	75.7	155014	91.0
Yes	1599	10.4	3356	21.9	10374	67.7	15329	9.0
Crash Attributes								
Curve Related								
No	6852	4.5	30946	20.4	113693	75.0	151491	88.9
Yes	1299	6.9	3505	18.6	14048	74.5	18852	11.1
Crash type								
Angle	2476	7.0	9581	27.0	23415	66.0	35472	20.8
Fixed Object	1881	9.3	3534	17.4	14875	73.3	20290	11.9
Others	1908	6.0	4165	13.2	25469	80.7	31542	18.5
Rear-end	1388	2.2	14973	24.2	45561	73.6	61922	36.4
Sideswipe	498	2.4	2198	10.4	18421	87.2	21117	12.4
Driver Characteristics								
Alcohol or Drugged Driver Involved								
No	7608	4.5	33809	20.2	126259	75.3	167676	98.4
Yes	543	20.4	642	24.1	1482	55.6	2667	1.6
Distracted Driver Involved								
No	4112	5.4	14732	19.3	57333	75.3	76177	44.7
Yes	4039	4.3	19719	20.9	70408	74.8	94166	55.3
Cell Phone in Use								
No	8038	4.8	33986	20.2	126556	75.1	168580	99.0
Yes	113	6.4	465	26.4	1185	67.2	1763	1.0

As noted before, 21 input variables were selected from the initial data analysis. Further, exploratory data analysis was performed to examine their detailed characteristics. It is worth mentioning that the performance of ML models depends on the quality and quantity of the data used for model development. Usually, big data without quality control can worsen model performance. An extensive exploratory data analysis (EDA) can help improve quality control by removing redundant information. This study conducted an extensive EDA and performed feature importance to select suitable variables for analysis. Notably, all the categorical and continuous variables are assigned numeric values while using them as input for the ML models.

3.2 Study Design

For this study, four years (2016-2019) of young driver-involved crash data were gathered and cleaned. Later, three feature importance techniques were applied to find the rank of the feature in the decision tree. Depending on the collinearity of variables, some features were excluded, and the final dataset was prepared. A 70:30 ratio between training and testing data was used for this investigation. Four ML models were performed on the dataset. The model performance was evaluated based on accuracy, precision, and recall values. Later, interpretable ML models like sensitivity analysis and Shapley values were applied to interpret the models. Finally,

countermeasures were suggested based on the model evaluation and model interpretation. The study design is sketched in Figure 1.

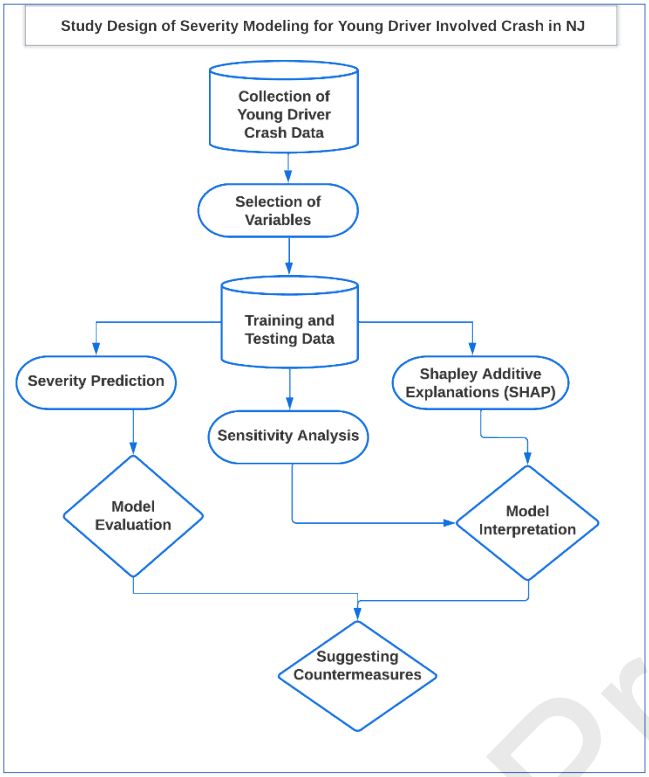


Figure 1. Study Design of Severity Modeling for Young Driver Crash Data in NJ

3.3 Feature Importance

Insignificant factors in the dataset often increase the noise in the model and deteriorate the prediction performance (C. Chen et al., 2016). Variable selection methods, such as variable importance ranking, are utilized to address these noises. The most used variable selection methods include CART, Discrete choice models, univariate selection, Random Forest, ExtraTree Classifier, XGboost, and other techniques. The relative importance of the contributing variables is determined by the ML models based on their impact on the crash severity prediction (C. Chen, Zhang, Tarefder, et al., 2015; C. Chen, Zhang, Wang, et al., 2015). Various previous studies have used the CART technique, ExtraTree classifier, and XGboost, and found those to be effective ranking techniques (Ahmed et al., 2012; Delen et al., 2017; Hossain & Muromachi, 2013). This study employs these three techniques for determining the relative significance of variables related to crash severity. In these models, variable importance scores indicate the contributions of the variables as major splitters of the regression tree in improving crash severity predictions (Banerjee et al., 2008)(Banerjee et al., 2008). For the CART model, the ranking of the variables is computed based on the value of the Gini index, which indicates the level of impurity or entropy in the decision tree (T.-E. Wu et al., 2020). Similarly, using a set of randomized trees, an extra tree classifier computes the variable rankings based on the Gini Index (Abubaker et al., 2020). XGBoost employs a gradient-boosting decision tree (GBDT) that is applicable to both classification and regression (T. Chen & Guestrin, 2016). The greedy algorithm maximizes the maximum gain of the target variable (crash severity) throughout each tree layer's construction. The

purpose of the method is to grow a tree by continually adding trees and separating features. Every time a new tree is added, the algorithm learns a new function to fit the residual from the previous prediction (Huang et al., 2021).

3.4 Description of Different ML Algorithms

3.4.1 Random Forest (RF)

RF is a tree-based classifier. RF uses random feature selection and bagging, two distinct techniques for classification (Breiman, 2001). Bagging creates each tree separately, whereas random feature collection creates decision trees quickly. RF selects the features of the subgroups at random, as opposed to utilizing every feature in the decision trees. RF predicts the output of a new dataset by averaging the outputs of separate random bootstrap training data.

3.4.2 Boosting Methods

By combining several weak classifiers into a single strong classifier, the boosting approaches aim to improve prediction performance. Three boosting methods were used in the classification model: XGboost, LightGBM, and Catboost. At each iteration, the gradient boosting method adjusts the losses by regressing the gradient vector function (Friedman, 2001). A gradient boosting model modifies the order of each decision tree, starting with the weak Decision Tree that served as the basis for the base Decision Tree. XGboost is a slow-boosting approach that reduces misclassification errors at each iteration through sequential model training. Catboost is a boosting method that accepts numerical and category input variables. It handles the variables during the training period and saves preprocessing duration. LightGBM is a boosting method that builds more accurate and complex decision trees leaf by leaf.

Gradient-boosted decision trees are incorporated into XGBoost, also known as an ensemble technique. Friedman initially proposed this algorithm in 2002 (Friedman, 2002). In numerous areas of study, the XGBoost algorithm has produced encouraging outcomes (L. Zhang & Zhan, 2017; Y. Zhang & Xie, 2007). This algorithm consists of a set of decision trees, each of which learns from its predecessor and influences its successor. The formulation of XGboost with K tree functions is as follows:

$$\hat{q}_i^{(t)} = \sum_{k=1}^{t} f_k(p_i) = \hat{q}_i^{(t-1)} + f_t(p_i) \qquad [1]$$

where $\hat{q}_i^{(t)}$ is the estimated crash severity after t^{th} iterations,

k is the num. of the additive trees,

t is the num. of iterations,

$f_k(p_i)$ is the k^{th} tree function for variables p_i,

$\hat{q}_i^{(t-1)}$ is the predicted response value for the final iteration, and

$f_t(p_i)$ is the tree function of i^{th} iteration.

The objective function for minimizing the loss $l(q_i,\hat{q}_i)$ can be shown as follows:

$$Obj = \sum_{k=1}^{n} l(q_i,\hat{q}_i) + \sum_{k=1}^{t} \Omega(f_k) \qquad [2]$$

$$\Omega(f_t) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} \omega_j^2 \quad [3]$$

where $\Omega(f_t)$ is the regularization term for preventing overfitting and reducing the complexity, T is the num. of leaves, ω_j^2 is the L_2 norm of j^{th} leaf scores, and n is the total num. of crashes in sample data.

3.4.3 Model Evaluation

In this study, we assessed severity prediction performance with respect to accuracy, precision, and recall (Delen et al., 2017). The ratio of the correctly predicted severity class to the total number of crashes is called accuracy. The ratio of predicted crashes for a crash severity (injury, no injury, and possible injury) by the ML algorithm to the total number of expected crash severity for that class is known as precision. On the other hand, recall is defined as the proportion of accurately anticipated crash severity to the total crashes for that severity class. A higher value for these evaluation indicators indicates that the classification model correctly predicted the crash severities.

3.4.4 Sensitivity Analysis

Understanding the relationships between the explanatory and the target variables is crucial in crash severity analysis (X. Li et al., 2008). Mathematical explanation techniques for inferring ML algorithm estimations have been proposed recently (Aas et al., 2021; Olden et al., 2004). Some examples of these explanation techniques are the Explanation Vector (Baehrens et al., 2009), LIME (Local Interpretable Model-agnostic Explanations) (Ribeiro et al., 2016), and the Shapley value (Eriktrumbelj & Kononenko, 2010). Methods from the field of explainable AI are employed to explain the model thoroughly. These methods include permutation feature importance, sensitivity analysis, and partial dependence graphs (Christoph Molnar, 2019; Hasan et al., 2021). Sensitivity analysis is a way to assess the input-output relationship by changing the value of one variable while holding the others constant (Hasan et al., 2022; Strobl et al., 2008; Yu & Abdel-Aty, 2013). To what extent a variable's value affects the model's predictive ability is thus shown by its sensitivity. A feature has a large effect on model predictions if a change in the feature's value significantly affects the model's prediction. This study used sensitivity analysis to assess the impact of potential crash severity modifiers in crashes involving young drivers.

3.4.5 Shapley Values

The output of ML models could be interpreted using SHAP, which was developed by Lundberg and Lee (Lundberg et al., 2017). SHAP provides a way to estimate the contribution of each feature and is based on local explanations (Ribeiro et al., 2016) and game theory (Erikštrumbelj & Kononenko, 2010). Let's assume an XGBoost model that predicts an output N using a group N (with n features). According to each feature's marginal contribution, the contribution of each feature (φ is the contribution of feature i) on the model output v (N) is allocated in SHAP (Parsa et al., 2020). Shapley values are established based on several axioms that help fairly allocate the contribution of each feature using the following equation (Parsa et al., 2020) :

$$\emptyset_i = \sum_{S \subseteq N\{i\}} \frac{|S|!\,(n - |S| - 1)!}{n!} [v(S \cup \{i\}) - v(S)] \quad [4]$$

Where S is a subset of N. Based on the subsequent Additive feature attribution technique, a linear function of binary features g is defined as:

$$g(z')= \emptyset_o + \sum_{i=1}^{M} \emptyset_i + z_i' \quad [5]$$

Where $z' \in \{0, 1\}^M$ equals 1 when a feature is observed; otherwise, it equals 0, and M is the number of input features (Lundberg et al., 2017). The model was further improved by Lundberg et al. (Lundberg et al., 2020) to use the Tree Explainer methods to efficiently calculate the risk factors of a SHAP value globally and locally (Ayoub et al., 2021). A detailed explanation of SHAP interpretation could be found in previous studies (Wen et al., 2021).

4. RESULTS AND DISCUSSION

4.1 Exploratory Data Analysis

A correlation test was performed on the 21 input variables (with the encoded or assigned value) chosen to see how similar they are. The correlation coefficient shows how closely two variables are related. When the correlation coefficient is high (> 0.7), it means that there is a strong relationship, which can be positive or negative (based on the signs). On the other hand, a lower value means that the relationship between any two variables is weak. The correlation results show that both the time-light condition and the environment-surface condition have correlation values of more than 0.7. On the other hand, all of the other variables have a correlation coefficient of less than 0.7, meaning there is no strong link between them.

The feature importance analysis shows the ranking of the input variables. Figure 2 shows the relative importance of the input variables in ascending order from top to bottom for all three feature selection techniques. All three methods (XGBoost, CART, and ExtraTree Classifier) have identified crash type, season, light conditions, the total number of vehicles involved, and median type as the most influential contributing factors. This variable importance also identifies which correlated variables need to be excluded from the study, as correlated variables have a similar influence in the ML model. Based on the analysis from the correlation matrix and Figures 2, time and surface condition variables are omitted from the study. Finally, 19 input variables were selected for developing an ML model where crash severity is the target variable.

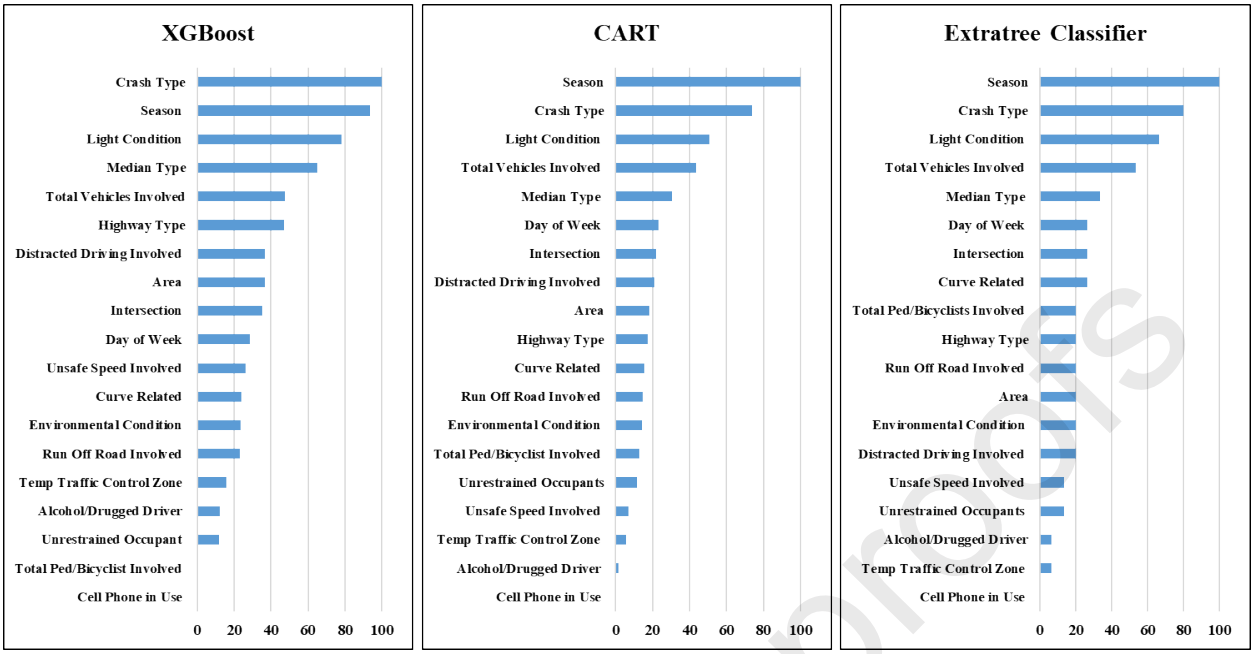


Figure 2. Feature Importance of input variables

4.2 Model Performance

A classification model based on ML is developed using the four ML algorithms described in the preceding section. The codes for the classification model are generated using an open-source sci-kit learn python package and the variables stated in the previous section (Pedregosa et al., 2011). As specified, a 70-30 split of the complete dataset is considered for the train and test set, with 70% of the data used to train the classification models and 30% used to evaluate the produced classification models (Friedman, 2001). Random splitting of the train and test sets is performed to ensure the performance of the test set on an entirely unknown data set. K-fold cross-validation (n=5) was performed for each of the ML algorithms to ensure the reliability of the model evaluation. In addition, when multiple ML methods are applied to the classification model, the train and test sets remain consistent. It is noteworthy that hyperparameter tuning was performed for the algorithms, and the hyperparameter combination with the best output for each model was chosen for the crash severity prediction. A description of the hyperparameters used for this study is mentioned in Table 2.

Table 2. Parameter values used for ML algorithms

Algorithms	Hyperparameters
RF	n_estimators=100, criterion='gini', random_state=0, verbose=0, class_weight=None, max_samples=None, max_depth=None, min_samples_split=2, min_samples_leaf=1, max_features='auto', max_leaf_nodes=None, ccp_alpha=0.0, bootstrap='True'
XGBoost	n_estimators=200, min_samples_split=2, min_samples_leaf=1, max_features='auto', max_depth=50, bootstrap='False'
LightGBM	boosting_type='gbdt', num_leaves=31, max_depth=-1, learning_rate=0.1, n_estimators=100, n_jobs=-1, subsample_for_bin=200000, class_weight=None, min_child_weight=0.001, min_child_samples=20, subsample=1.0, colsample_bytree=1.0, random_state=0

Algorithms	Hyperparameters
Catboost	iterations=500, learning_rate=0.03, depth=6, l2_leaf_reg=3.0, loss_function= 'Logloss', verbose=None, boosting_type=None, random_state=0

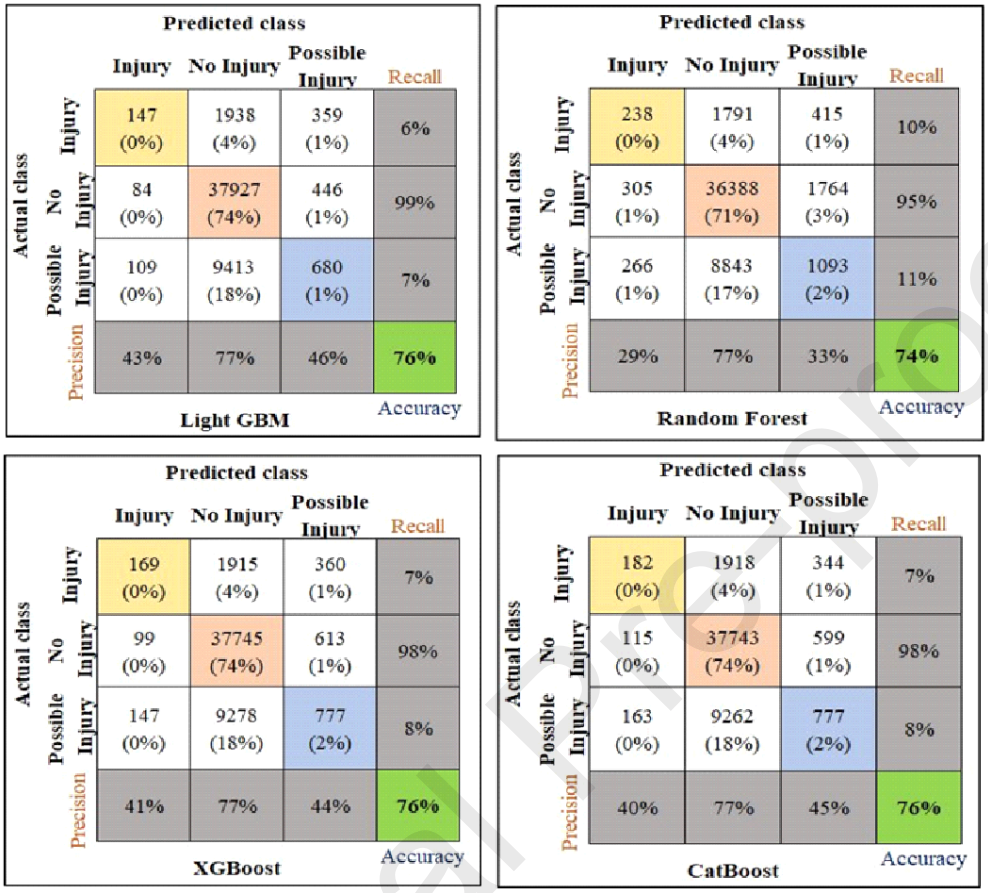


Figure 3. Performance evaluation of different ML algorithms

In Figure 3, we see how a confusion matrix may be used as a simple tool to display the model's performances in terms of the original class against the expected class. The accuracy of predictions made using a classification model is summarized in a confusion matrix table, where diagonal components represent accurate predictions and non-diagonal elements reflect inaccurate ones. As shown in Figure 3, the overall accuracy of the test dataset is 76% for all applied ML algorithms except for RF (74%). Since only the overall accuracy does not reflect the overall performance of the classification model, a further comparison is considered with the help of recall and precision values. Also, it is important to identify which class type is of interest for a particular classification model. For crash analysis, injury is considered the most important type of crash severity. Based on this consideration, XGBoost and CatBoost have identical values for overall accuracy (76%) and recall (7%). However, XGBoost (41%) outperformed the CatBoost model (40%) by a very small margin with the precision value for the injury class. Notably, LightGBM had the highest precision of all the models (41%) in predicting injury severity. However, it also had the lowest recall value (6%). Similarly, RF had the highest recall values (10%) and the least precision (29%).

For the model's overall performance, a combination of accuracy, precision, and recall is more important than having the most precision or recall values. Therefore, CatBoost and XGBoost were the best performers in predicting the severity of young driver-involved crashes for the selected years in New Jersey. Hence, the boosting methods had an overall higher performance than the Random Forest Model.

Notably, the selected models have shown higher accuracy than existing models. For example, some previous studies have used SVM and found accuracies like 48.8% and 55.58% (Chen et al., 2016; Li et al., 2012). Some other studies on neural networks also found crash severity prediction accuracies within 57-60% from various algorithms (Amiri et al., 2020). Compared to the previous studies, the chosen models have shown higher accuracy in the crash severity prediction.

4.3 Sensitivity Analysis of Important Variables

Previous research has employed ML models to assess the sensitivity of key predictors (X. Li et al., 2008; Z. Li et al., 2012; Yu & Abdel-Aty, 2013). Six variables (crash type, season, intersection involved, light conditions, median type, and total vehicles involved) were used in the feature selection analysis for this research. To determine the degree of association between the variables, a correlation test was used. The Pearson correlation test states that a value of 0.3 or below indicates a weak relationship, while a value above 0.7 indicates a strong one (Ratner, 2009). The correlation results show that there was little to no association between the variables. It is also obvious that the variables themselves are not significantly connected since all of them maintained a low correlation among themselves. These parameters are suitable for use in sensitivity analysis due to their minimal correlation and high significance scores.

Figure 4 illustrates how alterations to a single variable affect the crash severity predictions of multiple models. Each explanatory variable was modified by an amount specified by the user, while the remaining variables remained unaltered. Only the top six contributing factors were considered for this study. As a result of transforming every variable in our dataset to dummy variables, the variable input values changed from 0 to 1. In order to calculate the variable impact on crash severity, the proportion of each severity level before and after a variable perturbation was then recorded. (J. Zhang et al., 2018). The colored bars in Figure 4 represent changes in the predicted proportion of the reported severity classes as a result of changing one variable. The results of the sensitivity analysis are explained below:

4.3.1. Crash Type (Fixed Object)

Due to fixed-object crashes, all models showed an increase in the proportion of fatal and injury crashes and a decrease in the proportion of 'no injury' crashes (J. Zhang et al., 2018). For fixed-object crashes, the proportion of 'no injury' crashes decreased by 2% to 5%, while the proportion of fatal and injury crashes increased (1% to 2%). Liu and Subramanian discovered that, when compared to other ages, young drivers have the highest proportion of fatal fixed object crashes– particularly those involved in run-off road crashes (Liu & Subramanian, 2009).

4.3.2. Number of Vehicles Involved (Single Vehicle)

Single vehicle crashes contribute to an increase in the proportion of no injury crashes. The proportion of possible injury decreased to 1.5%, while the proportion of the no injury class increased to 2% due to involvement of single-vehicle crashes in young drivers. Previous researchers also found that the fatal or injury type of crash severity increases with the presence of multiple vehicles during the crash (Lin et al., 2020).

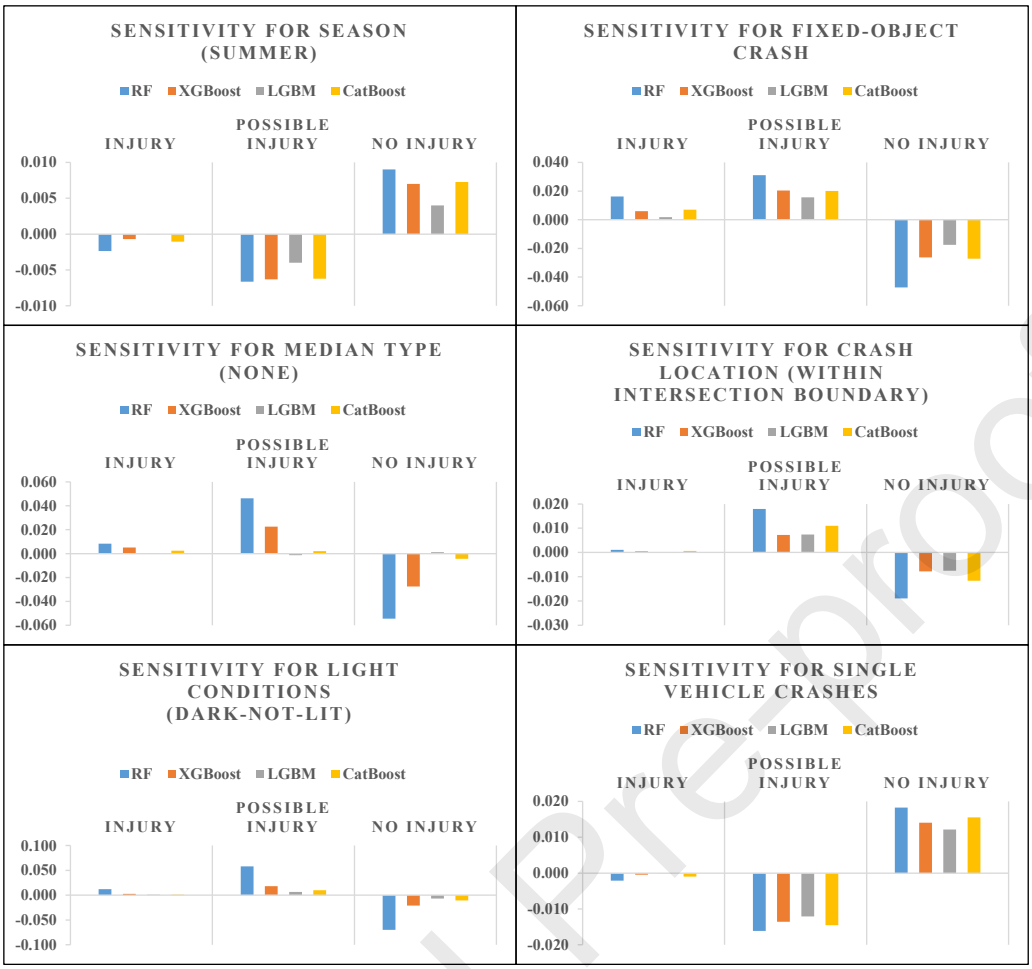


Figure 4. Sensitivity of the top contributing factors in predicting crash severity

4.3.3. Type of Median (No Median)

All the models showed an increase in the proportion of injury and possible injury crashes and a decrease in the no injury crashes due to crashes on roads without a median. The proportion of no injury crashes decreased by 1% to 4%, while the proportion of injury crashes (0.1%~0.2%) and possible injury crashes (0.1%- 4%) increased for crashes on no median roadway. Neyens and Boyle demonstrated that young drivers are more prone to angular crashes than other types of crashes, especially when distracted (Neyens & Boyle, 2007).

4.3.4. Season (Summer)

All models indicated a drop in the proportion of the fatal and injury class for the young driver-involved crash in summer, with RF demonstrating the greatest decrease of 0.2%. No injury crashes increased (0.3%- 0.6%) while possible injury decreased (0.4%-0.6%) for the crashes during summer.

4.3.5. Light Conditions (Dark-not-lit)

The dark-lit condition increased the proportion of fatal and injury crashes (0.5%- 2%) and possible injury crashes (1% to 6%) while decreasing the proportion of no injury crashes (0.5% - 6%). This

finding is intuitive because the likelihood of a collision increases with vision impairment created by a dark-lit or dark not-lit roadway. Previous research on truck-related accidents yielded similar results (Lin et al., 2020).

4.3.6. Intersection Involved (Crash within intersection boundary)

The involvement of crashes within the intersection boundary demonstrated an increase in both injury and possible injury (0.5%-2%) crashes and a decrease in no injury (1%-2%) crashes. Because of complex conflicting movements, frequent stop-and-go traffic crashes within intersection boundaries are more severe than in other roadway facilities (Kidando et al., 2021). As per USDOT, more than 50% of fatal and injury crashes occur within intersection boundaries (United States Department of Transportation, 2022).

4.4 Interpretable ML (SHAP values)

The impact of contributing factors on the crash severity prediction was further explained using SHAP values. The literature shows that XGBoost was extensively used in explanatory AI to get the SHAP values (Guo et al., 2020; Parsa et al., 2021; Yang et al., 2021, Zhou et al., 2020;). In this study, we also investigated the impact of the variables on the XGBoost models crash severity prediction by using SHAP values. The average absolute Shapley values per feature across the entire dataset were estimated by the global importance of the input variables (Fig. 5a). The greater the mean SHAP value, the greater the significance of the contributing factor. It should be noted that the SHAP value's importance is path independent and does not change depending on which features are removed first. Additionally, the plot highlights how important each input variable is for the three crash severities: injury, no injury, and possible injury outcomes. Such a plot offers previously unconsidered new insights into the crash severity prediction. The global SHAP value shows that crash type has the most impact on predicting crash severity, followed by total vehicles involved, run-off, the road involved, intersection, and season. Although the variable selection techniques discussed in Figure 2 gave similar ranks, the impact provided by the features on severity prediction for each class was unobserved in the feature selection methods. This additive explanation helps to interpret the model better. For instance, the SHAP value of the injury class (0.42) is more for crash type than the no injury or possible injury. The same trend is followed for total vehicles involved and run-off the road involved, making them a critical factor to consider for injury crashes. Contrarily, the impact for the variable intersection is similar in all three crash severities.

Summary plots can be used to show the range and distribution of input variable impacts on failure mode prediction (Figure 5b). Each point represents a Shapley value for the input variables and an instance on the plots in Figure 5b. The input variables are shown on the y-axis in ascending order of significance. Each dot is colored based on the input variable's value, ranging from low (blue) to high (red). The density depicts how the data set's points are distributed (i.e., whether it contains a broad range of values or just a few carefully chosen ranges). For example, both the number of vehicles and run off the road involved show red plots on the right side, indicating a positive SHAP value for injury class. This implies that the greater the number of vehicles, the greater the SHAP value and the greater the impact on the injury severity. Similarly, run-off the road involved and speeding involved both generated a higher SHAP value and a high impact on injury type of crashes. Observing closely the plot of total vehicles involved and run-off the road involved, we can also see the less the number of vehicles, the less the SHAP value, and it is with a higher density. Similarly, the density of blue values is more towards negative SHAP values for

run-off the road crashes, meaning that crashes not involving run-off the road events have less propensity to injury.

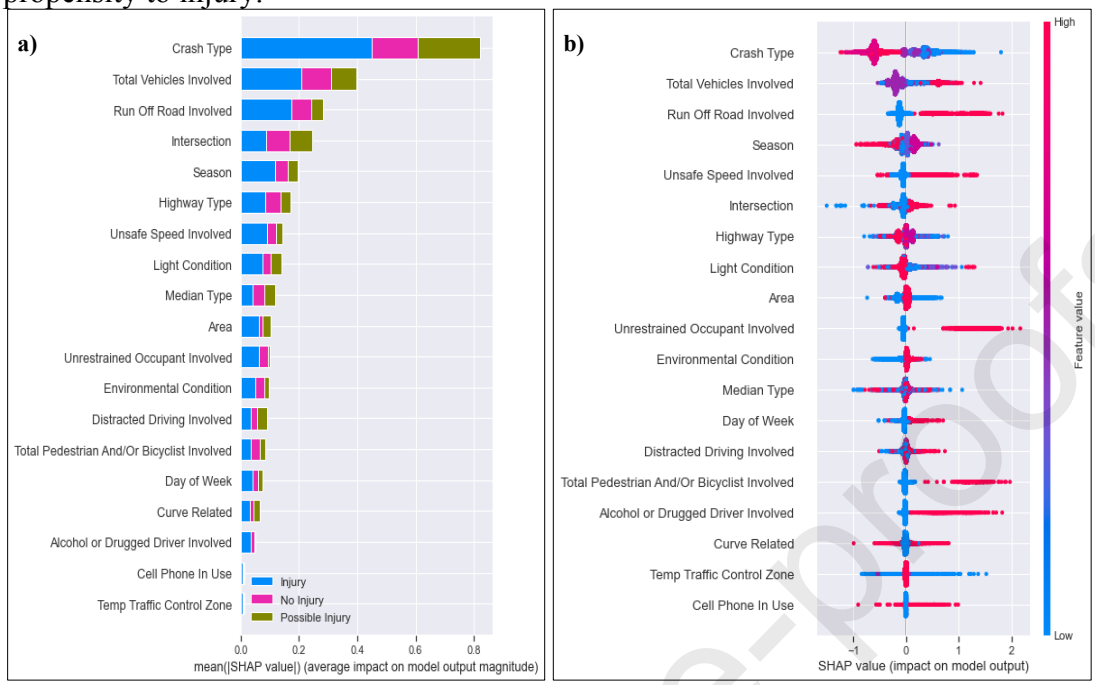


Figure 5. a) Global SHAP importance values and b) Summary plots for injury severity

Figure 6 illustrates the SHAP dependency for the injury severity, showing how the SHAP value changes as the variable input changes. The SHAP values shown in Fig.5 and 6 explain the impact of the features differently. Whereas Fig.5 is derived by averaging/summarizing the SHAP values (main effects of variables), Fig.6 shows the interaction effects of the variables. Figure 6 provides a more in-depth understanding of the spread and variation of SHAP values with the input variables. Figure 6(a) depicts the effect of crash type for a run-off road crash involved. It is seen that angular crash (crash type=1) has a higher SHAP value for run-off road crashes (run-off road involved=1), meaning that young drivers mostly get injured when run-off road crashes result in an angular crash. The increase in total vehicles involved in the crash experience has more SHAP values on the weekend (day of week=1), resulting in more injury crashes (Wundersitz, 2012). Speeding-related injury has smaller SHAP values in adverse weather, meaning that people avoid speeding in adverse weather, which results in less severe injury (Lin et al., 2020). Multiple vehicle crashes within intersection boundaries generate a high SHAP value; thus, those crashes experience high severity (Kidando et al., 2021; Lin et al., 2020).

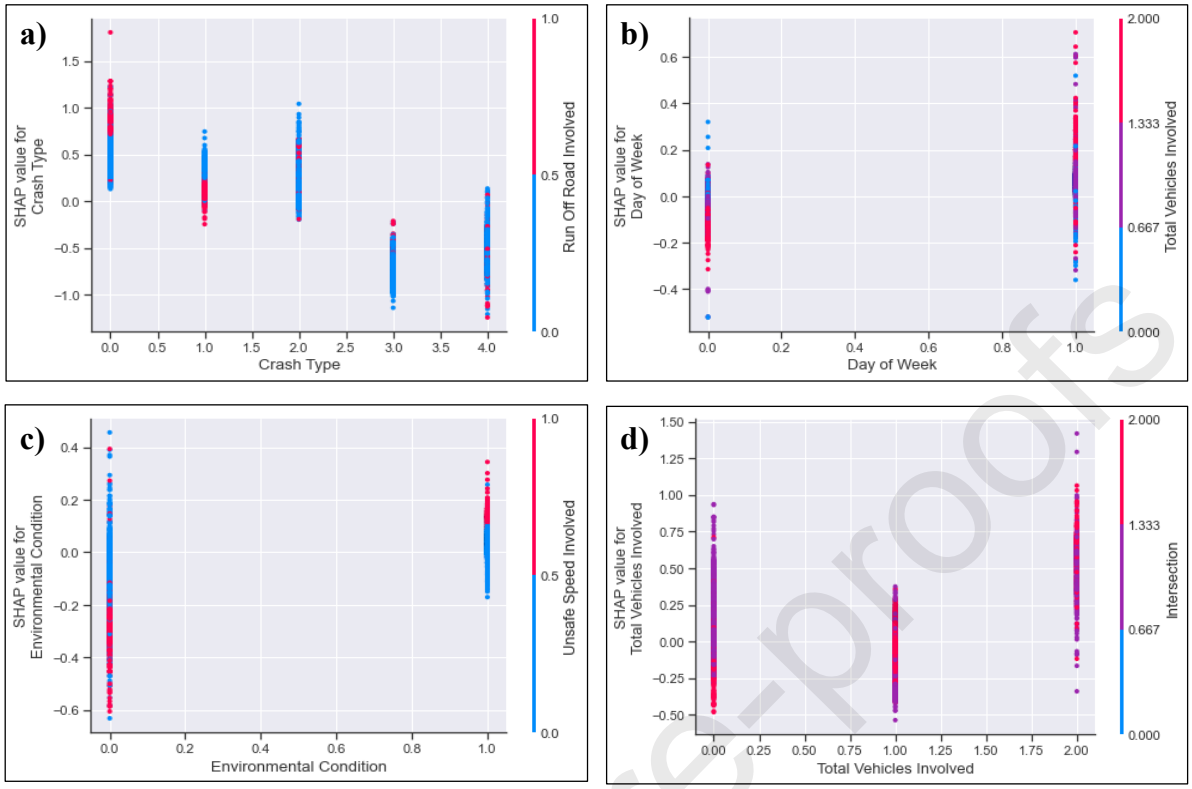


Figure 6. SHAP dependence plots of the top contributing factors for injury crashes

5. CONCLUSIONS

This study explored the performance of data-driven ML techniques for predicting the crash severity of young driver-involved crashes from different input variables. First, an extensive database was prepared from the raw database, which accumulated 170,343 crash data of three different crash severity: injury, no injury, and possible injury. Exploratory data analysis was performed to check correlation, and one of the correlated variables was excluded based on the ranking obtained from feature importance analysis. Finally, a total of 19 input variables were selected for developing the ML model, where crash severity is the target variable. Four different ML algorithms were employed to establish the classification model. The performance of the classification model was analyzed using a confusion matrix, where all of the evaluation metrics, such as overall accuracy, precision, and recall, were considered. XGBoost and CatBoost provided the best results when all three evaluation metrics were considered among the four ML models. Hence, the boosting methods had an overall higher performance than the Random Forest Model.

Third, the sensitivity analysis of the contributing factors indicated that the factors like median type, single vehicle crashes, fixed-object crashes, intersection, and dark-not-lit conditions significantly contributed to the severity of the young driver-involved crashes. The factors like summer and single-vehicle crashes decrease the proportion of injury or possible injury crashes. Contrarily, the factors like fixed-object crash, dark-not-lit conditions, crashes within intersection boundaries, and the presence of no-median increases the proportion of injury or possible injury crashes.

Finally, the SHAP dependence plots demonstrated that speeding with clear weather had a higher likelihood of injury crashes, and multi-vehicle crashes at the intersection had more injury

crashes. These outcomes would benefit engineers, practitioners, and policymakers in taking appropriate countermeasures to stop young driver crashes. For instance, strict law enforcement could stop speeding by young drivers. Special attention could be given to improving the safety features of undivided highways, intersections, and medians. Also, emphasis should be given to improving the lighting conditions of roads. During snowfalls or heavy rainfall, seasonal countermeasures like roadside warning signs should also be taken. The most important factors differentiating the severity of crashes involving young drivers are not behavior-related. These findings suggest that, despite a higher propensity of young driver crashes from driving behavior-related issues, reducing the severity of young driver crashes may be possible from countermeasures associated with roadway and crash environment-related factors.

This study has some limitations. First, it analyzed the crash data for four years only. Second, the dataset had missing values that needed to be removed before analysis, reducing the total data used. Third, 'fatal', 'major injury', and 'minor injury' crash severity classes were merged due to low frequency. Future studies on young driver-involved crashes can focus on crash data of longer periods to address the aforementioned limitations.

AUTHOR CONTRIBUTIONS

The authors confirm their contribution to the paper as follows: study conception and design: Mohammad Jalayer, Ahmed Sajid Hasan; data collection: Ahmed Sajid Hasan, Mohammad Jalayer; analysis and interpretation of results: Ahmed Sajid Hasan, Mohammad Jalayer, Subasish Das and Asif Bin Kabir; draft manuscript preparation: Ahmed Sajid Hasan, Mohammad Jalayer, Subasish Das, and Asif Bin Kabir. All authors reviewed the results and approved the final version of the manuscript.

DECLARATION OF COMPETING INTEREST

The authors declare no conflict of interest.

REFERENCES

Aas, K., Jullum, M., & Løland, A. (2021). Explaining individual predictions when features are dependent: More accurate approximations to Shapley values. *Artificial Intelligence*, *298*, 103502. https://doi.org/10.1016/j.artint.2021.103502

Abdel-Aty, M., Pande, A., Lee, C., Gayah, V., & Santos, C. dos. (2007). Crash risk assessment using intelligent transportation systems data and real-time intervention strategies to improve safety on freeways. *Journal of Intelligent Transportation Systems*, *11*(3), 107–120. https://doi.org/10.1080/15472450701410395

Abubaker, H., Ali, A., Shamsuddin, S. M., & Hassan, S. (2020). Exploring permissions in android applications using ensemble-based extra tree feature selection. *Indonesian Journal of Electrical Engineering and Computer Science*, *19*(1), 543–552. https://doi.org/10.11591/ijeecs.v19.i1.pp543-552

Ahmed, M. M., Abdel-Aty, M., & Yu, R. (2012). Bayesian updating approach for real-time safety evaluation with automatic vehicle identification data. *Transportation Research Record: Journal of the Transportation Research Board*, *2280*(1), 60–67. https://doi.org/10.3141/2280-07

Amiri, A. M., Sadri, A., Nadimi, N., & Shams, M. (2020). A comparison between artificial neural network and hybrid intelligent genetic algorithm in predicting the severity of fixed object

crashes among elderly drivers. *Accident Analysis & Prevention, 138*, 105468. https://doi.org/10.1016/j.aap.2020.105468

Andrey, J., Hambly, D., Mills, B., & Afrin, S. (2013). Insights into driver adaptation to inclement weather in Canada. *Journal of Transport Geography*, *28*, 192–203. https://doi.org/10.1016/j.jtrangeo.2012.08.014

Ayoub, J., Yang, X. J., & Zhou, F. (2021). Modeling dispositional and initial learned trust in automated vehicles with predictability and explainability. *Transportation Research Part F: Traffic Psychology and Behaviour*, *77*, 102–116. https://doi.org/10.1016/j.trf.2020.12.015

Baehrens, D., Schroeter, T., Harmeling, S., Kawanabe, M., Hansen, K., & Mueller, K.-R. (2009). *How to explain individual classification decisions*. http://arxiv.org/abs/0912.1128

Baireddy, R., Zhou, H., & Jalayer, M. (2018). Multiple correspondence analysis of pedestrian crashes in rural Illinois. *Transportation Research Record*, *2672*(38), 116-127. https://doi-org /10.1177/0361198118777088

Banerjee, A. K., Arora, N., & Murty, U. S. N. (2008). Classification and Regression Tree (CART) analysis for deriving variable importance of parameters influencing average flexibility of CaMK Kinase family. In *eJ JB Bi io o Electronic Journal of Biology* (Vol. 4, Issue 1). http://expasy.org/tools/

Bates, L. J., Davey, J., Watson, B., King, M. J., & Armstrong, K. (2014). Factors contributing to crashes among young drivers. *Sultan Qaboos University Medical Journal*, *14*(3), 297.

Bingham, C. R., Simons-Morton, B. G., Pradhan, A. K., Li, K., Almani, F., Falk, E. B., Shope, J. T., Buckley, L., Ouimet, M. C., & Albert, P. S. (2016). Peer passenger norms and pressure: Experimental effects on simulated driving among teenage males. *Transportation Research Part F: Traffic Psychology and Behaviour*, *41*, 124–137. https://doi.org/10.1016/j.trf.2016.06.007

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Center of Disease Control. (2017). *Leading causes of death by age group United States 2017*.

Chen, C., Zhang, G., Qian, Z., Tarefder, R. A., & Tian, Z. (2016). Investigating driver injury severity patterns in rollover crashes using support vector machine models. *Accident Analysis & Prevention*, *90*, 128–139. https://doi.org/10.1016/j.aap.2016.02.011

Chen, C., Zhang, G., Tarefder, R., Ma, J., Wei, H., & Guan, H. (2015). A multinomial logit model-Bayesian network hybrid approach for driver injury severity analyses in rear-end crashes. *Accident Analysis & Prevention*, *80*, 76–88. https://doi.org/10.1016/J.AAP.2015.03.036

Chen, C., Zhang, G., Wang, H., Yang, J., Jin, P. J., & Michael Walton, C. (2015). Bayesian network-based formulation and analysis for toll road utilization supported by traffic information provision. *Transportation Research Part C: Emerging Technologies*, *60*, 339–359. https://doi.org/10.1016/J.TRC.2015.09.005

Chen, H. Y., Ivers, R. Q., Martiniuk, A. L. C., Boufous, S., Senserrick, T., Woodward, M., Stevenson, M., Williamson, A., & Norton, R. (2009). Risk and type of crash among young drivers by rurality of residence: Findings from the DRIVE Study. *Accident Analysis & Prevention*, *41*(4), 676–682. https://doi.org/10.1016/j.aap.2009.03.005

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794. https://doi.org/10.1145/2939672.2939785

Christoph Molnar. (2019). *Interpretable Machine Learning - A guide for making black box models explainable*. https://www.iihs.org/topics/fatality-statistics/detail/teenagers

Das, S., Bibeka, A., Sun, X., Zhou, H. T., & Jalayer, M. (2019). Elderly pedestrian fatal crash-related contributing factors: applying empirical Bayes geometric mean method. *Transportation Research Record*, *2673*(8), 254-263. https://doi.org/ 10.1177/0361198119841570

Das, S., Dutta, A., Avelar, R., Dixon, K., Sun, X., & Jalayer, M. (2019). Supervised association rules mining on pedestrian crashes in urban areas: identifying patterns for appropriate countermeasures. *International Journal of Urban Sciences*, *23*(1), 30-48. https://doi.org/10.1080/12265934.2018.1431146

Das, S., Dutta, A., Dey, K., Jalayer, M., & Mudgal, A. (2020). Vehicle involvements in hydroplaning crashes: applying interpretable machine learning. *Transportation Research Interdisciplinary Perspectives*, *6*, 100176. https://doi.org/10.1016/j.trip.2020.100176.

Dahlen, E. R., Martin, R. C., Ragan, K., & Kuhlman, M. M. (2005). Driving anger, sensation seeking, impulsiveness, and boredom proneness in the prediction of unsafe driving. *Accident Analysis & Prevention*, *37*(2), 341–348. https://doi.org/10.1016/j.aap.2004.10.006

Deffenbacher, J. L., Deffenbacher, D. M., Lynch, R. S., & Richards, T. L. (2003). Anger, aggression, and risky behavior: a comparison of high and low anger drivers. *Behaviour Research and Therapy*, *41*(6), 701–718. https://doi.org/10.1016/S0005-7967(02)00046-3

Delen, D., Tomak, L., Topuz, K., & Eryarsoy, E. (2017). Investigating injury severity risk factors in automobile crashes with predictive analytics and sensitivity analysis methods. *Journal of Transport & Health*, *4*, 118–131. https://doi.org/10.1016/j.jth.2017.01.009

Dong, C., Dong, Q., Huang, B., Hu, W., & Nambisan, S. S. (2017). Estimating factors contributing to frequency and severity of large truck–involved crashes. *Journal of Transportation Engineering, Part A: Systems*, *143*(8). https://doi.org/10.1061/JTEPBS.0000060

Erikštrumbelj, E. E., & Kononenko, I. (2010). An efficient explanation of individual classifications using game theory. In *Journal of Machine Learning Research* (Vol. 11). http://www.ailab.si/orange/datasets.psp.

Ferguson, S. A. (2013). *Speeding-related fatal crashes among teen drivers and opportunities for reducing the risks*. www.ghsa.org

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, *29*(5). https://doi.org/10.1214/aos/1013203451

Friedman, J. H. (2002). Stochastic gradient boosting. *Computational Statistics & Data Analysis*, *38*(4), 367–378. https://doi.org/10.1016/S0167-9473(01)00065-2

Gershon, P., Sita, K. R., Zhu, C., Ehsani, J. P., Klauer, S. G., Dingus, T. A., & Simons-Morton, B. G. (2019). Distracted driving, visual inattention, and crash risk among teenage drivers. *American Journal of Preventive Medicine*, *56*(4), 494–500. https://doi.org/10.1016/j.amepre.2018.11.024

Gershon, P., Zhu, C., Klauer, S. G., Dingus, T., & Simons-Morton, B. (2017). Teens' distracted driving behavior: Prevalence and predictors. *Journal of Safety Research*, *63*, 157–161. https://doi.org/10.1016/j.jsr.2017.10.002

Groeger, J. A. (2006). Youthfulness, inexperience, and sleep loss: the problems young drivers face and those they pose for us. *Injury Prevention*, *12*(suppl_1), i19–i24. https://doi.org/10.1136/ip.2006.012070

Guo, M., Yuan, Z., Janson, B., Peng, Y., Yang, Y., & Wang, W. (2021). Older pedestrian traffic crashes severity analysis based on an emerging machine learning XGBoost. *Sustainability, 13*(2), 926. https://doi.org/10.3390/su13020926

Hasan, A. S., Kabir, M. A. bin, Jalayer, M., & Das, S. (2022). Severity modeling of work zone crashes in New Jersey using machine learning models. *Journal of Transportation Safety & Security*, 1–32. https://doi.org/10.1080/19439962.2022.2098442

Hasan, A. S., Kabir, Md. A. bin, & Jalayer, M. (2021). Severity analysis of heavy vehicle crashes using machine learning models: A case study in New Jersey. *International Conference on Transportation and Development 2021*, 285–296. https://doi.org/10.1061/9780784483534.025

Hasan, A. S., Orvin, M. M., Jalayer, M., Heitmann, E., & Weiss, J. (2022). Analysis of distracted driving crashes in New Jersey using mixed logit model. *Journal of Safety Research*, *81*, 166–174. https://doi.org/10.1016/j.jsr.2022.02.008

Hossain, M., & Muromachi, Y. (2013). Understanding crash mechanism on urban expressways using high-resolution traffic data. *Accident Analysis & Prevention*, *57*, 17–29. https://doi.org/10.1016/j.aap.2013.03.024

Huang, Y.-C., Li, S.-J., Chen, M., Lee, T.-S., & Chien, Y.-N. (2021). Machine-learning techniques for feature selection and prediction of mortality in elderly CABG patients. *Healthcare*, *9*(5), 547. https://doi.org/10.3390/healthcare9050547

IIHS. (2021). *Fatality Facts 2019: Teenagers*. https://www.iihs.org/topics/fatality-statistics/detail/teenagers

Iranitalab, A., & Khattak, A. (2017). Comparison of four statistical and machine learning methods for crash severity prediction. *Accident Analysis & Prevention*, *108*, 27–36. https://doi.org/10.1016/j.aap.2017.08.008

Islam, M., Hosseini, P., & Jalayer, M. (2022). An analysis of single-vehicle truck crashes on rural curved segments accounting for unobserved heterogeneity. *Journal of Safety Research*, *80*, 148-159. https://doi.org/10.1016/j.jsr.2021.11.011

Jalayer, M., & Zhou, H. (2016). A multiple correspondence analysis of at-fault motorcycle-involved crashes in Alabama. *Journal of Advanced Transportation*, *50*(8), 2089-2099. https://doi.org/10.1002/atr.1447

Jalayer, M., Pour-Rouholamin, M., & Zhou, H. (2017). *Multiple correspondence approach to identifying contributing factors regarding wrong-way driving crashes* (No. 17-01182).

Jalayer, M., Pour-Rouholamin, M., & Zhou, H. (2018). Wrong-way driving crashes: A multiple correspondence approach to identify contributing factors. *Traffic Injury Prevention*, *19*(1), 35-41. https://doi.org/10.1080/15389588.2017.1347260

Jalayer, M., Pour-Rouholamin, M., Patel, D., Das, S., & Parvardeh, H. (2021). A penalized-likelihood approach to characterizing bridge-related crashes in New Jersey. *Traffic Injury Prevention*, *22*(1), 63-67. https://doi.org/10.1080/15389588.2020.1842379

Jalayer, M., Shabanpour, R., Pour-Rouholamin, M., Golshani, N., & Zhou, H. (2018). Wrong-way driving crashes: A random-parameters ordered probit analysis of injury severity. *Accident Analysis & Prevention*, *117*, 128-135. https://doi.org/10.1016/j.aap.2018.04.019

Keating, D., & Halpernfelsher, B. (2008). Adolescent drivers: A developmental perspective on risk, proficiency, and safety. *American Journal of Preventive Medicine*, *35*(3), S272–S277. https://doi.org/10.1016/j.amepre.2008.06.026

Kidando, E., Kitali, A. E., Kutela, B., Ghorbanzadeh, M., Karaer, A., Koloushani, M., Moses, R., Ozguven, E. E., & Sando, T. (2021). Prediction of vehicle occupants injury at signalized intersections using real-time traffic and signal data. *Accident Analysis & Prevention*, *149*, 105869. https://doi.org/10.1016/j.aap.2020.105869

Lee, C., & Abdel-Aty, M. (2008). Presence of passengers: Does it increase or reduce driver's crash potential? *Accident Analysis & Prevention*, *40*(5), 1703–1712. https://doi.org/10.1016/j.aap.2008.06.006

Lee, D., Warner, J., & Morgan, C. (2019, April 9). Discovering crash severity factors of grade crossing with a machine learning approach. *2019 Joint Rail Conference*. https://doi.org/10.1115/JRC2019-1231

Li, D., Ranjitkar, P., Zhao, Y., Yi, H., & Rashidi, S. (2017). Analyzing pedestrian crash injury severity under different weather conditions. *Traffic Injury Prevention*, *18*(4), 427–430. https://doi.org/10.1080/15389588.2016.1207762

Lin, C., Wu, D., Liu, H., Xia, X., & Bhattarai, N. (2020). Factor identification and prediction for teen driver crash severity using machine learning: A case study. *Applied Sciences*, *10*(5), 1675. https://doi.org/10.3390/app10051675

Liu, C., & Subramanian, R. (2009). *Factors related to fatal single-vehicle run-off-road crashes (No. HS-811 232)*.

Li, X., Lord, D., Zhang, Y., & Xie, Y. (2008). Predicting motor vehicle crashes using Support Vector Machine models. *Accident Analysis & Prevention*, *40*(4), 1611–1618. https://doi.org/10.1016/j.aap.2008.04.010

Li, Z., Liu, P., Wang, W., & Xu, C. (2012). Using support vector machine models for crash injury severity analysis. *Accident Analysis & Prevention*, *45*, 478–486. https://doi.org/10.1016/j.aap.2011.08.016

Lundberg, S. M., Allen, P. G., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *31st Conference on Neural Information Processing Systems (NIPS 2017*, 1–10. https://github.com/slundberg/shap

Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., & Lee, S.-I. (2020). From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence*, *2*(1), 56–67. https://doi.org/10.1038/s42256-019-0138-9

Mafi, S., AbdelRazig, Y., & Doczy, R. (2018). Machine learning methods to analyze injury severity of drivers from different age and gender groups. *Transportation Research Record: Journal of the Transportation Research Board*, *2672*(38), 171–183. https://doi.org/10.1177/0361198118794292

Micucci, Mantecchini, & Sangermano. (2019). Analysis of the relationship between turning signal detection and motorcycle driver's characteristics on urban roads; a case study. *Sensors*, *19*(8), 1802. https://doi.org/10.3390/s19081802

Mokhtarimousavi, S., Anderson, J. C., Azizinamini, A., & Hadi, M. (2019). Improved support vector machine models for work zone crash injury severity prediction and analysis. *Transportation Research Record: Journal of the Transportation Research Board*, *2673*(11), 680–692. https://doi.org/10.1177/0361198119845899

National Highway Traffic Safety Administration. (2019). *Young driver survey (Traffic Tech Technology Transfer Series Report No. DOT HS 812 744)*.

Neyens, D. M., & Boyle, L. N. (2007). The effect of distractions on the crash types of teenage drivers. *Accident Analysis & Prevention*, *39*(1), 206–212. https://doi.org/10.1016/j.aap.2006.07.004

NJDHTS. (2018). *Young Driver Crashes in New Jersey, 2013-2017*. https://www.nj.gov/oag/hts/downloads/NJ-Young-Drivers-2013-2017.pdf

Olden, J. D., Joy, M. K., & Death, R. G. (2004). An accurate comparison of methods for quantifying variable importance in artificial neural networks using simulated data. *Ecological Modelling*, *178*(3–4), 389–397. https://doi.org/10.1016/j.ecolmodel.2004.03.013

Parsa, A. B., Movahedi, A., Taghipour, H., Derrible, S., & Mohammadian, A. (Kouros). (2020). Toward safer highways, application of XGBoost and SHAP for real-time accident detection and feature analysis. *Accident Analysis & Prevention*, *136*, 105405. https://doi.org/10.1016/j.aap.2019.105405

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V. and Vanderplas, J., 2011. Scikit-learn: Machine learning in Python. *the Journal of machine Learning research*, *12*, pp.2825-2830.

Penmetsa, P., & Pulugurtha, S. S. (2018). Modeling crash injury severity by road feature to improve safety. *Traffic Injury Prevention*, *19*(1), 102–109. https://doi.org/10.1080/15389588.2017.1335396

Racioppi, F., Eriksson, L., Tingvall, C., & Villaveces, A. (2004). *Preventing road traffic injury : a public health perspective for Europe*. World Health Organization Regional Office for Europe.

Rahman, M. A., Hossain, M. M., Mitran, E., & Sun, X. (2021). Understanding the contributing factors to young driver crashes: A comparison of crash profiles of three age groups. *Transportation Engineering*, *5*, 100076. https://doi.org/10.1016/j.treng.2021.100076

Ratner, B. (2009). The correlation coefficient: Its values range between +1/−1, or do they? *Journal of Targeting, Measurement and Analysis for Marketing*, *17*(2), 139–142. https://doi.org/10.1057/jt.2009.5

Rhodes, N., & Pivik, K. (2011). Age and gender differences in risky driving: The roles of positive affect and risk perception. *Accident Analysis & Prevention*, *43*(3), 923–931. https://doi.org/10.1016/j.aap.2010.11.015

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why Should I Trust You?" *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1135–1144. https://doi.org/10.1145/2939672.2939778

Rolison, J. J., & Moutari, S. (2020). Combinations of factors contribute to young driver crashes. *Journal of Safety Research*, *73*, 171-177. https://doi.org/10.1016/j.jsr.2020.02.017

Roque, C., Jalayer, M., & Hasan, A. S. (2021). Investigation of injury severities in single-vehicle crashes in North Carolina using mixed logit models. *Journal of Safety Research*, *77*, 161–169. https://doi.org/10.1016/J.JSR.2021.02.013

Sameen, M., & Pradhan, B. (2017). Severity prediction of traffic accidents with recurrent neural networks. *Applied Sciences*, *7*(6), 476. https://doi.org/10.3390/app7060476

Simons-Morton, B., Lerner, N., & Singer, J. (2005). The observed effects of teenage passengers on the risky driving behavior of teenage drivers. *Accident Analysis & Prevention*, *37*(6), 973–982. https://doi.org/10.1016/j.aap.2005.04.014

Strobl, C., Boulesteix, A.-L., Kneib, T., Augustin, T., & Zeileis, A. (2008). Conditional variable importance for random forests. *BMC Bioinformatics*, *9*(1), 307. https://doi.org/10.1186/1471-2105-9-307

Taylor, S. G., Russo, B. J., & James, E. (2018). A comparative analysis of factors affecting the frequency and severity of freight-involved and non-freight crashes on a major freight corridor freeway. *Transportation Research Record: Journal of the Transportation Research Board*, *2672*(34), 49–62. https://doi.org/10.1177/0361198118776815

United States Department of Transportation. (2022). *Intersection safety*. https://highways.dot.gov/research/research-programs/safety/inte rsection-safety

Wang, X., & Kim, S. H. (2019). Prediction and factor identification for crash severity: Comparison of discrete choice and tree-based models. *Transportation Research Record: Journal of the Transportation Research Board*, *2673*(9), 640–653. https://doi.org/10.1177/0361198119844456

Wen, X., Xie, Y., Wu, L., & Jiang, L. (2021). Quantifying and comparing the effects of key risk factors on various types of roadway segment crashes with LightGBM and SHAP. *Accident Analysis & Prevention*, *159*, 106261. https://doi.org/10.1016/j.aap.2021.106261

Williams, A. F. (2003). Teenage drivers: patterns of risk. *Journal of Safety Research*, *34*(1), 5–15. https://doi.org/10.1016/S0022-4375(02)00075-0

Williamson, M., Jalayer, M., Zhou, H., & Pour Rouholamin, M. (2015). A sensitivity analysis of crash modification factors of access management techniques in highway safety manual. In *Access Management Theories and Practices* (pp. 76-88). https://doi.org/10.1061/9780784413869.008

Wundersitz, L. N. (2012). *An analysis of young drivers involved in crashes using in-depth crash investigation data*. Centre for Automotive Safety Research, University of Adelaide.

Wu, Q., Chen, F., Zhang, G., Liu, X. C., Wang, H., & Bogus, S. M. (2014). Mixed logit model-based driver injury severity investigations in single- and multi-vehicle crashes on rural two-lane highways. *Accident Analysis & Prevention*, *72*, 105–115. https://doi.org/10.1016/j.aap.2014.06.014

Wu, T.-E., Chen, H.-A., Jhou, M.-J., Chen, Y.-N., Chang, T.-J., & Lu, C.-J. (2020). Evaluating the effect of topical atropine use for myopia control on intraocular pressure by using machine learning. *Journal of Clinical Medicine*, *10*(1), 111. https://doi.org/10.3390/jcm10010111

Yang, C., Chen, M., & Yuan, Q. (2021). The application of XGBoost and SHAP to examining the factors in freight truck-related crashes: An exploratory analysis. *Accident Analysis & Prevention, 158*, 106153. https://doi.org/10.1016/j.aap.2021.106153

Yu, R., & Abdel-Aty, M. (2013). Utilizing support vector machine in real-time crash risk evaluation. *Accident Analysis & Prevention*, *51*, 252–259. https://doi.org/10.1016/j.aap.2012.11.027

Zeng, Q., Huang, H., Pei, X., & Wong, S. C. (2016). Modeling nonlinear relationship between crash frequency by severity and contributing factors by neural networks. *Analytic Methods in Accident Research*, *10*, 12–25. https://doi.org/10.1016/j.amar.2016.03.002

Zhang, J., Li, Z., Pu, Z., & Xu, C. (2018). Comparing prediction performance for crash injury severity among various machine learning and statistical methods. *IEEE Access*, *6*, 60079–60087. https://doi.org/10.1109/ACCESS.2018.2874979

Zhang, L., & Zhan, C. (2017). Machine learning in rock facies classification: an application of XGBoost. *International Geophysical Conference, Qingdao, China, 17-20 April 2017*, 1371–1374. https://doi.org/10.1190/IGC2017-351

Zhang, Y., & Xie, Y. (2007). Forecasting of short-term freeway volume with v-support vector machines. *Transportation Research Record: Journal of the Transportation Research Board*, *2024*(1), 92–99. https://doi.org/10.3141/2024-11.

Zhou, B., Wang, X., Zhang, S., Li, Z., Sun, S., Shu, K., & Sun, Q. (2020). Comparing factors affecting injury severity of passenger car and truck drivers. *IEEE Access, 8*, 153849-153861. https://doi.org/10.1109/ACCESS.2020.3018183

Application of Machine Learning Models and SHAP to Examine Crashes Involving Young Drivers in New Jersey

Ahmed Sajid Hasan
Graduate Research Fellow, Department of Civil and Environmental Engineering, CREATES
Rowan University, Glassboro, New Jersey, 08028
Email: hasana26@rowan.edu

Mohammad Jalayer, Ph.D. (Corresponding Author)
Associate Professor, Department of Civil and Environmental Engineering, CREATES
Rowan University, Glassboro, New Jersey, 08028
Email: jalayer@rowan.edu

Subasish Das, Ph.D.
Assistant Professor, Department of Civil Engineering
Ingram School of Engineering, 327 W Woods St, San Marcos, TX 78666
Email: subasish@txstate.edu

Md. Asif Bin Kabir
Assistant Professor, Department of Civil Engineering
United International University, United City, Madani Ave, Dhaka, 1212
Email: asif@ce.uiu.ac.bd

Declaration of Competing Interest
The authors declare no conflict of interest.