International Journal for Traffic and Transport Engineering, 2021, 11(3): 341 - 358
DOI: http://dx.doi.org/10.7708/ijtte2021.11(3).01
UDC: 656.1.084
614.86

EVALUATION OF MULTI-CLASS MULTI-LABEL MACHINE LEARNING METHODS TO IDENTIFY THE CONTRIBUTING FACTORS TO THE SEVERITY OF ANIMAL-VEHICLE COLLISIONS

Kian Moghaddam[1], Vahid Balali[2], Prateechi Singh[3], Majid Khalilikhah[4]
[1,2] *Department of Civil Engineering and Construction Engineering Management, California State University, Long Beach, CA 90840, USA*
[3] *Department of Computer Engineering and Computer Science, California State University, Long Beach, CA 90840, USA*
[4] *C&M Associates, Dallas, TX 75248, USA*

Received 28 December 2020; accepted 24 March 2021

Abstract: Transportation is a fundamental tool to develop communities, cities, and countries on a larger scale, and more extensive transportation networks have developed ubiquitously. However, it is needed to consider the fact that animals also live in the same environment without using the same means, and there is always a chance of colliding with them while driving vehicles. Animal-Vehicle Collision (AVC) is a principal concern for transportation agencies and roadway hazards that influences human safety, property, and wildlife. State of Tennessee animal crash data has been collected for 23 years containing different types of information for each collision. This paper presents and evaluates the performance of five machine learning-based prediction models for animal collisions in the presence of both categorical and non-categorical features. These five models are developed using Logistic Regression, Random Forest, CatBoost, eXtreme Gradient Boosting (XGBoost), and Light Gradient Boosting Machine (LGBM). The CatBoost model has the highest accuracy level at 78.52%. Therefore, it seems to be the most suitable model to predict animal collisions based on 23-year data from Tennessee. The experimental results demonstrate the potential of leveraging categorical data with CatBoost as a viable solution for creating up-to-date and complete analysis for animal-vehicle collision data.

Keywords: animal-vehicle collision, machine learning, road safety, data analytics.

1. Introduction

Animal-Vehicle Collision (AVC) is a principal concern for transportation agencies and roadway hazards that influences human safety, property, and wildlife. AVCs have an economic effect on individuals and public agencies, and especially in rural districts, they become challenging issues (Rodríguez-Morales *et al.,* 2013). Each year due to AVCs, there are 200 human deaths, 29,000 injuries, and $1.1 billion in property damage in the United States (Conn *et al.,* 2004). Over 40 types of mitigation measures aimed to reduce collisions with large ungulates have been described in (Hedlund *et al.,* 2004; Huijser *et al.,* 2007; Knapp *et al.,* 2004). Examples of mitigation methods are to alert drivers of potential animal crossings by warning signs, wildlife warning reflectors or mirrors (Reeve and Anderson, 1993; Ujvari *et al.,* 1998), wildlife fences (Clevenger *et al.,* 2001),

[2] Corresponding author: Vahid.Balali@csulb.edu

and animal detection systems (Huijser *et al.,* 2006). Hence, exploring the various contributing factors that point to such collisions has become a necessity.

There are two strategies to lessen the numbers of AVC: 1) attempting to limit such crashes from the human end; and 2) restraining animals from becoming close enough to create crashes. The first type of mitigation can be conducted by having smart detective sensors on cars or by making drivers conscious of the likelihood of animal crossing areas. In a study by (Jeihani *et al.,* 2019), it has been revealed how people can be distracted while driving, which causes them to reduce their speed, change lanes, and deviate from the center of the road. In this case, keeping drivers focused on the road can play a significant role. In the second kind, there are multiple recommendations to keep animals off the roadways by fencing roads or constructing bridges on animal crossing routes. This second class of AVC reduction cannot be applied thoroughly to all highways and roads due to the high expenses of the method implementation. Moreover, the second method has shown less effective (Hedlund *et al.,* 2004).

This paper is an upgrade of previous research (Moghaddam *et al.,* 2020) that the contributing factors to the severity of animal-related crashes were identified using logistic classifier. The contribution of this study is to present and evaluate the performance of five machine learning-based prediction models of logistic regression, Random Forest, CatBoost, eXtreme Gradient Boosting (XGBoost), and Light Gradient Boosting Machine (LGBM) for animal collisions in the presence of both categorical and non-categorical features. By identifying these significant features and generating reliable prediction models, the number and severity of AVCs can recede from the human end. To achieve this purpose, the crash data are obtained from the Tennessee Department of Transportation (TDOT's) TRIMS (Tennessee Roadway Information Management System) database. Out of all features of this data set, first, the most crucial ones are found to have enriched impacts on accidents. Then, those selected features are utilized in various machine learning techniques to generate prediction models. The essential elements and the most accurate machine learning model for predicting AVCs are presented.

2. Related Background

The animal and non-animal-related collision in the United States were compared by (Langley *et al.,* 2006). The results showed that fatalities of nonanimal-related crashes stayed relatively stable with a 2.5% increase during 1995 to 2004. However, animal-related events remain significantly high at 78%. The major distinctions were rural roads, darkness, roads less than 4-lanes, between 6:00 PM to 6:00 AM, typical weather, and dry surface conditions. (Bartonička *et al.,* 2018) studied AVCs data in the Czech Republic using the precise binomial method demonstrated about 79% of AVCs occur at night. Furthermore, the study by (Haikonen and Summala, 2001) shows that the height of the sun above or below the horizon is a very active factor for AVCs. In the New South Wales of Australia, the most AVCs occurred on weekends (Ramp and Roger, 2008). They concluded that animals are the primary object of hit with a likelihood of 11.9%. About 81.7% of AVCs happened during fine weather conditions, and 86.7% of those AVCs had experienced a dry surface condition. (Hothorn *et al.,* 2012) investigated Deer-Vehicle Collisions

(DVCs) in south-eastern Germany, initially for its type of roads to show that the number of DVCs is significantly affected by road length. (Lao *et al.,* 2011) used the MP model of (Wang, 1998) and the Vehicle-Animal Interaction-based Probability (VAIP) model on three types of data that this model relates to the driver's ineffective response and the leading vehicle's presence. The results present that the probability of AVC is higher in rural areas than urban areas due to its population of animals, drivers' ineffective response while driving at a speed of 50 mph or higher, and the high probability of male animals escaping from AVCs.

(Khalilikhah and Heaslip, 2017) studied how to improve the performance of animal crossing warning signs for mitigating the AVCs in the state of Utah. They defined and analyzed the traffic sign recognition distance to investigate whether the crashes happened within or out of that distance. Out of all the recorded crashes, 4.5% of accidents were related to animal collisions. Although it is believed that AVCs occur mostly in non-high functional classes of roadway systems, about 58% of the AVCs took place in the highest functional classes of roadway systems, which was generally considered to contribute towards AVCs. To determine whether AVCs occurred within or outside sign recognition distance, specific algorithms were developed to obtain the corresponding results. The study also wanted to find out the hotspots where the greatest number of AVCs took place. To find this, Kernel Density Estimation (KDE) technique using ArcGIS was implemented. Then, KDE was applied to the animal-vehicle collision data, and the regions with a higher likelihood of AVCs were found out so that necessary countermeasures could be implemented in such places.

(Ha and Shilling, 2018) studied the AVCs and its relationship with most important features such as environmental factors, human population density, and spatial patterns change among different taxonomic groups (e.g. medium mammals, small mammals, avian, and ungulate). The development of novel machine-learning Species Distribution Models (SDMs) such as Maxent that use presence-only data are potentially well-suited for AVC modeling where non-linear relationships are likely, and many potential hotspots are possible. The Maxent model utilizes a deterministic machine learning algorithm to optimize environment-species connections based on maximum entropy and then projects those connections across geographic space to map other similar locations (Phillips *et al.,* 2009; Phillips and Dudík, 2008). The study was conducted on three state highways in California. The data regarding the AVCs were obtained from the California Roadkill Observation System (CROS). From the National Land Cover Database (NLCD, USGS) and the National Elevation Dataset (NED, USGS), the land cover data of 2006, and the digital elevation of 2010 were extracted to use. A list of factors within environmental factors and human factors were considered for the Maxent modeling purpose, and the contribution of these variables to the model was analyzed. For ungulate carcass occurrences, there are four dominant factors, such as three environmental variables, total forest area, road density, and elevation. Similar to the ungulate AVC contributing factors, two environmental variables, total forest area and road density within 1640 ft buffer, were important contributing factors to the rest of taxonomic groups. The results obtained from this study showed that Maxent modeling could be used in situations where the factors being considered for the analysis might be

complicated and non-linear. Hence, utilizing machine-learning algorithms such as Maxent is ideally suited for such circumstances.

A variety of factors have high impacts on elevating the risk of AVCs. Another study conducted by (Grace *et al.,* 2017) aimed to reduce AVCs by installing Roadside Animal Detection Systems (RADS) to assist drivers in making better decisions rather than attempting to keep animals off the road. In RADS, animals are detected using sensors, and this triggers signals on the roadside to flash. Consequently, the driver is warned to be careful as animals are nearby. This study was implemented in Florida to see how good RADS performed over one year. By sampling throughout an entire year, they aimed to assess not only whether the RADS was successful at reducing driver speed, but whether or not its effect varied with a seasonal influx of tourists. To assess the effect of RADS on driver speed, analysis of speed data was done using Analysis of Variance (ANOVA). The relationship between the speed of the participants in mile per hour (mph) and whether or not they crashed was modeled using binomial logistic regression in R. Their study concluded that RADS was useful to reduce the number of wildlife collisions, especially the collisions with Florida panthers. Also, they suggested that this system is far more effective when there is a higher risk for AVCs and more suitable when not used throughout the year to reduce acclimation to RADS by local drivers.

Found and Boyce (2011)'s focus is on Deer-Vehicle Collision (DVC) based on the carcasses data from 2002 to 2007 within and beyond the Edmonton City limits. They established definitions for hotspots and coldspots to narrow down their analysis locations to a reasonable number. Their roadside and landside-based models show different guides for predicting DVCs. Their landscape-based models that are based on location and frequency both prove that a higher speed limit has more probability of DVCs. The location-based landscape model demonstrates a higher risk of DVC where there are more forest and non-forest vegetation and where the interaction between the amount of forest and the distance to the nearest non-forest vegetation is decreasing. On the other side, the frequency-based landscape model also presents more chance of DVC where the distance to the most adjacent non-forest vegetation is shorter, the landscape is more heterogeneous, and the road density is lower.

Another study concluded by (Huijser *et al.,* 2009) compared the cost-benefit analysis of various mitigation measures aimed at reducing collisions with deer, mule deer, elk, and moose. They also addressed the importance of safe crossing options for animals by reviewing each mitigation measure. For their study, a cost-benefit analysis was conducted on 13 types of mitigation measures. Furthermore, they determined the cost of each mitigation measure over a 75-year period. They estimated the costs for the average collision with a deer, an elk, or a moose per kilometer per year for ten road sections in the US and Canada. The study concluded that the cost-benefit analysis was a beneficial tool, which can be utilized by transportation agencies when deciding which mitigation measure could be used to reduce AVCs. The study suggested that mitigation measures that include safe crossing opportunities for wildlife might not only substantially reduce road mortality, but also allow for wildlife movements across the road. This

connectivity is essential to the survival probability of the fragmented populations for some species in some regions.

To the best of our knowledge, there is not an analysis of the implementation of different machine learning techniques to evaluate and utilize AVC features that can demonstrate the most useful way of mitigation in the severity and numbers of such accidents. Hence, this paper implements and evaluates the performance of five machine learning-based prediction models for in presence of both categorical and non-categorical features. These five models are Random Forest, Logistic Regression, CatBoost, XGBoost, and LGBM.

3. Method

In this research, as shown in Fig. 1, after cleaning a very large dataset, two methods are implemented to identify the most critical features in AVCs, then five machine learning techniques are developed to create prediction models based on the essential features. Each AVC can have numerous features in the reports that usually are obtained from law enforcement agencies. Out of all recorded characteristics, only a few can contribute to building outstanding prediction models. Weight of Evidence (WOE) and Information Value (IV) techniques are applied to distinguish the most appropriate features. After obtaining these characteristics, the provided TRIMS data from TDOT is used to create prediction models. These models are generated by employing Logistic Regression, Random Forest, XGBoost, CatBoost, and LGBM machine learning techniques. After models are generated, they will be assessed with the testing data that was not utilized previously in the training stage in order to find the factor of accuracy for each method. Based on the result of each method, a ranking system established to measure accuracy within each technique.

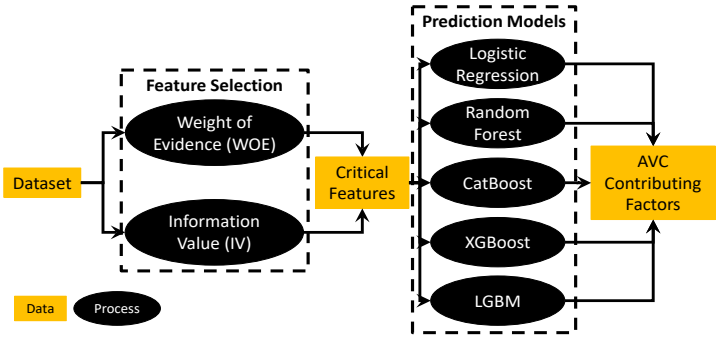

Fig. 1.
An Overview of the Proposed Pipeline

3.1. Data Cleaning

In this very large-scale dataset with more than 100,000 AVCs collected over 20 years, some AVCs lacked features. Therefore, those AVCs missing information were removed from dataset and the number of items were reduced to 82,000. This cleaning was performed to make a strong dataset with appropriate and complete features. Then,

this set is being utilized in the following sections to first find the most powerful features for prediction, and next, featured AVCs are used in prediction models.

3.2. Features Selection

3.2.1. Principal Component Analysis and Support Vector Machine

Two techniques of Principal Component Analysis (PCA) and Support Vector Machine (SVM) are applied in this research. PCA is a statistical method to transform a set of features of possibly correlated variables into a set of values of linearly uncorrelated variables that practices an orthogonal transformation. PCA is used widely in many applications, mostly for eliminating noisy, less informative data before doing regression/classification. PCA is designed for continuous variables, and it works to minimize variance whereas our data contains categorical variables. The concept of squared deviations breaks down when there are binary variables. Therefore, even when this method converts the data to binary using one-hot encoding, it does not guarantee it will work properly and strongly. Support-Vector Machines (SVM) have supervised learning models with associated learning algorithms that analyze data used for classification and regression analysis. Even SVM would not have worked as it works on Euclidean distance, and the categorical features in our dataset are nominal in nature instead of ordinal (Bechra and Kazi, 2017).

3.2.2. Weight of Evidence and Information Value

The WOE indicates an independent variable concerning the dependent variable in the power of prediction. This method has emerged from the world of credit scoring, and it is considered as a measure of depicting the difference between good and bad independent (output) variables. WOE assists in transforming continuous independent variables into a set of bins or groups based on the similarity of dependent variable distribution, which can be categorized as the number of events and non-events. There is no requirement for dummy variables because of this transformation that manipulates categorical features. WOE transformation accommodates to create a strictly linear relationship with log odds, and the value of it can be obtained from Eq. (1):

$$WOE = \ln\left(\frac{\%\ of\ non{-}events}{\%\ of\ events}\right) \quad (1)$$

In continuation of the WOE, IV is to achieve a better result to understand the power of prediction in each feature in the dataset, and the value for IV is derived from Eq. (2):

$$IV = \sum(\%\ of\ non - events - \%\ of\ events) \times WOE \quad (2)$$

By using IV based on the outcomes from WOE, a ranking table for variables can be imported based on their importance. This ranking demonstrates the ability of predictors for modeling. In Table 1, IV statistics are categorized for prediction liability (Bhala, 2015).

Table 1

Information Value Categories and the Feature Strength of Prediction

IV Value	Feature Evaluation
0.3<IV<0.5	Robust
0.1<IV<0.3	Medium
0.02<IV<0.1	Weak
IV<0.02 or IV>0.5	Low

3.3. Prediction Models

This section explains machine learning techniques that are executed for the prediction models of AVC.

3.3.1. Logistic Regression

In this section, the classifier model fits the effect of various independent variables on the dependent variable which is the probability of the crash resulting in a towed vehicle. The developed model finds whether each independent variable significantly affects the likelihood of the vehicle being towed or not. The Akaike Information Criterion (AIC) metric is used to measure the goodness of fit between the models. The modeling process starts with adding independent variables to the model and comparing the AIC of various models with each independent variable, the lower the AIC, the better the fit. The AIC assesses the goodness of the model's fit with consideration of the number of variables (i.e., the degree of freedom) used in the model.

Using the logistic classifier, the effects of different independent variables on the probability of vehicle being towed in the AVC are studied. It is assumed that if the vehicle is being towed, most likely the passengers are injured in the crash and it represents the severity of the incident. Thus, the effect of multiple attributes on the severity is investigated in this research. Although the logistic regression method makes no assumption for outliers, equal variances, and normality, there are independent assumptions and design considerations that apply to the method. Between popular choices of this method, the Gaussian Kernel is used (Kondor and Vert, 2004), which is formulated as follows:

$$y = a \times b^{(x-c)^2} \tag{3}$$

Where y is the response value for the prediction value x. a, b, and c are relative features of the Gaussian curve as a maximum response, standard deviation (between 0 and 1), and mean of the curve, respectively. Eq. (3) resembles a normal distribution where the coefficients should be the most accurate based on the goodness of fit feature. In order to achieve the precise coefficients Eqs. (4), (5), and (6) are utilized:

$$G_{1\,D}(x_1;\,\sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{{x_1}^2}{2\sigma^2}} \tag{4}$$

$$G_{2\,D}(x_1, x_2;\,\sigma) = \frac{1}{(\sqrt{2\pi}\sigma)^2} e^{-\frac{{x_1}^2+{x_2}^2}{2\sigma^2}} \tag{5}$$

$$G_{N\,D}(x_1, x_2, \ldots, x_N;\,\sigma) = \frac{1}{(\sqrt{2\pi}\sigma)^N} e^{-\frac{{x_1}^2+{x_2}^2+\cdots+x_N^2}{2\sigma^2}} \tag{6}$$

Where x_i is an independent variable that is taken into consideration, and σ is the standard deviation that demonstrates the width of the Gaussian kernel distribution. N is the number of independent variables.

3.3.2. Random Forest

Random Forest was introduced by (Breiman, 2001), and it implements the bagging method (Breiman, 1996) to ensemble decision trees. The main reason for selecting this approach is that categorical variables with many values are easier to deal with. Random Forest, in addition, provides handful internal estimates of error, correlation, and strength. It is faster than boosting/bagging and more uncomplicated and comfortable to parallelize.

In bagging, successive trees and earlier trees are independent, and the prediction is based on the majority vote. Random Forest is generated by growing trees depending on a random sample. Each tree results in an estimate of probability at the point that we

want to make a prediction and the mean of the probabilities is used for predicting. In another word, many trees grow to bootstrap samples, and then, the average is calculated and uses that average as the predictor. Random Forest is employed at each classification task, and it delivers accuracy that demonstrates high performance. This method is a refinement of bagged trees that improve on bagging by de-correlating the trees and reducing the variance. The essential steps to creating prediction models are as follow:

1. To select the training data set, the column sampling technique is used to grow the tree. The bootstrap is utilized due to the inability to use all samples fit into the tree growing.
2. The row sampling technique is used in order to select n features from all N features, and because of the Law of Large Numbers, they do not overfit.
3. Without any pruning, each tree in the forest grows thoroughly.

3.3.3. CatBoost

This section develops a model for AVC prediction using CatBoost technique. CatBoost is a gradient boost algorithm that uses machine learning to obtain robust predictions. Gradient boosting works by combining weak models to boost and improve them to robust models. Gradient boosting is a supervised method; this means it takes a set of labeled training instances as input. Then, it builds a model that the label of new unseen examples based on provided features can be correctly predicted. These robust predictions are derived from a combination of weaker models by greedily scoring these features. Created prediction models based on categorical features are not trustworthy in some cases, while CatBoost is primarily used for categorical features. These categorical features are a discrete set of values that are not necessarily comparable with each other but could be converted to numbers before training. The advantage of this machine learning technique is the ability to successfully handle categorical features and take advantage of them during the training time as opposed to pre-processing time.

In this method, the categorical features are converted to numerical values in order to be interpreted by the algorithm. Categorical features are dealt with during the processing time, and each category for every feature is substituted with one or several numerical values. Due to having a discrete set of values by categorical features, they are not comparable, and they cannot be used without having them in terms of numbers. There are different techniques for interpreting the data. One-hot encoding is the most common approach for low-cardinality categorical features, which is used in this section as a key element of having strong converted categorical features.

One-hot encoding presents categorical data expressively by having them in binary vectors. In this way, the number of exiting categories in that feature is counted; then, each category is numbered uniquely and sequentially. The highest number for each feature is used for producing binary vectors. In a vector, digit one is put in respect to the label number of the category, and the rest of the numbers are zeros. In this method, the number of exiting categories in the feature is counted; then, each category numbered uniquely and sequentially — the most number of categories for each feature used for producing binary vectors. In a vector,

digit one is put in respect to the label number of the category, and the rest of the numbers are zeros. That vector delivers the categorical feature but in terms of a numerical feature. As an illustration, the light conditions feature and the formed binary vector for each category are summarized in Table 2. In this feature, there are seven different categories, and the number of columns in the binary vector is seven.

Table 2
Different Categories of Light Conditions with Their Binary Vectors

Number	Category	Binary Vector
1	Day Light	[1,0,0,0,0,0,0]
2	Dark-Not Lighted	[0,1,0,0,0,0,0]
3	Dark-Lighted	[0,0,1,0,0,0,0]
4	Dark-Unknown Lighting	[0,0,0,1,0,0,0]
5	Dawn	[0,0,0,0,1,0,0]
6	Dusk	[0,0,0,0,0,1,0]
7	Other	[0,0,0,0,0,0,1]

CatBoost provides some knobs to tune the predictions, and those parameters make the CatBoost classification more accurate and reliable. First, it starts to optimize the border count and leaf reg independently while iterations and learning rate should be searched together. CatBoost uses a novel schema for calculating leaf values which allows for many permutations without overfitting. A combination of two features is used to allow for new information to be obtained, and no information will be lost. On the other side, iterations have a linear relation with model overfitting which means higher numbers of iteration have a higher chance. However, due to our large dataset for training, it is difficult to have an overfitted model. At the last point, it tries to find the best depth for the decision tree-based model. Moreover, CatBoost adopts a greedy method when splitting trees. For the first split, no combinations are considered. For the second split, all combinations are combined and categorical features present in the current tree with all the categorical features in the dataset. Combination values are converted into numerical values by considering all the splits in the tree as categorical with two values and uses as combinations in the same way as categorical ones.

3.3.4. XGBoost (eXtreme Gradient Boosting)

XGBoost (eXtreme Gradient Boosting) is a gradient boost algorithm that uses sequential decision trees to enhance its efficiency (Chen and Guestrin, 2016) for fitting a prediction model. The two primary assets of XGBoost are execution speed and model performance. The algorithm starts with high bias and low variance, and by the end of improvement, it will have low bias. The method forms fixed-sized trees (same depth) based on the previous tree's residual. Also, it scales all trees by the same amount. As this method builds another tree based on the errors made by the previous trees, this fashion will continue until the number of iterations is met, or additional trees fail to improve the fit. By the end of producing trees, all of them are scaled and will be used to make the prediction model.

3.3.5. LGBM (Light Gradient Boosting Method)

This section explains another type of gradient boost decision tree that can be utilized for making a prediction model for the AVC dataset. LGBM is well known for its speed of model training process when it becomes a large-scale data (Ke *et al.,* 2017). LGBM focuses on reducing the time consumption of regular gradient boost algorithm by using two novel techniques: (1) Gradient-based One-Side Sampling (GOSS); and (2) Exclusive Feature Bundling (EFB). These two techniques are to select a small size sample and fewer features from the data without impacting the accuracy of the overall classification. GOSS technique is to find the instances with a higher contribution to the information gain, and only randomly decline the ones with low contribution impact (low contribution to the information gain). For EFB, first, those mutually exclusive features are found and merged safely, and for other features (not mutually exclusive), they can be bundled by a greedy algorithm. By using these two innovative branches of science, the volume of data is deduced, and it assists the gradient boost to be applied in a shorter time with smaller memory for processing.

4. Case Study

The crash data were obtained from the Tennessee Department of Transportation (TDOT's) TRIMS (Tennessee Roadway Information Management System) database (Moghaddam *et al.,* 2020). The maintenance strategies within TDOT are directed at the region level, consisting of four administrative regions. Each region is subdivided into maintenance stations at the local level and has a carcass removal contract, different than others. Thus, we used the TDOT Risk Management crash data that is based on the accident reports filed by law enforcement officers. These crashes occurred in the years 1994 to 2017 captured the features as shown in Table 3.

Table 3
All Features in Raw Data Set

Number of Tennesee County	Number of Route	log_mle	Case Number
TDOT Location	Date of Crash	Time of Crash	Type of Crash
Total Killed	Total Injured	Total Incapicitating Injured	Total Other Injured
Total Vehicles	First Harmful Event	Weather Conditions	Light Conditions
ID Number	Reporting Agency Type	GPS Coordinate Latitude	GPS Coordinate Longitude
Vehilce Number	Body Code	Driver Factors Actions	Driver Factors Condition
Vehicle Most Harmful Event	Roadway Surface Conditions	Trafficway Hazards	Vehicle Body Type
Vehicle Going on Direction	Driver Vehicle Maneuver	Distraction Code	Alcohol Involved
Areas of Vehicle Damage	Driver Charges	Driver Presence	Driver Violations
Driver License Restriction	Driver Residence	Extent of Damage	Fire in Vehicle
First Impact Point	Number of Travel Lanes	Officer Damage Estimate	Registration State
Roadway Character Alignment	Roadway Character Profile	Roadway Route Signing	Roadway Surface Type
Rollover	Speed Limit	Total Occupants	Traffic Control Devision Function
Trafficway Flow	Truck Bus Supplement	Vehicle Color	Vehicle Defects
Vehicle Going on Highway	Vehicle Make	Vehicle Model	Vehicle Model Year
Vehicle Special Use	Vehicle Towed	Vehicle Trailer	Event sequence

Such data provide crucial information regarding crash severity, time of the crash, weather condition, vehicle age, vehicle speed, and type of animal involved in the accident. The recorded collision attributes include:

- Crash date and time (day, month, year, hour, and minute);
- Route name and direction;
- Roadway type (major, ramp, rest area, and turnaround);
- TDOT's mile point;
- Crash Universal Transverse Mercator (UTM) coordinates;
- Crash severity (no injury/property damage, possible injury, non-incapacitating injury, incapacitating injury, and fatal);
- Crash type (angle, front to rear, head-on, sideswipe same/opposite direction, parked vehicle, rear to side, rear to rear, and single-vehicle);
- Driver age category (teenager, adult, senior);
- Driver condition (normal, aggressive, drowsy, distracted, and DUI);
- Weather condition (clear, cloudy, rainy, snowy, sleet/hail, fog/smog, and blowing sand/soil/dirt);
- Animal-related (yes and no) and animal type (wild and domestic);
- Number of vehicles involved in the crash;
- Roadway posted speed limit;
- Estimated vehicle speed.

First, the raw data are cleaned by removing crashes that do not have all measured features. Then, a group of elements is selected based on their contribution to machine learning techniques that were discussed in the Method section. By finding the best parameters, the models are developed using 80 percent of cleaned data for training. The final step of the generated models is to assess the testing data in order to produce results of prediction models.

There are assumptions and dividing criteria in each feature. Car age is calculated from two different features from the raw data, and it is conducted by subtracting the year of the crash from the model year of the car. As a result, each crash has its car age feature that allows comparing crashes of different years to each other. In addition, the feature of crash time is developed from two important features, the time of the crash in a day and the month of a year. The time zone dividing is defined in Table 3. Table 4 shows the selected parameters in the prediction models of CatBoost, XGBoost, and LGBM.

Table 4

Considered Time Zones of a Day in the Model

Zone Number	Period within
Time Zone 1	12 AM to 3:59 AM
Time Zone 2	4 AM to 7:59 AM
Time Zone 3	8 AM to 11:59 AM
Time Zone 4	12 PM to 3:59 PM
Time Zone 5	4 PM to 7:59 PM
Time Zone 6	8 PM to 11:59 PM

Table 5

Parameters for Different Models

CatBoost Model				
Border Count	Leaf Reg	Iterations	Learning Rate	Depth
20	10	1800	0.08	7
XGBoost Model				
Base Score	No. Estimator	Sub Sample	Learning Rate	Max Depth
0.5	100	1	0.1	3
LGBM Model				
No. Leaves	No. Estimator	Sub Sample	Learning Rate	Max Depth
15	1000	0.9	0.1	1

5. Results and Discussions

5.1. Features Selection

Based on the IV ranking, as depicted in Table 5, a total of 16 features are considered for generating the prediction model.

Table 6

IV Ranking Table

No.	Category	Feature Name	IV
1	Robust	Roadway Surface Type	0.503
2	Robust	Route	0.320
3	Medium	Driver Actions	0.0299
4	Medium	Weather Condition	0.0258
5	Medium	Light Condition	0.0258
6	Medium	Crash Month	0.0242
7	Medium	Location	0.0129
8	Medium	Time of Crash	0.0105
9	Weak	Number of Travel Lanes	0.0087
10	Weak	Driver Factors Actions	0.0051
11	Weak	Driver Factors Condition	0.0031
12	Weak	Roadway Surface Condition	0.0027
13	Low	Body Code	0.0013
14	Low	Car Age	0.0006
15	Low	Driver Vehicle Maneuver	0.0003
16	Low	Speed Limit	0.0001

Table 6 shows those 16 features and the label feature, vehicle towed, by organizing them in two types of categorical and numerical features. The label feature is categorical as a vehicle was towed or not. This table shows the selected features in this research in order to classify the crashes based on the label. The label for this matter is whether a vehicle has towed or not after the collision. In the case of the vehicle is towed, it is deemed a significant crash and high severity. Also, this categorical data has been added to our chosen features as it is the base of our models. It is important to note that the PCA method did not provide results as it showed some features important such as the ID number of a crash.

Table 7
Selected Characteristics of AVCs Data

Type of Feature	Feature Description
Numerical	Number of Travel Lanes
	Speed Limit
	Car Age
Categorical	Vehicle Towed
	Time of Crash
	Month of Crash
	Driver Factors Actions
	Driver Vehicle Maneuver
	Driver Factors Condition
	Driver Actions
	Body Code
	Location
	Roadway Surface Type
	Weather Condition
	Light Condition
	Roadway Surface Condition

5.2. Prediction Models

After cleaning data and creating new combined and serviceable features, prediction models have been created. By using 20 percent of all in use collisions, the prediction models for each machine learning method is evaluated. The value for accuracy is calculated by Eq. (7). In this equation, T_P, T_N, F_P, and F_N are True-positive, True-negative, False-positive, and False-negative respectively. The accuracy of each method is presented in Table 7. Even though CatBoost has the most prolonged duration to train the data, it is the most reliable technique in comparison to other methods. One of the most important reasons that make this method more concrete can be denoted as its high capability to confront categorical features. Moreover, Logistic Regression and XGBoost are second and third accurate techniques.

$$\boldsymbol{Accuracy} = \frac{T_p+T_N}{T_p+T_N+F_P+F_N} \quad (7)$$

Table 7
Accuracy Comparison

Method	Accuracy
Logistic Regression	73.93 %
Random Forest	71.05 %
CatBoost	78.52 %
XGBoost	73.74 %
LGBM	72.73 %

The confusion matrices for each method are shown in Fig. 2. According to our concern about predicting towed class collisions, the first three accurate methods are XGBoost, Logistic Regression, and LGBM, respectively. It means that if towed class collisions are

considered in the calculation of accuracy, the methods get values as shown in Table 8. The difference between this accuracy and the model's accuracy is summarized in considering just the towed class accidents. In another word, if the crashes that caused towing are separately investigated in the prediction models, it demonstrates how precise these models can forecast this category. In this case, XGBoost, Logistic Regression, and LGBM are the first three ranks for having accurate results of prediction.

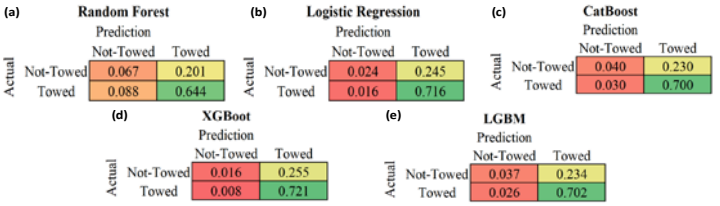

Fig. 2.
Confusion Matrices for Different Methods of (a) Random Forest; (b) Logistic Regression; (c) CatBoost; (d) XGBoost; and (e) LGBM.

Table 8
The Related Accuracy for Different Methods in Towed Class Prediction

Method	Accuracy
Random Forest	87.97%
Logistic Regression	97.82%
CatBoost	95.89%
XGBoost	98.89%
LGBM	96.38%

The correlation between the explanatory and response variables shows that the vehicle age, road speed limit, time of crash (day or night), road surface condition (dry or wet), and alcohol involvement in the collision are relevant factors in severe crashes (i.e., passengers end up being injured). In this dataset, only 26 records were fatal crashes. However, 2,665 injury crashes (5.3%) were observed (Moghaddam *et al.,* 2020). The frequency of crashes in the last six years is shown in Fig. 3.

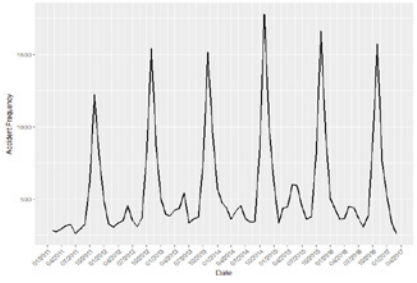

Fig. 3.
AVC Frequency in the Last Six Years
Source: (Moghaddam et al., 2020)

The Precision-Recall (PR) curves of developed models are depicted in Fig. 4. Precision (P) and Recall are determined by Eqs. (8) and (9) respectively. Average Precision (AP) as shown in Eq. (10) is considered the weighted mean of precisions achieved at each threshold. At the nth threshold, precision and recall are measured as P_n and R_n, respectively:

$$P = \frac{T_p}{T_p + F_p} \tag{8}$$

$$= \frac{T_p}{\mathrm{Tp} + F_n} \tag{9}$$

$$AP = \sum_{i=0}^{n} (R_n - R_{n-1}) P_n \tag{10}$$

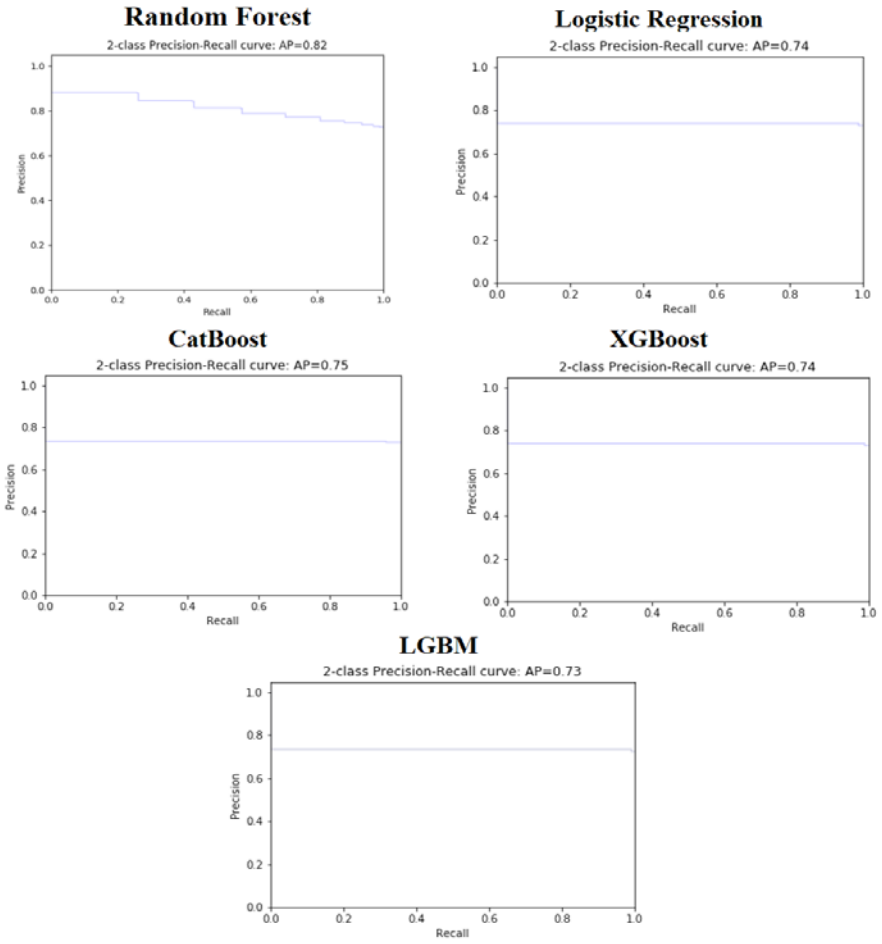

Fig. 4.
Precision-Recall Curves for All Methods Used

ROC curve depicts the True Positive Rate (Towed vs. Towed) against the False Positive Rate (Not-Towed vs. Not-Towed) at various threshold settings. In Fig. 5, the Receiver Operative Characteristic (ROC) curve is representing the slight difference in the area (variance 2%) between different methods.

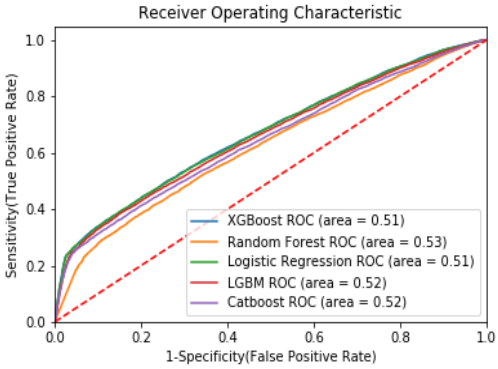

Fig. 5.
Receiver Operating Characteristic for all Methods Used

6. Conclusion

As the number of trips increases every day, the likelihood of accidents rises, and AVC becomes an immediate need to be considered. AVCs are concerned due to the massive amount of tolls on human, animal, and property damage. There are various sorts of mitigation to develop and lower the number and severity of animal-related crashes. This study aims to overcome these collisions by predicting them based on existing conditions and circumstances. Based on this approach, by utilizing the road, weather, and car conditions, WOE & IV can present an adequate number of predictors for AVCs. The prediction models are developed by five different machine learning methods of Random Forest, XGBoost, LGBM, CatBoost, and Logistic Regression. Then, by utilizing the most accurate machine learning techniques, CatBoost and Logistic Regression, a reliable prediction model can be developed. There are some limitations and weaknesses for CatBoost that can be concluded mostly in computational time. Also, this method converts the categorical features to numerical in order to have higher accuracy. After having a prediction model, it can be utilized to anticipate the time and place of AVCs. At the next step, there are some suggestions for road warning signs in case of a great possibility of a crash that can inform the drivers about it. On the other hand, car conditions such as age can warn the authorities for higher risk of AVCs. By combining all the information given to transportation agencies, they can choose the most useful approach based on their factors of decision making. The results of this research demonstrate the essential patterns that have a significant impact on this type of collision.

References

Bartonička, T.; Andrášik, R.; Duľa, M.; Sedoník, J.; Bíl, M. 2018. Identification of local factors causing clustering of animal-vehicle collisions, *The Journal of Wildlife Management* 82(5): 940-947.

Bhala, D. 2015. Weight of Evidence (WOE) and Information Value (IV) Explained. Available from Internet: <https://www.listendata.com>.

Breiman, L. 1996. Bagging Predictors, *Machine learning* 24(2): 123-140.

Breiman, L. 2001. Random Forests, *Machine learning* 45(1): 5-32.

Chen, T.; Guestrin, C. 2016, August. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785-794.

Clevenger, A.P.; Chruszcz, B.; Gunson, K.E. 2001. Highway Mitigation Fencing Reduces Wildlife-Vehicle Collisions, *Wildlife Society Bulletin (1973-2006)* 29(2): 646-653.

Conn, J.M.;Annest, J.L.; Dellinger, A. 2004. Nonfatal Motor-Vehicle Animal Crash-Related Injuries-United States, 2001-2002, *Journal of Safety Research* 35(5): 571-574.

Found, R.; Boyce, M.S. 2011. Predicting Deer–Vehicle Collisions in an Urban Area, *Journal of environmental Management* 92(10): 2486-2493.

Grace, M.K.; Smith, D.J.; Noss, R.F. 2017. Reducing the Threat of Wildlife-Vehicle Collisions during Peak Tourism Periods using a Roadside Animal Detection System, *Accident Analysis & Prevention* 109: 55-61.

Ha, H.; Shilling, F. 2018. Modelling Potential Wildlife-Vehicle Collisions (WVC) Locations using Environmental Factors and Human Population Density: A Case-Study from 3 State Highways in Central California, *Ecological Informatics* 43: 212-221.

Haikonen, H.; Summala, H. 2001. Deer-Vehicle Crashes: Extensive Peak at 1 Hour after Sunset, *American Journal of Preventive Medicine* 21(3): 209-213.

Hedlund, J.H.; Curtis, P.D.; Curtis, G.; Williams, A.F. 2004. Methods to Reduce Traffic Crashes Involving Deer: What Works and What Does Not, *Traffic injury Prevention* 5(2): 122-131.

Hothorn, T.; Brandl, R.; Müller, J. 2012. Large-Scale Model-Based Assessment of Deer-Vehicle Collision Risk, *PLoS One* 7(2): e29510.

Huijser, M.P.; Duffield, J.W.; Clevenger, A.P.; Ament, R.J.; McGowen, P.T. 2009. Cost–Benefit Analyses of Mitigation Measures Aimed at Reducing Collisions with Large Ungulates in the United States and Canada: A Decision Support Tool, *Ecology and Society* 14(2): 15.

Huijser, M.P.; Kociolek, A.V.; McGowen, P.T.; Ament, R.; Hardy, A.; Clevenger, A.P. 2007. *Wildlife-Vehicle Collision and Crossing Mitigation Measures: A Toolbox for the Montana Department of Transportation (No. FHWA/MT-07-002/8117-34)*. Montana. Dept. of Transportation. Research Programs. USA. 113 p.

Huijser, M.P.; McGowen, P.T.; Camel, W. 2006. *Animal Vehicle Crash Mitigation using Advanced Technology Phase I: Review, Design, and Implementation (No. FHWA-OR-TPF-07-01)*. Western Transportation Institute. USA. 271 p.

Jeihani, M.; Ahangari, S.; Hassan Pour, A.; Khadem, N.; Banerjee, S. 2019. *Investigating the Impact of Distracted Driving among Different Socio-Demographic Groups*. Morgan State University. USA. 54p.

Ke, G.; Meng, Q.; Finley, T.; Wang, T.; Chen, W.; Ma, W.; Ye, Q.; Liu, T.Y. 2017. LightGBM: A Highly Efficient Gradient Boosting Decision Tree. In *Proceedings of the Thirty-first Annual Conference on Neural Information Processing Systems (NeurIPS)*, 3146-3154.

Khalilikhah, M.; Heaslip, K. 2017. Improvement of the Performance of Animal Crossing Warning Signs, *Journal of Safety Research* 62: 1-12.

Knapp, K.K.; Yi, X.; Oakasa, T.; Thimm, W.; Hudson, E.; Rathmann, C. 2004. *Deer-Vehicle Crash Countermeasure Toolbox: A Decision and Choice Resource*. Midwest Regional University Transportation Center, Deer-Vehicle Crash Information Clearinghouse, University of Wisconsin-Madison. USA. 260 p.

Kondor, R.; Vert, J.P. 2004. Diffusion Kernels. In *Kernel Methods in Computational Biology*. Cambridge (Massachusetts): MIT Press, 171-192.

Langley, R.L.; Higgins, S.A.; Herrin, K.B. 2006. Risk Factors Associated with Fatal Animal-Vehicle Collisions in the United States, 1995–2004, *Wilderness & Environmental Medicine* 17(4): 229-239.

Lao, Y.; Zhang, G.; Wu, Y.J.; Wang, Y. 2011. Modeling Animal–Vehicle Collisions Considering Animal–Vehicle Interactions, *Accident Analysis & Prevention* 43(6): 1991-1998.

Moghaddam, K.; Balali, V.; Khalilikhah, M.; Rad, A.A. 2020. Identifying the Contributing Factors to the Severity of Animal-Vehicle Collisions. In *Proceedings of the Construction Research Congress 2020: Infrastructure Systems and Sustainability*, 819-826.

Bechra, K.N.; Kazi, A.R. 2017. Survey on Car Detection in Live Video incorporated with Machine Intelligence, *International Journal of Advance Research and Innovative Ideas in Educatio*n 3: 1488-1492.

Phillips, S.J.; Dudík, M. 2008. Modeling of Species Distributions with Maxent: New Extensions and a Comprehensive Evaluation, *Ecography* 31(2): 161-175.

Phillips, S.J.; Dudík, M.; Elith, J.; Graham, C.H.; Lehmann, A.; Leathwick, J.; Ferrier, S. 2009. Sample Selection Bias and Presence-Only Distribution Models: Implications for Background and Pseudo-Absence Data, *Ecological Applications* 19(1): 181-197.

Ramp, D.; Roger, E.; 2008. Frequency Of Animal-Vehicle Collisions In NSW. In Book *Too Close for Comfort: Contentious Issues in Human–Wildlife Encounters*. Royal Zoological Society of New South Wales. Australia. 118-126.

Reeve, A.F.; Anderson, S.H. 1993. Ineffectiveness of Swareflex Reflectors at Reducing Deer-Vehicle Collisions, *Wildlife Society Bulletin (1973-2006)* 21(2): 127-132.

Rodríguez-Morales, B.; Díaz-Varela, E.R.; Marey-Pérez, M.F. 2013. Spatiotemporal Analysis of Vehicle Collisions Involving Wild Boar and Roe Deer in NW Spain, *Accident Analysis & Prevention* 60: 121-133.

Ujvari, M.; Baagøe, H.J.; Madsen, A.B. 1998. Effectiveness of Wildlife Warning Reflectors in Reducing Deer-Vehicle Collisions: A Behavioral Study, *The Journal of Wildlife Management* 62(3): 1094-1099.

Wang, Y. 1998. *Modeling Vehicle-to-Vehicle Accident Risks Considering the Occurrence Mechanismat Four-Legged Signalized Intersections*. Doctoral Dissertation, University of Tokyo.