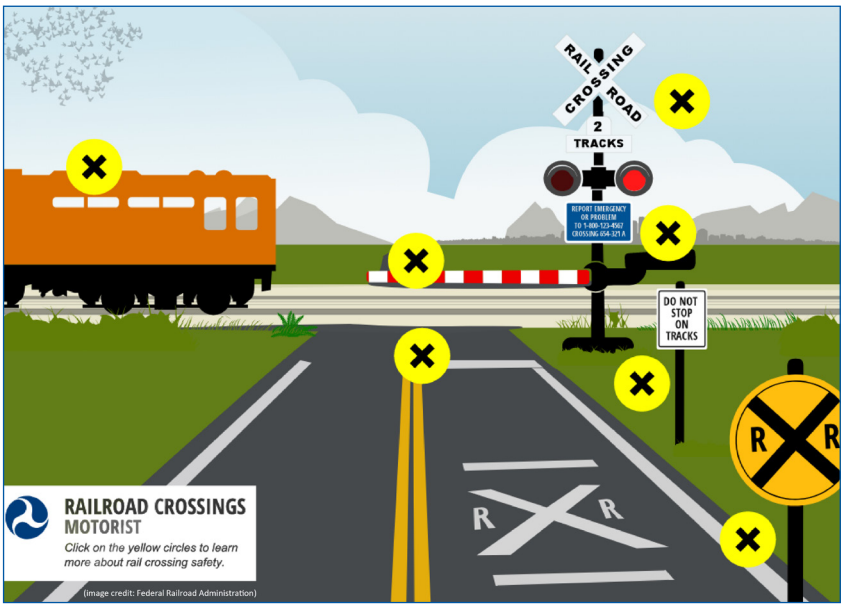
MOUNTAIN-PLAINS CONSORTIUM

MPC 18-354 | P. Lu, D. Tolliver, and Z. Zheng

Highway-Rail Grade Crossing Traffic Hazard Forecasting Model



MOUNTAIN-PLAINS CONSORTIUM REGION 8

A University Transportation Center sponsored by the U.S. Department of Transportation serving the Mountain-Plains Region. Consortium members:

Colorado State University
North Dakota State University
South Dakota State University
University of Colorado Denver
University of Denver
University of Utah
Utah State University
University of Wyoming

Highway-Rail Grade Crossing Traffic Hazard Forecasting Model

Prepared by

Dr. Pan Lu
Associate Professor/ Associate Research Fellow
Pan.lu@ndsu.edu
Tel: 701-212-3795

Dr. Denver Tolliver
Director
Denver.Tolliver@ndsu.edu
Tel: 701-231-7190

Zijian Zheng
Graduate Assistant
Zijian.Zheng@ndsu.edu

Department of Transportation and Logistics/College of Business
Mountain-Plains Consortium/Upper Great Plains Transportation Institute
North Dakota State University

September 2018

Acknowledgements

This research was made possible with funding supported by the U.S. Department of Transportation though the Mountain-Plains Consortium (MPC) Transportation Center. The authors express their deep gratitude to the U.S. DOT and MPC.

Disclaimer

The authors alone are responsible for the preparation and the accuracy of the information presented in this report. The document is disseminated under the sponsorship of the Mountain-Plains Consortium in the interest of information exchange. The U.S. Government assumes no liability for the contents or use thereof.

Two journal publications from the research were published before this report was finalized, and they are as follows:

1. Pan Lu, and Denver Tolliver. "Accident Prediction Model for Public Highway-Rail Grade Crossings." *Accident Analysis & Prevention*, 90, 73-81, 2016
2. Zijian Zheng, Pan Lu, and Denver Tolliver. "Decision Tree Approach to Accident Prediction for Highway-Rail Grade Crossings: Empirical Analysis." *Transportation Research Record*, 2545, 115-122, 2016

Two more journal publications out of the research were under review.

ABSTRACT

The highway-rail crossing accident has been long recognized as a priority safety concern for worldwide rail industries and researchers because traffic crashes at highway-rail crossings are often catastrophic with serious consequences, which include fatalities, injuries, extensive property damage, and delays in both railway and highway traffic. Relatively few studies have focused on investigating accidents at highway-rail crossings. Salmon et al. (2013) indicated that because of limited research efforts, various aspects of highway-rail crossing safety performance remain poorly understood. Therefore, a safety evaluation (i.e., accident frequency prediction) of highway-rail crossings is needed to re-examine both prediction methods and contribution factors (Austin & Carson, 2002).

Generalized linear models (GLMs) have been frequently used in highway safety studies to explore the relationship between crash likelihood and contributors and to forecast future highway rail grade crossing accident likelihood because they are believed to be better suited for discrete and non-negative crash frequency data. However, GLMs have several limitations, such as a pre-defined underlying relationship between target variable and predictors and their limitations to fit dynamic non-linear relationships. Non-parametric data mining methods are gaining popularity because they are not required to pre-define the underlying relationship between dependent and independent variables. They also model non-linear relationships among variables with missing data and between contributor variables and predictors.

This research seeks to investigate highway rail grade crossing (HRGC) crash predicting models and contributing factors by exploring the application of GLM and data mining models. In summary, data mining models can serve as great alternative modeling tools to perform crash forecasting with relatively accurate forecasting power and strong ability to model non-linear relationships between contributors and crash likelihood. All the models will provide different sets of contributors. However, decision tree models may be hard to apply due to their large tree structure. Since GLM models are parametric, they tend to pick a limited number of explanatory variables; data mining algorithms, also considered as non-parametric algorithms, tend to select more contributor variables. However, the top contributors identified by all the methods agree with each other on traffic exposure variables, such as highway traffic volume, rail traffic volumes, and their travel speed, and also on some crossing characteristics such as warning devices.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1. INTRODUCTION

1.1 Background

A highway-rail grade crossing (HRGC) is the intersection where a highway and a railway cross at the same elevation or grade. HRGCs are critical spatial locations of utmost importance for transportation safety because traffic crashes at HRGCs are often catastrophic with serious consequences, including fatalities, injuries, extensive property damage, and delays in railway and highway traffic (Raub, 2009.) The consequences can be exacerbated if collisions involve freight trains carrying hazardous materials, which could spill and create an environmental disaster and increased danger to those nearby. From 1996 to 2014, 26% of RGC accidents in North Dakota involved hazardous material. The need to improve traffic safety has been a major social concern in the United States for decades. Transportation agencies and other stakeholders must identify the factors that contribute to the likelihood of an RGC collision to better predict crash probability and provide direction for RGC designs and policies that will reduce crash numbers.

1.2 Research Objectives

The primary objective of this project is to explore forecasting models that can identify impacted grade crossings in terms of future safety upgrades.

The following major tasks have been included in the scope of the study:

1. Explore and compare HRGC crash forecasting models, including statistical models and data mining models
2. Develop, demonstrate, and validate the crash forecasting models and conduct contributor variable analysis
3. Train one Ph.D. student on the various theoretical and applicable methods employed
4. Develop publications and associated reports

These research objectives will further the overall goals of promoting economic development, safety, interdisciplinary education, workforce development, and technology transfers that serve the critical needs of the Mountain-Plains Region.

1.3 Report Organization

This section introduces the report's organization, which is as follows:

- Section 2 conducts a complete literature review on the crash forecasting models.
- Section 3 introduces the data used in this research.
- Section 4 summarizes the application findings with statistical analysis.
- Section 5 summarizes the application findings with data mining analysis.
- Section 6 summarizes the conclusions and recommendations from the study.

2. LITERATURE REVIEW

GLM models all require pre-defining the underlying distribution relationship between contributors and predictors. Poisson regression has been commonly used to model crash frequency because of the discrete and non-negative nature of crash data. However, Lord and Mannering (2010) pointed out that, although the GLMs possess desirable elements for describing accidents, these models face various data challenges and are shown to be a potential source of error in terms of incorrectly specifying statistical models that can result in incorrect predictions and explanatory factors. The most common crash data underlying distribution are over- or under-dispersion. Over-dispersion is where sample variance is greater than sample mean. In many collision data bases, the variance in accident frequency is higher than the mean. Over-dispersion arises from the unmeasured uncertainties associated with the observed or unobserved variables (Lord & Park, 2008). On rare occasions, crash data can display under-dispersion where the sample variance is less than the sample mean (Oh, Washington, & Nam, 2006). These issues are problematic (Lord & Mannering, 2010), because the most common count-data modeling approach requires that the variance be equal to the mean. Over- and under-dispersed data would lead to inconsistent standard errors for parameter estimates when using the traditional Poisson distribution (Cameron & Trivedi, 1998). Because of this, Poisson regression is usually a good modeling starting point (Oh, Washington, & Nam 2006). When data show over-dispersion, some modifications to the standard Poisson model are available to account for over-dispersion, such as Poisson-gamma or the negative binomial (NB) model (Lord & Mannering, 2010). When under-dispersion arises, less common models, such as the gamma probability count model, is believed to be capable of handling under-dispersion issues (Oh, Washington, & Nam, 2006). Poisson and NB or Poisson-gamma models have been shown to have great limitations when applied to under-dispersed crash data (Oh, Washington, & Nam, 2006).

The statistical analysis part of this research will explore the potential GLM model options to handle under-dispersed HRGC crash data by 1) demonstrating the general forms of various models and 2) Investigating and comparing models that may handle under-dispersed data with public North Dakota HRGC crash data analysis in this research. That is followed by a data mining algorithm analysis to explore the non-linear relationship and forecasting power.

2.1 Literature Review on Selected GLM Algorithms

2.1.1 Poisson Model

Non-negative integer count data are often approximated well by the Poisson regression model. In a Poisson regression model, the probability of HRGC i having y_i crash (where y_i is the expected number of 0, 1, 2,…) is given by:

$$P(y_i) = \frac{e^{(-\lambda_i)}(\lambda_i{}^{y_i})}{y_i!} = \frac{e^{-\mu}(\mu^{y_i})}{y_i!} \quad (1)$$

Where, λ_i is the predicted count or Poisson parameter for HRGC i, which is equal to HRGC i's expected number of crashes per year, E[y_i] or μ. The Poisson model is specifying the Poisson parameter λ_i as a function of explanatory variables, and the most commonly selected functional form (or function link) is in log-linear form:

$$\log(\lambda_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_m x_{im} \quad (2)$$

where βs are the estimated regression coefficients and xs are the explanatory variables. One important property of the Poisson distribution model is that it restricts equal mean and variance of the distribution:

$$Var[Y] = \mathrm{E}[\mathrm{Y}] = \mu \quad (3)$$

If the mean is not equal to the variance of the crash counts, then the data are said to be either over- or under-dispersed, and the resulting parameter estimate will be biased (Cameron & Trivedi, 1998). Crash data have been found to often exhibit over-dispersion due to unmeasured variances associated with the observed or unobserved variables (Lord & Park, 2008).

2.1.2 Negative Binomial Model

The negative binomial or Poisson-gamma mixture model is a variant of the Poisson model designed to deal with over-dispersed data. The negative binomial model relaxes the constraint of equal mean and variance. The model assumes that the Poisson parameter λ_i follows a gamma probability distribution as:

$$\log(\lambda_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_m x_{im} + \varepsilon_i \quad (4)$$

where $\mathrm{Exp}(\varepsilon_i)$ is a gamma distributed error term with other variables as defined earlier. The addition of this term allows the sample mean to differ from the sample variance such as:

$$Var[Y] = E[Y](1 + kE[Y]) = E[Y] + kE[Y]^2 = \mu + k\mu^2 \quad (5)$$

where k is a second ancillary or heterogeneity parameter and often refers to the dispersion parameter (Saha & Paul, 2005), if k equals 0, the negative binomial model reduces to the Poisson model. As an NB model, counts are gamma distributed as they enter into the model, and k also enters into the model as a measure of over-dispersion in the data.

The NB model is probably the most frequently used model in traffic crash data analysis; however, the NB models suffer limitations to handle under-dispersed data. If the dispersion parameter, k, is set as a negative value to try to handle the under-dispersion issue, it would make the count no longer gamma distributed and lead to unreliable parameter estimates, especially when sample mean is low and sample size is small (Saha & Paul 2005; Oh, Washington, & Nam, 2006).

2.1.3 The Gamma Model

The gamma model has been proposed by Oh et al. (2006) to handle under-dispersed HRGC crash data. The gamma count model for count data is given as:

$$\Pr(y_i = j) = Gamma(\alpha j, \lambda_i) - Gamma(\alpha j + \alpha, \lambda_i) \quad (6)$$

where $\lambda_i = \exp(\beta X_i)$ and λ_i is the mean of the crashes, μ.

$$Gamma(\alpha j, \lambda_i) = \begin{cases} 1, & if\ j = 0 \\ \frac{1}{\Gamma(\alpha j)} \int_0^{\lambda_i} u^{\alpha j - 1} e^{-u} du & if\ j > 0 \end{cases} \quad (7)$$

where α is the dispersion parameter, if α <1, there is over-dispersion; if α>1, there is under-dispersion; and if α =1, the gamma model reduces to the Poisson model, which would indicate the model can handle both under- and over-dispersion. Although the model is flexible enough to handle crash data, its

limitations constrain its applications (Lord & Mannering, 2010), these limitations include a time-dependent observation assumption (Cameron & Trivedi, 1998) and a two-state characteristic to handle zero observations separately (Lord & Mannering, 2010), because general gamma distribution restricts discrete responses to positive values.

2.1.4 The Conway-Maxwell-Poisson Model

The Conway-Maxwell-Poisson (CMP) is a generalization of the Poisson distribution that enables it to model both under- and over-dispersed data. The CMP is defined to be the distribution with probability mass function:

$$P(y_i) = \frac{1}{Z(\lambda_i, v_i)} \frac{{\lambda_i}^{y_i}}{(y_i!)^{v_i}} \qquad for\ y_i = 0, 1, 2, ... \tag{8}$$

Note that, except for the normalization factor, $Z(\lambda_i, v_i)$, which equals to $\sum_{n=0}^{\infty} \frac{\lambda_i^n}{(n!)^i}$, the CMP distribution is very similar to Poisson distribution with an extra parameter, v_i, which can take any non-negative value. The CMP distribution overcomes the requirement that mean and variance are equal by introducing υ to allow flexibility in modeling the tail behavior of the distribution. If $\upsilon = 1$, the distribution is Poisson distribution. If $\upsilon < 1$, the distribution will have longer tails than the Poisson distribution and can model over-dispersed data. A special case in this situation is, when $\upsilon = 0$ and $\lambda < 1$, the distribution is geometric distribution, an extreme over-dispersion. If $\upsilon > 1$, the distributions have shorter tails than the Poisson distributions and can model under-dispersed data. Another special case in this situation is, when $\upsilon \to \infty\ and\ \lambda < 1$, the distribution is Bernoulli distribution, an extreme under-dispersion, and the data can only take the values of 0 and 1.

The CMP distribution does not have closed-form expressions for its moments in terms of the parameters $\upsilon\ and\ \lambda$, approximated by Shmueli et al. (2005), and mean and variance are estimated by:

$$E[Y] \approx \lambda^{1/\upsilon} + \frac{1}{2\upsilon} - \frac{1}{2} \tag{9}$$

$$Var[Y] \approx \frac{1}{\upsilon} \lambda^{1/\upsilon} \tag{10}$$

Guikema and Coffelt (2008) proposed a re-parameterization of the Shmueli et al. (2005) model in which and mean and variance can be approximated as:

$$E[Y] \approx \mu + \frac{1}{2}\upsilon - \frac{1}{2} \tag{11}$$

$$Var[Y] \approx \frac{\mu}{\upsilon} \tag{12}$$

The dispersion is defined as:

$$D[Y] = \frac{Var(Y)}{E(Y)} \approx \frac{1}{\upsilon} \tag{13}$$

2.1.5 The Bernoulli Model

The Bernoulli distribution model is a logistic model that restricts responses and follows the Bernoulli distribution and only has two possible outcomes, "failure" or "success" (0 or 1). The distribution is given as:

$$P(y_i) = \mu^{y_i}(1-\mu)^{(1-y_i)} \qquad for\ y_i = 0,1,2,\ldots. \tag{14}$$

For the Bernoulli distribution, the response is binomial events. The variance is:

$$Var[Y] = \mu(1-\mu) \tag{15}$$

The Bernoulli model only can handle under-dispersed data based on the relationship between sample variance and mean.

2.1.6 The Hurdle Poisson Model

The hurdle Poisson model allows for a systematic difference in the statistical process governing observations with zero counts and those with a positive number of counts. One part of the hurdle model is a binary outcome model (logistic) governing observations with zero and positive counts; the second part of the model is a truncated-at-zero Poisson count model for observations with positive counts. The hurdle model is not only flexible enough to handle both under- and over-dispersion but also can account for "excess zeros." The probability distribution of the hurdle model is given as:

$$f(y_i) = \begin{cases} f_1(z_i\gamma) & y_i = 0 \\ \frac{1-f_1(z_i\gamma)}{1-f_2(0)}\, f_2(y_i) = \phi\, f_2(y_i) & y_i = 1,2,\ldots.. \end{cases} \tag{16}$$

Where $f_1(z_i\gamma)$, the probability of extra zeros, π, is the density function of the logistic model with explanatory variables z_i and parameters γ. $f_2(y_i)$ is the probability density function of a truncated Poisson regression model. The numerator of ϕ is the probability of crossing the hurdle and if the numerator is the same as the denominator, which is the sum of probabilities of positive counts, ϕ=1, which will reduce the hurdle model back to a one-stage parent model. When excess zeros exist, ϕ>1. The variance and mean are defined as:

$$Var[Y] = \phi\lambda_2(\lambda_2+1) - (\phi\lambda_2)^2 \tag{17}$$

$$Var[Y] = E[Y] + \frac{1-\phi}{\phi}(E[Y])^2 \tag{18}$$

where λ_2 is the expected value of the un-truncated parent distribution. Under-dispersion is obtained for $1< \phi < \frac{\lambda_2+1}{\lambda_2}$ and when $0< \phi < 1$ over-dispersion is obtained.

2.1.7 Zero-inflated Poisson Model (ZIP)

Zero-inflated models are theorized to account for "excess zeroes," which means zeroes observed in the data base are beyond the number of zeroes predicted by Poisson models. The model is a mixed model with two components and is a dual-state model like the hurdle model. The difference between the ZIP and hurdle models is that only one component of the hurdle model corresponds to the zero-generating process, but both components in the ZIP model govern the zero-generating process.

One part of the ZIP model is a binary outcome model (logistic) that governs observations with excess zeros and non-zero counts. The second part of the model is governed by a Poisson distribution that generates counts, some of which can be zero. ZIP assumes that events, Y = $\{y_i\}$, are independent, and the probability distribution of the hurdle model is given as:

$$f(y_i) = \begin{cases} f_1(z_i\gamma) + (1 - f_1(z_i\gamma))f_2(0) & y_i = 0 \\ \left(1 - f_1(z_i\gamma)\right) f_2(y_i) & y_i = 1, 2, \ldots\ldots \end{cases} \quad (19)$$

where $f_1(z_i\gamma)$ is the probability of extra zeros, π, $f_2(y_i) = \frac{\lambda^{y_i} e^{-\lambda}}{y_i!}$ is the probability density of Poisson, and λ_i is the expected Poisson count for the ith individual. The variance and mean are defined as:

$$E[Y] = \lambda(1 - \pi) \quad (20)$$

$$Var[Y] = \lambda(1 - \pi)(1 + \lambda\pi) \quad (21)$$

2.2 Literature Review on Data Mining Algorithms

2.2.1 Decision Tree Model

A decision tree (DT), a non-parametric data mining method without any requirement to pre-define the underlying relationship between dependent and independent variables, has a powerful capability for detecting patterns in a large data set. Unlike GLM models, it substitutes surrogate splitters for missing primary splitters. The surrogate splitters mimic the primary splitters (SAS Institute Inc.). Although the DT method is relatively new in HRGC crash rate studies, it shows a strong predictive ability and is widely used in economy, business, agriculture, and other fields (Raorane & Kulkarni, 2012).

A DT is a hierarchical tree-based prediction model. There are two types of DT models: classification tree and regression tree. A classification tree is developed for categorical target variables; whereas, a numerical target variable will be fitted with a regression tree. The target variable in this study is discrete with two outcome levels: crossings with a crash and crossings without a crash. Thus, a classification tree will be generated.

Generally, DT development involves three steps. The first is tree growth. As shown in Figure 2.1, at the beginning, all data are concentrated in the root node.

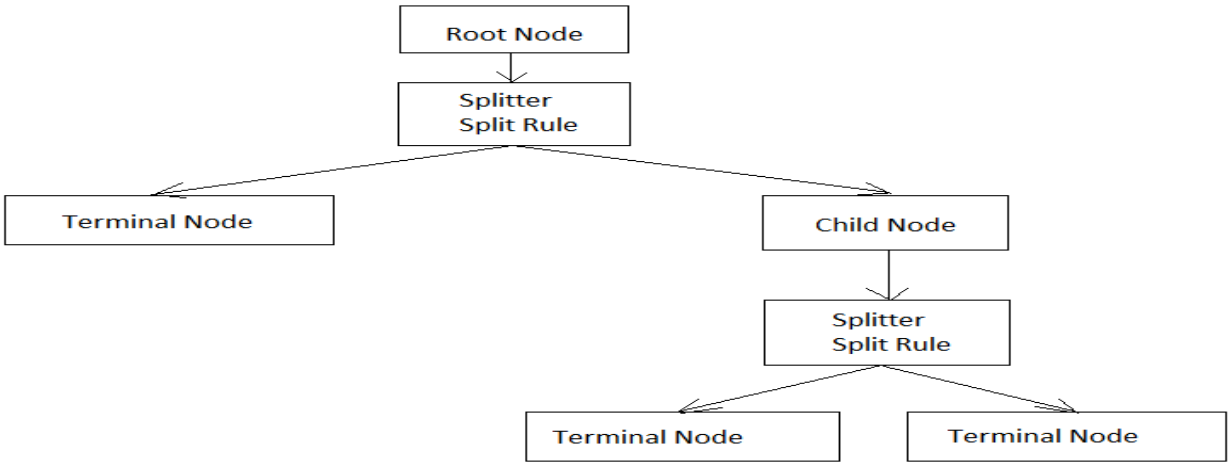


Figure 2.1 Structure of a Typical Decision Tree

The data set is then broken down into child nodes by applying a series of splitting variables (splitters). Each child node will be treated as a parent node for further splitting. The principle behind splitting is to ensure each child node is as homogeneous as possible after splitting. The ID3 algorithm measures entropy, expected entropy, and information gained to decide if a variable should be chosen as the splitter, and whether or not the node can be further split (Sayad, 2010). Entropy measures the amount of unpredictability in an event. The higher the entropy value, the harder it is to predict the outcome of an event. If a sample is completely homogeneous, the entropy value should be zero. For a variable *S* with *c* distinct values, the entropy *E*(*S*) of *S* is calculated as Equation (22): (Freitas, 2013)

$$E(S) = \sum_{i=1}^{c} - p_i \log_2 p_i \tag{22}$$

Where p_i is the probability of taking a certain value, *I* is the index number of options.

If variable *S* is divided into subsets: S_1 , … S_c by certain splitters, the expected entropy (*EH*) measures the expected unpredictability of these *c* outcomes of variable *S* after splitting, and calculated as:

$$EH = \sum_{i=1}^{c} \frac{a_i}{a} \times (-p_i \log_2 p_i) \tag{23}$$

Where: a_i is the number of observations in each subset S_1 , … S_c , and *a* is the total number of observations in parent node *S*.

The difference between *EH* and *E*(*S*) is called the reduction in entropy or information gain (*I*), shown in Equation (24). Information gain measures how much a splitter can help predict the outcomes. The variable that generates the highest information gain discriminates the parent node into the most homogeneous child nodes. Thus, after computing the information gain for candidate variables, the one with the highest information gain will be selected as the splitting variable.

$$I = E(S) - EH \tag{24}$$

A node with an information gain 0 is considered as a terminal node, which means no further splitting can be performed, and the data in each terminal node will be the most homogenous. After applying the steps above recursively, a saturated tree is obtained. The saturated tree provides a best fit to the training data,

but also ends up over-fitting the data set. Thus, the data set is divided for training and validating. The training data are used for splitting the nodes, and validating data are for measuring the misclassification rate in the pruning step. After a sequence of pruned trees are established, the last step is to select the optimal one from the sequence of pruned trees, based on a measurement of the misclassification rate of validation data.

2.2.2 Gradient Boosting Model

The gradient boosting (GB) method is an extension of DT algorithm, which is also known as multiple additive trees. The GB method theoretically extends and improves the simple DT model using stochastic gradient boosting (Friedman, 2001). A GB model can be viewed as a series expansion approximating the true functional relationship (Salford-Systems). In general, the GB model starts by fitting the data with a simple DT model, which has a certain level of error in terms of fitness with the data. The simple DT model is referred to as a weak learner. Considering the errors having the same correlation with outcome value, the GB model then develops another DT model on the errors or the residuals of the previous tree. The detailed algorithm of GB is described as follows (De'ath, 2007; Hastie, Tibshirani, & Friedman, 2009):

$$f(x) = \sum_n f_n(x) = \sum_n \beta_n g(x, \gamma_n) \tag{25}$$

where x is a set of predictors, and $f(x)$ is the approximation of the response variable. $g(x,\gamma_n)$ are single decision trees with the parameter γ_n indicating the split variables. β_n (n=1,2,…,n) are the coefficients and determine how each single tree is to be combined.

The iterative tree-building process keeps adding trees until all observations are perfectly fitted, and iterative training will stop when the performance of the model reaches a point where the model predicts well for both the training and testing data sets.

2.2.3 Neural Network Model

Neural Network (NN) is a progressive learning process inspired by the biological neural network of animal brains. Information is processed and passes through the NN by a group of connected units or nodes called neurons (analogous to biological neurons of animal brains). A typical NN structure is shown in Figure 2. A typical NN contains three layers: input, intermediate (also called the hidden layer), and output. Each neuron in the input layer is one predictor, denoted as Xi in Figure 2.2. A hidden layer is a layer of neurons transferring information from inputs into outputs. Several hidden layers can be placed between the input and output layers. The value of a neuron in the input layer is transferred into hidden layers through a transformation function. The weight (Wij) represents the ratio of transformed value to the value of the input variable. The downstream is computed as the summation of the values of neurons in the upstream layer multiplied by the corresponding connection weights (W in Figure 2.2). Information transfers from hidden layers to the output layer through an activation function. In this study, the target variable, crashes, is defined as a two-level variable: 0 and 1 indicating non-event level and event level, respectively. Thus, in this research, the binary step function is suitable for activation function and expressed as Equation (26) (McCulloch & Pitts, 1943):

$$f(x) = \begin{cases} 0 & for \quad x < threshold \\ 1 & for \quad x > threshold \end{cases} \tag{26}$$

where x is predicted value. When x is greater than a defined threshold, the predicted output is 1, otherwise, 0.

The NN will be initialized with random weights and run through the model for the first time. This run is very unlikely to result in the optimal solution. Thus, in the following iterations, the model will change the weights to get a smaller error. This process will repeat numerous times until the desired output agrees within some predetermined tolerance. The entire procedure is called back propagation.

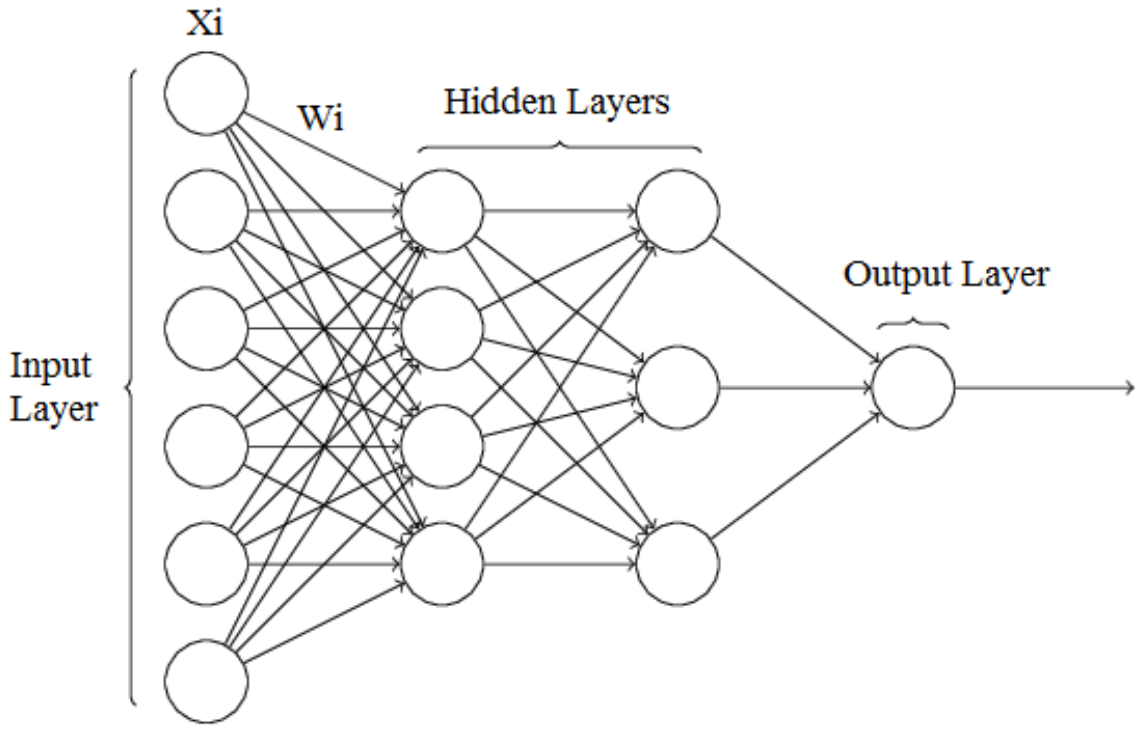


Figure 2.2 Structure of Neural Network

3. DATA

3.1 Data Sources

With consideration given to data size and needs for local RGC crash analysis for North Dakota, data for this investigation were extracted from public RGC data in the state from 1996 to 2014. In this research, 344 public highway-rail grade crossing accidents occurring from 1996 to 2014 were selected from a sample of 5,551 highway-rail grade crossings records.

Data to support the research came from two resources: (1) the FRA's Office of Safety accident/incident database and (2) the FRA's Office of Safety highway-rail crossing inventory. The accident/incident database provides accident-related information for each accident occurrence. The highway-rail crossing inventory database describes each crossing's location, traffic conditions, infrastructure equipment, and historical accident information. A new data set was generated by using the highway-rail grade crossing identification number in both data sets to include data elements in both data sets for each crossing.

3.2 Data Analyzed

All explanatory variables, including warning devices, highway pavement condition, appearance of pavement markings, appearance of interconnection/pre-emption, smallest crossing angle, appearance of pullout lane, functional classifications of highway, train traffic density, highway user types, weather conditions, track conditions, highway traffic density, maximum train speed, location, accident history, and commercial power availability were investigated and tested and are shown in Table 3.1.

A binary target variable (ACCIDENT) is defined with two classes: a value of 1 indicates that there was a crash, while value of 0 represents a crossing with no crash. These variables can be divided into three categories: traffic characteristics, highway characteristics, and crossing characteristics. Traffic characteristics record traffic information at crossings. These characteristics describe highway and railway traffic volume and traffic speed. Highway characteristics contain highway design information at crossings, such as number of highway lanes, pavement, and highway system levels. Crossing characteristics describe warning devices and other crossing related characteristics.

3.3 Data Management Plan

Detailed metadata and archiving information regarding the data used in this analysis are documented in Appendix A.

Table 3.1 Input Variable Description

Variable	Property	Description
ACCIDENT	Target	1= crash happened, 0=no crash
ID	ID variable	Crossing identification
Traffic Characteristics		
AADT_N	Numeric	Annual average daily traffic
AVERAGE_TRAIN_SPEED	Numeric	Average train speed
DAYSWT	Numeric	Day switching train movements
DAYTHRU	Numeric	Day through-train movements
NGHTSWT	Numeric	Night switching-train movements

NGHTTHRU	Numeric	Night through-train movements
SCHLBUS	Numeric	Average number of school bus passing over the crossing on a school day
Highway Characteristics		
Highway_Paved	Category	Is highway paved or not? 1=yes, 0=no
Highway_Stop	Category	Highway stop sign presence: 1=yes, 0=no
HWYSYS	Category	Highway system: 1=interstate national highway system, 2=other national highway system, 3=federal-aid highway system, 4=non-federal-aid highway system
TRUCKLN	Category	Are truck pull-out lanes present? 1=yes, 0=no
Crossing Characteristics		
ADVWARN	Category	Railroad advance warning signal presence: 1=yes, 0=no
COMPOWER	Category	Commercial power availability: 1=yes, 0=no
DOWNST	Category	Does track run down a street? 1=yes, 0=no
FLASHMAS	Numeric	Number of mast mounted flashing lights in pairs
FLASHNOV	Numeric	Number of cantilevered flashing light not over traffic lane
FLASHOV	Numeric	Number of cantilevered flashing light over traffic lane
FLASHPAI	Numeric	Number of flashing light in pairs
GATES	Numeric	Number of gates
Near_City	Category	In or near city? 1=near city, 0=in city
PAVEMRK	Category	Pavement markings: 1=less than 75 ft., 2=75 to 200 ft., 3=200 to 500 ft., 4=N/A
SGNLEQP	Category	Is track equipped with train signals? 1=yes, 0=no
SPSEL	Category	Train detection: 1=constant warning time system (CWT), 2=Direct current audio frequency overlay (DC/AFO), 3=N/A
STOPSTD	Numeric	Number of highway stop signs
TOTAL_NUMBER_TRACK	Numeric	Number of rail tracks
TRAFICLN	Numeric	Number of traffic lanes crossing railroad
WHISTBAN	Category	Quiet zone: 1=24 hours, 2=partial, 3=unknown, 4=no
WIGWAGS	Numeric	Number of wigwags
XANGLE	Category	Smallest crossing angle: 1=0-29, 2=30-59, 3=60-90
XBUCK	Numeric	Number of cross buck
No-Historical Accident	Category	1=with historical accident and 0=no historical

4. GLM ANALYSIS RESULTS AND DISCUSSIONS

4.1 Result Analysis

The estimation results of the saturated Poisson model are presented in Table 4.1. Seven variables are found to be statistically significant in determining accident likelihood, and under-dispersion may exist according to fractural lack-of-fit value, Value/DF=0.69, which is significantly different from 1. This reveals the data may be under-dispersed.

Table 4.1 Poisson Estimation Results

Variable	**Estimated**	**P-Statistic>ChiSq**	**Type 3 Analysis P**
Intercept	-1.3266	0.0001	<0.0001
Cross Buck	-0.6576	0.024	<0.0001
Gate	-0.2664	0.3895	
No Control	-2.9525	0.0046	
Flashing	-1.2687	0.057	
Stop Sign	-	-	
AADT	0.0001	<0.0001	0.0004
Train per Day	0.0267	<0.0001	<0.0001
Track Numbers	0.1763	0.0079	0.01
Paved Highway	0.5764	0.0001	0.0002
Max Train Speed	0.0149	<0.0001	<0.0001
No Historical Accident	-2.4058	<0.0001	<0.0001
AIC	1939		
BIC	2012		
Pearson Chi-Square	3888.66		
Value/DF	0.69		
Log likelihood	-958.579		

As shown in Table 4.1, highway traffic, train traffic, number of tracks, and max train speed all significantly and positively influence accident likelihood. Certain types of warning devices will decrease the accident likelihood compared with “stop sign” warning devices only. When the highway is paved, the accident likelihood increases compared with an unpaved highway. Likewise, when a crossing has no historical accident record, the likelihood of an accident decreases. All the results indicate that the Poisson model is a good first choice for investigating crash data. However, as mentioned earlier in this report, the negative binomial model, which is suggested by many researchers as a modeling method for crash data, is not appropriate to handle the under-dispersed data, which may occur in this research data set. To continue to improve the crash frequency regression model while suspecting possible under-dispersion, attempts were made to analyze accident frequency from under-dispersed data with Zero-Inflated-Poisson (ZIP), NB, Poisson Hurdle (PH), Bernoulli, and Conway-Maxwell-Poisson models using least squares regression techniques. Where ZIP and NB models are selected for their popularity in the literature, PH, Bernoulli, and CMP are selected to handle under-dispersed data.

4.2 Model Comparison

Table 4.2 compares statistically significant contributing variables and model goodness of fit (GOF) statistics, including AIC, BIC, Pearson chi-square statistics, degree of freedom (DF), and log likelihood (LL) statistics.

Smaller AIC and BIC values indicate a better fit. If the model fits the data perfectly without any dispersion, the Pearson chi-square is roughly equal to the model degree of freedom. In other words, the closer the Pearson chi-Square is to the DF, the better the model fits the data (SAS, 2011). The LL statistic is calculated by taking the logarithm of the estimated likelihood for each observation and summarizing those log-likelihoods. In the closer-to-zero sense, the larger LL indicates the better model (UCLA, 2011).

Table 4.2 GOF Model Comparison Results

Variable	Poisson	CMP	Bernoulli	PH	ZIP	NB
Intercept	X	X	X	X		X
Cross Buck						X
Gate						
No Control	X	X	X	X		X
Flashing		X	X	X		
Stop Sign	-	-	-	-	-	-
AADT	X	X	X	X	X	X
Train per Day	X	X	X	X	X	X
Track Numbers	X	X	X	X	X	X
Paved Highway	X	X	X	X		X
Max Train Speed	X	X	X	X	X	X
No Historical Accident	X	X	X	X		X
AIC	1939(5)	1609(2)	1601(1)	1623(3)	1825(4)	1941(6)
BIC	2012(5)	1709(2)	1674(1)	1769(3)	1971(4)	2021(6)
Pearson Chi-Square	3889(5)	4732(3)	4753(1)	4753(1)	4482(4)	3889(5)
DF	5609	5605	5605	5605	5609	5609(5)
Log likelihood	-958.58(5)	-789.49(3)	-789.5(1)	-789.5(1)	-890.7(4)	-958.58(5)

"X" indicates significant parameter and "-" indicates that the variable is a reference variable for warning devices.

As indicated from Table 4.2, all Poisson, CMP, Bernoulli, NB, and Poisson hurdle models provide consistent significant contributory variables, except ZIP model identifying less contributory variables is significant.

4.3 Model Result Discussions

Findings thus far include the following: 1) All GOF variables provide consistent model preferences; for AIC and BIC, the Bernoulli model is the first model preference followed by CMP, PH, ZIP, Poisson, and NB models; for Pearson Chi-Square/DF and LL, both the Bernoulli and PH models are tied as the first preference then followed by CMP, ZIP, Poisson, and NB models. 2) All three proposed models, CMP, Bernoulli, and PH, which potentially can handle under-dispersed data, do fit better than the Poisson model. The three models perform almost equally well for under-dispersed data since the GOF variables' values generated by those three models are very close. 3) If a wrong model, such as the NB model, was selected to fit under-dispersed data, the GOF criteria values will indicate poor model fit compared with the Poisson model. 4) If a model, such as the ZIP model, is selected to solve other issues but not under-

dispersion, the GOF criteria values may show improved model fit compared with the Poisson model as the potential improvement of unnoticed issues, but only with limited improvement since the under-dispersed issue is not solved by ZIP.

Note that the nu parameter, υ, in the CMP model is estimated as 32 for the data set. Recall that if $\upsilon > 1$, CMP can model under-dispersed data, and when $\upsilon \to \infty$, the distribution is reduced to Bernoulli. Observation of nu equals to 32, and the close GOF variables indicate that the data set used in the research may be an extreme under-dispersion and the CMP model performs equally as well as the Bernoulli model.

Of the six tested prediction models, the Poisson is suggested to be the starting model because its equal-dispersion assumption and the values of the crash data are represented by a discrete, non-negative integer. The Poisson model's equal-dispersion assumption can help assess data dispersion.

The Convey-Maxwell-Poisson, Bernoulli, and Poisson hurdle models are suitable models for assessing ND RGC accident data, which exhibited under-dispersion with respect to the Poisson distribution. The Convey-Maxwell-Poisson and Poisson hurdle are flexible models that can accommodate both under- and over-dispersion. The Bernoulli distribution regression model is only appropriate for under-dispersion. All three proposed models have rarely been used in transportation safety literature.

Data mining models will be explored in regard to model forecasting power and contributor variable nonlinear relationships with predictor in the following section.

5. DATA MINING ANALYSIS RESULTS AND DISCUSSIONS

5.1 Decision Tree Structure

Decision tree algorithm results will serve as reference for forecasting power comparison among gradient boosting and neural network algorithms. The final tree structure has 44 terminal nodes and can be viewed in Figure 5.1. Decision tree algorithm is the only method that can produce an interpretable structure among the selected three algorithms, so the further contributing variable sensitivity analysis will be conducted with gradient boosting and neural network. However, due to the tree's unstable structure, such an analysis cannot be conducted with decision tree method.

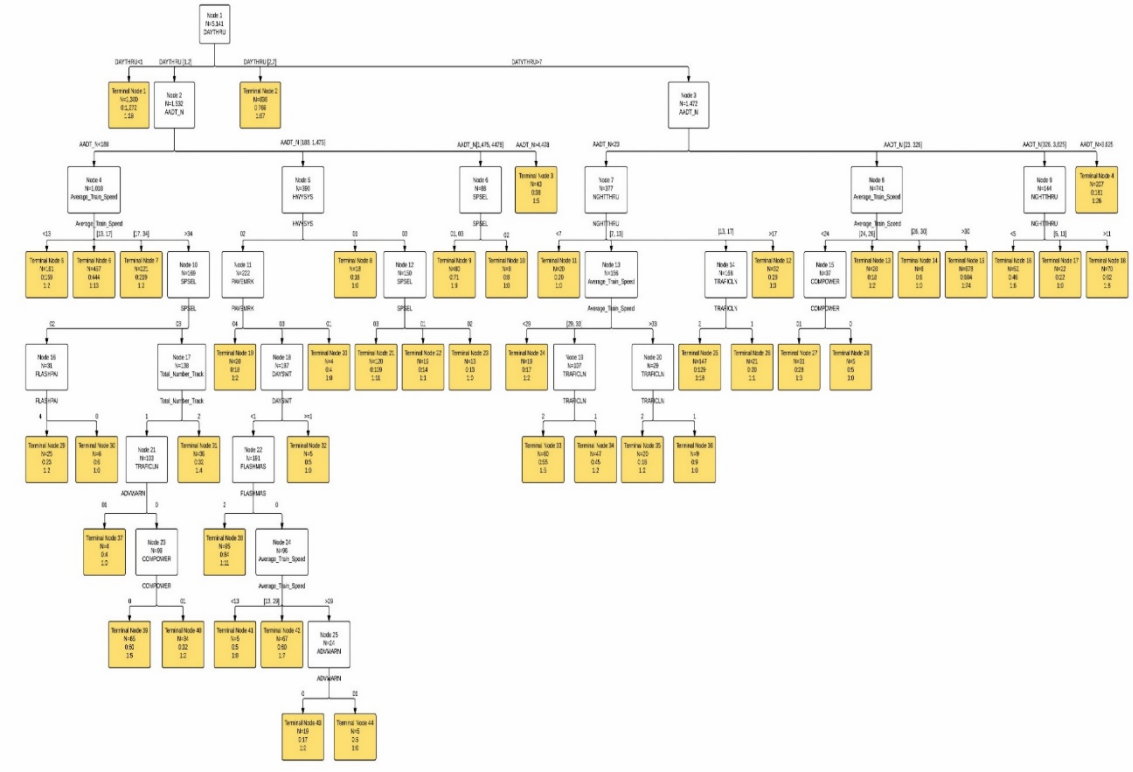


Figure 5.1 Decision Tree Output Structure

5.2 Contributing Variable Importance Analysis

All three algorithms take in the contributing variables for forecasting crash likelihood, and researchers can identify how importantly each variable was used to contribute to the final predicting crash likelihood. This section summarizes the variables important in each method.

As stated earlier, the importance of a variable in a simple single tree is measured by the number of times the variable is used as a splitter and the squared error improvement attributed to the tree due to the splits by the variable. The average value of the summation of those two values is used as the measurement of variable importance in the model. The GB model is an extension of a simple decision tree model so it uses the same algorithm to measure variable importance. Variable importance in a neural network can be measured by different criteria; to be consistent with the other two algorithms, the mean-square-error method is selected to calculate variable importance for NN. This algorithm focuses on predictive

importance since it eliminates one variable at a time and measures the change in mean square error. A variable makes more change in a mean square error, which indicates it is more important.

Table 5.1 summarizes the variables by their importance in predicting crash frequency for all three algorithms. Looking at Table 4, one can tell that three methodologies identify different sets of important variables, and their relative importance levels as measured by how important each variable counted toward forecasting crash likelihood in each of the three methods. However, as one can also tell from Table 4, some top contributors commonly agree with the three methods; those contributors include AADT, train volume, train speed, number of highway traffic lanes, total number of tracks, and crossing controls.

"Important" contributors are calculated based on each algorithm's theory and how they contribute to predicting crash likelihood. In the next section, the researchers conducted a prediction accuracy comparison analysis to demonstrate each model's prediction power. That is followed by a contributor variable sensitivity analysis to demonstrate how each contributor quantitatively contributes to forecasting crash likelihood.

Table 5.1 Variable Importance

Decision Tree Importance Analysis		
Variable	Relative Importance	
Night through train	1	1
Average Train Speed	0.9295	2
Day through train	0.8221	3
AADT	0.737	4
Highway system	0.6703	5
Signal equipped or not	0.6472	6
Highway Paved or not	0.6456	7
Train Detection System	0.6099	8
Number of traffic lane	0.4382	9
Gates	0.4283	10
HMAS Flash	0.4094	11
Total_Number_Track	0.3916	12
Advanced waring system	0.2328	13
Run down street or not	0.2313	14
Number of flashing lights	0.204	15
Cross bucks	0.2021	16
Night switching train	0.1877	17
Day switching train	0.1705	18
Pavement mark	0.0987	19
Commercial power	0.086	20
Truck lane or not	0.0809	21
School bus route or not	0.0734	22
Crossing angle	0.0435	23
Gradient Boosting Importance Analysis		
Variable	Relative Importance	
AADT	1	1
Day through train	0.7333	2
Train Detection System	0.6773	3
Night through train	0.5828	4
Average Train Speed	0.5493	5

Number of traffic lane	0.4562	6
Highway system	0.3798	7
Advanced waring system	0.3719	8
Total_Number_Track	0.3575	9
Pavement mark	0.3415	10
Crossing angle	0.3285	11
Number of flashing lights	0.3124	12
Commercial power	0.2752	13
Highway stop sign	0.2736	14
Near city or not	0.2631	15
Cross Bucks	0.2573	16
Highway Paved or not	0.2193	17
Mounted flashlight	0.2142	18
School bus route or not	0.1978	19
Train signal	0.1945	20
Gates	0.1800	21
Truck lane or not	0.1670	22
highway stop signs	0.1424	23
Run down street or not	0.1358	24
Whistle ban or not	0.1346	25
Day switching train	0.1319	26
Night switching train	0.1180	27
Cantilevered flashing lights	0.0937	28
Neural Network Importance Analysis		
Variable	Relative Importance	
AADT	1	1
Number of traffic lane	1	2
Number of tracks	0.6	3
Mounted flashlight	0.55	4
Day through train	0.54	5
Day switching train	0.53	6
Night through train	0.53	7
Cantilevered flashing lights	0.53	8
Cross Bucks	0.52	9
Advanced waring system	0.51	10
School bus route or not	0.5	11
Gates	0.49	12
highway stop signs	0.44	13
Train speed	0.43	14
Number of flashing lights	0.4	15
Night switching train	0.3	16
Highway Paved or not	0.28	17
Commercial power	0.27	18
Run down street or not	0.26	19
WIGWAG	0.24	20

5.3 Prediction Accuracy Analysis

For crash forecasting, prediction accuracy can be explained in Table 5 and equations (27) to (29). As indicated in Table 5.2, the letters a to d represent the number of the corresponding cases.

Table 5.2 Accuracy Description Table

		Observed Condition	
		Present	Absent
Predicted Condition	Positive	True Positive (a)	False Positive (b)
	Negative	False Negative (c)	True Negative (d)

Overall forecasting accuracy can be described with equation (27). Basically, it described total true estimates (a+d) out of total conditions (a+b+c+d). However, how accurate the models are in terms of crash forecasting is a more critical indicator to measure model performance. Thus, many researchers reported accuracy for crash (event) and non-crash (non-event), and they are described with equations (28) and (29).

$$\text{Accuracy}_{\text{overall}} = \frac{a+d}{a+b+c+d} \quad (27)$$

$$\text{Accuracy}_{\text{Crash}} = \frac{a}{a+c} \quad (28)$$

$$\text{Accuracy}_{\text{non\_crash}} = \frac{d}{b+d} \quad (29)$$

As one can tell from equations (28) and (29), those measurements measure the model forecasting accuracy given the number of observed conditions; but note, they ignore the false positive when accounting for crash accuracy and ignore the false negative when accounting for non-crash accuracy.

All accuracy measurements for both the training and testing data sets for three selected models are shown in Table 5.3. The three models perform very closely to each other. However, one can tell that the gradient boosting model outperformed the other two models in all three measurements, followed by the neural network and the decision tree model for the training data set. Regarding the testing data set, the gradient boosting model outperforms other two models for overall and non-crash accuracy; but for crash accuracy, the decision tree model is the best performer, followed by gradient boosting and neural network.

Table 5.3 Comparison Model Predictive Accuracy

		Overall	Crash	Non-Crash
Training Data set	Decision Tree	77.7%	84.1%	77.2%
	Gradient Boosting	84.4%	88.6%	84.1%
	Neural Network	82.6%	84.8%	82.5%
Testing Data set	Decision Tree	77.1%	88.9%	76.3%
	Gradient Boosting	81.91%	86.1%	81.6%
	Neural Network	79.5%	83.3%	79.2%

As stated earlier, the importance of a variable will be a bit different based on different algorithms and their final optimal model. In this report, the researchers pay more attention to the recommendations provided by the gradient boosting model due to its superior forecasting power.

Recall from Table 4, AADT, daily through train traffic, train detection type, nightly through-train traffic, train speed, and the number of traffic lanes are the top six contributors to crash prediction, with individual influence percentages greater than 5%.

5.4 Variable Sensitivity Analysis

To better understand contributor variables, conducting further sensitivity analysis is needed to indicate their directional and quantitative effects.

A general method to evaluate the effect of explanatory variables on the response variable is to describe the relationship between the predictor variable and the studied contributor variable with partial dependent plots. Figure 5.2 indicates partial plot analysis results for the gradient boosting model with x-axis indicating the contributor variable and y-axis indicating the likelihood of the crash.

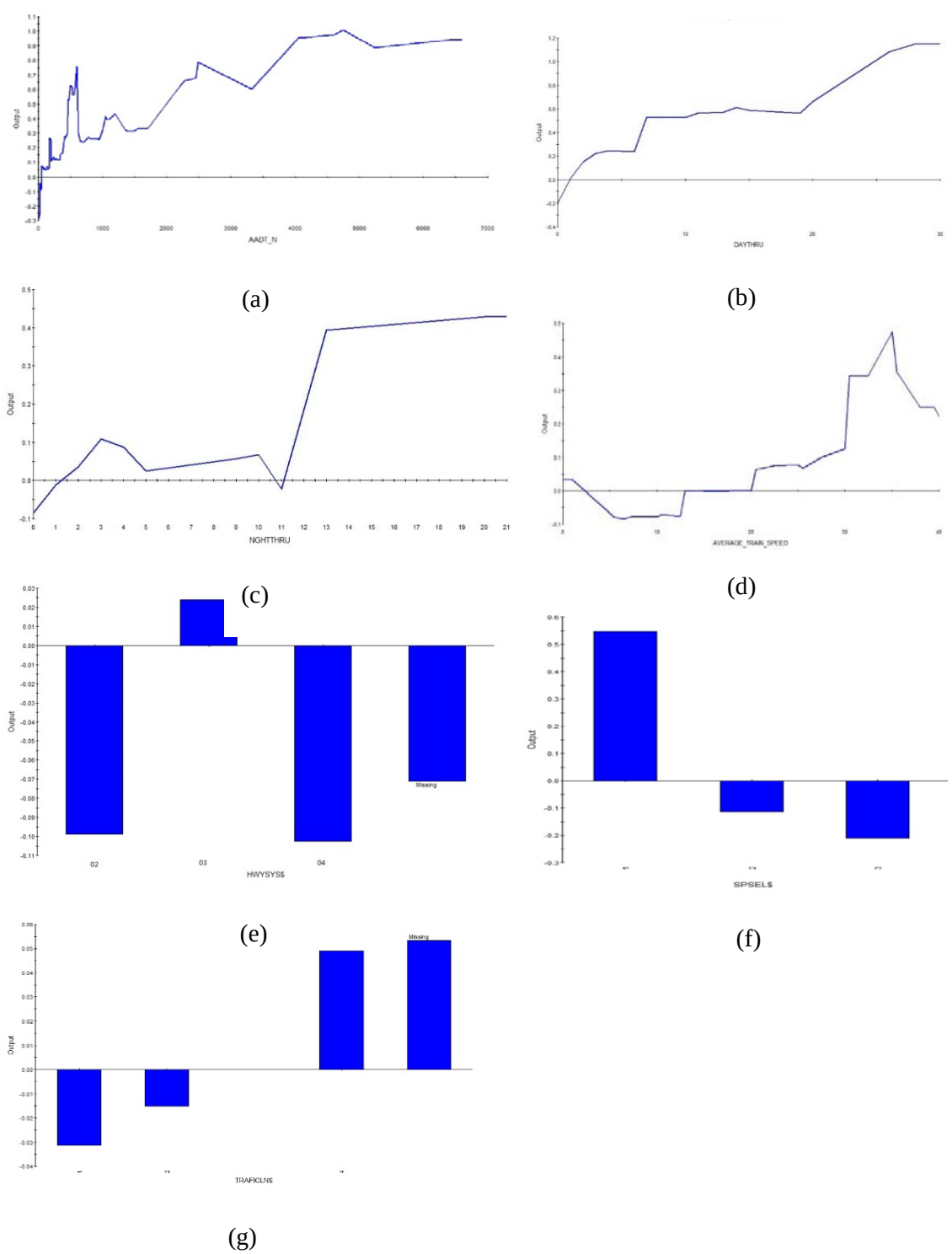


Figure 5.2 Partial dependent plots

One can tell from Figure 5.2(a) that all the contributors exhibit a complex and nonlinear relationship with crashes. However, in general, a positive (increasing) relationship is observed for traffic exposure variables, such as AADT, daily through train traffic, nightly through train traffic, and average train speed. For character variables, the analysis indicates their relative relationship with crashes. For example, for HWYSYS, Figure 5.2(e) shows that incidents tend to occur at crossings with federal-aid highways, and not to occur at crossings with non-interstate highways and non-federal-aid highways. For SPSEL, crashes tend not to happen at SPSEL=2, which is equipped with direct current audio frequency overlay, but tend to happen at SPSEL=1, which is equipped with constant warning time systems. For TRAFICLN, it is suggested that crossings intersect with highways with no more than two lanes as those are less likely to have crashes than highways with four lanes.

6. SUMMARY AND FUTURE RESEARCH

This research explores the modeling options for highway rail grade crossing crash analysis with an empirical analysis of accidents in North Dakota. Six statistical models and three data mining models are explored. Each model has its own advantages and challenges. For data mining models, further prediction accuracy, contributor variable importance analysis, and contributor variable sensitivity analysis are also conducted and compared.

In summary, data mining models can serve as great alternative modeling tools to perform crash forecasting with relatively accurate forecasting power and strong ability to model non-linear relationships between contributors and crash likelihood. All the models will provide different sets of contributors. However, a decision tree model may be hard to apply due to its large tree structure. Since GLM models are parametric, they tend to pick a limited number of explanatory variables; data mining algorithms, also considered as non-parametric algorithms, tend to select more contributor variables. However, the top contributors identified by all the methods agree with each other on traffic exposure variables, such as highway traffic volume, rail traffic volumes and their travel speed, and also on some crossing characteristics such as warning devices.

Throughout the literature search, we found little research conducted on highway rail grade crossing safety analyses compared with highway roadside crash analyses. Highway rail grade crossing safety is a major safety and economic focus in the United States; however, it received limited research effort, and its safety performance remains poorly understood. Therefore, this research re-examined both prediction methods and how their identified contribution factors affect HRGC crashes.

Further research will focus on developing 1) an agency-friendly user tool that allows agencies to conduct contributor sensitivity analyses for various “what-if” scenarios, and 2) an integrated forecasting model for agencies that accounts for both crash likelihood and severity. Moreover, a decision-making tool will also be developed for agencies to allocate their limited safety improvement resources.

REFERENCES

Austin, R., & Carson, J. (2002). "An Alternative Accident Prediction Model for Highway-rail Interfaces." *Accident Analysis and Prevention*, Vol. 34, pp. 31–42.

Cameron, A.C., & Trivedi, P.K. (1998). *Regression Analysis of Count Data*. Cambridge University Press, Cambridge, UK

De'ath, G. (2007). "Boosted trees for ecological modeling and prediction." *Ecology*, 88(1), 243–251. https://doi.org/10.1890/00129658(2007)88[243:BTFEMA]2.0.CO;2.

Freitas, N. (2013). Decision Trees. University of British Columbia. https://www.youtube.com/watch?v=-dCtJjlEEgM. Accessed at Jun. 20, 2015.

Friedman, J. H. (2001). "Greedy function approximation: A gradient boosting machine." *The Annals of Statistics*, 1189–1232. Salford Systems (d). SPM user guide: Introducing TreeNet. http://media.salford-systems.com/pdf/spm7/IntroTN.pdf.

Guikema, S., D., and Coffelt, J., P. (2008). "Modeling count data in risk analysis and reliability engineering." In K. B. Misra. Handbook of Performability Engineering. London, Springer UCLA, 2007. Regression Models with Count Data. http://www.ats.ucla.edu/stat/stata/seminars/count_presentation/count.htm. Accessed June 12, 2015.

Hastie, T., Tibshirani, R. J., & Friedman, J. H. (2009). *The elements of statistical learning* (2nd ed.). NY: Springer-Verlag.

Lord, D., & Mannering, F. (2010). "The Statistical Analysis of Crash-Frequency Data: A Review and Assessment of Methodological Alternatives." *Transportation Research Part A: Policy and Practice*, Vol. 44(5), pages 291-305.

Lord, D., & Park, P., Y. (2008). "Investigating the effects of the fixed and varying dispersion parameters of Poisson-gamma models on empirical Bayes estimates." *Accident Analysis and Prevention*, 40: 1441–1457

McCulloch, W., & Pitts, W. (1943). "A Logical Calculus of the Ideas Immanent in Nervous Activity." *Bulletin of Mathematical Biophysics.* Vol. 5, pp. 115-133.

Oh, A., Washington, S., P., & Nam, D. (2006). "Accident Prediction Model for Railway-highway Interfaces." *Accident Analysis and Prevention*, Vol. 38, pp. 346-356.

Raorane, A., A., and Kulkarni, R., V. (2012). "Data Mining: an Effective Tool for Yield Estimation in the Agriculture Sector." *International Journal of Emerging Trends and Technology in Computer Science*, Vol. 1. Issue 2.

Raub, R.A. (2009). "Examination of Highway–Rail Grade Crossing Collisions Nationally from 1998 to 2007." *Transportation Research Record* 2122 63–71.

Saha, K., & Paul, S. (2005). "Bias-corrected maximum likelihood estimator of the negative binomial dispersion parameter." *Biometrics* 61 (1), 179-185.

Salmon, P.M., Read, G.J.M., Stanton, N. A., & Lenné, M.G. (2013). “The crash at Kerang: Investigating Systemic and Psychological Factors Leading to Unintentional Non-Compliance at Rail Level Crossings.” *Accident Analysis and Prevention*, Vol. 50: 1278–1288.

SAS Institute Inc. (d). Getting Started with SAS Enterprise Miner 14.1. http://support.sas.com/documentation/cdl/en/emgsj/67981/PDF/default/emgsj.pdf

SAS, user’s guide, The GENMOD procedure. http://support.sas.com/documentation/cdl/en/statug/63033/HTML/default/viewer.htm#genmod_toc.htm. Accessed December 7th, 2015.

Sayad, S. (2010). Decision Tree – Classification. An Introduction to Data Mining. http://www.saedsayad.com/decision_tree.htm. Accessed at Jun. 10, 2015.

Shmueli, G., Minka ,T., P., Kadane, J., B., Borle, S., & Boatwright, P. (2005). “A useful distribution for fitting discrete data: revival of the Conway–Maxwell–Poisson distribution.” *Journal of the Royal Statistical Society* C 54 (1), 127–142.

UCLA, 2007. Regression Models with Count Data. http://www.ats.ucla.edu/stat/stata/seminars/count_presentation/count.htm. Accessed June 12, 2015.

APPENDIX A. DATA

The data used in this analysis is archived in SAS data format and the detail about the data metadata is documented in the following tables.

Variable	Label	Variable Description	Variable Values and Description
AADT		Annual average daily traffic	
Average_Train_Speed		Average train travel speed	
Highway_Paved		Is highway paved?	0=Yes; 1=No
ID		Index ID	
Near_City		In or near city	0=In city; 1=Near city
TRUCKLN	TRUCKLN	Number of truck lanes	
Total_Number_Track		Total number of track	
accident		Does the highway-rail grade crossing have an accident record before?	0=No; 1=Yes.
advance_warning	ADVWARN	Does the HRGC equip with advance warning signs?	0=Missing; 1=Yes; 2=No
crossing_angle	XANGLE	Smallest crossing angle.	0=Missing; 1=0-29; 2=30-59
highway_system	HWYSYS	Highway systems	0=Missing 1= Interstate National Highway System; 2= Other National Highway System; 3= Other Federal-Aid Highway-Not NHS; 4= Non Federal-Aid
illumination	COMPOWER	Commercial power available?	0=Yes; 1=No; 2=Missing
lights_in pairs	FLASHPAI	Number of flashing light in pairs	
lights_not_over_lane	FLASHNOV	Number of levered (or bridged) flashing lights not over traffic lane.	
lights_over_lane	FLASHOV	Number of Flashing lights over traffic lane.	
mast_mounted_lights	FLASHMAS	Number of mast mounted flashing lights	
Gates	GATES	Presents of gates.	0=No; 1=Yes.
number_of_crossbucks	XBUCK	Number of crossbucks.	
number_of_trafficlane	TRAFICLN	Number of traffic lanes.	
number_of_wigwags	WIGWAGS	Number of wigwags.	
pavemark_distance	PAVEMRK	Pavemark distance.	0=Missing 1=<75ft; 2=75 to 200ft; 3=200 to 500ft; 4=NA
quiet_zone_hour	WHISTBAN	Is HRGC in quite zone?	0=Yes; 1=No

schoolbus_traffic	SCHLBUS	Number of schoolbus passing over the crossing on a school day.	
stop_sign	STOPSTD	Does highway stop sign present?	0=No; 1=Yes.
track_run_down_street	DOWNST	Does track run down a street?	0=Missing; 1=Yes; 0=No
train_detection_type	SPSEL	Train detection type.	1=constant warning time system; 2=Direct current audio frequency overlay; 3=NA
DAYTHRU	DAYTHRU	Day through train movements	
DAYSWT	DAYSWT	Day switch train movements	
NGHTTHRU	NGHTTHRU	Night through train movements	
NGHTSWT	NGHTSWT	Night switch train movements	

accident	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	5359	93.80	5359	93.80
1	354	6.20	5713	100.00

stop_sign	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	5621	98.39	5621	98.39
1	92	1.61	5713	100.00

Highway_Paved	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0=Yes	4660	81.57	4660	81.57
1=No	1053	18.43	5713	100.00

FLASHOV				
lights_over_lane	**Frequency**	**Percent**	**Cumulative Frequency**	**Cumulative Percent**
0	5652	98.93	5652	98.93
1	7	0.12	5659	99.05
2	49	0.86	5708	99.91
4	3	0.05	5711	99.96
8	2	0.04	5713	100.00

FLASHNOV				
lights_not_over_lane	**Frequency**	**Percent**	**Cumulative Frequency**	**Cumulative Percent**
0	5700	99.77	5700	99.77
1	3	0.05	5703	99.82
2	7	0.12	5710	99.95
4	3	0.05	5713	100.00

FLASHMAS				
mast_mounted_lights	**Frequency**	**Percent**	**Cumulative Frequency**	**Cumulative Percent**
0	5128	89.76	5128	89.76
1	6	0.11	5134	89.87
2	516	9.03	5650	98.90
3	12	0.21	5662	99.11
4	21	0.37	5683	99.47
5	4	0.07	5687	99.54
6	5	0.09	5692	99.63
8	21	0.37	5713	100.00

WIGWAGS				
number_of_wigwags	**Frequency**	**Percent**	**Cumulative Frequency**	**Cumulative Percent**
0	5712	99.98	5712	99.98
2	1	0.02	5713	100.00

COMPOWER				
illumination	**Frequency**	**Percent**	**Cumulative Frequency**	**Cumulative Percent**
0=Yes	87	1.52	87	1.52
1=No	3730	65.29	3817	66.81
2=Missing	1896	33.19	5713	100.00

SPSEL				
SPSEL	**Frequency**	**Percent**	**Cumulative Frequency**	**Cumulative Percent**
1=constant warning time system	411	7.24	411	7.24
2=Direct current audio frequency overlay	1056	18.59	1565	27.56
3=NA	4114	72.44	5679	100.00

DOWNST				
track_run_down_street	**Frequency**	**Percent**	**Cumulative Frequency**	**Cumulative Percent**
0=Missing	89	1.56	89	1.56
1= Track Run Down a Street	176	3.08	265	4.64
2=Track does not run down a street	5448	95.36	5713	100.00

PAVEMRK				
pavemark_distance	**Frequency**	**Percent**	**Cumulative Frequency**	**Cumulative Percent**
0=Missing	89	1.56	89	1.56
1=<75ft	68	1.19	157	2.75
2=75 to 200ft	34	0.60	191	3.34
3=200 to 500ft	5340	93.47	5531	96.81
4=NA	182	3.19	5713	100.00

ADVWARN				
advance_warning	**Frequency**	**Percent**	**Cumulative Frequency**	**Cumulative Percent**
0=Missing	89	1.56	89	1.56
1=Yes	2044	35.78	2133	37.34
2=No	3580	62.66	5713	100.00

XANGLE				
crossing_angle	**Frequency**	**Percent**	**Cumulative Frequency**	**Cumulative Percent**
0=Missing	89	1.56	89	1.56
1=0-29	112	1.96	201	3.52
2=30-59	1318	23.07	1519	26.59
3=60-90	4194	73.41	5713	100.00

TRAFICLN				
number_of_trafficlane	**Frequency**	**Percent**	**Cumulative Frequency**	**Cumulative Percent**
0	89	1.56	89	1.56
1	1284	22.48	1373	24.03
2	4290	75.09	5663	99.12
3	5	0.09	5668	99.21
4	44	0.77	5712	99.98
5	1	0.02	5713	100.00

TRUCKLN				
TRUCKLN	**Frequency**	**Percent**	**Cumulative Frequency**	**Cumulative Percent**
0	89	1.56	89	1.56
1	154	2.70	243	4.25
2	5470	95.75	5713	100.00

HWYSYS				
highway_system	**Frequency**	**Percent**	**Cumulative Frequency**	**Cumulative Percent**
0=Missing	88	1.54	88	1.54
1= Interstate National Highway System	1	0.02	89	1.56
2= Other National Highway System	197	3.45	286	5.01
3= Other Federal-Aid Highway-Not NHS	773	13.53	1059	18.54
4= Non Federal-Aid	4654	81.46	5713	100.00

SCHLBUS				
schoolbus_traffic	**Frequency**	**Percent**	**Cumulative Frequency**	**Cumulative Percent**
0	5556	97.25	5556	97.25
1	1	0.02	5557	97.27
2	35	0.61	5592	97.88
4	73	1.28	5665	99.16
5	1	0.02	5666	99.18
6	35	0.61	5701	99.79
8	9	0.16	5710	99.95
10	1	0.02	5711	99.96
13	1	0.02	5712	99.98
36	1	0.02	5713	100.00

WHISTBAN				
quiet_zone_hour	**Frequency**	**Percent**	**Cumulative Frequency**	**Cumulative Percent**
0=Yes	5684	99.49	5684	99.49
1=No	29	0.51	5713	100.00

XBUCK				
number_of_crossbucks	**Frequency**	**Percent**	**Cumulative Frequency**	**Cumulative Percent**
0	708	12.39	708	12.39
1	830	14.53	1538	26.92
2	4114	72.01	5652	98.93
3	35	0.61	5687	99.54
4	24	0.42	5711	99.96
6	2	0.04	5713	100.00

GATES				
Gates	**Frequency**	**Percent**	**Cumulative Frequency**	**Cumulative Percent**
0	5058	88.53	5058	88.53
1	6553	11.47	5713	100.00

FLASHPAI				
lights_inpairs	**Frequency**	**Percent**	**Cumulative Frequency**	**Cumulative Percent**
0	5194	90.92	5194	90.92
1	2	0.04	5196	90.95
2	4	0.07	5200	91.02
4	299	5.23	5499	96.25
5	112	1.96	5611	98.21
6	58	1.02	5669	99.23
7	10	0.18	5679	99.40
8	12	0.21	5691	99.61
9	4	0.07	5695	99.68
10	9	0.16	5704	99.84
11	8	0.14	5712	99.98
14	1	0.02	5713	100.00

AADT

Basic Statistical Measures			
Location		**Variability**	
Mean	351.0004	**Std Deviation**	1388
Median	30.0000	**Variance**	1925729
Mode	20.0000	**Range**	25600
		Interquartile Range	105.00000

Average_Train_Speed

Basic Statistical Measures			
Location		**Variability**	
Mean	18.78741	**Std Deviation**	11.32215
Median	13.00000	**Variance**	128.19110
Mode	13.00000	**Range**	62.50000
		Interquartile Range	17.50000

Total_Number_Track	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	34	0.60	34	0.60
1	4637	81.17	4671	81.76
2	764	13.37	5435	95.13
3	209	3.66	5644	98.79
4	56	0.98	5700	99.77
5	11	0.19	5711	99.96
6	2	0.04	5713	100.00

Basic Statistical Measures			
Location		**Variability**	
Mean	1.239804	**Std Deviation**	0.59695
Median	1.000000	**Variance**	0.35635
Mode	1.000000	**Range**	6.00000
		Interquartile Range	0

Near_City	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	1192	20.86	1192	20.86
1	4521	79.14	5713	100.00

DAYTHRU

Basic Statistical Measures			
Location		**Variability**	
Mean	2.668300	**Std Deviation**	5.5014072
Median	1	**Variance**	30.26548
Mode	0.00000	**Range**	34
		Interquartile Range	1

DAYSWT

Basic Statistical Measures			
Location		**Variability**	
Mean	0.15315	**Std Deviation**	0.65049
Median	0	**Variance**	0.42314
Mode	0.00000	**Range**	30

		Interquartile Range	0

NGHTTHRU

Basic Statistical Measures			
Location		**Variability**	
Mean	1.75056	**Std Deviation**	3.63952
Median	0	**Variance**	13.24607
Mode	0.00000	**Range**	32
		Interquartile Range	2

NGHTSWT

Basic Statistical Measures			
Location		**Variability**	
Mean	0.05968	**Std Deviation**	0.52438
Median	0	**Variance**	0.27497
Mode	0.00000	**Range**	30
		Interquartile Range	0