COMPARATIVE EVALUATION OF DISTINCT TREE-BASED MODELS FOR PEDESTRIAN-RELATED CRASH INJURY SEVERITY ANALYSIS

Edward Clay, College of Engineering, California State Polytechnic University, Pomona, CA 91768, 909-749-7358, erclay@cpp.edu
Wen Cheng, College of Engineering, California State Polytechnic University, Pomona, CA 91768 909-869-2957, wcheng@cpp.edu
Yasser Salem, College of Engineering, California State Polytechnic University, Pomona, CA 91768 909-869-4312, ysalem@cpp.edu
Long N. Truong, College of Science, California State Polytechnic University, Pomona, CA 91768 626-554-6182, intruong@cpp.edu
Sheng Tan, College of Engineering, California State Polytechnic University, Pomona, CA 91768 626-319-1874, shengtan@cpp.edu

ABSTRACT

Pedestrians are among the most unsafe and vulnerable road users in terms of traffic crashes. Consequently, many methods including various tree-based models are used frequently the explore variables surrounding pedestrian safety. Unfortunately, little research is dedicated to evaluating tree-based models' comparative performances. To this end, five years of pedestrian-related crash data were collected to compare the predictive capabilities of injury severity between four distinct tree-based models: Bagging, Boosting, Random Forest, and the relatively new Rotation Forest; using alternative evaluation criteria, including in-sample and out-of-sample performance evaluations. The results indicate each tree-based model possesses a unique set of benefits and drawbacks.

Keywords: Pedestrian safety, Injury severity, Tree-Based Model, In-Sample, Out-of-Sample

INTRODUCTION

Pedestrians are considered to be the most unsafe and vulnerable road users from the perspective of traffic crashes [29]. In the United States, 6590 pedestrian fatalities and around 70,000 injuries were reported in 2019 [10]. Furthermore, the number of pedestrians deaths in 2019 was projected to reach the highest level since 1988. These statistics demonstrate the apparent need to further explore the effects of numerous different factors on pedestrian crashes to properly implement the appropriate policies and strategies to improve pedestrian safety. Therefore, a plethora of literature has been published focusing on the various factors that affect pedestrian injury severities such as roadway geometry [23] [34], pedestrian behavior [6], as well as social and demographic features [17] [37] to name a few.

Numerous statistical models have been utilized to evaluate the significant factors of pedestrian-related crash severities due to the wide selection of models to better analyze different types of data, the ability to overcome missing data, as well as the numerous model criteria available [1] [24] [39]. Different models will vary in terms of scale through the use of nominal and ordinal scales [4] [8], logit and probit regression [9] [25], binary and multinomial categorical predictions [3] [35], univariate and multivariate numerical predictions [12] [15] and immeasurably more. Despite the benefits that statistical models offer, it often lacks predictive capabilities, limitations in the amount of data that can be processed at any given time, and its inability to differentiate between relevant and irrelevant data [11] [36]. To combat these shortcomings, Machine Learning (ML) has been employed more often in recent years.

As previously stated, ML can handle large sets of data, sort through data more efficiently, and be more accurate than its statistical modeling counterpart. In addition, one key difference between ML and statistical modeling is the automation that ML offers due to its ability to learn as it goes on [40]. Many ML algorithms have been used in safety literature to evaluate the contributing factors in crash injury severities. For example, K-Nearest Neighbor (KNN) has shown promise in real-time detection of a vehicle fall event using a smartphone [18]. Another ML algorithm is Naïve Bayes. When paired with other algorithms, Naïve Bayes has been proven to increase the overall predictive accuracy [2], as shown in a study in 2019 that utilized several ML algorithms along with Naïve Bayes and reported an overall accuracy between 92-98% [16]. Furthermore, principal component analysis (PCA) and support vector machine (SVM) algorithms are commonly used together to provide more accurate results than anyone alone can produce [21] [26]. These previously mentioned ML algorithms have exhibited highly accurate predictive results. However, these ML algorithms do not demonstrate the process they undergo to achieve said results. For this reason, the use of decision trees is preferred.

Decision Trees have been used extensively as not only does this algorithm aid in understanding the process the algorithm utilizes within its process. Decision Trees produce results comparable to SVM's predictive capabilities and are better than statistical modeling [13]. Several distinct tree-based models are commonly used in the field of traffic safety. The first is known as Random Forest (RF) and has been employed to be compared to various other ML algorithms and has been shown to exhibit better accuracy [41]. Second, boosting (BOO), both in the form of Adaptive Boosting and Gradient Boosting Decision Tree, has shown promise against RF, SVM, and Multi-Layer Perceptron [28] [38]. The Third is known as Bootstrap Aggregating or Bagging (BAG) and has been used in systems for real-time automated crash notifications and has shown promise in these systems [20]. These decision tree algorithms all work great with larger sets of data. However, BAG shortcomings become apparent when being used to analyzed smaller sets of data. In 2006, a new algorithm, Rotation Forest, was developed to work more efficiently based on feature extraction [27].

Rotation Forest is relatively new; therefore, its applications have yet to be fully understood. In 2007, Rotation Forest (ROTF) was tested against the previously mentioned decision trees with a small set of data and had revealed itself to produce the highest prediction accuracy [19]. Since then, ROTF has been employed in diverse fields of study, including the medical field in cancer research [22] and the film industry in developing cameras capable of capturing more details and higher resolutions [40]. Despite the benefits ROTF offers, its use in the transportation safety field is minimal to the authors' best knowledge. In 2017, a study was conducted to compare the prediction accuracies of ROTF with several other ML algorithms and concluded that ROTF is the most accurate [33].

The primary goal of this study is to compare the performances of RF, BOO, BAG, and the relatively new ROFT. The data utilized were collected from the Highway Safety Information System (HSIS) and consist of crash data for pedestrians across five years (2010-2014) in California. The performance of each tree-based model is assessed by evaluating the sensitivity, specificity, positive-predictive-value, and negative-predictive-values for both In-Sample (IS) and Out-Of-Sample (OOS) forecasts. It is anticipated that the results of this study will provide crucial insights into the performance of distinct tree-based models.

DATA DESCRIPTION

The data used for this study were obtained from HSIS, which collected the data in different raw files from California TASAS (Traffic Accident Surveillance and Analysis System). Five years of available pedestrian crash data (2010 to 2014) were used to evaluate various tree-based models for crash severity analysis. In this study, the pedestrian-related crash data were obtained from three different files linked with road, vehicle, and crash characteristics. The data collected from these files have crash number along with other factors which include geometric (number of lanes, median type, etc.), traffic (Average

Annual Daily Traffic, Design Speed, average lane length, etc.), and driver attributes (race, sex and alcohol consumption), and so on. A total of 2869 pedestrian crashes and 53 variables were selected, of which 11 are numerical variables, and 42 are categorical variables. The numerical variables and categorical variables are listed separately under each characteristic, as shown in **Table 1**.

TABLE 1. DESCRIPTIVE STATISTICS OF ALL VARIABLES USED IN THE DATASET

Numerical Variables					
Variables	**Description**	**Minimum**	**Maximum**	**Mean**	**S.D.**
		Roadway Level			
aadt	Annual average daily traffic	180.00	365000.00	56140.00	66918.94
desg_spd	Design Speed	25.00	70.00	55.20	12.19
lanewid	Average Lane Width	8.00	30.00	12.02	1.27
lshldwid	Left Shoulder Width Roadway 1 of separate Highway	0.00	80.00	3.29	4.10
no_lanes	Total Number of Lanes	2.00	16.00	4.66	2.37
no_lane1	Number of Lanes Roadway 1 of separate Highway	1.00	12.00	2.00	1.20
no_lane2	Number of Lanes Roadway 2 of separate Highway	0.00	8.00	2.33	1.20
pav_wdl	Left Paved Shoulder Width Roadway 1 of separate Highway	0.00	22.00	2.80	3.70
rshldwid	Right Shoulder Width Roadway 1 of separate Highway	0.00	20.00	7.05	3.53
surf_wid	Traveled Way Width Roadway 1 of separate Highway	11.00	144.00	31.78	12.99
		Crash Level			
numvehs	Total number of vehicles involved in the crash	1.00	8.00	2.15	0.54
		Vehicle Level			
drv_age	The age of the driver of the vehicle involved in the crash	5.00	95.00	42.77	17.27
Categorical Variables					
Variables	**Description**	**Details of categories (frequency, percentage)**			

		Roadway Level
access	Access Control	1- No (2073, 72.26%); 2- Partial (117, 4.08%); 3- Full (679, 23.66%).
curb1	Curb and Landscape	C- Curb M- Median T- Trees S- Shrub 1-C. M. (313, 10.91%); 2-C. M. W/T. (198, 6.90%); 3-C. M. W/S. (115, 4.01%); 4-Raised Traffic Bar (5, 0.17%); 5-M. W/T. (7, 0.24%); 6-M. W/S. (143, 4.98%); 7-N/Curbs/Shrubs (2088, 72.78%).
divided	Divided Highway	1-Not (775, 27.01%); 2-Divided (2094, 72.99%).
feat_rg	Right Road Border Special Feature	L-Lane M-Median A- Auxiliary 1-L. Transitions (32, 1.12%); 2-Passing or Truck Climbing L. (3, 0.10%); 3-A. L. (Included in No. Lanes Field) (34, 1.19%); 4-A. Lanes (Included in No. Lanes Field) (2, 0.07%); 5-Toll Plaza and Approaches (1, 0.03); 6-M. L. Is HOV L. (92, 3.21%); 7-M. Lanes Are HOV Lanes (11, 0.38%); 8-No Special Feature (2694, 93.90%).
feat_lf	Left Road Border Special Feature	L-Lane M-Median A- Auxiliary 1-L. Transitions (39, 1.36%); 2-Passing or Truck Climbing L. (3, 0.10%); 3-A. L. (Included in No. Lanes Field) (43, 1.49%); 4-A. Lanes (Included in No. Lanes Field) (5, 0.17%); 5-Toll Plaza and Approaches (1, 0.03%); 6-M. L. Is HOV L. (93, 3.24%); 7-M. Lanes Are HOV Lanes (16, 0.56%); 8-No Special Feature (2669, 93.03%).
med_var	Median Variance	1-Median 100'+, No Variance (25, 0.87%); 2-Element Median Width (306, 10.67%); 3-Median Constant Width - No Variance (2538, 88.46%).
rururb	Rural/Urban	1-Rural (576, 20.08%);2-Urban (1964, 68.46%); 3-Invalid (329, 11.47%).
surf_typ	Surface Type Roadway 1 of separate Highway	1-PCC, Bridge Deck (51, 1.78%); 2-PCC, Concrete (506, 17.64%); 3-Unp- Undetermined (4, 0.14%); 4-AC, Base & Surface 7" (2184, 76.12%); 5-AC, Base & Surface < 7"(112, 3.90%); 6-AC, Oiled Earth-Gravel (12, 0.42%).
medbarty	Median Barrier Type	B- Barrier C- Concrete G-Guardrail R- Roadway M-Median 1-Cable B. (4, 0.14%); 2-Metal Beam B. (21, 0.70%);3-Metal Beam B. Glare Screen (36, 1.25%) ; 4-C. B. (265, 9.24%); 5-C. B. Glare Screen (107, 3.73%); 6-Bridge B. Railing (15, 0.52%); 7-Chain Link Fence (7, 0.24%); 8-G. in M. , Both R. (18, 0.63%); 9-G. in M. , Left R. ; 10-G. in M. , Right R. (9, 0.31%); 11-Thrie Beam B. (7, 0.24%); 12-C. B. , Both Ways, Both Shoulders (86, 3.00%); 13-C. B. ,Shoulder of Left R. (23, 0.80%); 14-C. B. ,Shoulder of Right R.(3, 0.10%) ;15-No B. (2257, 78.67%).
med_type	Median Type	U- Undivided D- Divided L-Lane S- Separated M- Median 1-U, Not S.(2, 0.07%); 2-U, S.(773, 26.94%); 3-D, Two-Way L-Turn L. (338, 13.52%) ; 4-D, Continuous L-Turn L. (416, 14.50%); 5-D, Paved M. (692, 24.20%); 6-D, Unpaved M. (526, 18.22%); 7-D, S. Grades (25, 0.87%); 8-D, S. Grades With Retaining Wall (1, 0.03%); 9-D, Sawtooth (Unpaved) (6, 0.21%); 10-D, S. Structure (11, 0.38%); 11-D, Railroad(10, 0.35%); 12-D, Occasional L. (3, 0.10%); 13-D, Railroad, Bus L. (2, 0.07); 14-D, Peak-Hour L.(S) (2, 0.07); 15-D, Other (12, 0.42%).

terrain	Terrain	1-Flat (1781, 62.08%); 2-Mountainous (242, 8.43%);3-Rolling (846, 29.49%).
func_cls	Functional Class	0-Not(10, 0.14%); 1- Rural Principal Arterial With Extension Into Urban Area Principal Arterial (2310, 33.40%); 2-Rural Principal Arterial With Extension Into Urban Area Minor Arterial(1325, 19.16%); 3-Principal Arterial Lying Entirely In Urban or Rural Area(2135, 30.87%); 4-Minor Arterial(990, 14.31%); 5-Major Collector(133, 1.92%); 6-Minor Collector(12, 0.17%)
		Crash Level
acctype	Type-of-Collision	1-Head-On(61, 2.13%);2-Sideswipe(296, 10.32%); 3-Rear-End(170, 5.93%); 4-Broadside(546, 19.03%); 5-Hit Object(19, 0.66%); 6-Overturned(4, 0.14%); 7-Auto-Pedestrian (1352, 47.12%); 8-Other(421, 14.67%).
cause1	Primary Collision Factor (DOT)	1-Influenced of Alcohol (89, 3.10%);2-Following Closely (21, 0.73%); 3- Failure to Yield (409, 14.26%);4-Improper Turn (338, 11.78%); 5-Speeding (263, 9.17%); 6-Other Violations (1612, 56.19%);7-Other Improper Driving (22, 0.77%); 8-Other Than Driving (52, 1.81%); 9-Unknown (63, 2.20%).
hit_run	Hit and Run	1- Not Hit and Run (2659, 92.68%); 2- Hit and Run (210, 7.32%)
hour	Crash hours	1- night-time (504, 17.57%); 2-Daytime (2365, 82.43%).
hwy_grp	Highway Group	1-Divided Highway (2094, 72.99%); 2-Undivided Highway (775, 27.01%).
inter	Intersection Crash	1-Not Intersection (2491, 86.82%); 2-Intersection (378, 13.18%).
light	Light Condition	1-Daylight (1626, 56.67%); 2-Dusk - Dawn (86, 3.00%);3-Dark - Street Lights (614, 21.40%); 4-Dark - No Street Lights (527, 18.39%); 5-Dark - Street Lights Not Functioning (16, 0.56%).
ped_actn	Pedestrian Action	Cing- Crossing Cwalk- Crosswalk I- Intersection 1-No Pedestrian (1486, 51.80%); 2- Cing. in Cwalk. at Intersection (265, 9.24%); 3- Cing. Not in Cwalk. (424, 14.78%); 4-In Road, Including Shoulder (661, 23.04%); 6-Not in Road (33, 1.15%).
pop_grp	Population Group	1- Less Than 2500 (2, 0.07%); 2-2500 To 10000 (125, 4.36%); 3-10000 To 25000 (199, 6.94%); 4- 25000 To 50000 (338, 11.78%); 5- 50000 To 100000 (446, 15.55%); 6- 100000 To 250000 (459, 16.00%); 7- Greater Than 250000 (403,14.05%); 8-Unincorporated (Rural) (897,31.27%)
rdsurf	Road Surface	1-Dry (2686, 93.62%); 2-Wet (166, 5.79%); 3-Snowy, lcy(17, 0.59%)
rd_def1	Roadway Condition	1- Holes (10,0.35%);2-Loose Material (8, 0.28%);3-Obstruction (18, 0.63%); 4-Construction (52, 1.81%); 5-Other (18, 0.63%); 6-No Unusual (2763, 96.31%)
rodwycls	Roadway Classification	U- Urban R- Rural M- Multilane N.F-Non-Freeways D-Divided 1-U. F. (605, 21.09%); 2-U. F. < 4 Lanes (1, 0.03%);3-U. Two Lane Roads (327, 11.40%); 4-U. M. D. N. F. (1210, 42.17%); 5-U. M. Undi. N. F. (146, 5.09%);6-R. F. (72, 2.51%); 7-R. Two L. Roads (401, 13.98%);8-R. M. D. N. F. (78, 2.72%); 9-R. M. Undi. N. F. (25, 0.87%);10-Others (4, 0.14%)

severity	Collision Severity	1-Fatal (657, 22.90%); 2-Severe Injury (401, 13.98%); 3-Other Visible (895, 31.20%); 4-Complaint of Pain (623, 21.71%); 5-PDO (293, 10.21%)
veh_invl	Involved in The Accident	1-Pedestrian (1476, 51.45%); 2-Bicycle (1393, 48.55%)
weather1	Weather	1-Clear (2401,83.69%); 2-Cloudy (359, 12.51%); 3-Raining (76, 2.65%); 4-Snowing (8, 0.28%); 5-Fog (23, 0.80%); 6-Other (1,0.03%); 7-Wind (1,0.03%)
weekday	Day of Week	1- (405, 14.12%); 2- (380, 13.25%); 3- (380, 13.25%); 4- (378, 13.18%) 5- (409, 14.26%); 6- (487, 16.97%); 7- (430, 14.99%)
		Vehicle Level
celphone	Usage of cellphone in the vehicle	1- Handheld (8, 0.28%); 2- Hands Free (31,1.08%); 3- Not (2789, 97.21%); 4- In Use (1, 0.03%); 5-No Cell Phone/Unknown (40, 1.39%)
contrib1	First Associated Factor	1-Vehicle Code Violation (262, 9.13%); 2-Vision Obscurement (61, 2.13%); 3-Inattention (88, 3.07%); 4-Stop and Go Traffic (22, 0.77%); 5-Enter/Leave Ramp (17, 0.59%); 6-Previous Collision (37, 1.29%); 7-Unfamiliar With Road (1, 0.03%); 8-Defect Vehicle Equipment (9, 0.31%); 9-Uninvolved Vehicle (6, 0.21%); 10-Other ((16, 0.56%); 11-None Apparent (2347, 81.81%); 12-Runaway Vehicle (3, 0.10%)
drv_race	Driver Race	1-Asian (193, 6.73%); 2-Black (144, 5.02%); 3-White (740, 25.79%); 4-Hispanic (191, 6.66%); 5-Other (1601, 55.80%)
drv_sex	Driver Sex	1-Male (1994, 69.50%); 2-Female (875, 30.50%).
insur	Insurance	1-No (87, 3.03%); 2-Yes (2003, 69.82%); 3-Not Applicable (778, 27.12%);4-Used (1, 0.03%)
loc_typ1	First Collision Location	1-Beyond Median- Driver's Left (130, 4.53%); 2-Beyond Shoulder - Driver's Left (141, 4.91%); 3-Left Shoulder Area (27, 0.94%); 4-Left Lane (419, 14.60%); 5-Interior Lanes (274, 9.55%); 6-Right Lane (1339, 46.67%); 7-Right Shoulder Area (233, 8.12%); 8-Beyond Shoulder - Driver's Right (139, 4.84%); 9-Other (152, 5.30%); 10-Not Stated (15, 0.52%)
miscact1	Movement Preceding Accident	1-Crossing In Crosswalk at Intersection (1, 0.03%);2-Crossing - Not In Crosswalk (1, 0.03%); 3-In Roadway - Include Shoulder (3, 0.10%); 4-Stopped (67, 2.34%); 5-Proceeding Straight (1944, 67.76%); 6-Ran Off Road (22, 0.77%); 7-Making Right Turn (193, 6.73%); 8-Making Left Turn (166, 5.79%); 9-Making U Turn (3, 0.10%); 10-Backing (14, 0.49%); 11-Slowing, Stopping (57, 1.99%); 12-Passing Other Vehicle (31, 1.08%); 13-Changing Lanes (51, 1.78%); 14-Parking Maneuver (3, 0.10%); 15-Entering Traffic From Shoulder, Median, Parking Strip Or Private Drive (109, 3.80%); 16-Other Unsafe Turning (61, 2.13%); 17-Crossed Into Opposing Lane (13, 0.45%); 18-Parked (8, 0.29%); 19-Merging (6, 0.21%); 20-Traveling Wrong Way (58, 2.02%) ; 21-Other (58, 2.02%)

object1	First Object Struck	1-Side of Bridge Railing (1, 0.03%); 2-Traffic Sign or Sign Post (3, 0.10%) 3-Guardrail (1, 0.03%); 4-Barrier (2, 0.07); 5-Dike or Curb (Including Curb Of Median & A.C. Berm) (8, 0.28%); 6-Guidepost, Culvert or Mile Post Marker (1, 0.03%); 7-Cut Slope or Embankment, Struck From Below (7, 0.24%); 8-Drainage Ditch (With or Without Water) (1, 0.03%); 9-Temporary Barricades, Cones or Signs (1, 0.03%); 10-Other Object On Road (7, 0.24%); 11-Other Object Off Road (3, 0.10%);12-Overturned (2, 0.07%); 13-No Object Involved (28, 0.96%); 14-Vehicle 1 to 9(2804, 97.73%)
spec_inf	Special Information	1-Cell Phone in Use (40, 1.39%); 2- Not in Use (2789, 97.21%); 3-None/Unknown (40, 1.39%)
sobriety	Sobriety of the driver of this vehicle	1-Not(2624, 91.46%); 2-Drinking, Under Influence (97, 3.38%); 3-Drinking, Not Under Influence (43, 1.50%); 4-Drinking, Impairment Unknown (20, 0.70%); 5-Impairment Unknown (74, 2.58%); 6-Not Applicable (11, 0.38%)
vehtype	Vehicle Type	1-Truck with 2 Trailers (3, 0.10%); 2-Truck with Tank Trailer (2, 0.07%); 3-Passenger Car (1564, 54.51%); 4-Passenger Car With Trailer (4, 0.14%); 5-Motorcycle (33, 1.15%); 6-Pickup or Panel Truck (335, 11.50%); 7-Pickup or Panel Truck With Trailer (16, 0.56%); 8-Truck or Truck Tractor (32, 1.12%); 9-Truck With 1 Trailer (68, 2.37%); 10-School Bus (2, 0.07%); 11-Other Bus (21, 0.7%); 12-Emergency Vehicle (13, 0.45%); 13-Bicycle (765, 26.66%); 14-Other Motor Vehicle (6, 0.21%); 15-Pedestrian (5, 0.17%)

Note: S.D. represents standard deviation.

It is important to note that while the categories for severity presented in **Table 1** are part of the initial dataset, these categories were reorganized to improve overall model performance. A severity of 1 and 2 are organized into a single category as more severe, “1” whereas severity levels 3, 4, and PDO are classified as less severe, “0”. These new severity levels will be used to evaluate each models’ performance. A correlation analysis was performed to determine which variables were closely correlated, as shown in **Figure 1**.

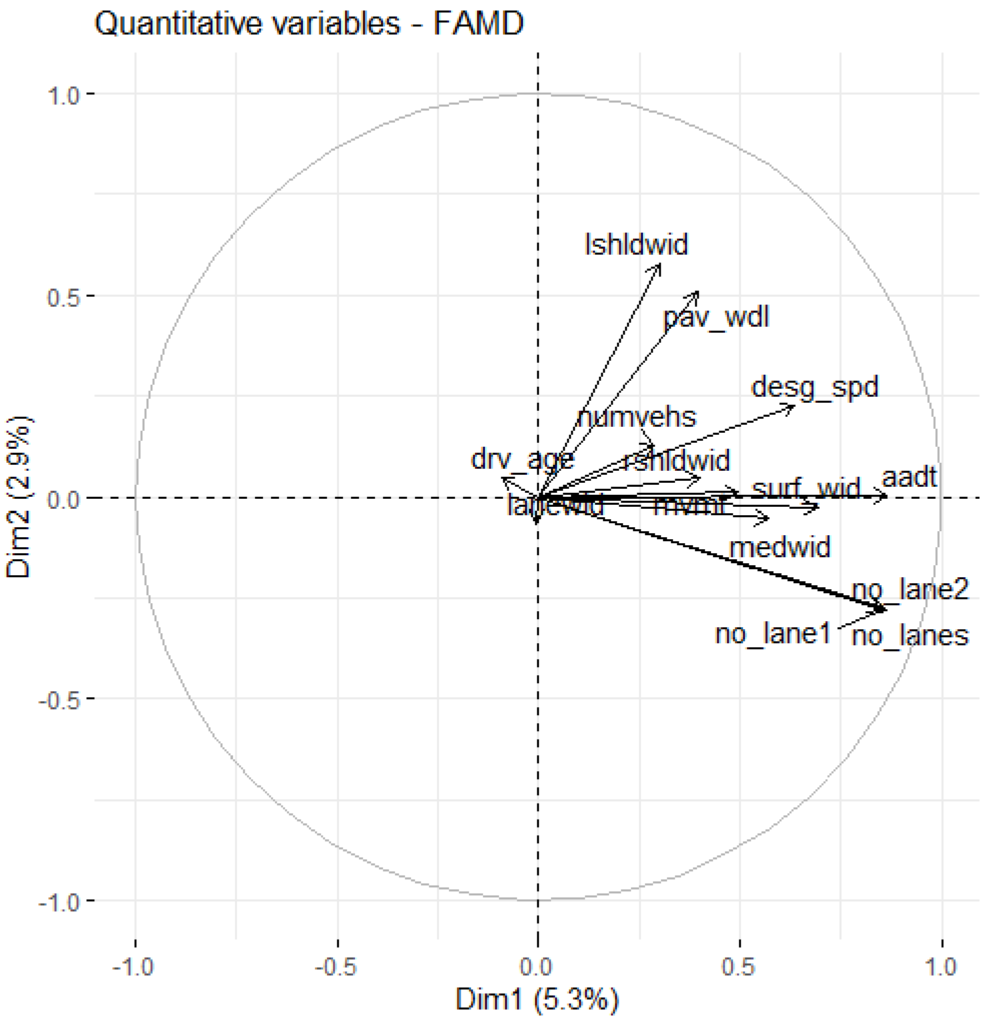


FIGURE 1. CORRELATION PLOT OF NUMERICAL VARIABLES IN THE CRASH DATA USING PRINCIPAL COMPONENT ANALYSIS

Figure 1 illustrates the plot for displaying the correlation between numerical variables by using PCA. Only one variable, driver age (drv_age), is negatively correlated, which is positioned opposite to other variables. All other variables (except drive age) are positively correlated. The variables that are away from the origin show a high significance, and the grouped variables indicate a strong correlation with one other.

Upon closer inspection, it appears that the total number of lanes(no_lanes), number of numbers RD1 (no_lane1), and number of lanes RD 2 (no_lane2) are closely compacted to each other and are away from the origin. On the other hand, other variables such as the number of vehicles (numvehs), left shoulder width RD1(ishldwid), and traveled way width (surf_wid) are group together and are near to the origin.

METHODOLOGY

This study utilizes machine learning to generate several tree-based models to analyze accident severity involving pedestrian casualties. Following the generation of these tree-based models, several criteria were implemented to evaluate the efficacy of each model including sensitivity, specificity, positive-predictive-value, and negative-predictive-values. In an effort to reduce bias, these values were generated a total of ten times using both In-Sample and Out-Of-Sample forecasting methods.

Tree-Based Methods

Most tree-based models begin with a single node that splits off into two branches; the directions of each branch are determined by a single variable's criteria that the model itself determines. This process repeats itself until a terminal node is reached, representing the model's decision itself. As a result, tree-based models provide accurate and easy to understand results. However, as data become more complex with more variables, decision trees become biased and inaccurate. A simple tree for the data used is shown in **Figure 2**, where each branch represents true and false statements. At each node, the resulting branch that goes left is false, while the branch that goes right is true.

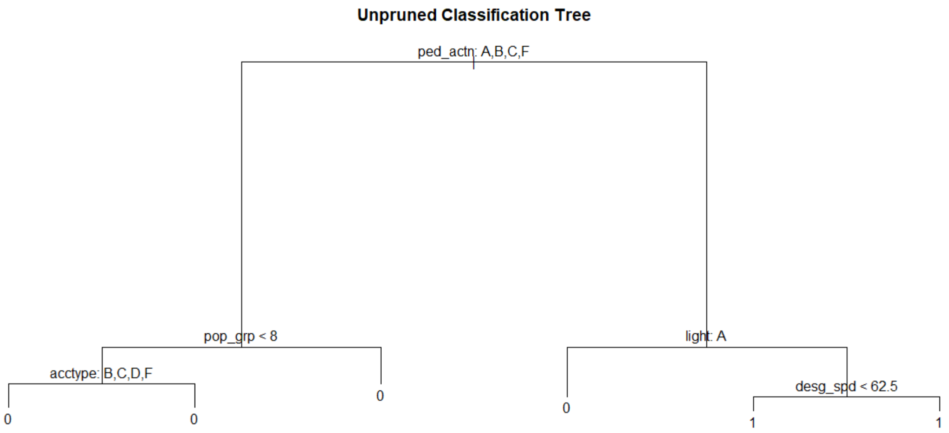


FIGURE 2. ILLUSTRATION OF AN UNPRUNED CLASSIFICATION TREE FOR THE CRASH DATASET

Specific pruning methods are used to prevent decision trees from becoming biased by reducing the level of variance. Pruning is the process that reduces the size of decision trees by removing non-significant variables to reduce the complexity of the tree. Boosting, also known as Gradient Boosting, is a method of pruning that combines a large number of decision trees by slowly learning to narrow down significant variables. A process of cross-validation determines the number of trees produced (B) to ensure that overfitting does not occur. In this experiment, B was chosen to be 5000. The shrinkage factor (λ) determines the rate at which the model narrows down significant variables, thus directly controlling the rate at which boosting learns. The number of branches that expel out of each node is set to two for this experiment, and the remaining tree-based models also utilized two branches at each node. Equation 1 describes the learning model that boosting uses to make predictions. Note that b represents the ordinal number of the boosted tree generated [14].

$$\hat{f}(x) = \sum_{b=1}^{B} \lambda \hat{f}^{b}(x) \tag{1}$$

Bagging, also known as bootstrap aggregation, is similar to boosting in which it also combines a large number of decision trees. Bagging reduces the variance by averaging a set of observations by generating a new prediction model and then averaging the resulting prediction. The number of bags generated in the model divides the data into subsets, after which the model runs; the number of bags generated in this experiment is set to 13. Like Boosting, the number of trees must also be specified; 25

trees were generated in the investigation. The equation representing the learning model that bagging uses to make predictions is shown in Equation 2 [14].

$$\hat{f}_{bag}(x) = \frac{1}{B}\sum_{b=1}^{B} \lambda \hat{f}^{*b}(x) \tag{2}$$

Random Forest is another tree-based model that is similar to bagging, with one notable exception. Random Forest decorates the trees generated by only considering a set number of predictors, or variables, at each tree. Therefore, instead of developing trees that look similar to one another, it generates trees by being forced to use a random set of predictors thus, decorrelating each tree and making its predictions more accurate [14]. Not including a set number of "bags" or trees that the model must generate uses a variety of predictors when creating each tree. The number of predictors used at each tree split can be calculated using Equation 3:

$$m = \sqrt{p} \tag{3}$$

Where *m* is the number of predictors within the dataset and *p* is the total number of predictors utilized within each tree generated by Random Forest. The new model then generates a *random* decision based on what each tree suggests.

Rotation forest is a relatively new ensemble method used to formulate accurate, efficient, and diverse classifiers compared to various existing algorithms [31]. Rotation forest promotes diversity by employing PCA that performs feature extraction for each base classifier. Assuming that x = $[x_1,\ldots,x_n]^T$ be a data point, and X be the data set containing N training samples and n features in the form of N X n matrix. Consider Y = $[y_1,\ldots,y_N]^T$ as a vector for class labels where y_j obtains values from the set of class labels $\{w_1, w_2,\ldots, w_c\}$. The various classifiers in the ensemble are denoted by $\{D_1, D_2,\ldots,D_L\}$ and feature set by *F*. For most of the ensemble techniques, *L* needs to be selected first. To build the training set for the classifier D_i, the executed steps are shown as follows:

1. *F* (feature set) will be divided into *K* subsets, where *K* represents the algorithm's parameter. To maximize the chance for high diversity, disjoint subsets were chosen such that each feature subset contains M = n/K features.
2. For each subset (denoted by $F_{i,j}$, the j^th subset for the training set of classifier D_i), a randomly nonempty subset of classes was selected bootstrap sample of objects were drawn. The coefficients of principal components, $a_{i,j}^{(1)}, a_{i,j}^{(2)}, \ldots, a_{i,j}^{(M_i)}$ was stored for each size M X 1.
3. Arranging the vectors with coefficients in a rotation matrix R$_i$ as shown in Equation 4:

$$R_i = \begin{bmatrix} a_{i,1}^{(1)}, a_{i,1}^{(2)} \ldots a_{i,1}^{(M_1)} & [0] & \cdots & [0] \\ [0] & a_{i,2}^{(1)}, a_{i,2}^{(2)} \ldots a_{i,2}^{(M_2)} & \cdots & [0] \\ \vdots & \vdots & \ddots & \vdots \\ [0] & [0] & & a_{i,2}^{(1)}, a_{i,2}^{(2)} \ldots a_{i,2}^{(M_2)} \end{bmatrix} \tag{4}$$

The Rotation matrix will have dimensionality n× $\sum_j M_j$. PCA on the subset of classes was performed to extract the significant features. Equation 5 demonstrates how rotation forest classifies each prediction [30].

$$\mu_j(x) = \frac{1}{L}\sum_{i=1}^{L} d_{i,j}(xR_i^a), j = 1, \ldots, c. \tag{5}$$

It is important to note that L represents the number of classifiers in the ensemble generated in the induvial decision tree.

Evaluation Criteria

To effectively compare the predictive capabilities of each tree-based model, the Sensitivity, Specificity, Positive Predictive Values (PPV), and Negative Predictive Values (NPV) are calculated. This is accomplished by first recording the total number of true positives (TP), true negatives (TN), false positives (FP), false negatives (FN) are recorded following the results generated from each tree-based model. A summary of how TP, TN, FP, and FN are calculated are shown in **Table 2**.

TABLE 2. SUMMARY OF HOW TRUE POSITIVE, TRUE NEGATIVE, FALSE POSITIVE, AND FALSE NEGATIVES ARE CALCULATED

	Actual Outcome	Predicted Outcome
TP	1	1
TN	0	0
FP	0	1
FN	1	0

The total number of TP, TN, FP, and FN are recorded and used to calculate the Sensitivity, Specificity, Positive Predictive Values (PPV), and Negative Predictive Values (NPV) using Equations 6-9 [32].

$$Sensitivity = \frac{TP}{TP + FN} \tag{6}$$

$$Specificity = \frac{TN}{TN + FP} \tag{7}$$

$$PPV = \frac{TP}{TP + FP} \tag{8}$$

$$NPV = \frac{TN}{TN + FN} \tag{9}$$

To compare each model and to reduce bias, each tree-based model was generated 10 times. During each round of model generation, a random half of the data is used. This allows for each model to generate 10 rounds for In-Sample (IS) and Out-Of-Sample (OOS) forecasting separately.

RESULTS

As previously stated, each tree-based model had been generated 10 times to compensate for bias in the data. The total number of true positives, true negatives, false positives, and false negatives were recorded. For each time or round (R), a random half of the data were used in distinct techniques, and the same half data were used to estimate the IS to self-validate the model. The remaining unused half data were used to cross-validate the model by OOS testing. The results provided in **Tables 3-4** illustrate the probabilities of crash severities. With the cut-off value of 0.50, which means a probability of 0.50 or higher would result in a severity level of 1. In contrast, a probability less than 0.49 would result in a severity level of 0. The sensitivity, specificity, PPV, and NPV of each round with IS testing are shown in **Table 3**.

TABLE 3. RESULTS OF SENSITIVITY, SPECIFICITY, POSITIVE PREDICTIVE VALUES, AND NEGATIVE PREDICTIVE VALUES BASED ON IN-SAMPLE PERFORMANCE EVALUATION OF DIFFERENT ROUNDS OF MODEL-RUNNING AND AVERAGES AMONG DIFFERENT ROUNDS

	R1	R2	R3	R4	R5	R6	R7	R8	R9	R10	AVE
sen_rotf_in	0.69	0.67	0.7	0.76	0.71	0.69	0.74	0.71	0.66	0.65	0.70
spc_rotf_in	0.89	0.89	0.9	0.88	0.9	0.89	0.86	0.89	0.89	0.89	0.89
ppv_rotf_in	0.22	0.21	0.19	0.2	0.19	0.2	0.23	0.21	0.21	0.22	0.21
npv_rotf_in	0.84	0.82	0.83	0.86	0.83	0.82	0.84	0.83	0.81	0.82	0.83
sen_ranf_in	0.99	0.99	1	0.99	0.99	1	0.99	1	0.99	1	0.99
spc_ranf_in	1	1	1	1	1	0.99	1	0.99	1	1	1.00
ppv_ranf_in	0	0	0	0	0	0.01	0	0.01	0	0	0.00
npv_ranf_in	1	1	1	1	0.99	1	0.99	1	1	1	1.00
sen_boo_in	0.84	0.83	0.85	0.85	0.83	0.85	0.83	0.86	0.84	0.83	0.84
spc_boo_in	0.93	0.93	0.92	0.93	0.94	0.93	0.93	0.93	0.92	0.93	0.93
ppv_boo_in	0.13	0.13	0.13	0.12	0.11	0.11	0.12	0.11	0.14	0.13	0.12
npv_boo_in	0.91	0.9	0.91	0.91	0.9	0.91	0.9	0.91	0.9	0.91	0.91
sen_bag_in	0.99	0.98	0.98	0.98	0.98	0.98	0.97	0.98	0.99	0.98	0.98
spc_bag_in	1	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99
ppv_bag_in	0.01	0.02	0.02	0.01	0.01	0.02	0.01	0.02	0.01	0.01	0.01
npv_bag_in	0.99	0.99	0.99	0.99	0.99	0.99	0.98	0.99	0.99	0.99	0.99

Notes: (1) R1 to R10 represents the 10 rounds that each model had ran.
(2) AVE represents the average value of all 10 rounds.
(3) Prefixes "sen", "spc", "ppv", and "npv" represent Sensitivity, Specificity, Positive Predictive Values, and Negative Predictive Values, respectively.

Upon closer examination of **Table 3**, several points of interest are revealed. For starters, all models falter in their capabilities in terms of PPV, with Rotation Forest having the highest score at 0.21 and Boosting being the second highest at 0.12. In contrast, all models' sensitivity, specificity, and negative predictive values show significant improvements among all models. This is especially true with Random Forest as it scores a 0.99 with sensitivity and a 1.00 with both Specificity and NPV. Following these values, Bagging scores just under Random Forest in these values, followed by Boosting and then Rotation Forest scoring the lowest. The OOS testing results with sensitivity, specificity, PPV, and NPV of each round are shown in **Table 3**.

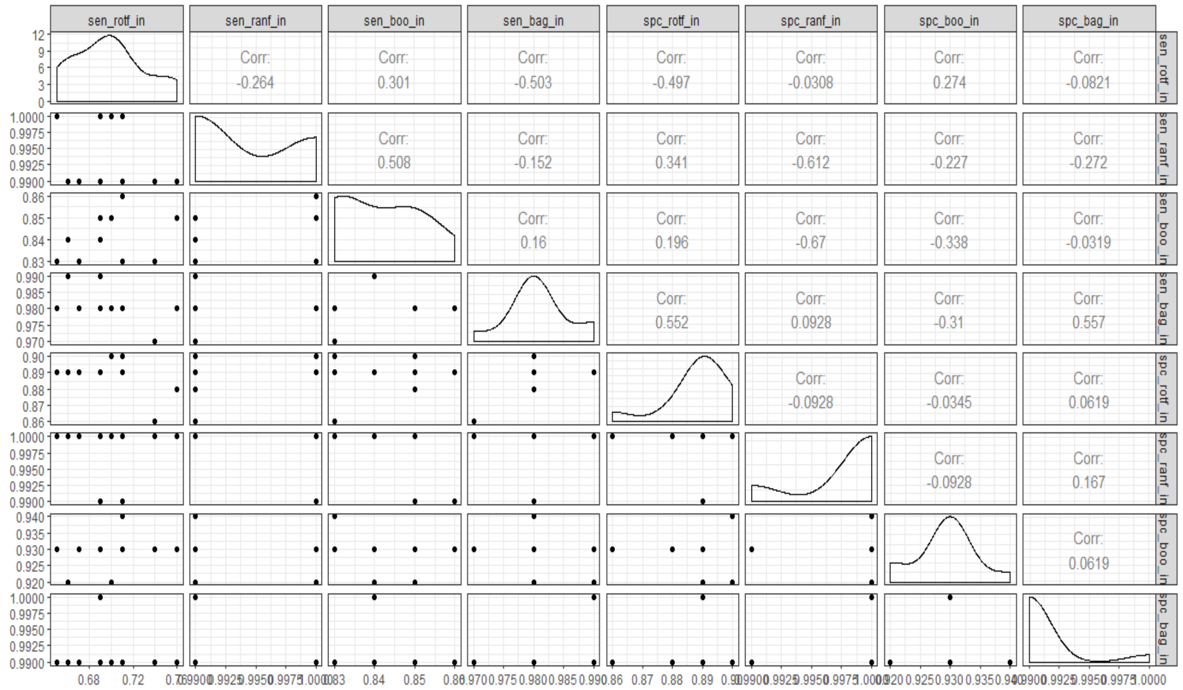
TABLE 4. RESULTS OF SENSITIVITY, SPECIFICITY, POSITIVE PREDICTIVE VALUES, AND NEGATIVE PREDICTIVE VALUES BASED ON SAMPLE CROSS-VALIDATION AND AVERAGES AMONG DIFFERENT ROUNDS

	R1	R2	R3	R4	R5	R6	R7	R8	R9	R10	AVE
sen_rotf_out	0.61	0.65	0.65	0.68	0.68	0.63	0.69	0.67	0.67	0.65	0.66
spc_rotf_out	0.88	0.86	0.89	0.86	0.87	0.88	0.87	0.86	0.88	0.89	0.87
ppv_rotf_out	0.24	0.28	0.23	0.26	0.26	0.27	0.25	0.27	0.24	0.23	0.25
npv_rotf_out	0.79	0.81	0.82	0.83	0.83	0.81	0.84	0.82	0.83	0.81	0.82
sen_ranf_out	0.69	0.68	0.7	0.69	0.71	0.71	0.7	0.72	0.74	0.71	0.71
spc_ranf_out	0.88	0.88	0.89	0.87	0.86	0.88	0.89	0.86	0.88	0.88	0.88
ppv_ranf_out	0.22	0.25	0.22	0.25	0.27	0.24	0.23	0.26	0.23	0.22	0.24

npv_ranf_out	0.82	0.83	0.84	0.83	0.84	0.85	0.84	0.85	0.86	0.84	0.84
sen_boo_out	0.64	0.66	0.7	0.66	0.69	0.66	0.68	0.69	0.69	0.68	0.68
spc_boo_out	0.85	0.82	0.83	0.87	0.83	0.82	0.83	0.82	0.84	0.84	0.84
ppv_boo_out	0.28	0.32	0.31	0.26	0.31	0.33	0.31	0.32	0.29	0.28	0.30
npv_boo_out	0.8	0.81	0.83	0.82	0.83	0.82	0.83	0.83	0.83	0.82	0.82
sen_bag_out	0.69	0.68	0.7	0.69	0.71	0.7	0.68	0.68	0.71	0.69	0.69
spc_bag_out	0.89	0.86	0.88	0.87	0.85	0.88	0.88	0.87	0.86	0.88	0.87
ppv_bag_out	0.21	0.27	0.23	0.25	0.28	0.24	0.24	0.26	0.25	0.22	0.25
npv_bag_out	0.82	0.83	0.84	0.83	0.84	0.85	0.83	0.83	0.84	0.83	0.83

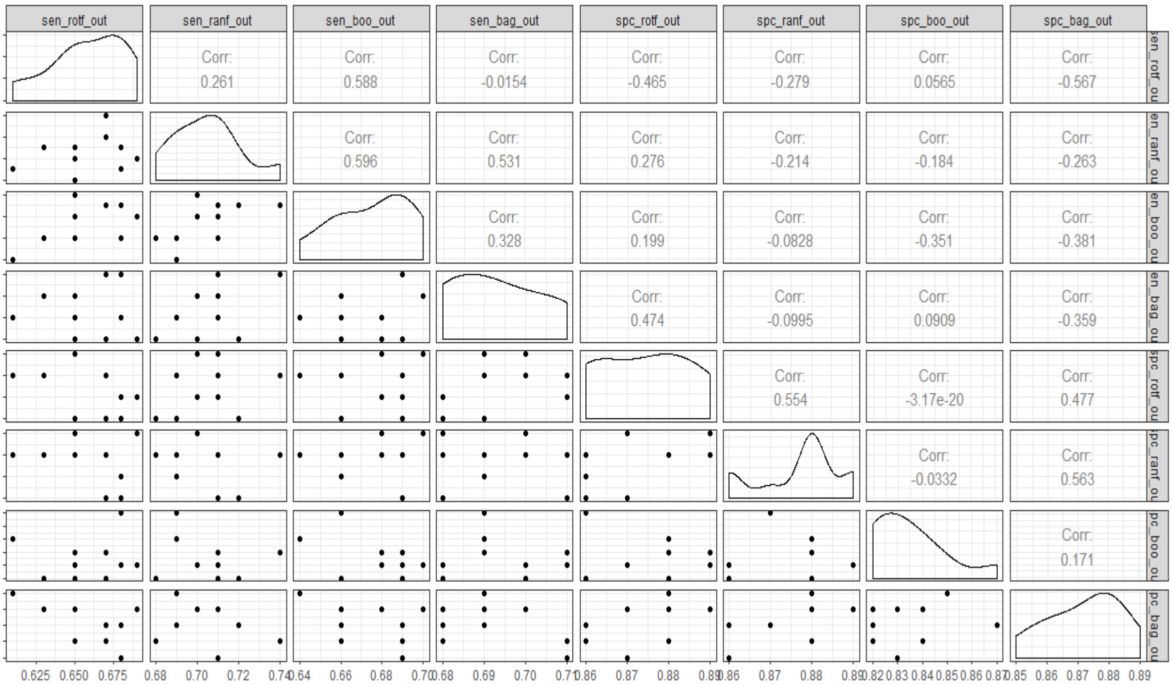
Notes: (1) R1 to R10 represents the 10 rounds that each model had ran.
(2) AVE represents the average value of all 10 rounds.
(3) Prefixes "sen", "spc", "ppv", and "npv" represent Sensitivity, Specificity, Positive Predictive Values, and Negative Predictive Values respectively.

When reviewing **Table 4**, compared to the PPV present in **Table 3**, all models perform better, with Boosting having the highest score at 0.30 and Rotation Forest and Bagging having a score of 0.25. Unfortunately, despite the increased scores of PPV, all other scores are decreased compared to **Table 3**. Much like the results of **Table 3**, Random Forest maintains the highest scores for Sensitivity, Specificity, and NPV with 0.71, 0.88, and 0.84, respectively. In terms of sensitivity, Bagging presents the second-highest score at 0.69, followed by Boosting at 0.68. Specificity scores present a different order, with both Rotation Forest and Bagging having scores at 0.87. Finally, Bagging shows an NPV score of 0.83, and both Rotation Forest and Boosting showcasing a score of 0.82. When looking at the values of both **Table 3** and **Table 4**, it is clear that each tree-based model provides their unique set of benefits and drawbacks. To further understand the similarities and dissimilarities between each tree-based model, a scatter plot of the average sensitivity and specificity of both IS and OOS are shown in **Figures 3-4,** respectively.



Notes: (1) sen-sensitivity; spe-specificity; roft-Rotation Forest; ranf-Random Forest; boo-Boosting; bag-Bagging
(2) The numbers above the main diagonal indicate the correlation coefficients.

FIGURE 3. SCATTERPLOT FOR SENSITIVITIES AND SPECIFICITIES OF VARIOUS METHODS BASED ON IN-SAMPLE DATA EVALUATION



Notes: (1) sen-sensitivity; spe-specificity; roft-Rotation Forest; ranf-Random Forest; boo-Boosting; bag-Bagging
(2) The numbers above the main diagonal indicate the correlation coefficients.

FIGURE 4. SCATTERPLOT FOR SENSITIVITIES AND SPECIFICITIES OF VARIOUS METHODS BASED ON OUT-OF-SAMPLE DATA EVALUATION

Figures 3-4 clearly demonstrate the correlation between the Sensitivity and Specificity of each model. The diagonal graphs in **Figures 3-4** show the overall accuracy of the specific model. The graphs above the diagonal represent the correlation between the two models in question. The correlation ranges from 1 to -1; the closer to 1 that a correlation is, the more correlated the two models were. Likewise, the closer to -1, the more negatively correlated the two models are. Finally, the graphs under the diagonal graphs represent the significant values of both models in question. Thus, it is possible to understand the results of each model and how it relates to the results of other models as well. In addition to using OSS and IS to determine the prediction accuracy, the true positive rate vs. the false positive rate can be seen in the ROC of **Figure 5.**

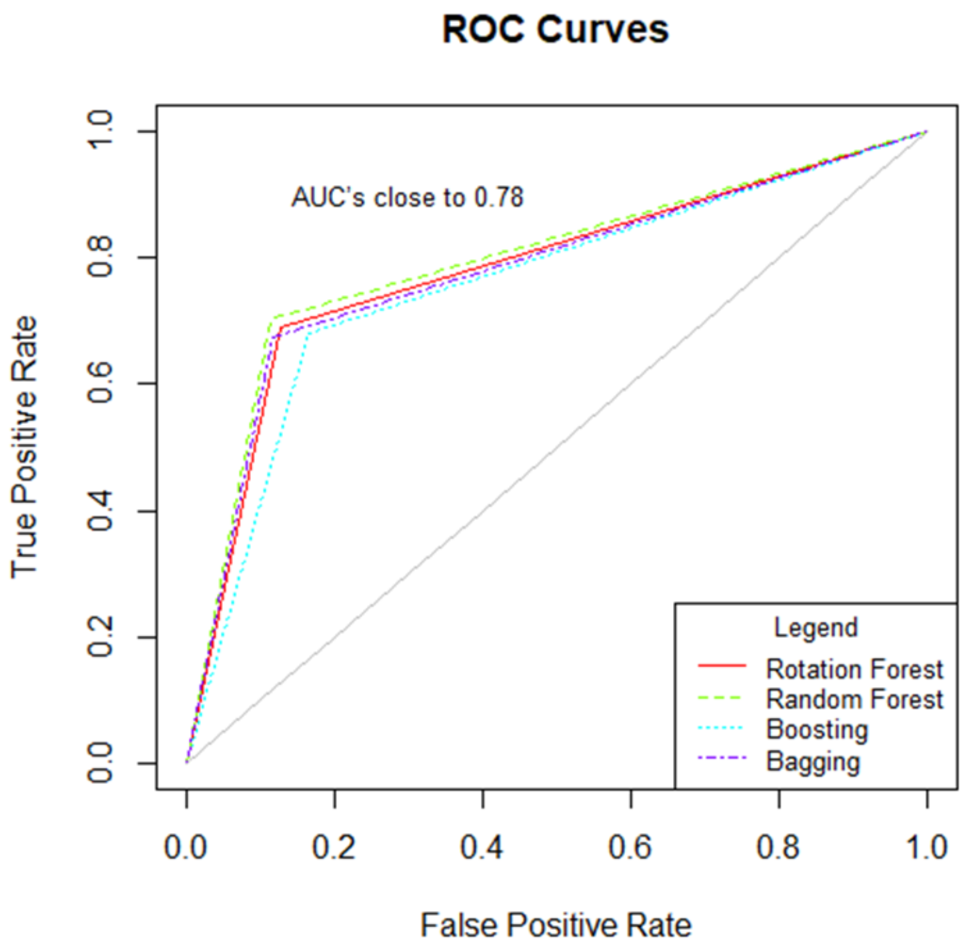


FIGURE 5. ROC CURVES AND AVERAGE CALCULATED AUC FOR ROTATION FOREST, RANDOM FOREST, BOOSTING, AND BAGGING ANALYSIS

The ROC curve aids in the calculation of determining the optimal cut-off rate for a true or false value. Since the variable being test is binary, the baseline prediction is 50% which, on a ROC curve, this baseline is represented as a straight line with a slope of 1 and an intercept of 0. Upon closer examination of **Figure 5**, it becomes apparent that the accuracy of rotation forest is on par with random forest and bagging. It also shows that rotation forest is more accurate than Boosting. This difference, although minor, will create a greater deviation when dealing with larger data sets. Despite this, the average AUC has been calculated to be approximately 0.78. This means that the prediction accuracy of all tree-based methods combined is roughly 78%, statistically more significant than the baseline prediction accuracy that all binary systems have, 50%. Therefore, it is viable to state that Rotation Forest can model generation in traffic safety.

CONCLUSIONS

The primary objective of this study was to perform a comparative evaluation between several distinct tree-based models in traffic safety, specifically in pedestrian crash severity prediction. Additionally, this study proved to be an adequate opportunity to explore the capabilities of a relatively new tree-based model, Rotation Forest. The pedestrian crash data utilized in this study were obtained from HSIS over five years (2010-2014). Four tree-based models, rotation forest, random forest, bagging, and boosting, were employed to predict crash severity, with each model running ten times. In each round, the model was used to self-validate in IS testing and cross-validate in OOS testing. After each round, the Sensitivity, Specificity, PPV, and NPV are recorded and used to compare the different tree-based models' efficacy. Upon closer examination of the results that were described previously, the following conclusions and recommendations can be made:

1) Random Forest is the most accurate when comparing the Sensitivity in both IS and OOS. However, when consulting the PPV, Rotation Forest has the highest accuracy for IS forecasting.
2) When consulting the Sensitivity of both IS and OOS, Rotation Forest is closer related across both IS and OOS. This is likely attributed to the size of the data used in the experiment.
3) According to the ROC curve and the resulting average AUC in **Figure 5**, it is apparent that the average accuracy of each tree-based model is 78%, significantly more accurate than the binary prediction baseline of 50%.

The results here show that each tree-based model has their respective benefits and drawbacks in their performances. Additionally, the relatively new tree-based model, Rotation Forest, shows promise in its ability to predict pedestrian crash severities accurately. Unfortunately, due to the lack of studies that utilize Rotation Forest in traffic safety, its potential benefits are still unclear compared to other tree-based models. For this reason, it is recommended that future studies compare Rotation Forest, as well as Boosting, Bagging, and Random Forest, to other machine learning models such as Classification and Regression Trees (CART), Chi-square Automatic Interaction Detection (CHAID), or Multivariate Adaptive Regression Spline (MARS). It is also worth noting that modifications to these values could yield vastly different results; attention should be paid to modifying the parameters that Bagging and Boosting utilize in Equations 1-2. Furthermore, future studies should consider the limitations provided by Rotation Forest as the dataset used was heavily filtered to remove all observations with missing variables. Since Boosting, Bagging, and Random Forest can operate despite these missing variables within observations, a method that allows Rotation Forest to run with these impurities with the data could change the Sensitivity, Specificity, PPV, and NPV in both IS and OOS validation methods. Finally, future studies should also perform a multi-class prediction rather than a binary prediction as it could provide more insight into each tree-based model's benefits and drawbacks.

REFERENCES

[1]Arms, L., Cook, D., & Cruz-Neira, C. (1999, March). The benefits of statistical visualization in an immersive environment. In Proceedings IEEE Virtual Reality (Cat. No. 99CB36316) (pp. 88-95). IEEE.

[2]Chen, C., Zhang, G., Yang, J., & Milton, J. C. (2016). An explanatory analysis of driver injury severity in rear-end crashes using a decision table/Naïve Bayes (DTNB) hybrid classifier. Accident Analysis & Prevention, 90, 95-107.

[3]Chen, Z., & Fan, W. D. (2019). A multinomial logit model of pedestrian-vehicle crash severity in North Carolina. International journal of transportation science and technology, 8(1), 43-52.

[4]Chica-Olmo, J., Gachs-Sánchez, H., & Lizarraga, C. (2018). Route effect on the perception of public transport services quality. Transport Policy, 67, 40-48.

[5]Dommes, A., Granié, M. A., Cloutier, M. S., Coquelet, C., & Huguenin-Richard, F. (2015). Red light violations by adult pedestrians and other safety-related behaviors at signalized crosswalks. Accident Analysis & Prevention, 80, 67-75.

[6]Ezabadi, S. G., Jolfaei, A., Kulik, L., & Kotagiri, R. (2019, August). Differentially private streaming to untrusted edge servers in intelligent transportation system. In 2019 18th IEEE International Conference On Trust, Security And Privacy In Computing And Communications/13th IEEE International Conference On Big Data Science And Engineering (TrustCom/BigDataSE) (pp. 781-786). IEEE.

[7]Forbes, J. J., & Habib, M. A. (2015). Pedestrian injury severity levels in the Halifax regional municipality, Nova Scotia, Canada: hierarchical ordered probit modeling approach. Transportation Research Record, 2519(1), 172-178.

[8]Governors Highway Safety Association (GHSA). 2019 Bicyclist and Pedestrian Safety.

[9]Hale, A. T., Stonko, D. P., Brown, A., Lim, J., Voce, D. J., Gannon, S. R., ... & Shannon, C. N. (2018). Machine-learning analysis outperforms conventional statistical models and CT classification systems in predicting 6-month outcomes in pediatric patients sustaining traumatic brain injury. Neurosurgical focus, 45(5), E2.

[10]Huang, H., Zhou, H., Wang, J., Chang, F., & Ma, M. (2017). A multivariate spatial model of crash frequency by transportation modes for urban intersections. Analytic methods in accident research, 14, 10-21.

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

[12]James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An introduction to statistical learning (Vol. 112, p. 18). New York: springer.

[13]Kadiyala, A., & Kumar, A. (2016). Univariate time series based radial basis function neural network modeling of air quality inside a public transportation bus using available software. Environmental Progress & Sustainable Energy, 35(2), 320-324.

[14]Kakhki, F. D., Freeman, S. A., & Mosher, G. A. (2019). Evaluating machine learning performance in predicting injury severity in agribusiness industries. Safety science, 117, 257-262.

[15]Kim, J. K., Ulfarsson, G. F., Shankar, V. N., & Kim, S. (2008). Age and pedestrian injury severity in motor-vehicle crashes: A heteroskedastic logit analysis. Accident Analysis & Prevention, 40(5), 1695-1702.

[16]Kumar, N., Barthwal, A., Lohani, D., & Acharya, D. (2020, January). Vehicle Fall Severity Modeling using IoT and K-Nearest Neighbor Algorithm. In 2020 International Conference on COMmunication Systems & NETworkS (COMSNETS) (pp. 105-109). IEEE.

[17]Kuncheva, L. I., & Rodríguez, J. J. (2007, May). An experimental study on rotation forest ensembles. In International workshop on multiple classifier systems (pp. 459-468). Springer, Berlin, Heidelberg.

[18]Kusano, K., & Gabler, H. C. (2014). Comparison and validation of injury risk classifiers for advanced automated crash notification systems. Traffic injury prevention, 15(sup1), S126-S133.

[19]Liao, Y., Zhang, J., Wang, S., Li, S., & Han, J. (2018). Study on crash injury severity prediction of autonomous vehicles for different emergency decisions based on support vector machine model. Electronics, 7(12), 381.

[20]Liu, K. H., & Huang, D. S. (2008). Cancer classification using rotation forest. Computers in biology and medicine, 38(5), 601-610.

[21]Ma, Z., Lu, X., Chien, S. I. J., & Hu, D. (2018). Investigating factors influencing pedestrian injury severity at intersections. Traffic injury prevention, 19(2), 159-164.

[22]Neumann, C., Mateos-Garcia, I., Langenburg, G., Kostroski, J., Skerrett, J. E., & Koolen, M. (2011). Operational benefits and challenges of the use of fingerprint statistical models: a field study. Forensic science international, 212(1-3), 32-46.

[23]Nijkamp, P., Reggiani, A., & Tsang, W. F. (2004). Comparative modelling of interregional transport flows: Applications to multimodal European freight transport. European Journal of Operational Research, 155(3), 584-602.

[24]Parikh, R., Mathai, A., Parikh, S., Sekhar, G. C., & Thomas, R. (2008). Understanding and using sensitivity, specificity and predictive values. Indian journal of ophthalmology, 56(1), 45.

[25]Radzi, N. H. M., Gwari, I. S. B., Mustaffa, N. H., & Sallehuddin, R. (2019, August). Support Vector Machine with Principle Component Analysis for Road Traffic Crash Severity

Classification. In IOP Conference Series: Materials Science and Engineering (Vol. 551, No. 1, p. 012068). IOP Publishing.

[26]Rahman, M. S., Abdel-Aty, M., Hasan, S., & Cai, Q. (2019). Applying machine learning approaches to analyze the vulnerable road-users' crashes at statewide traffic analysis zones. Journal of Safety Research.

[27]Retting, R. A., Ferguson, S. A., & McCartt, A. T. (2003). A review of evidence-based traffic engineering measures designed to reduce pedestrian–motor vehicle crashes. American journal of public health, 93(9), 1456-1463.

[28]Rodriguez, J. J., Kuncheva, L. I., & Alonso, C. J. (2006). Rotation forest: A new classifier ensemble method. IEEE transactions on pattern analysis and machine intelligence, 28(10), 1619-1630.

[29]Sarang Narkhede. (2018, June 26). Understanding AUC - ROC Curve. Retrieved from Medium website: https://towardsdatascience.com/understanding-auc-roc-curve-68b2303cc9c5

[30]Satu, M. S., Akter, T., Arifen, M. S., & Mia, M. R. (2017). Predicting accidental locations of dhaka-aricha highway in bangladesh using different data mining techniques. International Journal of Computer Applications, 165(12).

[31]Schuurman, N., Cinnamon, J., Crooks, V. A., & Hameed, S. M. (2009). Pedestrian injury and the built environment: an environmental scan of hotspots. BMC public health, 9(1), 233.

[32]Sharp, E. (2020, July 21). Statistics: Sensitivity, specificity, PPV and NPV. Geeky Medics. Retrieved October 15, 2021, from https://geekymedics.com/sensitivity-specificity-ppv-and-npv/.

[33]Shen, G., & Wang, J. (2012). A freight mode choice analysis using a binary logit model and GIS: The case of cereal grains transportation in the United States. Journal of transportation technologies, 2(02), 175.

[34]Shirzadi, A., Shahabi, H., Chapi, K., Bui, D. T., Pham, B. T., Shahedi, K., & Ahmad, B. B. (2017). A comparative study between popular statistical and machine learning methods for simulating volume of landslides. Catena, 157, 213-226.

[35]Tabibi, Z., Pfeffer, K., & Sharif, J. T. (2012). The influence of demographic factors, processing speed and short-term memory on Iranian children's pedestrian skills. Accident Analysis & Prevention, 47, 87-93.

[36]Tang, J., Cheng, G., Feng, S., Zhao, X., Zhang, Z., Ju, X., & Yang, H. (2019). Boosting performance and safety of energetic materials by polymorphic transition. Crystal Growth & Design, 19(8), 4822-4828.

[37]Wagenmakers, E. J., Morey, R. D., & Lee, M. D. (2016). Bayesian benefits for the pragmatic researcher. Current Directions in Psychological Science, 25(3), 169-176.

[38]Wuest, T., Weimer, D., Irgens, C., & Thoben, K. D. (2016). Machine learning in manufacturing: advantages, challenges, and applications. Production & Manufacturing Research, 4(1), 23-45.

[39]Xia, J., Du, P., He, X., & Chanussot, J. (2013). Hyperspectral remote sensing image classification based on rotation forest. IEEE Geoscience and Remote Sensing Letters, 11(1), 239-243.

[40]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.