Accident Analysis and Prevention 128 (2019) 65–77

Contents lists available at ScienceDirect

Accident Analysis and Prevention

journal homepage: www.elsevier.com/locate/aap

A Comprehensive Railroad-Highway Grade Crossing Consolidation Model: A Machine Learning Approach

Samira Soleimani[a,*], Saleh R. Mousa[b], Julius Codjoe[c], Michael Leitner[a]

[a] Geography & Anthropology Department, Louisiana State University, Baton Rouge, LA 70802, United States
[b] Civil & Environmental Engineeing Department, Louisiana State University, LA 70808, United States
[c] Louisiana Transportation Research Center, Louisiana State University, LA 70808, United States

ARTICLE INFO

Keywords:
Highway-Rail Grade Crossing Consolidation
Closure Rating Formula
Safety
XGboost
Machine Learning

ABSTRACT

In the United States, there are approximately 212,000 highway-rail grade crossings, some of which experience vehicle-train incidents that often cause a massive financial burden, loss of life, and injury. In 2017, there were 2,108 highway-rail incidents resulting in 827 injuries and 307 fatalities nationwide. To eliminate collision risks, crossing grade separation and active alarm improvement are commonly used. Moreover, crossing closures are considered to be the most effective safety improvement program. While it may be very difficult, and in some cases impossible to close crossings, there are some incentive programs that facilitate the closure process. One of these programs is a working consolidation model that aims to determine crossing closure suitability. Using details of highway-rail grade crossings in the United States and applying an eXtreme Gradient Boosting (XGboost) algorithm, this paper proposes a data-driven consolidation model that takes into consideration a number of engineering variables. The results indicated an overall accuracy of 0.991 for the proposed model. In addition, the developed XGboost consolidation model reported the relative importance of the variables input to the model, offering an in-depth understanding of the model's behavior. Finally, for the practical implementation of the model, a simplified version containing fewer variables was developed. A sensitivity analysis was performed considering the aggregate gain and the different correlation threshold values between variables. This analysis developed a simplified model utilizing 14 variables, with aggregated gain values of 75% and a correlation threshold of 0.9 which would perform similarly to the full model. Based on this model, 62% of current highway-rail grade crossings should be closed.

1. Introduction

In the United States, highway-rail incidents at both public and private crossings are a major concern in regard to the fatalities and injuries that result from such incidents. Moreover, these incidents place a massive financial burden on state agencies and railroad administrators due to delays in services as well as damage to trains, tracks, and other equipment. The amounts of such damages can be captured by applying statistical methods to highway-rail grade crossing data. In this study, the authors used data obtained from the Rail Inventory Management System (RIMS) database. The RIMS database is a web-based service developed in 2008 and was later complied with the 2015 National Highway Rail Crossing inventory reporting requirement. RIMS is designed not only to meet specific crossing safety program requirements, but also to perform faster and more efficient management jobs (Rail Inventory Management System: RIMS, 2018b). Based on the RIMS database for the year 2016, there were approximately 212,000 highway-rail grade crossings (Rail Inventory Management System (RIMS), 2018a) which at which 2,025 highway-rail incidents occurred including 798 injuries and 265 fatalities nationwide. Then in 2017, there were 2,108 highway-rail incidents including 827 injuries and 307 fatalities nationwide. The Federal Railroad Administration (FRA) requires that ten states which experience a high number of safety challenges in developing a grade crossing Safety Action Plan (SAP) to address the safety issues, including Alabama, California, Florida, Georgia, Illinois, Indiana, Iowa, Louisiana, Ohio, and Texas. (TxDOT, 2013; FHWA, 2016) There is ultimately a need for identifying ways to improve the safety of highway-rail grade crossings, one of which is to close redundant grade crossings via a consolidation program.

The grade crossing closure program was implememted much earlier than the grade crossing consolidation program. Since the terms consolidation program and the closure program have the same objectives,

* Corresponding author.
E-mail address: ssolei1@lsu.edu (S. Soleimani).

https://doi.org/10.1016/j.aap.2019.04.002
Received 5 September 2018; Received in revised form 26 March 2019; Accepted 1 April 2019
Available online 10 April 2019

they are used interchangeably in this paper. The consolidation program seeks to investigate a group of crossings in an area by locating redundant ones. Due problems with crossing distribution, redundant crossings are defined as those which are hazardous and not inherent. An inherent crossing is one that its closure cannot be avoided since it may result in link removal between places. It would be an issue of optimization to find and properly distribute inherent crossings among an area since they can be moved (Song and Wang 1999). Through the consolidation program, redundant crossings are closed while alternative routes are used to offset these closures. However, closing all grade crossings in general may be very difficult, and in some cases impossible. Several incentive programs facilitate the closure process including: cash incentives, nearby crossing and road improvements, quiet zone establishments, and track relocations (Soleimani et al., 2018). One of the recently used incentive programs is a rating formula that aims to generate a candidate list of crossings for closure (Johnson, 2015). Several rating formulas have already been developed in the following states: Iowa (Hans et al., 2015), Kansas (Russell and Mutabazi, 1648), California (CPUC, 2013), and Texas (TxDOT et al., 1998). The states with high amount to safety challenges at crossings and which are pioneers in using these prioritization (rating) formulas are California, New Hampshire, North Carolina, Texas, Oregon, Washington, Iowa, Arkansas, Kansas, Missouri, Illinois, Connecticut, and Florida (Weissmann et al., 2013; Johnson, 2015; Ogden, 2007). However, previous rating formulae only considered limited numbers of variables. The importance of the variables used was based on expert opinion an judgement. These above incentive programs, however, are sometimes inapplicable due to time constraint and budget.

As mentioned before, previous consolidation/closure models were based on a selection method or weighting methods by safety personnel (experts) (Russell and Mutabazi, 1648; Hans et al., 2015; TxDOT et al., 1998; Russell and Mutabazi, 1648). Through the selection method, various groups of variables are used with defined values to create consolidation models. If a crossing met the warranted thresholds, it would be chosen for further closures. The North Carolina DOT (NCDOT), for example, chose consolidation candidates using variables such as crossings being within a quarter mile of one another, crossings where traffic congestion could be safely redirected, crossings with a high number of crashes, and crossings with reduced sight distance (TxDOT, 2013). The variable weighting approach seeks to define a particular rating for each variable, then aims to rank each crossing appropriate for closure. For example, the Florida DOT considered a list of variables including: the number of vehicles per day (less than 2000), the number of trains per day (more than 2), the amount of nearby alternative crossings (within 1300 ft), skewed crossing angles, and crossings on routes not commonly used by emergency vehicles (TxDOT, 2013). Any crossing with multiple characteristics is a candidate for crossing closure. However, the variables investigated in these studies were few and limited to the ones previously mentioned.

Another method that has not been discussed before is integrating machine learning and data-driven approaches with consolidation models to facilitate solving time-consuming and costly problems. Machine learning can be used to determine the relationships of several entities using various classifiers including Decision Trees (DT), Logistic Regressions, Random Forest (RF), and eXtreme Gradient Boosting (XGboost). XGboost is a highly effective technique that commonly displays a high order of accuracy (Omar, 2018; Chen & Guestrin, 2016). Using XGboost, the importance of the applied variables can be obtained as a particular gain value (relative importance) (Friedman et al., 2000; Chen & Guestrin, 2016). A large number of crossing data is available including closed crossings and newly opened crossings (Rail Inventory Management System (RIMS), 2018a). Rather than using the judgment of safety expert, this paper attempts to implement machine-learning techniques to rank the importance of variables in the closure decision process and to develop a consolidation model.

In this context, the paper aims to bridge the existing gap in previous research efforts regarding the limited number of variables considered in the closure program of grade crossings, as well as offers a methodology of identifying the relative importance of the variables used by the suggested consolidation model. First, the paper identifies the required variables for the consolidation of highway-rail grade crossings. Second, it ranks the variables based on their importance values for closure. Third, it reports results on survey findings from safety experts who use their knowledge to rank variables and compares it with the results from the previous section. Fourth, it describes the creation of different models using the selected variables to classify categories for crossings to open and close based on various machine learning algorithms. Finally, the paper outlines a comparative evaluation of the different models to select the most accurate one for highway-rail grade crossing problems.

2. Literature Review

The United States has many safety challenges involving collisions between trains and vehicles, where in 2017 almost one collision occurred every 4 hours (FRA dataset, Accessed Feb 2018). Possible solutions to reduce the number of collisions at highway-rail grade crossings are active road alarms, auditory alarms, in-vehicle alarms, visibility improvements, gates, corridors, grade separations, and grade crossing consolidation and closures. Given the fact that having no highway-rail grade crossing results in having no collision between train and vehicle, consolidation/closure of crossings are the most effective and efficient railroad crossing safety solution (De Gruyter & Currie, 2016).

2.1. The grade crossing consolidation program

The consolidation program tries to close a few unneeded crossings that are near each other to reduce the number of crossings in an area. The closure program also tries to close crossings whit one crossing in an area. Road consolidation and closure are known as the cheapest and most effective way to assure no future collisions. Though community cohesion (Taylor & Crawford, 2009) and land-use applicability may be inversely affected by this process, crossing consolidation/closure programs are still worth completing if there is an alternative route to offset the closures. Moreover, community agreement to close the crossing is also difficult to secure due to assumptions that residents have about the loss of property or advantage (De Gruyter & Currie, 2016). Strong justification is needed to investigate the relationship between the availability of redundant crossings and variables such as safety (De Gruyter & Currie, 2016), pollution, economy, community cohesion, and quality of life. This justification can be used as an incentive to encourage people to cooperate in crossing consolidation (Chadwick et al., 2014).

The main problem of consolidation program is the applicability of its associated incentive programs. The common incentive programs are cash incentives, track relocation, quiet zone establishment, nearby crossing improvement, nearby crossing grade separation, and road improvement. The Louisiana Transportation Research Center distributed a survey among state DOT's and railroad agencies across the United States to ask about the effectiveness and popularity of incentive programs. The survey was distributed among 240 railroad company experts, as well as 52 verified experts working in DOT's nationwide. Overall, there were 60 completed responses obtained, which included 33 responses from railroad companies, and 28 from state DOTs.

Based on the survey results, the available incentive programs are either too expensive or not effective enough to be employed for the consolidation program (Soleimani et al., 2018). Demonstrated in Fig. 1 are the effectiveness and popularity of each currently used incentive program as ranked by the safety personnel survey respondents. These incentive programs are Track Relocation (TR), Nearby Crossing Separation (NCS), Nearby Road Improvement (NRI), Nearby Crossing Improvement (NCI), and Cash Incentive (CI). Fig. 1(a) describes the popularity of incentive programs based on counting the number of

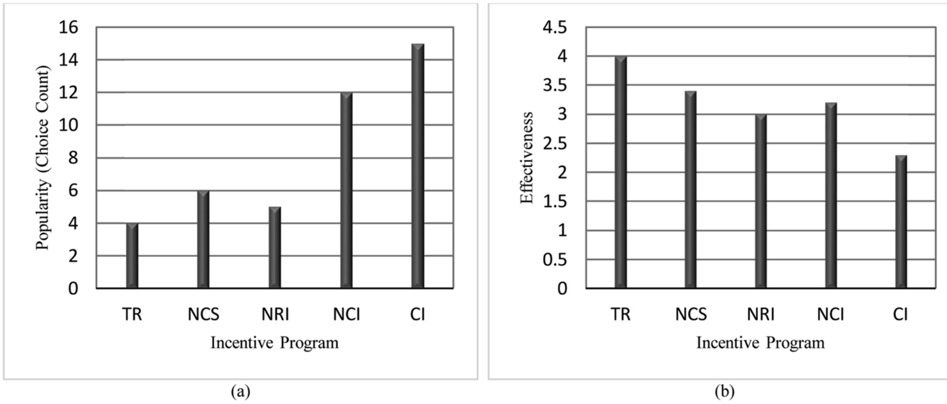

Fig. 1. (a) The popularity of incentive programs; (b) The average effectiveness of the available incentive programs (TR: Track Relocation; NCS: Nearby Crossing Separation; NRI: Nearby Road Improvement; NCI: Nearby Crossing Improvement; CI: Cash Incentive)

times survey responders selected these programs. Fig. 1(b), shows the mean effectiveness of the incentive programs (from 0 = not effective to 4 = highly effective) regarding what safety personnel responded at the survey. In some states the particular laws for consolidation/closure programs, if any exist, are not effective or reliable enough for future safety plans. The need, therefore, arises for a new program to convince crossing owners of further closure. A comprehensive tool is required to calculate the priority of crossing closure based on the variables that matter for all residences, railroad agencies, and DOTs.

2.2. Highway-rail crossing variables

According to the Guidelines for Highway-Rail Grade Crossings (FRA, 2009), the states that had the most highway-rail grade crossing collisions during 2006-2008 should implement their own model with the aim of removing redundant or unneeded crossings while simultaneously improving the safety of crossing and budget objectives. A number of grade crossing variables (factors) are being used by different states for the consolidation program, such as economic and transportation variables (collision history, vehicle delay, operating cost, road traffic, train traffic, type and size of train, grade separation cost, accessibility/connectivity, crossing angle, topography, sight distance, and construction cost, vegetation, development level), social variables (land use and type of property, community cohesion, visual severance (Taylor & Crawford, 2009), geographic distribution, noise, crime, visual amenity (underpass, overpass), site of social significance), and environmental variables (air and water quality, site of environmental significance) (Fakhrhosseini et al., 2015; Hans et al., 2015; De Gruyter & Currie, 2016; Landry et al., 2016).

In one of the holistic studies undertaken, Hans et al. (2015) worked on prioritizing grade crossings for the consolidation program by using the six quantitative variables of traffic volume, heavy-truck traffic volume, road system, proximity to schools, proximity to emergency medical services, and out-of-distance travel. These variables were weighted differently for crossings located in either urban or rural area. In this study however, the authors did not consider the possible correlation impact between variables. For instance, the road system has a direct effect on traffic volume. Likewise, the proximity to particular land use may change the traffic volume at a different time. Also due to lack of information, several variables (such as humped crossing, crime, noise and visual amenity, land use, community cohesion, etc.) were not considered.

In another study, Arellano et al. (2017) considered corridor-levels when prioritizing crossings. However, the study considered each crossing separately when rating them for closure. In the study, for a corridor with n total crossings, the average probability of having m crashes is calculated as follows:

$$CP_n^m = \sum_{i_1=0}^{1}\sum_{i_2=0}^{1}\ldots\sum_{i_n=0}^{1}(1-p_1)^{(1-i_1)}{p_1}^{i_1}(1-p_2)^{(1-i_2)}{p_2}^{i_2}\ldots (1-p_n)^{(1-i_n)}{p_n}^{i_n} \mid \sum_{j=1}^{n} i_j = m \tag{1}$$

Where P_i is the probability of having a crash on crossing i and CP_n^m is the probability of having m crashes on a corridor with n crossings. This safety related equation was used for measuring rail system reliability and optimizing safety improvements in a corridor. By doing this, the mobility, safety variable, and rail system reliability would be increased, though only the sub-variable crash rate is used to calculate the safety variable. In another study, the accessibility variable was calculated by detour distance, which uses just one nearest grade separated distance. Likewise, the safety variable was obtained using the sub-variables of AADT, peak train per day, speed, number of main tracks, and accident history (Schrader & Hoffpauer, 2001). It is beneficial to use a corridor approach for crossings, as well as to utilize safety and mobility variables in one equation to reduce the correlation of different variables. However, the problems of variables correlation and reliability could be considered as the research gap of corridor-based researches.

In a study conducted in Australia, a country that experiences many conflicts at highway-rail grade crossings (Taylor & Crawford, 2009), the correlation between variables was reduced by using Multi-Criteria Assessment (MCA). For variables such as noise and visual amenity for which the appropriate data might not be easily available, the authors used the qualitative indicators regarding the objective rating of the relevant effects. The crossings in this assessment, whether local or urban and private or public, are considered under the same variable sensibility. This means that the result may have errors depending on the development context of each crossing.

In another study in Europe, (Ćirović and Pamučar, 2013) created a neuro-fuzzy decision support system using twenty experts' knowledge on road and traffic safety to select the best alternative crossing for closure. However, in this study, the correlation between the eight used variables was not considered.

2.3. Highway-rail crossing formula

One of the recently used incentive programs is the use of a rating formula that aims to generate a candidate list of the crossings appropriate for the closure program (Johnson, 2015). Taking into account certain variables (e.g., safety, economic, environmental, social, and engineering), every state defines its own rating formula to rank their grade crossings. It is advantageous for every state to consider as many variables as possible when making any investment on crossing control programs; therefore it would be better to develop an equation rather than simply mention individual variables. Rating formulas were initially created at Kansas State University in 1998, where variables such as crossing angle, sight distance, approaching grade, a number of through trains per day, and number of trucks were used (Russell and Mutabazi, 1648). The consolidation model was based on variable correlation weights that were determined by consensus of expert opinion of the advisory crossing safety committee. Finally, a list of good closure candidates was obtained.

The traffic control devices handbook recommends a few variables for closure such as having a value less than 2,000 for average daily traffic (ADT), running more than two trains per day, and having alternate crossing within 0.25 mile that has an ADT less than 5,000 if two lanes or less than 15,000 if four lanes (TxDOT, 2013); FHWA, 2007). These variables are necessary, but not however for a consolidation analysis. The strategies for the consolidation of every state are declared in the Highway-Rail Grade Crossing Handbook (2007, 2016).

To get a bigger picture of consolidation requirements, every state usually depends on crash prediction and severity formulas. Many crash prediction equations raise attention to the effective variables also involved in crossing controls (Austin & Carson, 2002). TxDOT (2013) retrieved information on the thirteen existing priority formulas (indices) of Texas, USDOT, Peabody Dimmick, New Hampshire, Florida, Ohio, Mississippi, Wisconsin, North Dakota, Missouri, City of Detroit, Coleman-Stewart, and National Cooperative Highway Research Program. The most examined variables in the previous formulas were traffic volume, train volume, warning devices, accident, number of tracks, sight distance, train type, bus/special vehicle, train speed, approach grade, crossing angle, pedestrian volume, crossing condition, road/track alignment, highway type/lanes, and highway speed. All of these variables are directly involved in crossing consolidation models (CPUC, 2013; Ogden, 2007; TxDOT, 2013).

In one of the most recent studies in Iowa, Hans et al. (2015) demonstrated a weighted-index approach considering variables including traffic volume, truck traffic volume, proximity to emergency medical services, proximity to schools, road system, and out-of-distance travel. This study is more accurate than previous works since the crossings were segregated into the urban and rural crossings in order to design different models for each. However, as in the previous works, this study also revolved around expert judgment and used a limited number of effective variables.

It is worth mentioning that some of the studies focused on grade crossing variables for highway-rail crossing grade separation that may also be remarkable for a consolidation program as well. In central Arkansas, Schrader and Hoffpauer (2001) applied seven variables of noise, community cohesion, delay, accessibility, connectivity, safety, and geographic distribution. This study used normalized raw weights of the variables to yielded factors of approximately equal weight. In another study, Taylor and Crawford (2009) used a multi-criteria assessment (MCA) tool to identify different weights for variables including economic, social, and environmental variables.

Overall, a limited number of rating formulae for crossing closure is used by some states including Carolina, Texas, Oregon, Washington, Iowa, Arkansas, Kansas, Missouri, Illinois, Connecticut, and Florida (Weissmann et al., 2013; Johnson, 2015; Ogden, 2007). They simply extract information from expert judgment with a limited number of variables. Knowing that various ranges of grade crossing variables affect the closure program, it is sometimes beyond human ability to consider all these variables for final analysis. This is where the machine learning approach could be used to train the available data and apply this trained model on the target data. This procedutre improves both the accuracy, certainty, and applicability of consolidation models.

2.4. The XGboost machine learning method

Previously, Multi-Criteria Assessments (MCA) were used for identifying the significance of important variables which needed expert judgment. Another approach for identifying the importance of the variables without depending on expert knowledge is data mining and machine learning techniques. Since the behavior and interaction among variables for the grade crossing closure program are not already defined, machine learning helps to train the available grade crossing data to develop a holistic consolidation model. Among the supervised machine learning algorithms, XGboost demonstrated its unique capabilities in solving various classification problems and became widely recognized among researchers for its accuracy, simplicity, and interpretability (Mousa et al., 2018). Accordingly, it was implemented by the research group for developing the consolidation model.

XGboost is a supervised algorithm that is an advanced implementation of the original model of gradient boosted trees presented in Friedman (2001). Supervised algorithms need sufficient training data to retrieve information. The larger the training data, the better the accuracy of the trained model. After training, the prediction model would then be applied to the test data to calculate the outcome y_i. The model can be trained to solve either regression or classification problems. To estimate how well the model is working, an objective function (L), containing two terms of loss function (l) and regularization term (Ω), is also needed (Chen & Guestrin, 2016).

$$L(\varnothing) = l(\varnothing) + \Omega(f) \tag{2}$$

The loss function measures the difference between prediction and target functions, so the most common loss functions are the mean square error for regression and the logistic loss classification. The regulation term, however, is required to avoid unnecessary complexity and overfitting (Escabias, 2017). Eq. (3) describes the regularization term within the XGboost model.

$$\Omega(f) = \gamma t + 1/2\lambda\sum_{j=1}^{t} w_j^2 \tag{3}$$

Where t is the total number of leaves, W_j is the weight score on the j^{th} leaf, γ is the minimum split loss reduction, and λ is a regularization parameter. The λ parameter defines the complexity of the model in a way that the higher is it, the higher is the shrinkage of parameters towards 0 (Escabias, 2017). A detailed analytical overview for the XGB algorithm training can be found in (Chen & Guestrin, 2016; Escabias, 2017).

3. Data Acquisition

To develop a consolidation model for highway-rail grade crossings, the crossing data available in the Rail Inventory Management System (RIMS) database was used. The top 18 highway-rail grade crossing safety-challenged states were selected for further analysis including Alabama, Arkansas, California, Florida, Georgia, Illinois, Indiana, Louisiana, North Dakota, Michigan, Minnesota, Mississippi, Missouri, Ohio, Oklahoma, Texas, Washington, and Wisconsin. To avoid sampling bias, the authors decided to overlook the crossing data of other states that had fewer problems with their highway-rail grade crossings. After pre-processing the data by removing either null or missing cells, the extracted data included 18,485 rows of crossings. The data were categorized into 12,741 closed crossings, 424 newly opened crossings, and 5,320 existing (open) crossings. The information of closed crossings as well as the newly opened crossings were used to define a consolidation model. Afterward, by applying the final model on the 5,320 open crossings the percentage of crossings that can be closed was identified. The descriptive information of the used numeric and nominal data are illustrated in Tables 1 and 2 respectively.

As for the model training step, the variable “reason” in RIMS database was considered as the class variable for the model. The variable “reason” consisted of several states: “closed crossing”, “new crossing”, and “existing open crossing” (including: administrative correction, change in primary operating railroad, change in crossing inventory data, no train traffic, reopened crossing). The objective of the paper was

Table 1
Descriptive information of numeric variables used in this study.

Numeric Variable	Min	Max	Mean	SD	Numeric Variable	Min	Max	Mean	SD
Day thru train movements (6 am to 6 pm)	0	99	2.15	5.320	Roadway gate arms	0	7	0.20	0.671
Average no of school buses passing over the crossing on a school day	0	250	0.37	4.698	Cantilevered (or bridged) flashing Light structures over traffic lane	0	8	0.12	0.510
Other flashing lights or warning devices	0	9	0.03	.294	Mast-mounted flashing lights: Mast (post) count	0	9	0.44	0.983
Total trains	0	200	5.46	10.927	Total count of flashing light pairs	0	44	0.19	1.224
Maximum timetable speed (mph)	0	90	23.42	17.104	Bells	0	9	0.27	0.658
Typical minimum speed (mph)	0	74	9.18	11.175	Total switching trains	0	120	1.34	3.140
Typical maximum speed (mph)	0	90	21.8	16.494	Number of traffic lanes Crossing track	0	9	2.08	0.815
Main tracks	0	7	0.78	0.529	Aadt	0	295000	2516.57	7202.550
Cross buck assemblies	0	13	1.36	0.938	Estimated percent trucks	0	99	7.98	8.657
Night thru train movements (6 pm to 6 am)	0	189	1.95	5.246	Stop signs (r1-1)	0	5	0.13	0.491

training a model based on previously closed crossings as well as newly open crossings. Accordingly, the trained model would be applied on "existing open crossing" to classify them as "to be closed" and "to remain open." Then, only the "closed crossing" and "newly open crossing" were employed to train the model (Table 2).

4. Methodology

In this section, the steps for developing and evaluating the crossing consolidation model are discussed. While conducting the analysis, the following two assumptions were made regarding the closed and newly opened crossings. The authors assumed that the closed crossings were the best candidates for training the consolidation model. This is because the decision for closing these crossings was based upon the expert decision and there is no exact rule defining closed crossings. The authors also believed that the newly opened crossings were the best candidates for training the model on the crossings expected to be open. The reason is that the decision for opening a new crossing is based upon the law of supply and demand as well as expert decision. So, regarding the recent pressure for restricting highway-rail grade crossings, if a crossing was newly opened then there should be an urgent need for it. In other words, the newly opened grade crossings must contain almost all the required variables for having a new crossing.

The proposed model aimed to classify open crossings to the categories of "to be closed crossing" and "to remain open crossing." Once the model is trained, validated, and tested, it will be applied to the uncertain 5,320 rows of open crossings in Louisiana to identify the portion of crossings that can be closed. By doing so, the best crossing candidates for future crossing closure are defined.

Table 2
Descriptive information of nominal variables used in this study

Nominal Variable	Type	Frequency	Percentage	Nominal Variable	Type	Frequency	Percentage
Reason	Closed	12741	96.8	Does track run down a street (Y/N)?	No	11626	88.3
	Newly Open	424	3.2		Yes	1539	11.7
In city / near city	In city	9030	68.6	Is crossing illuminated?	No	12634	96
	N City	4135	31.4		Yes	531	4
Crossing type	Private	222	1.7	Crossing surface (main track)	Asphalt	6723	51.1
	Private/Pubic	24	0.2		Asphalt Timber	999	7.6
	Public	12919	98.1		Concrete Rubber	725	5.5
Crossing position	At grade	13144	99.8		Other	71	0.5
	RR Over	10	0.1		Timber	3317	25.2
	RR Under	11	0.1		Unconsolidated	1330	10.1
Type of Land use	Industrial/ Commercial/ Institutional	6938	52.7	Intersecting Roadway Within 500 Feet?	No	221	1.7
	Open/ Farm	3011	22.9		Yes	12944	98.3
	Recreation/RR yard	4	0.0	Smallest Crossing Angle	0-29	575	4.4
	Residential	3212	24.4		30-59	2176	16.5
Quiet zone	Chicago	5	0		60-90	10415	79.1
	No	13075	99.3	Commercial Power Available	No	1202	9.1
	Pa	1	0	Within 500 Feet (Y/N)?	Yes	11963	90.9
	Yes(24 h)	84	0.6	Functional Classification:	Rural	5828	44.3
Are There Signs or Signals?	Yes	1222	9.3	Development	Urban	7337	55.7
	No	11943	90.7	Functional Classification: Road	Local Access	9248	70.2
Low ground clearance signs	No	13141	99.8	Function	Major collector	1550	11.8
(w10-5) present?	Yes	24	0.2		Minor Arterial	1144	8.7
Pavement markings	Combination	3	0		Minor collector	431	3.3
	No Mark	9313	70.7		Other Freeway/ Expressway	53	0.4
	RR mark	356	2.7		Other Principal Arterial	739	5.6
	Stop RR	3028	23	Is Roadway/Pathway Paved?	No	3037	23.1
	Stop Line	465	3.5		Yes	10128	76.9
Advance warning signs	Yes	6785	51.5	Highway type and number	0	4575	34.7
	No	6380	48.5		1	8590	62.5

4.1. Model Performance Measures

The data used for training is considered highly imbalanced because of the disproportionate number of closed crossings to newly opened crossings. For such imbalanced models, the level of accuracy for trained models in predicting tends to be very high. To address this problem, other performance measures were also explored in addition to the accuracy levels for four different algorithms that were developed and evaluated in this study. To choose the most appropriate algorithm, there was a need for a performance measure to calculate and compare the efficiency of algorithms. The Area Under the Curve (AUC) metric, a common model performance measure, was used to evaluate the performance of the model over the training and validation data. The AUC metric refers to the areas under the Receiver Operating Characteristics (ROC) curve. The AUC value deals with the skewed sample situation to prevent overfitting. There is also always a trade-off between specificity (how correctly negative events are classified) and sensitivity (how correctly positive events is classified) in most classifiers. The AUC value attempts different thresholds to classify data and to plot specificity and sensitivity. A useless ROC curve has an area of 0.5, while a perfect ROC curve, which traverses the point correct positive fraction and correct negative fraction equal to 1, has an AUC value of 1 (Brown & Davis, 2006). The sensitivity and accuracy values are used as the performance measures over the testing data using the below equations.

$$Sensitivity = a/A \tag{6}$$

$$Specificity = b/B \tag{7}$$

$$Accuracy = a + b/A + B \tag{8}$$

Where a is the number of correctly classified closed crossings, b is the number of correctly classified newly opened crossings, A is the total number of actual closed crossings, and B is the total number of actual open crossings in the dataset. In this study 70% of the data was used for tuning hyper-parameters, validating and training the models, while 30% was kept away as a final test for the developed models.

4.2. Model Training

The available data suffered from highly imbalance classes (a large number of closed crossing compared to the small number of newly opened ones) might bias the performance of the model towards the frequently occurring instances. Accordingly, to avoid any bias, a sampling technique was applied while training the data. Traditionally, two approaches can be implemented, either “downsampling” the major class or “upsampling” the minor class of data set (Kuhn & Johnson, 2013). In this study, the upsampling technique was used for training the data. The upsampling technique randomly sampled from the available minor classes to set an approximately uniform distribution before the classification step began in each iteration (Zhang & Schuller, 2012).

The crossing dataset included 40 variables that may affect crossing closures. A few of these vairbles are: nearby intersection, train speed, amount and type of signals, road type, and area development. All of these variables could be classified into an engineering variable category. Though other variables such as environmental and social variables could also affect consolidation programs, this study just focused on engineering variables. The aim was to develop a highway-rail grade crossing consolidation model not only to predict whether a crossing should be closed but also to calculate the relative importance of variables in the decision. Several algorithms can be used to create a consolidation model. Since it is crucial to interpret the results correctly, the tree-based algorithms were preferable. Depending on the type of tree algorithms, one or more hyper-parameters should be tuned from the following hyper-parameters: Maximum tree depth (D), a subset of features (S), number of trees (T), and a learning rate (L). In this study, four models were trained and evaluated, and the best one was selected for developing the consolidation model. Specifically, three tree-based algorithms (decision tree, random forest, and XGboost) and one statistical logistic regression model were evaluated.

In traditional Decision Tree (DT) algorithms, each node is selected based on the best split among other values (Liaw & Wiener, 2002). To select a certain variable to split a node, the information gained by branching on that node is calculated (Quinlan, 1986). The gain value is measured by the changes in entropy due to splitting this node into two sub-nodes. The entropy is an indicator that computes the relative frequency of classes to measure the impurity of the classification on the sub-nodes. The decision tree prediction model only needs the D parameter to be defined. It denotes the number of successive nodes/splits in the tree. The higher the value of D, the larger the depth of the tree, the more interaction between variables, and the higher the complexity and accuracy of the tree, however, this may lead to overfitting.

Random Forest (RF) is another tree-based algorithm that selects each node based on the best split values among a subset of randomly chosen values rather than all values. RF combines predictions/classifications from a group of weak tree models to reach more accurate prediction/classification. To ensure that each tree is unique within the group of trees, a different subset sample (S) of the variables is used for growing each individual tree. The RF algorithms propose additional randomness layers to bagging of classification trees and change the structure of classification or regression trees (Liaw & Wiener, 2002). Before training the model using RF, a subset of features (S) should be defined as well as number of trees (T). There is no need to define the tree depth since for each tree the maximum possible depth of trees is considered. In this study, the S variable ranges from 1 to 40, including all the variables in the dataset. For T values, the higher the number of trees the higher the accuracy of the model. However, since the computational cost is increased, an optimal value is always selected by RF. The results of these algorithms in training the model are demonstrated in further sections.

XGboost is another implementation of a tree-based algorithm that is designed to improve the performance of models. XGboost requires the tuning of D and T, as well as the extra regularization parameters L, γ, and λ. It is worth mentioning that, unlike RF, XGboost requires tuning the D value due to the sequential process of growing the trees. The γ and λ are assigned a value of 1 while tuning the hyper-parameters. The role of the L value is to avoid overfitting by decreasing the contribution of each successive tree ($0 < L < 1$). Similar to the RF algorithm, the accuracy of the model is increased by increasing T while it may also cause an overfitting problem. To tune these hyper-parameters for different algorithms, a combination of ten-fold and grid search techniques is applied. Grid search, as an exhaustive search, works to define the optimal combination of hyper-parameters values. The different parameters spaces are defined as D ε [1, 2, …, 10], S ε [10%, 20%, 25%, 30%, 50%, 75%, 100%], T ε [1, 2, …, 4000], and L ε [0.0001, 0.0005, 0.001, 0.005, 0.008, 0.009, 0.01, 0.02, 0.03, 0.05, 0.1, 0.5]. While the learning rate values are commonly assumed to fall between 0.1 and 0.3, this study implemented a wider range of learning rate values due to the large number of trees (1-2000). The searched learning rate values and the varying step size was determined based on sensitivity analysis and preliminary investigation using different values.

The XGboost model possesses four hyperparameters (D, S, L, and T), each of them has at least eight different values; therefore it would be impossible to plot the effect of hyperparameters on the performance of the model in one figure. However, for a better understanding of the developed model, four subplots are presented each of them represents a specific D value (D equal to 3, 5, 8, 10) for a distinct S = 0.2 (Fig. 2). The subplots indicate the Area Under the ROC Curve (AUC) value for the different T values for four different learning rates (L equal to 0.01, 0.03, 0.001, and 0.0001). Illustrated in the four subplots are the changes for ROC with a different number of trees (T).

Plots show that the AUC value increases by any rise in the number of trees (T) until reaching a certain threshold. Beyond the nearby 1000

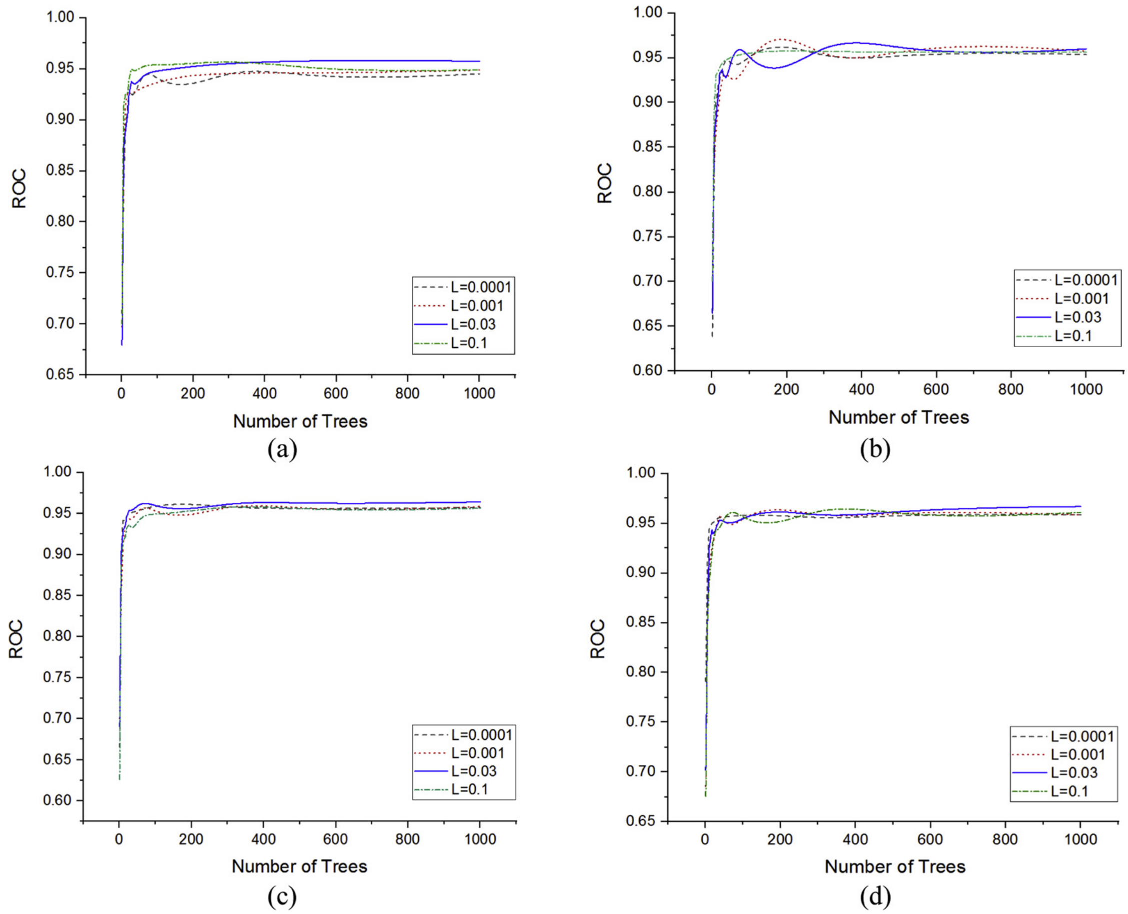

Fig. 2. The AUC for four different subsets of used hyperparameters with constant S value of 0.2. (a) D = 3, (b) D = 5, (c) D = 8, and (d) D = 10.

trees, the improvement in accuracy is negligible. Moreover, for the same D value, the slope of the AUC varies for the different L values. For example, in Fig. 2(a), the slope for L = 0.1 is higher than others, however, after about fifty trees it decreases. Generally, smaller L values reduce the contribution of each tree in XGB, hence it requires more trees to be added. Since the number of training data is ample, the AUC is quite similar to each other for a different combination of hyperparameters. The final best combination for hyperparameters was L = 0.03, D = 10, S = 20%, T = 1000.

Finally, Logistic Regression model is a widespread statistical model by which a categorical dataset with independent variables can be easily interpreted and explained in regression equations. In practice, multiple regression models may be used to predict a dichotomous outcome (e.g., closed crossing or open crossing), and in theory, this model is a linear combination of all variables (X_i), variable coefficients (B_i), and error (e) (Dayton, 1992).

$$logit\ (p) = b_0 + b_1X_1 + b_2X_2 + ...+b_kX_k + e \tag{4}$$

Where p is the probability of the outcome of interest:

$$logit\ (p) = \ln(p/1 - p) \tag{5}$$

It is not necessary to tune any hyper-parameter before training logistic models. Accordingly, no validation of data is needed, and 70% of the data was used only for training and 30% for testing. Additionally, the likelihood of a specific outcome can be calculated based on the existing data. The significance of input variables was also calculated based on the logistic regression model.

5. Results

5.1. Model Validation

Prior to evaluating the four models and selecting one of them, there is a need to tune the hyper-parameters. For each algorithm, the grid search was applied to select the optimal combination of the hyper-parameters. The grid search was guided by a ten-fold cross-validation technique. To perform the ten-fold cross-validation, the 70% training/validation dataset was divided into ten subsets. Then, the model training was performed using nine subsets and validation carried out using the remaining subset. This was repeated ten times by changing the validation subset. The AUC value which is the areas under the ROC curve was then used to evaluate the performance of the model over the training and validation data. For each trial, the AUC value was obtained, and the average AUC value was then obtained for the ten trials to evaluate the model performance. The AUC values for DT (J48), RF, Regression, and XGboost are 0.984, 0.987, 0.986, and 0.991, respectively. In reference to the AUC measurement, the closer the value is to 1, the higher the performance of the mode. Therefore, the RF model and the XGboost model seem to incorporate training data in a better way.

5.2. Model Performance

To boost the accuracy of the highway-rail grade crossing consolidation model, all of the algorithms above were implemented: DT, RD, Regression, and XGboost. To evaluate the performance of the tuned model and select the best one, the tuned remaining models were applied to the test dataset to calculate the sensitivity and accuracy measures. Table 3 summarizes the results for the model sensitivity and

Table 3
The performance of the developed model over the testing data

True Class	Predicted Class		
	Closed	Newly Opened	
Decision Tree (J48)			
Closed	3804	16	Sensitivity 0.996
Newly Opened	46	83	Specificity 0.643
Correctly classified instants: 98.43%			
Logit Regression Model			
Closed	3812	8	Sensitivity 0.998
Newly Opened	47	82	Specificity 0.635
Correctly classified instants: 98.61%			
Random Forest (RF)			
Closed	3819	1	Sensitivity 0.999
Newly Opened	47	82	Specificity 0.635
Correctly classified instants: 98.78 %			
XGboost Model			
Closed	3822	34	Sensitivity 0.991
Newly Opened	6	88	Specificity 0.936
Correctly classified instants: 99.11 %			

accuracy as well as the overall confusion matrices obtained for each applied algorithm of DT, RF, Regression, and XGboost. Comparing the actual with the predicted classes in a dataset, the confusion matrix states the performance of the classification algorithm (Jung et al., 2018).

For the DT algorithm and applying all variables into the model, the size of the pruned tree was 93, while the number of obtained leaves was 49. The confusion matrix for DT with 3,949 samples was illustrated in Table 3, from which 16 closed crossings were wrongly classified as newly opened crossings and 46 newly opened crossings were wrongly classified as closed crossings.

However, unlike DT, in RF, different subsets of variables should be tested to select the optimal subset. Different parameters were defined as S ε [20%, 50%, 80%, 100%] and the iteration size (size of the tree) was 100. Illustrated in Table 3 is the result of the optimal RF which was associated with the subset of 80% of the variables including variables with ID ranges from 1 to 40 described in Table 4. Based on the confusion matrix, 47 newly opened crossings were wrongly classified as closed crossings.

Logistic regression is often superior when it comes to yield the statistical significance of each variable. The impact of each variable on predicting the probability of the desired outcome is explored to determine their significance on the final model (Karp, 1998). The confusion matrix for this algorithm showed that 55 crossings from 3,949 samples were wrongly classified. Illustrated in Table 4 is the significance of each variable for the consolidation model based on Logistic Regression. The higher the significance value, the more critical the variable is for the consolidation model.

The last algorithm tested was XGboost, in which the importance of the variables in a model can also be identified. The confusion matrix represented 34 wrongly classified closed crossings. Six newly closed crossings were also wrongly classified as closed ones. Like the logistic regression model, all of the used variables were associated with a number indicating their level of importance in the consolidation model (Table 5).

Overall, the test results are consistent with the validation results as they confirm that the XGboost algorithm outperforms all other algorithms with a prediction accuracy of 99.11%. This project aimed to fit a model which truly classifies closed crossings (as a positive class) as well as newly opened crossings (as a negative class). Thus, it is necessary to take both sensitivity and specificity values into consideration as well. The XGboost model represented more balanced results for classifying true positive as well as true negative classes with a sensitivity of 99.1% and specificity of 93.6%. The RF algorithm comes second in performance with 98.78% accuracy followed by the Regression with 98.61% accuracy and DT (J48) algorithms with 98.43% accuracy.

This study implemented different algorithms to improve the consolidation model performance. Approximately 13,165 crossings (including 12,741 closed crossings and 424 newly opened crossings) were used to emerge data from the algorithms, taking advantage of machine learning and data mining skills to learn crossing consolidation models from raw crossing data. In Table 3, it was obvious from the confusion matrix that XGboost was superior to DT, RF, and even logistic regression, so the XGboost model is selected as the final algorithm for the consolidation model. However, to overcome any bias in the performance of XGboost, the upsampling technique was also used before training the data.

5.3. The Significance of Variables

Most of the machine learning models are difficult to interpret. From among all the machine learning techniques, the tree-based algorithms and Logistic Regression algorithm report the relative importance of all variables in addition to the parameters of the model performance. This provides better insight and understanding of the model as well as knowing the extent to which each variable effects the model. Among DT, RF, and XGboost, XGboost algorithm was selected to report the variable importance since it was superior to DT and RF. The normalized relative importance of the variables from Regression and XGboost are shown in Tables 4 and 5 respectively. These values provide technical guidance for developing a simplified model without compromising the detection accuracy.

Logistic models declare the regression coefficient, standard error coefficient (SE coefficient), P-value, and odd ratio. The comparison of these statistics together can help to determine significant variables in a model. It should be noted that, due to the different manner of variables, the scale of the regression coefficient of them would be different and which leads bias. To avoid any bias, the authors need to standardized continuous variables to put them in the same scale. To do so, the authors subtracted numeric values by their mean then divided them by their standard deviations (Minitab Blog Editor, 2016). The regression coefficient explains the positive and negative mathematical relationship between model variables and the response (dependent variable). The higher the coefficient, the greater the impact its associate variable would have on the response variable. Standard error coefficient estimates the accuracy and the model calculates coefficients. So, the smaller SE coefficient indicates the more precise estimation. The P value describes whether the relationship between each model variables and the response variable are statistically significant. A low p-value (smaller than 0.05) indicates that the variables are more likely to have a strong effect on the model. Finally, odd ratio describes the ratio of the probability of success to the probability of failure for a constant variable in a model. In this model, the authors choose closure of crossing as the success event.

Regarding the largest absolute regression coefficient values, the top 10 significant variables are: intersecting roadway within 500 ft., quiet zone, low ground clearance signs, crossing illumination, highway traffic signals, total number of trains, crossing position, night through train movements, availability of signs or signals, and either roadway is paved. However, SE coefficient statistic indicated that some of these variables are not sufficiently precise, including intersecting roadway within 500 ft., quiet zone, the total number of trains, crossing position, and night through train movements. In support of that, the latter

Table 4
The significance of variables based on Logistic Regression

ID	Variable	Coefficient	SE Coefficient	P Value	Odd Ratio
1	Intersecting Roadway Within 500	1.12E+12	67271.2	0	*
2	Quiet Zone	15.4774	723.384	0.983	5269250
3	Low Ground Clearance	−2.30435	0.655016	0	0.1
4	Is Crossing Illuminated	1.71725	0.245786	0	5.57
5	Hwy Traffic Signals	−1.67324	0.570281	0.003	0.19
6	Total Trains	−1.60835	7.91457	0.839	0.2
7	Crossing Position	1.55717	5.71216	0.785	4.75
8	Night Thru Train Movements	1.36254	3.80326	0.72	3.91
9	Are There Signals	−1.05747	0.308237	0.001	0.35
10	Is Roadway Pathway Paved	0.929018	0.305701	0.002	2.53
11	Functional Classification Development	−0.832269	0.204999	0	0.44
12	Highway Type Number	0.645284	0.173846	0	1.91
13	Crossing Type	−0.607634	0.219955	0.006	0.54
14	Does Track Run Down a Street	−0.463304	0.245321	0.059	0.63
15	Maximum Timetable Speed	0.460439	0.32754	0.16	1.58
16	Total Switching Trains	0.433313	2.27567	0.849	1.54
17	Mast Mounted Flashing	−0.414147	0.0754768	0	0.66
18	Crossbuck Assemblies	−0.396499	0.0775234	0	0.67
19	Commercial Power Available With	−0.357193	0.293151	0.223	0.7
20	Total Flashing Light pair	−0.350782	0.0510162	0	0.7
21	Number of Traffic Lanes	0.3166	0.105493	0.003	1.37
22	Advance Warning Signs	0.292077	0.190255	0.125	1.34
23	Estimated Percent Trucks	−0.289389	0.0474512	0	0.75
24	Day Thru Train Movements	0.278545	3.85433	0.942	1.32
25	Typical Minimum Speed	0.250903	0.131164	0.056	1.29
26	AADT	0.236287	0.110945	0.033	1.27
27	Smallest Crossing Angle	−0.235459	0.155883	0.131	0.79
28	In or N City	−0.22149	0.211189	0.294	0.8
29	STOP Signs	−0.214096	0.0590874	0	0.81
30	Pavement Markings	−0.17766	0.103261	0.085	0.84
31	Roadway Gate Arms	−0.172312	0.0785449	0.028	0.84
32	Avg School Buses	−0.136459	0.0321317	0	0.87
33	Bell	−0.135263	0.0939075	0.15	0.87
34	Main Tracks	0.120434	0.0907022	0.184	1.13
35	Crossing Surface	0.107234	0.0544439	0.049	1.11
36	Type of Land Use	0.0923095	0.109289	0.398	1.1
37	Other Flashing Lights	−0.0827276	0.0372573	0.026	0.92
38	Cantilevered Flashing Light	0.0493004	0.0739506	0.505	1.05
39	Functional Classification Road	0.0159315	0.084373	0.85	1.02
40	Typical Maximum Speed	0.0028403	0.334406	0.993	1

variables also have large p values (p-value > 0.05) when compared to those having the smallest SE coefficients. It shows that these variables are not significant for the consolidation model. To determine that, either larger coefficients or lower p values don't necessarily identify significant variables (Minitab Blog Editor, 2016). The comparison of these statistics helps to determine the significant variable. Based on the regression analysis, other significant variables are: mast-mounted flashing light, crossbuck assemblies, total flashing light pairs, number of traffic lanes, estimated percent trucks, AADT, stop signs, roadway gate arms, average school bus passing the crossing per day, crossing

Table 5
The significance of variables based on XGboost

ID	variable	Gain %	ID	Variable	Gain %
1	Intersecting roadway within 500 ft.	18.10	21	Type of Land Use	1.64
2	Estimated percent trucks	9.34	22	Roadway gate arms	1.48
3	AADT	7.76	23	Main tracks	1.36
4	Typical minimum speed	5.55	24	Number of Traffic Lanes Crossing Track	1.20
5	Typical maximum speed	5.12	25	Functional classification development	1.11
6	Average number of school buses passing	4.00	26	Smallest crossing angle	1.10
7	Total switching trains	3.81	27	Stop signs	1.06
8	Total count of flashing light pairs	3.78	28	Bells	0.99
9	Maximum timetable speed	3.50	29	Commercial Power Available Within 500 ft.	0.97
10	Day thru train movements	3.24	30	Is crossing illuminated	0.80
11	Total trains	2.82	31	Is roadway pathway paved	0.59
12	Crossbuck assemblies	2.81	32	Advance warning signs	0.55
13	Night thru train movements	2.26	33	Cantilevered or Bridged Flashing Light Structures Over Traffic Lane	0.42
14	Crossing surface	2.25	34	Crossing type	0.36
15	In or near city	2.03	35	Other Flashing Lights or Warning Devices	0.32
16	Does track run down a street	1.90	36	Are There Signs or Signals	0.23
17	Functional classification road function	1.90	37	Low ground clearance signs	0.16
18	Pavement markings	1.84	38	Quiet zone	0.07
19	Highway type number	1.79	39	Crossing position	0.07
20	Mast mounted flashing lights mast post count	1.71	40	Hwy traffic signals controlling crossing	0.00

surface, and other flashing lights. These variables are significant but, their level of importance is lower than the latter ones.

Regardless of the significant and insignificant variables extracted from logistic regression, one of the main variables that directly affects the closure of a crossing is AADT. However, through logistic regression, AADT does not select as a very significant factor comparing p-value and coefficient value. This is where the possibility of doubting the variable significance of logistic regression model arises. Furthermore, the authors tried the XGboost model to get the significance of variables as well (Table 5). Based on XGboost results, the top ten significant variables are intersecting roadway within 500 ft., estimated percent trucks, AADT, typical train speed, average number of school buses, total switching trains, the total count of flashing light, day thru train movements, and total train. The significant contribution of the variable “intersecting roadway within 500 ft.” was expected since the higher the number of intersections around the crossing is, the higher the probability of accident near the crossing would be.

Also the higher the number of nearby intersections to a crossing the higher the chance of that crossing to be either separated, upgraded, or closed. The high importance level of truck percentages indicates the importance of the crossing for transportation which makes the consolidation agreement between railroad agencies and crossing owners more difficult. Surprisingly, several variables that were expected to be significant for closures did not come up as significant ones including crossing illumination, crossing angle, and quiet zone. Accordingly, evaluation of expert knowledge was necessary to see what variables they believe to be the most important ones.

5.4. Experts’ Judgment

An online survey was distributed among railroad crossing safety experts to verify the results of variable significance obtained from XGboost. These variables were: annual average daily traffic (AADT), estimated percent trucks (EPT), number of school bus/EMS (NSBE), active alarms (AA), passive alarms (PA), typical train speed (TTS), crossing surface (CS), day through train movement (DTM), night through train movement (NTM), crossbuck assemblies (CA), gate arms (GA), crossing illumination (CI), development (Urban/ Rural) (D), land use (LU), road function (RF), smallest angle of road and rail (SARR), bell (B), crossing type (private/public) (CT), pavement condition (PC), crossing purpose (CP), disability access (DA), low ground clearance sign (LGCS), specific characteristics of locations (SCL), sight distance (SD), and crime pattern (CP).

The most important and the least important variables are illustrated in Fig. 3. Safety expert personnel considered some of the variables as critical ones including: AADT, development (whether urban or rural area), sight distance (visibility at the crossing), low ground clearance sign, number of school bus/EMS crossing per day, road function (which defines the number of lanes), smallest angle of road and rail, day/night through movement, active alarms, and typical train speed.

The experts also put low values for other variables such as crossing surface, crossing illumination, land use, bells, pavement condition, crime pattern, and the specific physical condition of the place (flood, ice, etc.). Several variables were not included in the models due to the lack of available data including crime pattern, sight distance, and specific physical condition of the place.

Generally, the results of logistic regression was not consistent with the results of the distributed survey among safety expert. For example, the logistic regression model indicated crossing surface as well as roadway pavement as important variables, while safety expert’s personnel did not consider these variables as an important one. As previously mentioned, while safety experts believed AADT as a most important variable for the closure of a crossing, logistic regression results suggested a zero importance for its contribution.

The survey was more in agreement with the XGboost results in regard to the importance of the variables. The variables such as AADT, road function, number of school/EMS bus, and day thru train movements were selected as important variables in both XGboost model and the experts’ survey. According to the model, crossing surface was an almost significant variable. However, safety experts did not consider it as important ones.

In short, among the XGboost and Logistic Regression models, the XGboost findings were more consistent with the previous research (Hans et al., 2015) as well as the experts’ judgment which was discussed here. Thus, comparing the obtained significant values of variables, the authors decided to use XGboost results in our further analysis.

5.5. Simplified Model Development

After a comparison of model performances, the XGboost algorithm was selected for the highway-rail grade crossing consolidation model. Despite selecting a model of good performance, it is also required to develop a model that is as simple as possible. The reason behind developing a simpler model was for the ease of practical implementation of the model. Additionally, a model with a large number of variables may statistically bias since the chance of parameters being randomly selected would be increased. Therefore, the XGboost algorithm was investigated for the possibility of developing a simpler model with fewer input variables without compromising accuracy. Various subsets of the variables in Table 5 were investigated. All variables were

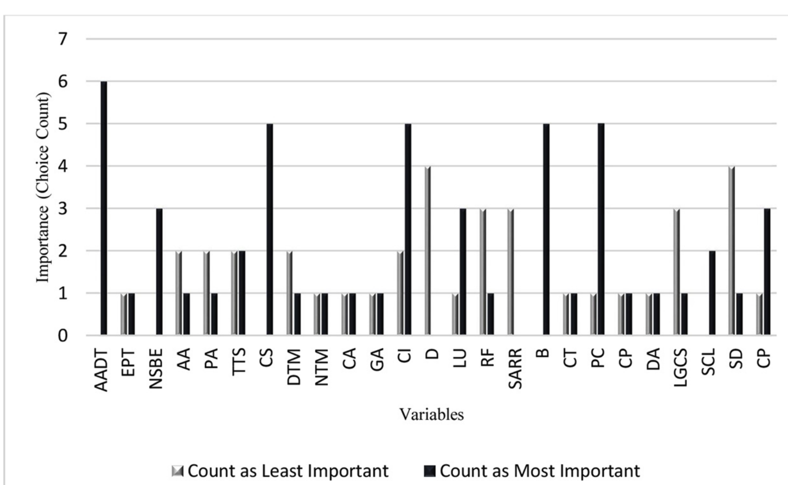

Fig. 3. Ranked variables by safety experts in Louisiana

Table 6
The correlation of removed variables

Thresholds of Correlation	Variable 1	Variable 2	Correlation Value
r > 0.9	Total trains	Day thru train movements (6 am to 6 pm)	0.9073
	Total trains	Night thru train movements (6 pm to 6 am)	0.9092
	Removed Variable(s):		
	1. Total train		
r > 0.8	Night thru train movements (6 pm to 6 am)	Day thru train movements (6 am to 6 pm)	0.8075
	Removed Variable(s):		
	1. Total train (from previous step)		
	2. Night thru train movements (6 pm to 6 am)		
r > 0.5	Type of land use	In or near city	0.5618
	Typical maximum speed	Day thru train movements (6 am to 6 pm)	0.5082
	Typical maximum speed	Night thru train movements (6 pm to 6 am)	0.5020
	Typical maximum speed	Typical minimum speed	0.5865
	Bells	Roadway gate arms	0.6457
	Bells	Mast mounted flashing lights: mast (post) count	0.6528
	Removed Variable(s):		
	1. Total train (from previous steps)		
	2. Night thru train movements (6 pm to 6 am) (from previous step)		
	3. Bells		
	4. Typical maximum speed (mph)		
	5. In or near city		

arranged in a descending order based on the gain value (relative importance) to select the deserved aggregated gain value. The higher the gain value of a variable, the more chance it has to be selected. Six different aggregated gain values were considered to remove the variables with low gain values. Moreover, before any variable selection, a correlation analysis was also applied to remove the highly correlated variables. Three different thresholds of correlation coefficient (R) were used before aggregating the gain values: thresholds of correlation larger than 0.9, thresholds of correlation larger than 0.8, and thresholds of correlation larger than 0.5.

To select appropriate variables, the relative importance values (gain value) of variables were aggregated. For each correlation threshold (0.5, 0.8, and 0.9), 6 different aggregated gain values (65%, 70%, 75%, 80%, 85%, and 90%) were considered. The correlated variables of each correlation coefficient ranges are shown in Table 6. Also, the performance of the simplified model for each aggregated gain and correlation threshold combination is demonstrated in Table 6.

In the various combination parameters based on aggregated gain values and correlations, the values of 0.05, 1, and 10 were considered for learning rate (η), model complexity (γ), and maximum tree depth respectively (Table 7). The aim was to select the most simplified yet optimal model among the eighteen obtained in the model above. In Table 7, model number sixteen is the one with aggregated gain values of 90% and a correlation coefficient bigger than 0.9. In this model, most of the variables have contributions. The simplest model, model number 3, had 65% aggregated gain values with a correlation coefficient bigger than 0.5. This model combined fewer variables in comparison with the other seventeen models.

Based on the performance measures, there is a need to trade-off between a smaller R and aggregated gain value, as well as a higher AUC value, Specificity, Sensitivity, and accuracy. In classification problems, there is a small difference between accuracy and AUC value measures. The accuracy depends on the power of the model to select a threshold to assign patterns to the closed crossing and the open crossing. The AUC value, however, depends on various thresholds to plot specificity and sensitivity.

High rates of sensitivity were found in all the used algorithms (DT, RF, Regression, and XGboost). However, when it came to specificity, the XGboost model was superior to the other models (Table 3). Generally, high sensitivity tests have low specificity. Because the objective was for the final model was to predict the future closed crossings from the currently open ones, the focus of the performance measure was also kept on specificity. Considering the specificity in particular, and the other performance measurement in general, the selected simplified model was model 7, with an aggregated value of 75% and a correlation coefficient of 0.9.

To find included variables in model 7 the correlation coefficient higher than 0.9 between variables were removed first. According to Table 6, there would be no 0.9 or higher correlation between variables by excluding the total train variable from the variable list in Table 5. Since the gain values in Table 5 are sorted, the authors started aggregating the gain values and stopped until the authors reached 75%. All the variables fell in this aggregation were counted in model 7. The selected variables in model 7 were: in or near city (gain value = 2.03), night thru train movement (6 pm to 6 am) (gain value = 2.26), crossing surface (main track) (gain value = 2.25), crossbuck assemblies(gain value = 2.81), day thru train movement (6 am to 6 pm) (gain value = 3.24), maximum timetable speed (gain value = 3.50), total count of flashing light pair (gain value = 3.78), total switching train (gain value = 3.81), average number of school buses passing over the crossing on a school day (gain value = 4.00), typical maximum speed (gain value = 5.12), typical minimum speed (gain value = 5.55), AADT (gain value = 8.76), estimated percent trucks (gain value = 9.34), and intersecting roadway within 500 ft. (gain value = 18.10).

5.6. Simplified Model Prediction

Using 14 variables mentioned above, the simplified XGboost model 7 is capable of representing these effects on each crossing to identify which crossing needs to be closed. By applying this model to the 5,320 current open crossings in Louisiana, the candidate list of the crossings for closure was generated. The simplified XGboost model 7 predicted that 3,307 (62%) currently open crossings need more investigation for further closures.

6. Conclusion

The consolidation of highway-rail grade crossings does not exist in a vacuum. Several variables differentially affect the closure of a grade crossing. Previous studies only considered a limited number of variables, but a crossing consolidation program is influenced by many variables of social, environmental, safety, and economic value. This study investigated the role of 40 variables and developed a highly accurate model to select candidate crossings for closure. Unlike previous studies that solely focused on expert judgment to select the most significant variables for the consolidation program, this study

Table 7
The performance of model based on defined correlation and aggregated gain values

Aggregate Gain	Model Parameters		Model performance (r > 0.9)		Model performance (r > 0.8)		Model performance (r > 0.5)
65%	AUC	Model 1	0.958111	Model 2	0.958111	Model 3	0.958988
	Sensitivity		0.993493		0.993493		0.982729
	Specificity		0.748449		0.748449		0.775185
	Accuracy		0.9853165		0.9853165		0.9858228
70%	AUC	Model 4	0.960211	Model 5	0.95323	Model 6	0.964304
	Sensitivity		0.993156		0.991358		0.995176
	Specificity		0.755083		0.745083		0.761683
	Accuracy		0.9860759		0.98557		0.9870886
75%	AUC	Model 7	0.968434	Model 8	0.96584	Model 9	0.966533
	Sensitivity		0.991923		0.995849		0.997083
	Specificity		0.803993		0.768317		0.751782
	Accuracy		0.9782278		0.9886		0.9883544
80%	AUC	Model 10	0.967699	Model 11	0.968183	Model 12	0.966533
	Sensitivity		0.995176		0.997307		0.997083
	Specificity		0.745116		0.765017		0.751782
	Accuracy		0.9886076		0.9896203		0.9883544
85%	AUC	Model 13	0.970006	Model 14	0.96715	Model 15	0.966635
	Sensitivity		0.997532		0.997532		0.996298
	Specificity		0.745248		0.735281		0.748482
	Accuracy		0.9901266		0.9898734		0.9891139
90%	AUC	Model 16	0.967746	Model 17	0.96314	Model 18	0.96706
	Sensitivity		0.997756		0.992441		0.991371
	Specificity		0.741815		0.72548		0.738548
	Accuracy		0.9911392		0.9893671		0.9898734

implemented the machine learning algorithms including decision tree, random forest, and XGboost as well as logistic regression to retrieve information from the highway-rail grade crossing data. Training these models with 70% of the data, and testing them with the remaining 30% of the data, revealed that XGboost was far superior to the others based on its high accuracy in prediction tests, sensitivity, and specificity. The results indicated an overall accuracy of 0.991 for the proposed model. Finally, for the practical implementation of the model, a simplified version containing fewer variables was developed which removed several highly correlated variables. Thus, the simplified model was performed similarly to the full XGboost model, but only containing 14 variables with aggregated gain values of 75%, and a correlation threshold of 0.9.

Based on this simplified model, 62% of current highway-rail grade crossings in Louisiana should be either closed or undergo safety improvements. It should be noted that the purpose of this study was to introduce a data driven approach and to demonstrate at what level the choice of variables using such approach will reflect those chosen by expert judgement. Evidently, the choice of variables to feed into any such model will depend on local conditions. Certain variables will reflect the future decisions of the locality. Our study only demonstrated the possibility of using a data driven approach and is by no measure suggesting that the developed model is ready to be used by all localities nationwide without first adjusting for localized characteristics

To continue improving the model in the future, more variables, including social, environmental, safety, and economic variables, will be used in addition to the engineering variables that the model currently utilizes. After identifying the potential crossing for closure, the authors should apply a micro-grid analysis to consider the location-based characteristics of each crossing and decide if the crossing needs to be completely closed or to simply improve the crossing's condition (e.g., crossing surface, warning devices, etc.) to some extent. It should be considered that some crossings may have several micro-level conditions which make them stand out from the rest of crossings. Some of these conditions are: the relative accessibility of neighborhood streets, the geometry of the crossing, the number of nearby intersections/crossings, the distance to nearby intersection/crossing, the number of crashes which occurred at the crossing, the cost of crashes, and the severity of crashes. The micro-grid analysis should always be considered in the final decision since the authors are also required to capture the influence of temporary variables on the conditions at crossings which relate to crashes. For example, the street work zone is found to have a significant effect on highway crashes, so it may have a significant influence on crossing performance as well. Though it is a temporary effect and it is one that should be studied further. As for future research, the authors are recommended to first continue training the model using augmented data, then to customize the consolidation model for a state employing effective variables on crash severity, and finally, to consider spatial characterizes of crossings to visualize the geometry of the road.

References

Arellano, J.R., Mindick-Walling, A., Thomas, A., Rezvani, A.Z., 2017. Prioritization of Infrastructure Investment for Rail Safety Projects: 1 A Corridor-Level Approach 2 (No. 17-04963).

Austin, R.D., Carson, J.L., 2002. An alternative accident prediction model for highway-rail interfaces. Accident Analysis & Prevention 34 (1), 31–42.

Brown, C.D., Davis, H.T., 2006. Receiver operating characteristics curves and related decision measures: A tutorial. Chemometrics and Intelligent Laboratory Systems 80 (1), 24–38.

Chadwick, S.G., Zhou, N., Saat, M.R., 2014. Highway-rail grade crossing safety challenges for shared operations of high-speed passenger and heavy freight rail in the US. Safety Science 68, 128–137.

Chen, T., Guestrin, C., 2016. XGboost: A scalable tree boosting system. Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining 785–794.

Ćirović, G., Pamučar, D., 2013. Decision support model for prioritizing railway level crossings for safety improvements: Application of the adaptive neuro-fuzzy system. Expert Systems with Applications 40 (6), 2208–2223.

CPUC (California Public Utilities Commission), 2013. Section 190 Grade Separation Program. California Public Utilities Commission Rail Crossings Engineering Section, Los Angeles, CA.

Dayton, C. Mitchell, 1992. Logistic regression analysis. Stat 474–574.

De Gruyter, C., Currie, G., 2016. Impacts of Rail-Road Crossings: International Synthesis and Research Gaps (No. 16-1588).

Escabias, Carlos Bort, 2017. Tree Boosting Data Competitions with XGboost. Interuniversity Master in Statistics and Operations Research UPC-UB. Accessed:https://upcommons.upc.edu/bitstream/handle/2117/100293/memoria.pdf?sequence=1&isAllowed=y.

Fakhrhosseini, S.M., Jeon, M., Lautala, P., Nelson, D., 2015. An investigation on driver behaviors and eye-movement patterns crossings using a driving simulator. 2015 Joint Rail Conference pp. V001T06A012-V001T06A012.

Federal Highway Administration, 2007. Highway-Rail Grade Crossing Handbook. Accessed:https://www.fra.dot.gov/Elib/Document/1464 Accessed by December 2018.

Federal Highway Administration, 2016. Highway-Railway Grade Crossing Action Plan

and Project Priorization. Accessed:. https://safety.fhwa.dot.gov/hsip/xings/fhwasa16075/fhwasa16075.pdf.
Friedman, J., Hastie, T., Tibshirani, R., 2000. Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors). The annals of statistics 28 (2), 337–407.
Friedman, J.H., 2001. Greedy function approximation: a gradient boosting machine. Annals of statistics 1189–1232.
Hans, Z., Albrecht, C., Johnson, P.M., Nlenanya, I., 2015. Development of Railroad Highway Grade Crossing Consolidation Rating Formula.
Johnson, P.M., 2015. An investigation of railroad-highway grade crossing consolidation rating in Iowa. Doctoral dissertation. Iowa State University.
Jung, M., Niculita, O., Skaf, Z., 2018. Comparison of Different Classification Algorithms for Fault Detection and Fault Isolation in Complex Systems. Procedia Manufacturing 19, 111–118.
Karp, Andrew H., 1998. Using logistic regression to predict customer retention. Proceedings of the Eleventh Northeast SAS Users Group Conference. Accessed: http://www.lexjansen.com/nesug/nesug98/solu/p095.pdf.
Landry, S., Jeon, M., Lautala, P., Nelson, D., 2016. Getting Active with Passive Crossings: Investigating the Use of In-Vehicle Auditory Alerts for Highway-Rail Grade Crossings. 2016 Joint Rail Conference pp. V001T06A019-V001T06A019.
Liaw, A., Wiener, M., 2002. Classification and regression by Random Forest. R news 2 (3), 18–22.
Blog, Minitab (Ed.), 2016. How to Identify the Most Important Predictor Variables in Regression Models. The Minitab Blog (Accessed December 2018). http://blog.minitab.com/blog/adventures-in-statistics-2/how-to-identify-the-most-important-predictor-variables-in-regression-models.
Mousa, S.R., Bakhit, P.R., Osman, O.A., Ishak, S., 2018. A Comparative Analysis of Tree-Based Ensemble Methods for Detecting Imminent Lane Change Maneuvers in Connected Vehicle Environments. Transportation Research Record.
Ogden, B.D., 2007. Railroad-highway grade crossing handbook (No. FHWA-SA-07-010). Federal Highway Administration, United States.
Omar, K.B.A., 2018. XGboost and LGBM for Porto Seguro's Kaggle challenge: A comparison.
Quinlan, J.R., 1986. Induction of decision trees. Machine learning 1 (1), 81–106.
Rail Inventory Management System (RIMS), Database. https://rims.tavlasolutions.com. (Accessed February 2018).
Rail Inventory Management System: RIMS, TAVLA Solutions. Accessed: http://www.tavlasolutions.com/#1. (Accessed December 2018).
Russell, E., Mutabazi, M., 1998. Rail-highway grade crossing consolidation in Kansas. Transportation Research Record: Journal of the Transportation Research Board (1648), 1–7.
Schrader, M., Hoffpauer, J., 2001. Methodology for evaluating highway-railway grade separations. Transportation Research Record: Journal of the Transportation Research Board (1754), 77–80.
Soleimani, S., Ledet, S.J., Codjoe, J., 2018. Incentive Programs for Closure of Grade Crossings in the United States: A State-of-the-Practice Synthesis (No. 18-04693).
Song, X., Wang, Y., 1999. On the crossing distribution problem. ACM Transactions on Design Automation of Electronic Systems (TODAES) 4 (1), 39–51.
Taylor, J., Crawford, R., 2009. September. Prioritising road-rail level crossings for grade separation using a multi-criteria approach, vol. 32 Australasian Transport Research Forum (ATRF), Auckland, New Zealand 32nd, 2009.
Texas Department of Transportation, 2013. Integrated Priorization Method for Active and Passive Highway-Rail Crossings. Accessed:. https://static.tti.tamu.edu/tti.tamu.edu/documents/0-6642-1.pdf.
TxDOT (Texas Department of Transportation), 1998. Railroad Operations Volume: Chapter 6, Section 4: The Texas Priority Index. Texas Department of Transportation, Austin, Texas (Accessed 9 May 2018). onlinemanuals.txdot.gov/txdotmanuals/tfe/the_texas_priority_index.htm.
US Department of Transportation, 2009. Federal Railroad Administration, Research Results, Crossing Consolidation Guidelines, Report # RR 09-12. July. .
Weissmann, A.J., Weissmann, J., Kunisetty, J.L., Warner, J., Park, E.S., Sunkari, S., Protopapas, A., Venglar, S., 2013. Integrated prioritization method for active and passive highway-rail crossings (No. FHWA/TX-12/0-6642-1). A&M Transportation Institute, Texas.