

APPLICATIONS OF DATA MINING TECHNIQUES IN TRANSPORTATION SAFETY
STUDY

A Dissertation
Submitted to the Graduate Faculty
of the
North Dakota State University
of Agriculture and Applied Science

By

Zijian Zheng

In Partial Fulfillment of the Requirements
for the Degree of
DOCTOR OF PHILOSOPHY

Major Program:
Transportation and Logistics

November 2018

Fargo, North Dakota

North Dakota State University
Graduate School

Title

APPLICATIONS OF DATA MINING TECHNIQUES IN
TRANSPORTATION SAFETY STUDY

By

Zijian Zheng

The Supervisory Committee certifies that this *disquisition* complies with
North Dakota State University's regulations and meets the accepted standards
for the degree of

DOCTOR OF PHILOSOPHY

SUPERVISORY COMMITTEE:

Pan Lu

Chair

Denver Tolliver

Joseph Szmerekovsky

Annie Tangpong

Approved:

11/13/2018

Date

Joseph Szmerekovsky

Department Chair

ABSTRACT

Most of current studies are based on Generalized Linear Models (GLMs), which require several assumptions. Those assumptions limit GLMs with the nature of data, and jeopardize models' performance when handling data with complex and nonlinear patterns, high missing values, and large number of input variables with various data types. Data mining models are famous for strong capability of extracting valuable information and detecting complex patterns from large noisy data. However, they are not popular in transportation safety research, because they are criticized to be unable to provide interpretable and practical outputs. In this study, data mining models are tested in transportation safety research to prove their feasibility to be served as alternative models in safety study. Influential variable importance, contributor variable marginal effect analysis, and model predicting accuracy are further conducted to identify complex and nonlinear patterns in study dataset, and to respond to the criticism that data mining models do not provide practical outputs.

Due to the high fatality rate, two types of crashes are selected as research areas: 1) predicting crashes at Highway Rail Grade Crossings (HRGCs); and 2) commercial truck involved crash injury severity.

In the HRGC crash likelihood study, three data mining models, Decision Tree (DT), Gradient Boosting (GB), and Neural Network (NN), are tested, and demonstrated to be solid in Highway Rail Grade Crossing (HRGC) crash likelihood study.

In the commercial truck involved crash injury severity study, the GB model identifies 11 out of 25 studied variables to be responsible for more than 80% of injury severity level forecasting, and their nonlinear impact on the severity level. Several factors such as trucking company attributes (e.g., company size), safety inspection values, trucking company commerce

status (e.g., interstate or intrastate), and registration condition are found to be significantly associated with crash injury severity. Even though most of the identified contributing factors are significant for all four levels of crash severity, their relative importance and marginal effect are all different. Findings in this study can be helpful for transportation agencies to reduce injury severity level, and develop efficient strategies to improve safety.

Keywords: safety, crash, prediction, big data, data mining, Decision Tree, Gradient Boosting, Neural Network.

ACKNOWLEDGEMENTS

First, I want to give my gratitude to Dr. Pan Lu, my advisor. Her utmost professionalism, considerable efforts, patient, wisdom and continuous encouragement have kept me moving forward during the study. What I learnt from her is going to be a fortune in my life, and I cannot express my gratitude enough. A very sincere thank you is due to my committee members for their guidance and support to improve the quality of this paper. Dr. Tolliver, Dr. Joseph, and Dr. Tangpong shared their experience and knowledge, and helped me to complete this thesis.

TABLE OF CONTENTS

ABSTRACT.....	iii
ACKNOWLEDGEMENTS.....	v
LIST OF TABLES.....	viii
LIST OF FIGURES.....	ix
CHAPTER 1. INTRODUCTION.....	1
1.1. Transportation Safety.....	1
1.2. Data Driven Decision Making.....	5
1.3. Transportation Safety Big Data.....	6
1.4. Current Research Limitations.....	9
1.5. Data Mining.....	11
1.6. Research Focus Area.....	14
CHAPTER 2. APPLYING DATA MINING TECHNIQUES IN HIGHWAY RAIL GRADE CROSSING ACCIDENT ANALYSIS.....	16
2.1. Introduction.....	16
2.2. Literature Review.....	20
2.3. Data Description.....	25
2.4. Methodology.....	27
2.4.1. Decision Tree.....	30
2.4.2. Gradient Boosting.....	33
2.4.3. Neural Network.....	35
2.4.4. Over-Fitting.....	36
2.4.5. Rare Event Predictions.....	39
2.4.6. Variable Importance Analysis.....	40
2.4.7. Model Prediction Accuracy.....	42
2.4.8. Marginal Effect of Influential Variables.....	45

2.5. Results Analysis	46
2.5.1. Decision Tree Results Analysis	46
2.5.2. Gradient Boosting Model Results Analysis	52
2.5.3. Neural Network Model Results Analysis	60
2.5.4. Model Prediction Accuracy	69
2.5.5. Section Summary	73
CHAPTER 3. APPLYING GRADIENT BOOSTING IN COMMERCIAL TRUCK CRASH SEVERITY LEVEL ANALYSIS	75
3.1. Introduction	75
3.2. Literature Review	76
3.3. Data Description	82
3.4. Result Analysis	87
CHAPTER 4. CONCLUSIONS	104
4.1. Summary and Conclusions	104
4.2. Limitation and Future Study	107
REFERENCES	111

LIST OF TABLES

<u>Table</u>	<u>Page</u>
1. Traffic Fatalities vs Age.....	4
2. Input Variable Description.....	27
3. Classification Table	43
4. Variable Importance.....	48
5. Effect of Factors on Crash Likelihood.....	50
6. Misclassification Rate vs Learn Rate and Complexity of Trees.....	54
7. Variable Importance Based on GB Model.....	56
8. Variable Importance (NN Model).....	63
9. DT, GB, NN Model Prediction Classification Table.....	70
10. Model Prediction Accuracy Summary.....	71
11. Summary of Unanalyzed Variables	83
12. Variable Description	86
13. Variable Importance under Each Level of Severity.....	89
14. Marginal Effect of Influential Variables.....	92

LIST OF FIGURES

<u>Figure</u>	<u>Page</u>
1. Injuries by Year, 1990-2016	1
2. Fatalities by Year, 1975-2016.....	2
3. HRGC Accident Count Nation Wide.....	17
4. HRGC Accident Count North Dakota	18
5. Structure of a Typical Decision Tree	30
6. Gradient Boosting Model Training Process (Alexander I).....	34
7. Structure of Neural Network.....	36
8. Problem of Over-fitting.....	37
9. Decision Tree Model Output	49
10. Partial Dependent Plots (GB Model)	59
11. Crash Likelihood vs Explanatory Variables	66

CHAPTER 1. INTRODUCTION

1.1. Transportation Safety

Transportation safety is vital to transportation system operation performance. Traffic accidents cause traffic delay, property loss, and even people's lives. In the U.S., there are more than 268 million registered vehicles, and 218 million people holding a valid driver license (Statista, 2017). The huge number of vehicles is one of the potential reasons leading to more traffic accidents. In 2015, there were approximately 6.3 million traffic accidents in the U.S. (Statista, 2017). On average, traffic accidents cost economy loss of 871 billion dollars each year (PBS, 2014), which is more than double of national public spending on transportation and water infrastructure at 416 billion dollars. (COB, 2015)

In addition to the high economy loss, traffic accidents also cause a huge number of injuries. As shown in Figure 1, traffic accidents caused more than 3 million injuries each year before 2002. Even though the number of injuries has decreased significantly since 2002, there are still more than 2 million of injuries each year. In addition, the trend shows the number of injuries started to rise after 2013.

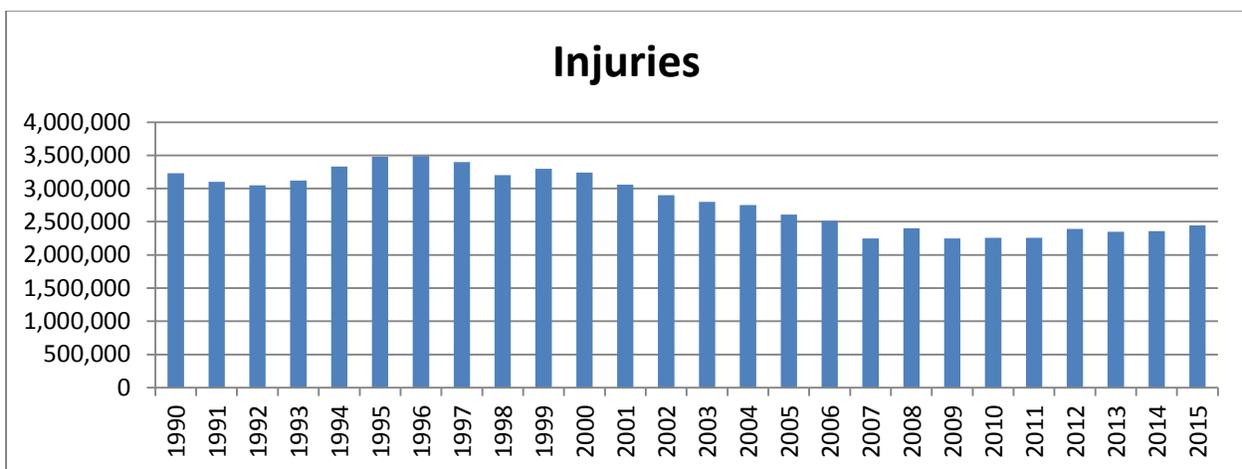


Figure 1. Injuries by Year, 1990-2016

Source: Federal Highway Administration, 2016

The most tragic fact about traffic accident is that it causes deaths. As automobiles became more popular in early twentieth century, traffic accident fatalities increased tremendously, and transportation safety did not attract enough public concern until late twentieth century, when more safety studies were conducted and more safety enhancements were implemented. Therefore, the number of people died in traffic accidents decreased significantly. However, nowadays, there are still a considerable number of deaths due to traffic accidents. As shown in Figure 2, traffic crashes took more than three million lives in the U.S. including more than 30,000 people killed on the roads of the United States each year (FHWA, 2016). In a typical month, traffic accident causes more death than the terrorist attack on New York and Washington on September 11th 2001. Every year traffic accident causes more than a million deaths around the world. It is estimated that as the total number of accidents increase due to the rapid growth of the number of motor vehicles in many formerly less-motorized countries, total death in traffic accident is likely to exceed 2 million by the year 2020 (WHO, 2001).

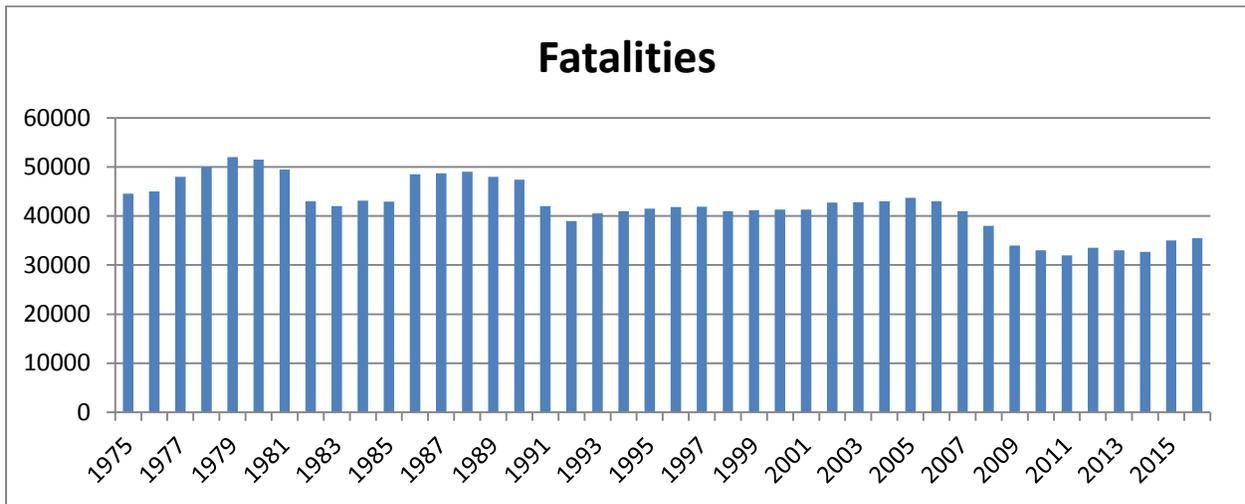


Figure 2. Fatalities by Year, 1975-2016

Source: Federal Highway Administration, 2016

In addition, traffic accident is the fourth leading of cause of death, and is considered as one of the world's largest public health problems. Unlike the other top three cause of death (heart disease, cancer, and respiratory disease) the victims are overwhelmingly young and healthy prior to their crashes. According to Table 1, more than 50% deaths from traffic accidents are younger than 45 years old. Especially, approximately 13% of fatalities are younger than 20 years old.

Table 1. Traffic Fatalities vs Age.

	2015	Cumulative percentage	2014	Cumulative percentage	2013	Cumulative percentage	2012	Cumulative percentage	2011	Cumulative percentage
<10	726	2.1%	691	2.1%	737	2.2%	755	2.2%	702	2.2%
10 to 20	3717	12.7%	3572	13.1%	3565	13.1%	3861	13.7%	4058	14.7%
21 to 24	3415	22.5%	3297	23.1%	3331	23.2%	3453	23.9%	3296	24.8%
25 to 34	6281	40.4%	5824	41.0%	5757	40.8%	5936	41.5%	5518	41.9%
35 to 44	4652	53.7%	4237	53.9%	4398	54.2%	4564	55.0%	4340	55.2%
45 to 54	5256	68.7%	4914	69.0%	4966	69.3%	5226	70.5%	5099	71.0%
55 to 64	4787	82.4%	4402	82.5%	4368	82.6%	4330	83.4%	3991	83.3%
65 to 74	3115	91.3%	2750	90.9%	2755	91.0%	2712	91.4%	2542	91.1%
> 74	3050	1	2976	1	2961	1	2895	1	2881	1

Since transportation safety drawn a lot of public attention, it has been improved with safety innovation implemented and drawn public awareness. However, during the past five years, the number of traffic accidents start back to increase. Thus, it is urgent to improve transportation safety in a further step. Traditionally, decision makings in transportation safety were relied on experience and limited number of available data. With developed computer technologies in early twenty-first century, data collection and storage techniques were innovated and improved, which enable data analysis and data-driven decision making more feasible and reliable than ever.

1.2. Data Driven Decision Making

Gilbreth et. al first started the study of scientific management, and contributed to the work that make decision making formalized and structured. Following their research, researchers were pursuing to model decision making mathematically and formulized. However, in 1970s, it was pointed out that not all the decision making problem can be quantitatively described in a mathematically formula. ([Barnat, 2014](#)). With the fast development and expansion of information technology in late 90s, society production efficiency has been significantly improved due to automation management. Meanwhile, large amount of raw data recording the system activities was generated and recorded. However, these raw data was not summarized, analyzed, and evaluated in an appropriate way, which fails to convert the data to valuable information for decision makers. Therefore, data analysis and data driven decision making starts to drawn the concerns of researchers.

In transportation safety study, aiming to reduce the vast losses caused by traffic accidents, studies are taken from many disciplines. Solutions are sought from basic physical principles, engineering, medicine, psychology, human behavior, law, mathematics, logic, and philosophy, where data driven analysis can provide convincible and reliable results for decision makers. In

transportation safety, data driven analysis can be used with a wide range of purposes, and can be summarized into three main groups: descriptive analysis, explanatory analysis and predictive analysis. As most of raw data does not offer a lot of value before it gets processed, by conducting the three main analyses, valuable insights can be extracted from the raw data. Descriptive analysis is usually the preliminary step to process the raw data to create and summarize, so that useful information can be provided and prepared for further analysis. Explanatory analysis can be used to better understand the data through a variety of algorithms. Explanatory analysis uses collected data, such as crash data, roadway data, and traffic data, to define crash related factors, and how these factors affect crash likelihood and outcomes. Unlike descriptive and explanatory analysis, predictive analysis focuses on what might happen in the future based on current research results.

In the past, crash and safety analysis were mostly relied on subjective or limited quantitative measures of safety performance, and limited number of data, which makes researchers have difficulty in accurately evaluating each factor' impact on safety when planning projects. Within the last decade, concept of big data gradually known by the public and its technology and application start to be mature and used in various fields. As data-driven analysis relies on real quality data, the big data provides foundation to data-driven analysis in transportation safety study.

1.3. Transportation Safety Big Data

With the rapid development of information and computer technologies, people have been increasingly relying on information network. Meanwhile, the terminology of “big data” is repeatedly mentioned and getting popular in various industries and in different situation. The term of “big data” is used to describe and represent the data with massive variety and volume

than traditional data, and can hardly be processed using old techniques and knowledge. An exact definition of "big data" is difficult to give because even within transportation safety field professionals use it differently. Generally speaking, safety big data is a large dataset that can hardly be reasonably processed and managed using traditional techniques and knowledge, is a digital asset with a rapidly increasing value and requires new methodologies to store, obtain and process, so that it could assist with system optimization and decision making (Arthur, 2013). To be more specific, safety big data is a database integrated with multi-dimension datasets recording crash records, traffic history, environment, transportation infrastructure status, etc., and the actual safety big data will vary according to the use of purpose. The challenge when facing the massive scale and heterogeneous data is to surface insights and connections, which would not be possible using conventional methods (Ellingwood, 2016).

Safety big data is not only a simple big database as it seems like. Huge volume is only one of its characteristics. Doug Laney (2001) from Gartner company first presented "three Vs of big data" to describe three most important characteristics that make big data different from other data processing: volume, velocity, and variety. Every day, there are billions of trips generated in the U.S. Information of each trip is recorded in a certain format, such as text, video, audio, etc. Road sensors collect traffic information in time. GPS and smartphone apps record the path for each trip. When accidents happen, accidents information is collected by officers. Smart city is a new proposed concept expressing a new urban area in the future that uses different types of electronic data collection sensors to supply information which is used to manage assets and resources efficiently. Transportation will be an important component, and the interaction data between vehicles and transportation facilities and among vehicles each other will expand the safety big data in respect of volume, velocity and variety. Transportation safety is a multi-

discipline field. The safety data always includes data from other fields: engineering, human behavior, environment, geography, etc. Therefore, the safety big data expands not only within transportation industry, but evolves with many other fields.

In General, safety big data can be classified into two groups: structured and unstructured data (Taylor, 2017). Structured data are those that can be summarized, analyzed, stored, and accessed in a fixed format. Structured data's format is always known in advance. For example, when analyzing effect of driver related factors on crash likelihood, the data can be structured as drivers with crash records and those without crash history. Computer languages and skills have been developed and achieved great success to process this kind of data. However, emerging issues are getting complicated and complex when such data extends to a huge size.

Unstructured data, on the other hand, are those without a pre-known form or structure. In an example of a group of data points representing drivers with and without traffic accident history, there are differences and features among these points to distinguish a certain group from the others. However, it is unknown that what the distinguished group is called, or what the rest of data is called. It could result from different genders or maybe age. Besides the huge size of unstructured data, deriving value from it through mathematical model is a huge challenge.

Safety big data provides foundation to enable decision making in transportation safety decided based on data and data analysis instead of experience and intuition. Nowadays, researches on big data help transportation planners with state and local projects. With the help of safety big data, historical traffic data, census data and geographically data are integrated and analyzed, based on which driver behaviors are analyzed, road network is designed to achieve the balance between demand and supply, and planners can allocate limited resources more efficiently.

Safety big data could find new solutions for improving transportation safety. With the safety big data concept, more achievements that are considered impossible or difficult can be made, such as real-time information interaction. More hidden trends and patterns are discovered and studied. With these findings, limited enforcement and management resources can be allocated and controlled in a more efficient way. Another improvement that the safety big data brings to transportation safety is on warning, including warning to cars conditions, traffic conditions, and drivers' conditions. For example, digital maps could show the path and locations of vehicles carrying hazard materials, so that other drivers can make their own decisions to avoid potential dangers. Approximately 90% of car accidents are caused by human errors. In the future of unmanned vehicles, there will be more data generated each day between vehicles and vehicles, between vehicles and infrastructures, from each facility, and from each person. Human behavior will be substituted with computer system setups, which makes "human behavior" easy to be controlled. By setting up a driving speed limitation in the system of unmanned vehicles, over speed could be controlled better and possible to be monitored.

Even though safety big data is so critical, it can hardly provide any information that can be directly used by decision makers. Researchers have developed numerous mathematically models and algorithms to discover information from the data.

1.4. Current Research Limitations

"Big data is the key to a business success, big data will change the world, and big data will do this and that (Ghodke, 2015)." Nowadays, statements like this are popular. Although safety big data is important, it is the value extracted from the data that decision makers really need. To discover and explore information and patterns from the data, numerous models have been developed. Traditional Generalized Linear Models (GLMs) are considered as the most

popular models in transportation safety research. By building up a direct generalized linear relationship between target variable and independent variables, user can easily interpret and use this understandable quantitative relationship. However, GLMs heavily rely on the following assumptions (PSECS, 2017):

- 1) Independent variables are independently distributed.
- 2) The dependent variable does NOT need to be normally distributed, but it typically assumes a distribution from an exponential family (e.g. binomial, Poisson, multinomial, normal, etc.)
- 3) GLM does not have to assume a linear relationship between the dependent variable and the independent variables, but it does assume linear relationship between the transformed response in terms of the link function and the explanatory variables.
- 4) Independent (explanatory) variables can be even the power terms or some other nonlinear transformations of the original independent variables.
- 5) The homogeneity of variance does NOT need to be satisfied. In fact, it is not even possible in many cases given the model structure, and overdispersion (when the observed variance is larger than what the model assumes) maybe present.
- 6) Errors need to be independent but NOT normally distributed.
- 7) It uses maximum likelihood estimation (MLE) rather than ordinary least squares (OLS) to estimate the parameters, and thus relies on large-sample approximations.
- 8) Goodness-of-fit measures rely on sufficiently large samples, where a heuristic rule is that not more than 20% of the expected cells counts are less than 5.

Once these assumptions are violated, it will generate numerous errors. In addition, when handling transportation safety big data, GLMs are limited by big data's features, especially the

complicated non-linear patterns. Therefore, it is necessary to apply other models that can handle the big data, especially explore the non-linear patterns.

1.5. Data Mining

While GLMs show limitations when handling the safety big data, data mining models, a type of non-parametric data analytical models have been developed. Data mining is the process of discovering patterns and correlations, and extracting valuable information within large data sets (SAS, 2017). Data mining research was triggered as the large amount of data was generated and stored (emerging of big data), and the need of utilizing these data becomes urgent, because information extracted from the data can guide decision makers and bring large amount of profits to business.

The common data mining tasks can be divided into association analysis, clustering, classification, and prediction.

Association analysis is a meant to find frequent patterns, correlations, associations, or causal structures from various kinds of databases, such as relational databases, transactional databases, and other forms of data repositories (Techopedia, 2017). Given a set of data sets, association rule mining aims to find the rules which enable researchers to predict the occurrence of a specific item based on the occurrences of the other items in the data sets. For example, a red light violation are often found with drinking driving, because alcohol makes drivers make wrong decisions and slow down their reaction. Clustering or cluster analysis is the task to identify the attribute that can be used to separate a set of objects in a certain way, so that the objects in the same group are more similar to each other than to those in other groups in terms of certain attribute (Stefanowski, 2008). For example, clustering can be used when locate locations with high risk of crashes. Classification is the derivation of a function or model which determines the

class of an object based on its attributes (Fu, NA). A set of objects is given as the training set in which every object is represented by a vector of attributes along with its class. A classification function or model is constructed by analyzing the attributes and the class of objects in the data set. Such a classification function or model can be used to classify future objects and develop a better understanding of the classes of the objects in the database. For example, from a set of crash record data, which serve as the training data, a classification analysis can be built, which concludes each crash's outcome (severity level). The generated classification model can be used to diagnose a new crash's outcome based on the training crash record data, such as weather condition, traffic volume, type of vehicles, etc. Prediction is to discover the trend in a dataset, and build up a model to predict features of future data.

There are various data mining algorithms, such as K-means, support vector machine, neural network, classification tree, etc. (Li, 2015). Statistics and data mining share the same goal: discover and identify structure of the data and turn the data into valuable information. Even though data mining relies a lot on statistics theories, it utilizes knowledge from other fields as well, including machine learning, computer science, and database technology (Priyadharshini, 2017). In a statistics study, a hypothesis needs to be proposed and mathematic function and models are built up to test the hypothesis. In data mining, no hypothesis is pre-required. The links between target variable and its associated factors are automatically established.

Generalized Linear Models (GLMs) are the most popular statistical models favored by researchers in transportation safety study. GLMs are capable to construct quantitative relationship between target variable and its contributor variables with mathematically equation. The inference reflects statistical hypothesis testing. The most favorable feature of GLMs is that the quantitative relationship is easy to interpret, so the outcome can be directly used by

researchers. It is easy to draw conclusions that can be used by decision makers. However, GLMs have several limitations. They can only handle structured data. When the data is complex, especially when it includes the mixture of interval, nominal, ordinal, and numerical variables, and there are a large number of redundant and irrelevant variables (Brusilovkey, NA), GLMs can low performance. GLMs' performance can also be affected when the data is highly heterogeneous and with high percentage of missing values, and outliers. As mentioned in previous sections, safety big data is usually noisy, and consists of a large number and various types of variables. In addition, the high percentage of missing values and a large number of redundant and irrelevant variables can all affect GLMs' performance. In addition, GLMs are heavily relied on the pre-defined assumptions. When handling safety big data, all the pre-defined assumptions can hardly be satisfied at the same time. Once they are violated, it can cause numerous errors.

Compared with GLMs, data mining models usually handle bigger size of data. In a GLM, a hypothesis has to be proposed before testing the model. Therefore, the noisy data type and redundant variables affect a lot on the model and variable selection. However, data mining models require no hypothesis. In fact, it can reveal the underlying pattern among variables. GLMs are mainly limited by the required pre-defined assumptions. On the contrary, data mining models are non-parametric, and they have no limitations on any pre-defined assumptions. Especially, data mining models are capable to discover non-linear complicated relationship between dependent variables and associated factors. In spite of the limitation of GLMs plus the strong capability of data mining models when handling big data, data mining models are not favored by researchers.

1.6. Research Focus Area

Even though data mining models are powerful when handling big data and have achieved success in many fields, application of data mining models in transportation safety research is still a gap. In a safety research, researchers still prefer to GLMs. The reason is that data mining outputs are hard to interpret. Data mining models can predict well but make a bad job when explaining the outcome (Shmueli, 2010). For example, one of the most popular data mining models, neural network, is always criticized because the blackbox in the model prevents people from understanding the model, even though the model can provide high prediction accuracy. Increasing model interpretability will trigger a better acceptance of the data mining models (Cortez & Embrechts, 2011). In this study, three popular data mining models, Decision Tree (DT), Gradient Boosting (GB), and Neural Network (NN) model, will be applied in transportation safety research. Their robustness in mining safety big data will be tested. A few more analysis, such as variable importance and marginal effect, are conducted to provide better understanding of data mining models' outputs.

Transportation safety is a big scope, and can be improved during processes, such as planning, design, facility construction, operation and infrastructure maintenance etc. This study will focus on two fields due to their severe influence on traffic and tragedy consequences: crash likelihood analysis at Highway-Rail Grade Crossings (HRGC) and injury severity level analysis of commercial truck involved crashes. The DT, GB, NN model Data mining models' robustness will be first tested in the HRGC crash likelihood analysis. Then a model will be selected based on prediction accuracy, be tested in the commercial truck crash injury severity study using a bigger data.

By conducting this study, the author aims to achieve:

- 1) Demonstrate the DT, GB, NN model feasibility in HRGC crash likelihood study.
- 2) Define the complicated and non-linear relationship between HRGC crash likelihood and related factors.
- 3) Show that the DT, GB, and NN model can provide practical outputs by conducting further analysis in the study of HRGC crash likelihood, including contributor variables' marginal effect analysis, variable importance identification, and prediction accuracy.
- 4) Prove the GB model robustness in explaining commercial truck crash injury severity levels.
- 5) Identify explanatory variables of commercial truck crash injury severity levels, especially truck company characteristic and driver characteristic variables. Explore the relationship between truck involved crash severities and influential variables, especially the nonlinear relationship that cannot be identified by a GLM.
- 6) Prove that data mining is a feasible tool in transportation safety study.

CHAPTER 2. APPLYING DATA MINING TECHNIQUES IN HIGHWAY RAIL GRADE CROSSING ACCIDENT ANALYSIS

2.1. Introduction

Highway-rail grade crossings (HRGCs) have long been recognized as critical transportation assets. The costs from disruptions to both the road and rail networks can be significant. In addition, the economic impact of those accidents is often compounded because of traffic delays on both the highway and railway. However, the high fatalities rate makes traffic accidents at these locations more catastrophic. From 1996 to 2014, there were 54,649 crashes at HRGCs across the United States where active warning devices (i.e. gates, lights, signs, bells, etc.) are in place (FRAOSA, 2015). About 12% of these crashes resulted in 6,527 fatalities (FRAOSA, 2015), while only 0.06% of all types of accidents lead to deaths. As shown in Figure 3, in the U.S., from 1996 to 2014, the number of HRGC accidents and resulted injuries and fatalities keep a decreasing trend with a little fluctuation in between. However, in North Dakota State, as shown in Figure 4, the number of accidents is not controlled well. Even though the number of accidents was low in 2006, it starts to increase since then. The fatality rate varies between 0% and 44%, and the fatality rate is higher than the national average for most of the time. Transportation agencies must identify the contributing factors to HRGC crashes to improve HRGC safety performance and reduce the number of crashes. A large volume of literature explores and evaluates the explanatory factors that contribute to the likelihood of HRGC collisions. Thus, an accurate HRGC accident prediction model is critical for HRGC safety improvement.

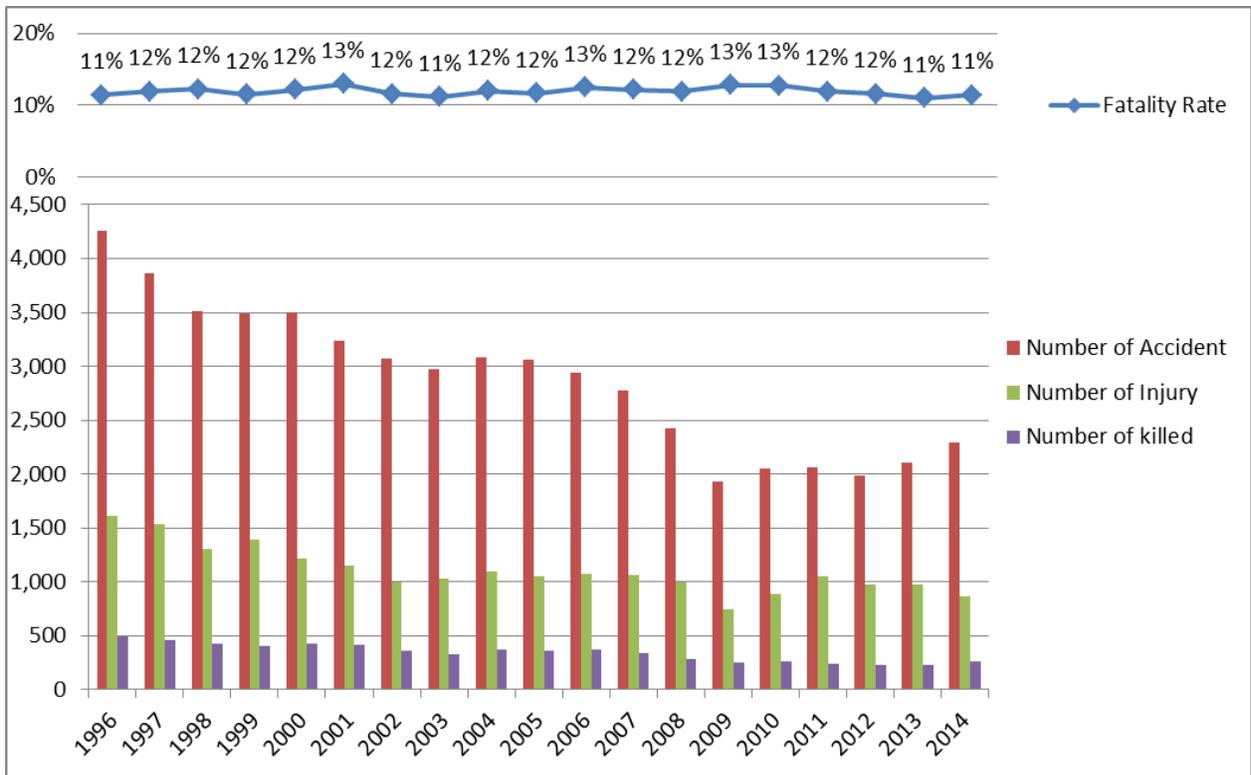


Figure 3. HRGC Accident Count Nation Wide

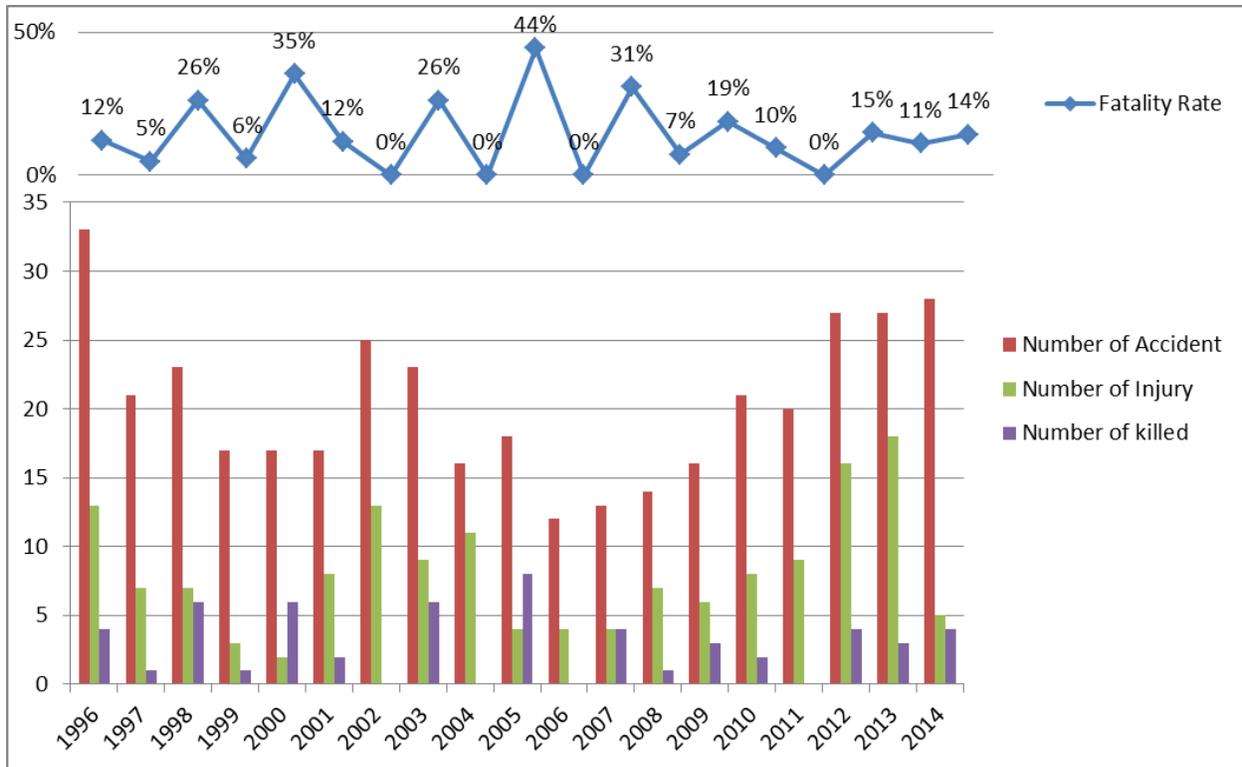


Figure 4. HRGC Accident Count North Dakota

Because crash accident data has random, discrete, and non-negative characteristics, Generalized Linear Models (GLMs) have been commonly selected to investigate the relationship between crashes and contributing factors. However, Lord and Mannering (2010) pointed out that, although the GLMs possess desirable elements for describing accidents, these models face various data challenges which stem from crash data distribution and inappropriately fitted GLMs. Fitting the GLMs with data that exhibits a different distribution than the assumed distribution of the model can result in incorrect prediction and explanatory factors. As pointed out by Lu and Tolliver (2016) and Oh et al. (2006), HRGC crash data often shows under-dispersion distribution and less common GLMs are suitable for such datasets. Moreover, the available crash dataset is often containing a large portion of missing data and outliers. GLMs are sensitive to this noise. Outliers and missing values are often either deleted or imputed with the mean value. Furthermore,

to fit a well-performed GLM, a considerable number of observations are required. However, crashes at HRGCs are rare events. Event-level data, positive for a crash, are only a small portion of the entire dataset. The majority of data are at non-event level, representing zero crash. Thus, although some studies achieved remarkable overall prediction accuracy, the model only explains well for a non-event situation but was less accurate at the event-level (Chang and Chen, 2005; Chang and Wang, 2006; Chang and Chien, 2013). In addition, as one of important model performance measurements, prediction accuracy is not analyzed thoroughly in previous studies.

Data mining techniques have proven their robust ability to explore large, noisy, and complex data sets in recent years. Several data mining models have been applied in vehicle accident studies, including the neural network (NN) model (Chiou 2006; Zeng and Huang 2014), and the classification and regression decision tree (DT) models (Kashani, Rabieyan and Besharati 2014; Yan, Richards and Su 2010). Applying data mining models to HRGC crash data modeling has received much less attention than GLMs. With the improvement of computing capability and software availability, data mining approaches are worth investigating with regard HRGC crash data modeling performance. Data mining approaches are used to find patterns in large datasets and relationships between target variables and predictors. Moreover, data mining approaches are a non-parametric method, which do not require any pre-defined underlying relationships between target variables and predictors, and the under-dispersed data does not affect model performance.

DT models are among the most favorable data mining models used in crash studies due to their capability to generate a visualized and easy-to-interoperate predictive-tree-based output and their effectiveness in handling non-linear interactions among variables with missing data. Using simple decision trees as fundamental components, the GB model improves the DT model in

terms of predicting capability while inheriting advantages of the DT model. NN models have a number of properties that make them very attractive over traditional GLMs. A NN is a non-linear processing system. Each layer in a NN represents a non-linear combination of non-linear functions from the previous layer and requires no underlying theory about the data like most GLMs do. In addition, it is strongly capable of exploring patterns and requires little data clearness (SAS Institute Inc. 2015).

In this chapter, the DT, GB, and NN model are applied to predict crashes at HRGCs. Model performance in respect of forecasting accuracy is compared. In addition, influential variables are identified and their importance is compared. Moreover, further efforts to explore marginal effects with several traffic exposure factors and warning devices are also conducted in this chapter.

2.2. Literature Review

Studies of crash frequency at highway-rail crossings dates back to 1940s when the Peabody Dimmick Formula, one of the earliest predicting models for HRGC accident, was used to estimate the expected number of accidents based on historical crash data (USDOT, 1986). This model is developed based on accident data from rural rail-highway crossings in 29 states (USDOT, 1986). It uses three factors to forecast crash rate: Average Annual Daily Traffic (AADT), average daily train traffic, and the presence of warning devices. The developed relationship is as follows (Federal Highway Administration, 1986):

$$A_5 = 1.28 \frac{V^{0.17} T^{0.151}}{P^{0.171}} + K \quad (\text{Equation 1})$$

Where A_5 is the estimated number of accidents in 5 years. V is the Average Annual Daily Traffic (AADT). T is the average daily train traffic. P is the protection coefficient indicative of warning devices present. K is the additional parameter.

Inspired by the Peabody Dimmick Formula, other models were developed based on similar theories and using similar factors, but with an adjustment of coefficient factors (Austin, 2002). Even though these models consider the major factors influencing crash rates, the resulting accuracy is still questionable, because of the limited number of explanatory variables. The U.S. Department of Transportation (USDOT) accidents prediction equations was a revolutionary model in this field and overcame previous shortcomings by taking more crossing design factors into account, such as type of warning devices, type of gate, and control type (USDOT, 1986). The USDOT accident prediction formula comprises of four equations (FRA, 1987; FRA, 1999):

$$a = (K)(EI)(DT)(MS)(MT)(HP)(HL)(HT) \quad (\text{Equation 2})$$

$$B = \frac{T_0}{T_0 + T}(a) + \frac{T_0}{T_0 + T}\left(\frac{N}{T}\right) \quad (\text{Equation 3})$$

$$T_0 = \frac{1.0}{0.05 + a} \quad (\text{Equation 4})$$

$$\begin{aligned} A &= 0.7159B \quad \text{For passive devices;} \\ A &= 0.5292B \quad \text{For flashing lights;} \end{aligned} \quad (\text{Equation 5})$$

$$A = 0.4921B \quad \text{For gates.}$$

Where a is predicted number of accidents; K is the Formula constant; EI is Exposure index factor; DT is Day through trains factor; MS is Maximum speed factor; MT is Main tracks factor; HP is Highway paved factor; HL is Highway lanes factor; HT is Highway type factor. N , is the number of observed accidents in T years at the crossing, and T_0 is the formula-weighting factor. a , is determined on the basis of the various crossing-specific characteristics. B adjusts the predicted number of accidents (a) to reflect the actual accident history at the crossing. A introduces a normalizing constant that is multiplied by the adjusted predicted accident value, B .

Although the USDOT accidents prediction equations were a great step forward in comprehensively explaining associated factors that impact crossing crash rate, they can hardly quantitatively measure the contribution that each factor makes to crash rate. This shortcoming is remedied by more recent regression models (Austin & Carson, 2002; Oh, Washington & Nam, 2006). Austin and Carson (2002) applied negative binomial accident prediction model in HRGC safety research. Compared to the USDOT accident prediction formula, their model made great improvement on interpretation of both of the magnitude and direction of the effect of significant contributor variables on HRGC accident frequencies. Their research defined traffic characteristics, roadway characteristics, and crossing characteristics that are significant on affecting HRGC accident frequencies. Traffic characteristics, including night through train volume, AADT in both directions, number of tracks and traffic lanes, and train speed are found to be positively related with HRGC accident frequencies. Only one roadway characteristic variable was proved to be significant: if a highway is paved, the likelihood of accident is higher. For crossing characteristics, gates were proved to be effective on preventing accidents at HRGCs. However, presence of stop signs, flashing lights, or bells were all found to increase accident risk. Oh, Washington, and Nam (2006) tested gamma probability model in HRGC crash study by comparing it with previous models: Peabody Dimmick Formula, New Hampshire Index, NCHRP Hazard Index, USDOT prediction formula. Their research found out that AADT, presence of commercial area and train detector distance are significantly positively related with HRGC frequencies, while presence of track circuit controller, presence of guide, and presence of speed hump have a negative effect on the crash frequency. Although road and crossing width and number of tracks were found to be not significant, the authors believed that the effects of these factors were likely to be captured by the correlated variables, such as traffic and train volume.

Several other researchers also adopt GLMs (Raub 2009; Hu, Li and Lee 2012). In Lu and Tolliver's study (2016), the different types of GLMs were compared and summarized, including Poisson model, negative binomial model, Conway-Maxwell-Poisson model, Bernoulli model, the hurdle Poisson model, and zero-inflated Poisson model. Poisson, Conway-Maxwell-Poisson, Bernoulli, and hurdle Poisson model were demonstrated to be able to handle under-dispersion issue. In a further comparison among these four models, the authors concluded that all models provided the same parameter estimates for all the studied variables, such as AADT, train volume, number of tracks, and max train speed. The Bernoulli model and hurdle Poisson model showed almost identical results on parameter estimates. The Conway-Maxwell-Poisson model generated most distinctive parameter estimates for explanatory variables compared with the Poisson regression model. Crossing warning devices, highway traffic, rail traffic, train speed, number of tracks, appearance of paved highway, are all significantly related with HRGC accident likelihood defined by all four models. AADT, trains per day, number of tracks, paved highway, and maximum train speed are all examined to be positive contributor variables.

Although GLMs are useful for predicting crash frequency and for interpreting relationships, the models often have low prediction accuracy and are often limited with underlying data assumptions. The specification of the functional form depends on the nature of the data and can significantly affect the goodness-of-fit of GLMs and result in erroneous parameter estimations and low prediction accuracy if the assumptions are violated (Xie, Lord and Zhang 2007).

With no pre-defined underlying data theory or assumptions, data mining algorithms start to gain popularity in accident frequency study. However, most of studies focus on highway safety and very limited on accident frequency at HRGCs (Chang & Chen, 2005; Li, Zhang & Xie,

2008; Chang, 2005). Within the limited number of studies on HRGC accident prediction based on data mining models, decision tree model is the most common selected method due to its ease-of-use tree structure result. Yan, Richards, and Su (2010) investigated crash frequency at HRGCs using two-stage classification and a regression decision tree model. Decision tree based results are generated and analyzed under three different scenarios. They find that the crash pattern and contributory factors are significantly different between cross-buck-only-controlled crossings and stop-sign-controlled crossings. However, DT method still contains few draw backs, even though the decision tree provides easy-to-view illustration, the tree structure is instable and sensitive to outliers. Moreover, DT method often produces a large and complex tree which still poses great presentation difficulties.

The GB model, an ensemble of simple decision trees, is extremely powerful in understanding the structure of complex datasets and exploring potential relationships between dependent variables and independent variables, and is believed to be superior to simple DT models because of its techniques for handling missing data, robustness with data noise and resistance to over-fitting (Trevor, 2014).

Different with tree-based data mining models, NN model is inspired by mimicking human brain learning process. Codur and Tortum (2015), and Abdulhafedh (2016) apply neural network models to analyze highway accident frequencies. They identify influential factors of highway accidents, evaluate their models' performance, and indicate that NN model do not impose the stringent distribution assumptions and can provide robust results with even chaotic input data such as a data with a lot of missing values. However, they fail to explore relationship between crash likelihood and influential factors because of the black box in the NN model. Fish and Glogett (2003) propose a method to look inside of the black box and analyze the effect of

input variables on target variable by keeping all other independent variables unchanged at certain level, and varying the studied independent variable within a range. Their method is used by Xie, Lord, and Zhang, and Chang's study on highway accident frequency prediction (Xie, Lord & Zhang, 2007; Chang, 2005; Zhang & Meng, n.d.). However, Gevrey, Dimopoulos, and Lek (2003) point out that it is arguable and questionable when keeping the unstudied variables at a random or meaningless value. In addition, the explored relationship could be different when the unstudied variables are kept at various levels. Thus, they proposed a method to generate the relationship by keeping the other variables at meaningful values, such as mean value of the variable, and all the explored relationships are recorded while the remaining variables are held at different values.

In the following chapter, the DT, GB, and NN algorithms are described, and HRGC accident likelihood is predicted with all three algorithms. The model predictive accuracies are observed to be comparable among all three models and furthermore, result of each model and its performance are analyzed and compared thoroughly.

2.3. Data Description

This study uses 19 years of accident/incident data merged from the FRA's Office of Safety accident/incident database and the FRA's Office of Safety highway-rail crossing inventory. The data was merged by using the HRGC identification number in both of the databases. The merged database contains all the historical crash information at HRGCs in North Dakota from 1996 to 2014, including crossing location, traffic condition, infrastructure equipment, accident information, time, and weather conditions. There were 5,713 HRGCs in North Dakota during that period, of which 354 have historical accident records. To study crash-associated factors and effectiveness of warning devices, 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. Table 2 lists all screened variables, including one target variable, one ID variable, and twenty-two input variables. These input variables can be divided into two categories: traffic characteristics, and crossing characteristics. Traffic characteristics record traffic information at crossings. These characteristics describe highway and railway traffic volume and traffic speed. Crossing characteristics describe presence of warning devices and other crossing related characteristics.

Table 2. Input Variable Description

Variable	Property	Description
ACCIDENT	Target variable	1= crash happened, 0=no crash
ID	ID variable	Crossing identification
Traffic Characteristics		
AADT	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
TOTAL_NUMBER_TRACK	Numeric	Number of rail tracks at crossings
TRAFICLN	Numeric	Number of traffic lanes crossing railroad
Crossing Characteristics		
Highway_Paved	Category	Is highway paved or not? 1=yes, 0=no
ADVWARN	Category	Presence of static advance warning Signs: 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	Category	Presence of mast mounted flashing lights: 1=yes, 0=no
FLASHNOV	Category	Presence of cantilevered flashing light not over traffic lane: 1=yes, 0=no
FLASHOV	Category	Presence of canti-levered flashing light over traffic lane: 1=yes, 0=no
FLASHPAI	Category	Presence of flashing light in pairs: 1=yes, 0=no
GATES	Category	Presence of Gates: 1=yes, 0=no
Near_City	Category	In or near city? 1=near city, 0=in city
STOPSTD	Numeric	Highway stop signs presence: 1=yes, 0=no
WIGWAGS	Numeric	Presence of wigwags: 1=yes, 0=no
XBUCK	Numeric	Presence of cross buck: 1=yes, 0=no

2.4. Methodology

Every day, the safety big data is expanding with new generated data. However, the pure data set can hardly provide any valuable information. In addition, traditional analytically GLMs

show limitations when handling big data. The concept of data mining starts to gain its popularity among various fields. With the goal to detect consistent patterns and relationships between variables, data mining is developed and defined as an analytic process to explore the huge, noisy, incomplete, and random data set. The first stage of data mining is initial exploration, which involves data cleaning, data transformation, data pre-screen, and problem definition. The second stage is model training and validation. In this stage, various data mining models are analyzed and an optimal model is selected based on model performance and assessment measurements. The third stage is deployment, where the optimal model in the second stage will be applied to a new data to generate predictions or estimations. The ultimate goal of data mining is prediction, which has the most direct business applications (Data Mining Techniques, 2018).

Statistics is quantitative analysis, interpretation, and summary of numbers or data. It provides fundamental definitions and concepts to data mining. However, even though both of data mining and statistics have the same goal: to discover and identify information from data, because of the 3 Vs characteristics of big data, traditional statistics models have limitations when handling big data. On the other hand, data mining uses scientific methods, processes, and systems to extract knowledge from nasty data in various forms, either structured or unstructured. In addition, data mining is a multi-disciplinary field, which grew out of database technology. It covers a variety of tasks over statistics, such as data preparation, data inspection, and data cleaning.

Data mining methods can be generally divided into five categories: classification, estimation, association rules, clustering, and mining complicated data types. Classification is to classify raw data based on pre-defined categories. For example, insurance companies can use classification method to classify applicants to into customers with high and low risk of traffic

accidents. Estimation is similar to classification. The difference between the two concepts is that classification has pre-defined categories, and the number of categories is fixed. However, estimation will output an unfixed number. For example, Walmart makes estimation on size of a household based on the family's purchasing frequency and amount. In this case, the size of households is not a fixed number, and it could be one, two, etc. Association rule is to analyze a group of events that tend to happen together. For instance, when people shop at supermarkets, there could be certain patterns existing. Such as people who purchase eggs could possibly buy milk as well. Supermarket will use this analysis to arrange the display of goods, and promote sales. Clustering is to group a particular set of observations based on their characteristics, and aggregate them based on their similarities. The group could be pre-defined and unknown. Mining complicated data types includes text mining, graphic mining, audio mining, etc. By training the model with a big certain type of data set, the model will learn to recognize certain patterns existing in the data set, such as to train a model learn to recognize people's hand writing.

Data mining has been successfully applied in various fields, such as industrial, business, education, military, etc. Nowadays, more innovative equipment and technologies make transportation data collection more easily, making more transportation data generated every day. In the future, as unmanned vehicles and intellectual transportation system are getting mature, with vehicle to vehicle data, vehicle to infrastructure data, and vehicle to passenger data coming in, the size of big data will expend tons of times as today. Thus, data mining starts to draw more concerns in transportation field, especially when decisions and strategies in transportation are transferring from experience driven to data driven.

In this study, three popular data mining models are introduced and applied in transportation safety study.

2.4.1. Decision Tree

A decision tree is a hierarchical tree-based prediction model. There are two types of decision tree 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 without a crash. Thus, a classification tree will be generated.

Generally, development of a decision tree involves three steps. The first step is tree growth. As shown in Figure 5, at the beginning, all data concentrates in the root node.

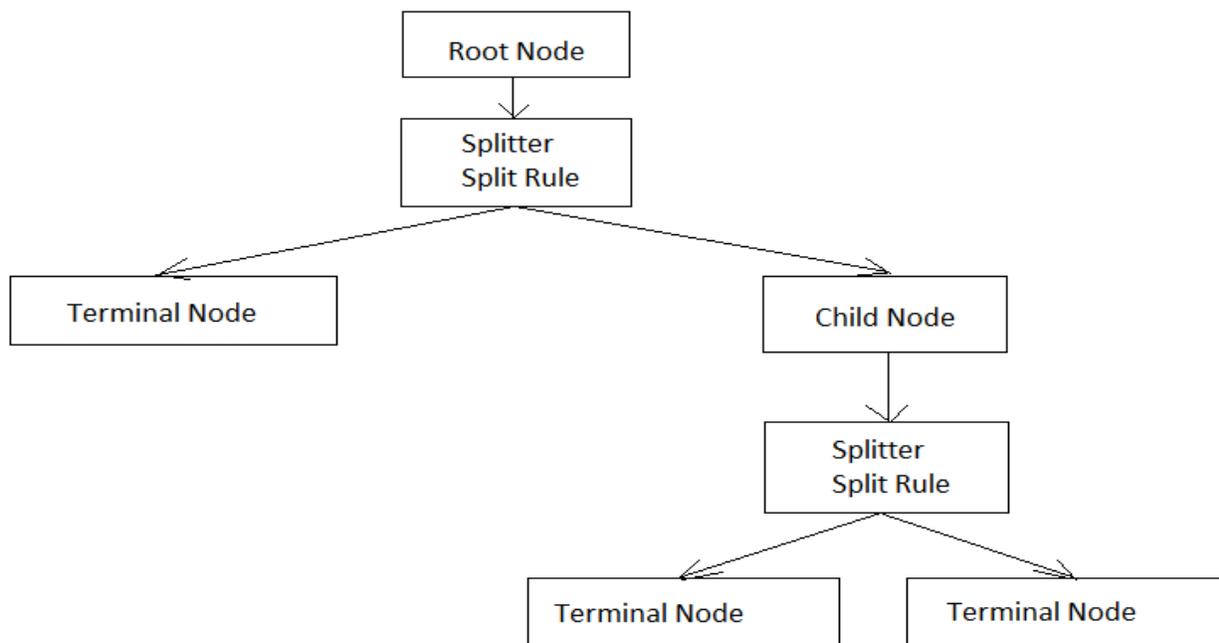


Figure 5. Structure of a Typical Decision Tree

Then, the dataset is broken down into child nodes by applying a series of splitting variables (splitters). Each child node will be treated as parent node for a 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 gain to decide if a variable should be chosen as the splitter, and whether the node can be further split or not

(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 is zero. For a variable S with c distinct values, the entropy $E(S)$ of S is calculated as Equation (6): (Freitas, 2013)

$$E(S) = \sum_{i=1}^c -p_i \log_2 p_i \quad (\text{Equation 6})$$

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, \dots, S_c by certain splitter, the expected entropy (EH) measures the expected unpredictability of these c outcomes of variable S after splitting, and calculated as: (25)

$$EH = \sum_{i=1}^c \frac{a_i}{a} \times (-p_i \log_2 p_i) \quad (\text{Equation 7})$$

Where: a_i is the number of observations in each subset S_1, \dots, 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 (8). Information gain measures how much a splitter can help predict the outcomes. The variable that generates the highest information gain discriminates 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.

$$I = E(S) - EH \quad (\text{Equation 8})$$

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 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 testing. The training data is used for splitting the nodes, and testing data is for measuring misclassification rate in optimal tree selection step.

The recursive algorithm behind the decision tree model keeps splitting the data until it ends up with pure sets. The decision tree always classifies the training data perfectly, and reaches an accuracy of 100% for the training data. However, as the decision tree keeps splitting the data, the tree gets bigger and bigger, and it becomes more and more accurate for the training data. But at some point, predictions become less accurate on the data that hasn't been used to train the tree. Thus, to avoid yielding a very large size tree, three parameters can be established. The first is the node size. If a node contains too few observations, splitting will not continue. The second is the number of nodes in the path between the root node and the given node. When this number equals the set-up value, splitting will be stopped. The third is to conduct significance test to test if all the observations in the node contain nearly the same target value or not. When the significance level value is equal to the set-up threshold, a further split will not be allowed.

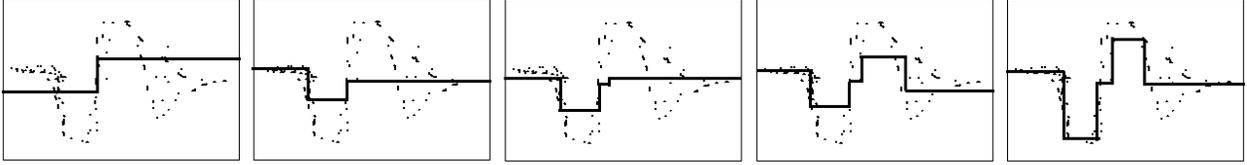
After a sequence of pruned trees are established, the last step is to select the optimal one from the sequence of pruned trees, based on a measurement of the misclassification rate of testing data, so that the training data will not be over-fitted. As the tree grows larger and larger, the misclassification rate of training data decreases monotonically, indicating that the saturated tree always fits best to the training data. However, the misclassification rate for the testing data decreases first to a minimum value, and then keeps increasing and approaching a certain value. The depth of an optimal tree is decided when misclassification rates reach a minimal value for both training and validation data.

2.4.2. Gradient Boosting

The gradient boosting method is also known as multiple additive trees (MAT), and is a machine-learning data-mining technique for regression and classification problems proposed by Friedman (2002, 2003) at Stanford University. GB method theoretically extends and improves the simple DT model using stochastic gradient boosting (Friedman, 2002). GB produces a predictive model in the form of an ensemble of several weak-prediction, simple, tree-based models (Schapire 1999; Monteiro 2004). Therefore, the GB model inherits all of the advantages of tree-based models while improving in other aspects, such as forecasting accuracy (Friedman, Meulman 2003). Moreover, several other features make the GB model, including its: handling of large datasets without pre-processing, resistance to outliers, handling of missing values, robustness to complex data, and resistance to over-fitting (Friedman, Meulman 2003; Salford-Systems)

A GB model can be viewed as a series expansion approximating the true functional relationship (Salford-Systems). In general, GB model starts by fitting the data with a simple decision tree model, which has certain level of error in terms of fitness with the data. The simple DT model is referred as a weak learner. A detail description of the algorithm of simple decision tree can be referred to section 2.4.2. Considering the errors having the same correlation with outcome value, the GB model then develops another decision tree model on the errors or the residuals of the previous tree. This sequential process will repeat itself until errors are minimized. This procedure is shown in Figure 6 (Alexander I, 2002).

Data and Prediction Function



Error Residual

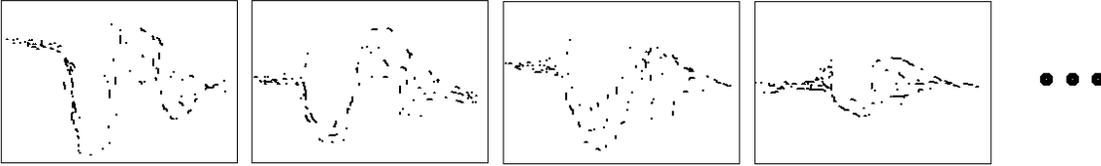


Figure 6. Gradient Boosting Model Training Process (Alexander I)

The detailed algorithm of GB is described as follows (De'ath 2007; Hastie et al. 2009):

$$f(x) = \sum_n f_n(x) = \sum_n \beta_n g(x, \gamma_n) \quad (\text{Equation 9})$$

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 ($n=1,2,\dots,n$) are the coefficients, and determine how each single tree is to be combined. Loss function measures prediction performance, such as deviance. Friedman (2001) proposed a numerical optimization method called functional gradient descent, which is summarized below:

1. Initialize $f_0(x)$;
2. For $n=1$ to n (number of trees)
 - a) For $i=1$ to m (number of observations), calculate the residuals

$$\tilde{y}_{in} = -\left[\frac{\partial L(y_i, f(x_i))}{\partial f(x_i)} \right]_{f(x)=f_{n-1}(x)}$$

- b) Fit a decision tree to \tilde{y}_{in} to estimate γ_n .
- c) Estimate β_n by minimizing loss function of $L(y_i, f_{n-1}(x_i) + \beta_n g(x, \gamma_n))$;

- d) Update $f_n(x) = f_{n-1}(x) + \beta_n g(x, \gamma_n)$
3. Calculate $f(x) = \sum_n f_n(x)$

2.4.3. Neural Network

A typical NN structure is shown in Figure 7. Note the three parts: input layer, intermediate (also called the hidden layer), and output layer. Each neuron in the input layer is one predictor, denoted as X_i in Figure 1. 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. However, there is no specific guidance to determine the number of hidden layers and neurons. A typical approach is to choose the average number of input and output nodes. The value of a neuron in the input layer is transferred into hidden layers through a transformation function. The weight (W_{ij}) 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 with the corresponding connection weights (W in Figure 7). Information transfers from hidden layers to the output layer through an activation function. Activation functions could be an identity function, binary step function, logistics function, ArcTan function et al. Different activation functions greatly impact the result and performance of entire model. In this study, the target variable 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 selected and expressed as Equation (10) (McCulloch & Pitts, 1943):

$$f(x) = \begin{cases} 0 & \text{for } x < \text{threshold} \\ 1 & \text{for } x > \text{threshold} \end{cases} \quad (\text{Equation 10})$$

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.

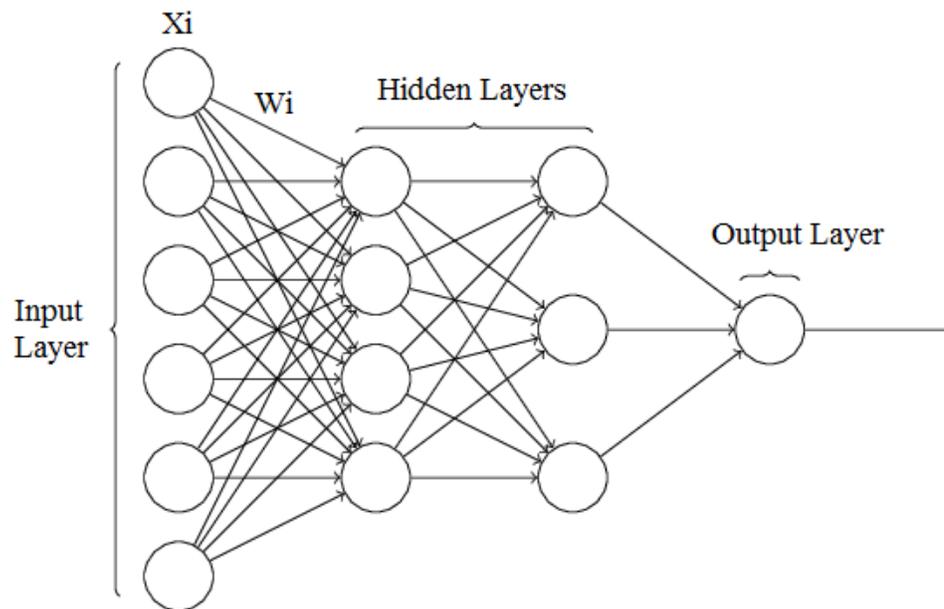


Figure 7. Structure of Neural Network

2.4.4. Over-Fitting

Over-fitting is a common problem in data mining. In predicting modeling, over-fitting happens when a model is too closely fit to limited set of data points or the true pattern of the data. However, this “perfect” model won’t perform well when fitting with other data sets. Over-fitting is more likely with nonparametric and nonlinear models that have more flexibility when learning a target function. As such, many nonparametric machine learning algorithms also include parameters or techniques to limit and constrain how much detail the model learns.

To prevent over-fitting, there are two main concepts. The first concept is to train the data mining model with different data set. This concept includes: cross-validation, training with more data, and removing features. Cross-validation is a powerful preventative measure against over-fitting. The idea is to use the initial training data to generate multiple mini train-test splits. Use these splits to tune the model. In standard k-fold cross-validation, the data is partitioned into k subsets, called folds. Then, the model is iteratively trained the algorithm on k-1 folds while using the remaining fold as the test set (called the “holdout fold”). Training with more data won’t work every time, but training with more data can help algorithms detect the signal better. The idea is to add more relevant data to the training data set. However, if more noisy data added, it will result in more errors. Thus, it is necessary to keep the data clean. Unlike adding more data, removing features is the opposite by manually removing irrelevant features to improve models’ generalizability.

The second concept is to set up proper parameters to the models, so that the model will stop before it gets to over-fitting. This concept includes: early stopping and regularization.

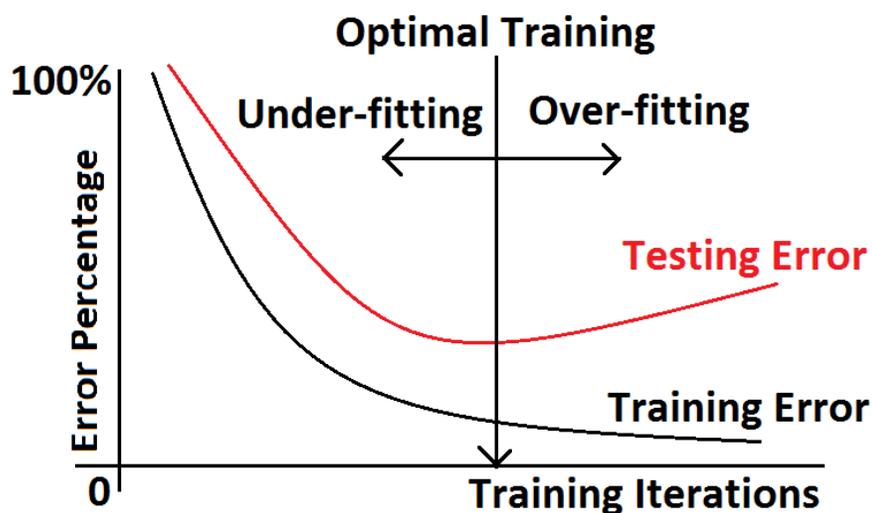


Figure 8. Problem of Over-fitting

In a data mining model, early stopping criteria represents when the model is set to stop training. The most common criteria is the number of training iterations. For a complex data mining model, such as the GB and the NN model, learning process is complicated and time consuming. When training with a large dataset, training time could be as long as a few hours or even days. However, as shown in Figure 8, it may not take a lot of training iterations or time to reach a certain desired training error percentage. Therefore, in most studies, the early stopping criteria is set up, so that the training will stop after reaching the desired error percentage instead of reaching to the 0% training error model generated. In a DT model, the number of training iterations is also called tree depth by some researchers.

Regularization parameters are critical for avoiding over-fitting and for improving model performance. For the selected three data mining models, popular regularization parameters are: learning rate for the GB and NN model, and tree complexity for the GB model. Learning rate is also called shrinkage rate (SPM User Guide 2013). It controls how fast the model updated or improved after each stage. The value of learning rate ranges from 0 to 1. A small value of learning rate yields great improvement and minimizes loss function, but requires more iterations and computational time (De'ath 2007). Higher values, close to 1, result in over-fitting and poor performance (SPM User Guide 2013). Tree complexity represents the number of nodes per single simple tree. A tree with two nodes is the simplest tree, which has only one split variable (Hastie et al., 2009). As the GB model's performance is controlled by both of learning rate and tree complexity, both learning rate and tree complexity rate must be balanced to avoid over-fitting. To detect interactions between variables, and to take full advantage of the GB model, a higher level of tree complexity and a low learning rate are suggested for experimentation (SPM User Guide, 2013). In this study, the model is tested under three values of learning rate: 0.05,

0.01, and 0.005, and five levels of tree complexity: 2, 4, 6, 8, and 15. The NN model is tested under three values of learning rate: 0.05, 0.01, 0.005, and 0.001, with the same purpose.

In this study, a combination of the two concepts is applied to avoid over-fitting in the study. This combined method starts with splitting the raw data into two: a training data and a testing data. Then, the models are tested under the training data and testing data, and under different regularization parameter setups at the same time. The optimal DT, GB, and NN are selected when both of training and testing errors are relatively low compared with all setup combinations. Both of the training and testing data are originated from the raw data, so it avoids involving irrelevant data to the model training. After training the model with training and testing data, a prediction error trend will generated like the one in Figure 8. The selection of training and testing data percentage is simulated by assigning the testing data percentage from 10% to 40% with 5% incremental. The optimal percentage is selected when both of predicting errors from training and testing data are relatively low. Early stopping parameter is set up; otherwise, the training process will keep running until the model fit the training data perfectly.

2.4.5. Rare Event Predictions

A data set, in which one class is exceptionally rare, is defined as unbalanced data (Cieslak and Chawla, 2015). When fitting the data in traditional predictive model, decisions will be biased towards the majority class. This will result in a good prediction for the majority class, but a relatively poor prediction for the minority class, which can be observed in previous studies (Chang and Chen, 2005; Chang and Wang, 2006; Pande, Aty and Das, 2010; and Chang and Chien, 2013). Some of them even failed to forecast rare events. Specifying correct prior probabilities and decision consequences is generally sufficient to achieve correct prediction results of a rare class in a predictive model. Usually, prior probability is represented by the

frequency of each class. Setting up an equal prior probability for both classes of target variable makes the rare class more over represented than before (SAS Institute Inc. Prior Probabilities, 2015). When increasing the prior probability of the rare event class, it increases the posterior probability of the class, which moves the classification boundary for that class so that more observations are classified into the class. Also, the benefit of choosing the best decision for a case from a rare class should be larger than that from a common class (SAS Institute Inc. Detecting Rare Classes, 2015). By doing so, the rare class is more heavily weighted, so that the model will treat both classes equally. The theory dealing with a rare event is complicated. A more detailed description of the theory can be found in (Wielenga, 2007).

2.4.6. Variable Importance Analysis

Data mining models are criticized by researchers due to failure to generate practical outputs. Thus, in this research, further variable importance analysis is conducted.

For decision tree and gradient boosting models, 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 improvement on mean squared error attributed to the tree due to the splits by the variable. After summing the importance score computed in a simple single tree over the ensemble of trees, the average value of the summation is scaled, so that the most important variable scores 100. Then, the scaled average value is regarded as the variable's importance in the model. A high value of variable importance indicates a high contribution that a variable makes to the prediction (Fridman and Meulman, 2003).

In a NN model, starting with storing all input variables in the input layer, information of input variables is transferred through hidden layers to the output later. Impact of each input variables on the outcome is defined as variable importance. Generally, variable importance in a

NN model can be defined by two types of ideas: connection-weights-based idea, such as the connection weights (Olden & Jackson, 2002), Garson's algorithm (Garson, 1991), and mean-square-error-based idea, such as forward stepwise addition, backward stepwise elimination, improved stepwise selection (Gevrey et al, 2003). Connection-weights-based idea identifies important variables as those with the greater value of connection weights. For example Garson's algorithm evaluates the variable importance of each variable by summing the absolute value of connection weights associated with each input neuron. Considering the huge difference in the variables' range, before measuring variable importance based on connection-weight-based ideas, all input variables have to be standardized into the same range to avoid connection weight biasing to variables with lower value scale. For example, in the studied data, AADT variable varies from 0 to more than 10,000 compared with variable ADVWARN (Table 2) with range of 0 to 1. Without standardization, even if they would equally impact the target variable in essence, AADT would be given an extremely small connection weight to fit the activation function, and AADT would be identified as a less important variable. Standardization methods include 0-1 scaling method, range-based method, Z-score scaling method, and standard deviation-based method (U.S. Patents, n.d.). However, errors are generated during standardization because an underlying relationship between input variables and the target variable is assumed before standardization no matter which standardization method is selected.

On the other hand, mean-square-error-based idea requires no assumption like connection-weights-based idea does. For example, the backward stepwise elimination algorithm by Gevrey, Dimopoulos & Lek (2003) measures the change in mean square error by sequentially removing input variables from the neural network. A more significant change in mean square error indicates a greater impact of an input variable on the target variable. In this study, NN variable

importance is measured based on both connection-weights-based and mean-square-error-based ideas.

A general method to evaluate the effect of explanatory variables on the response variable is to describe the relationship between the response variable and the studied variable across its range while holding the other variables consistent (Fish & Blodgett, 2003). However, Gevrey, Dimopoulos, and Lek (2003) pointed out that in the traditional method, a fair value could hardly be determined to keep the other variables consistent with, and the meaning of the value is arguable. For example, in the study of crash likelihood, when binary variables represent the presence of warning devices, to keep them at their mean values is absurd. In addition, in this case, the explored relationship between response variable and an explanatory is meaningless. Furthermore, the effect of an explanatory variable on the response variable is expected to change when the other variables are kept at different values. Thus, Gevrey, Dimopoulos, and Lek (2003) proposed a new method to explore and record all the generated relationships between the target variable and explanatory variables while holding the remaining variables at various and meaningful levels. In this study, the effects of explanatory variables on crash rates is conducted within an explanatory variable value range while holding all the other variables either at their highest level or at their lowest level.

2.4.7. Model Prediction Accuracy

Prediction results of a categorical prediction analysis can be summarized in a classification table (Table 3), based on which, model prediction accuracy measurements are computed. The observed event class is represented by present condition in Table 3, while the observed non-event class is represented by absent condition in Table 3. If an observation is predicted to be event class, it is indicated positive in Table 3, otherwise negative. The number of

true positive (a) and true negative (b) indicate the number of correct predictions when condition is present and absent, respectively. The number of false positive (c) and false negative (d) indicate the number of wrong predictions against observed conditions.

Table 3. Classification Table

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

Even though the prediction accuracy is critical indicator to measure model performance, only a limited number of researchers published their prediction accuracy results in their studies (Prati, Pietrantoni, Fraboni, 2017; McCollister and Pflaum, 2007; Saccomanno, Fu, Miranda-Moreno, 2004; Dhruvit, Varia, Shah, 2016). Prediction accuracy in all the studies above is evaluated based on equation (11), (12), and (13) for event class, non-event class, and overall prediction, respectively.

$$Accuracy_traditional_{event} = \frac{a}{a+c} \quad \text{(Equation 11)}$$

$$Accuracy_traditional_{non-event} = \frac{d}{b+d} \quad \text{(Equation 12)}$$

$$Accuracy_{overall} = \frac{a+d}{a+b+c+d} \quad \text{(Equation 13)}$$

In previous studies, what equation (11) and (12) actually measure can be considered as true alarm rate of event class and non-event class. In other words, $Accuracy_traditional_{event}$ and $Accuracy_traditional_{non-event}$ compute the number of correct predictions given the number of observed conditions. For instance, when evaluating event class prediction accuracy, Equation 11 indicates that among all the present conditions ($a+c$), the model makes a number of a correct

predictions. However, it ignores the number of false positive predictions (b). In fact, the model predicts ($a + b$) to be event class, among which, a of them are correct. Thus, the new measurement, prediction accuracy of event class, is proposed here to measure the ratio of the number of true positive to the total number of forecasted positive. It is easy to see that traditional prediction accuracy parameters only partially represent a model's prediction accuracy. To draw the full picture of a model's prediction accuracy, four measurements are proposed to evaluate model prediction accuracy: traditional prediction accuracy, proposed prediction accuracy, forecasted class rate, and observed class rate.

Proposed prediction accuracy is computed as equation (14) and (15) for event class ($Accuracy_proposed_{event}$) and non-event class ($Accuracy_proposed_{non-event}$), respectively. Forecasted class rate is calculated by equation (16) and (17) for event class ($Forecast_{event}$) and non-event class ($Forecast_{non-event}$), respectively. Different from traditional accuracy measurement, the proposed prediction accuracy measures the ratio of true prediction, given the actual predicting counts instead of the number of true event. Forecasted class rate indicates the predicted event and non-event ratio.

$$Accuracy_proposed_{event} = \frac{a}{a+b} \quad (\text{Equation 14})$$

$$Accuracy_proposed_{non-event} = \frac{d}{c+d} \quad (\text{Equation 15})$$

$$Forecast_{event} = \frac{a+b}{a+b+c+d} \quad (\text{Equation 16})$$

$$Forecast_{non-event} = \frac{c+d}{a+b+c+d} \quad (\text{Equation 17})$$

2.4.8. Marginal Effect of Influential Variables

To further facilitate the use of the GB and NN model and respond to the criticism that the GB and NN models are often unable to generate interpretable parameter for each influential variable, it is necessary to evaluate the magnitude of the relationship and determine the direction of the relationship between contributors and target variables. A general method to evaluate the effect of explanatory variables on the response variable is to describe the relationship between the response variable and the studied variable across its range while holding the other variables consistent (Fish & Blodgett, 2003). In this study, influential variables' marginal effect analysis based on the GB model will be conducted based on Fish and Blodgett's method. However, Gevrey, Dimopoulos, and Lek (2003) pointed out that in the traditional method, a fair value could hardly be determined to keep the other variables consistent with, and the meaning of the value is arguable. For example, in the study of crash likelihood, when binary variables represent the presence of warning devices, to keep them at their mean values is absurd. They argued that in the situation, the explored relationship between response variable and an explanatory is meaningless. Furthermore, the effect of an explanatory variable on the response variable is expected to change when the other variables are kept at different values. Thus in the NN model, the effects of any single explanatory variable on crash rates is conducted with holding all the other variables either at their highest levels or at their lowest levels. It should be noted that the NN models do not assume that explanatory variables may not be completely independent with each other. In fact, they may depend on each other. Thus, in this study we keep other explanatory variables kept at various levels. In a DT, when changing variables in a DT model, the generated tree structure will change accordingly. Thus, the marginal effect analysis is not applicable in a DT model.

2.5. Results Analysis

2.5.1. Decision Tree Results Analysis

A total of 30 predictors (independent variables) were tested as input variables in the model for crash frequency (target or dependent variable) prediction at HRGCs. The input variables represent crossing attributes, highway attributes, and both train and vehicle traffic characteristics. Description of input variables is displayed in Table 2. The results of the model, shown in Figure 9, are applied for crash frequency prediction at HRGCs. The final hierarchical tree structure for HRGC crash frequency shown in Figure 9 involves 14 splitters, includes but not limited to NGHTTHRU, average_train_speed, DAYTHRU and AADT_N. They are listed in Table 4 with the second column value greater than 0.

The first optimal split in node 1 is based on daytime through-train traffic (DAYTHRU), which classifies the crossings into four groups in nodes 2 and 3, and terminal node 1 and 2: DAYTHRU is less than 1 trains per day, DAYTHRU is greater than 1 trains per day and less than 2 trains per day, DAYTHRU is greater than 2 trains per day and less than 7 trains per day, and DAYTHRU is greater than 7 trains per day. Terminal node 1 and 2 indicate that the tree predicts 28 crashes per year and 67 crashes per year for them respectively. The tree further splits node 2 and 3 according to the same predictor variable, highway traffic (AADT_N) to form the second level of the tree. However, the splitting thresholds are different between the two DAYTHRU groups (nodes 1 and 2). The second level of the tree has total 8 nodes, of which 2 are terminal nodes: terminal 3 and 4. For terminal node 3 (DAYTHRU between 0.5 and 1.5 and AADT greater than 4477) the tree predicts 5 crashes/year. The data are further segmented into various subgroups through a total of eight levels of the tree with a total of eleven predictor variables until terminal node are reached. For detailed model results, please refer to Figure 9.

The tree has 44 terminal nodes and it can be observed that all of the splitter variables are associated with the terminal nodes. This implies that these variables are critical in predicting crash frequency at HRGCs. However, not all important variables are splitters. In contrast to a regression-based model, a variable in decision tree can be highly important, but never be identified as a node splitter. These variables are normally selected as surrogate splitters. When the value of a primary splitter is missing, a surrogate splitter will be selected instead of dropping the record. This is also why the decision tree is superior to regression-based models in handling missing values.

Table 4 summarizes the variables by their importance in predicting crash frequency in the tree. Variable importance is calculated based on sum of square error (SSE) of each variable. The importance value of the most important variable is 1. Then all other variables are assigned with a relative importance (SAS Institute Inc. 2016). In general, the identified critical variables in predicting crash frequency at crossings meet the expectation and are consistent with the findings of previous studies ((Austin & Carson, 2002), (Hu, Hsieh, & Lee, 2013)).

Table 4. Variable Importance

Variable	Times used as splitting rules	Times used as surrogate rules	Importance
NGHTTHRU	2	9	1
Average_Train_Speed	4	5	0.9295
DAYTHRU	1	0	0.8221
AADT_N	2	0	0.737
HWYSYS	1	3	0.6703
SGNLEQP	0	4	0.6472
Highway_Paved	0	3	0.6456
SPSEL	3	3	0.6099
TRAFICLN	3	2	0.4382
GATES	0	8	0.4283
FLASHMAS	1	2	0.4094
Total_Number_Track	1	3	0.3916
ADVWARN	2	3	0.2328
DOWNST	0	3	0.2313
FLASHPAI	1	3	0.204
XBUCK	0	3	0.2021
NGHTSWT	0	4	0.1877
DAYSWT	1	2	0.1705
PAVEMRK	1	0	0.0987
COMPOWER	2	1	0.086
TRUCKLN	0	2	0.0809
SCHLBUS	0	1	0.0734
XANGLE	0	1	0.0435

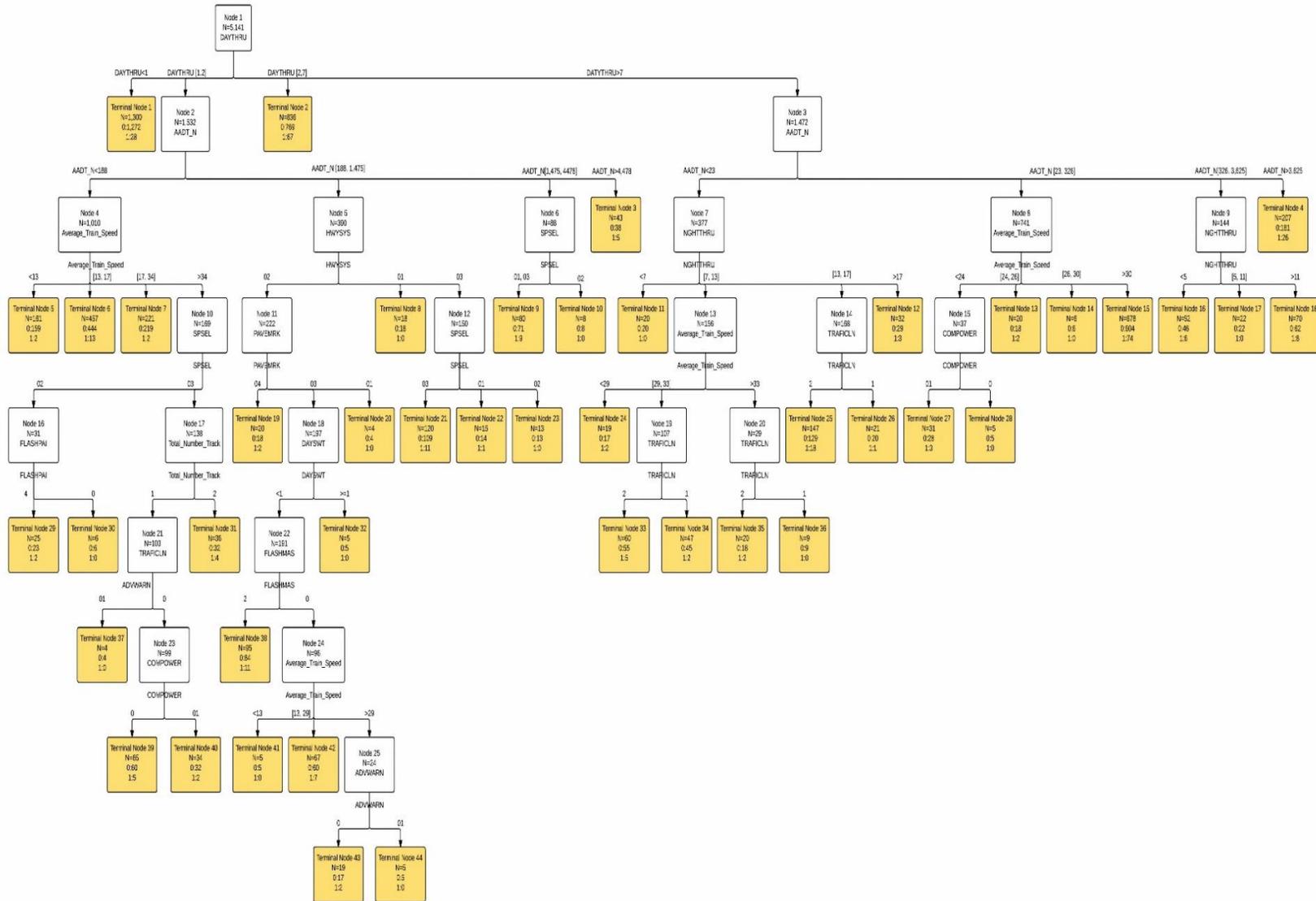


Figure 9. Decision Tree Model Output

The effect of splitter variables on HRGC crash likelihood can be observed in the decision tree model output summarized in Table 5. In Table 5, a positive effect of an influential variable on crash likelihood indicates that crash likelihood increases as the value of the influential variable increases. A negative effect indicates crash likelihood decrease as the value of the variable increase. An unclear effect indicates that contributor variables are categorical class variables such as HWYSYS where the numerical value of the variable only represents variable classes. Unclear can also mean the effect of the variable is truly unclear or the impact of variable can be positive or negative depending on the conditions such as COMPOWER. However, effect of variables used as surrogate rules is invisible through the result tree, although they also make great a contribution to predicting the target variable.

Table 5. Effect of Factors on Crash Likelihood

Variable Name	Effect on crash likelihood
NGHTTHRU	Positive
Average_Train_Speed	Positive
DAYTHRU	Positive
AADT_N	Positive
HWYSYS	Unclear (Categorical)
SPSEL	Unclear (Categorical)
TRAFICLN	Positive
FLASHMAS	Positive
Total_Number_Track	Positive
ADVWARN	Negative
FLASHPAI	Positive
DAYSWT	Negative
PAVEMRK	Unclear (Categorical)
COMPOWER	Unclear

The overall impact of traffic characteristics on crash rate is consistent and meets the expectation: for highway traffic, railroad traffic volume and train speed, HRGC crash frequency is increasing with the increasing variable values. For example, as indicated by DAYTHRU in terminal nodes 1 and 2, the number of crashes is 67 with a day through-train movement between

2 and 7, and it drops to 28 when day through-train movement is less than 1. It indicates that HRGCs with a high train traffic or highway traffic, or fast-moving trains have a high possibility of crashes.

By analyzing the relationship between crossing characteristics and crash rate, expected outcomes are achieved: 1) the presence of some warning devices, including advance train warning devices, and train detection systems, helps to reduce crash likelihood at crossings. Crash probability tends to be higher at crossings without advance train warning devices (ADVWARN=0). And 2) type of train detection system (SPSEL) also impacts the crash rate. Crossings with no train detection systems (SPSEL=3) have the higher the crash likelihood, compared with crossings with constant warning time system (SPSEL=1). Direct current audio frequency overlay system (SPSEL=2) has the least crash likelihood among all three train detection systems. As a splitting rule, six terminal nodes generated by TRAFICLN indicate that crossings with more traffic lanes have higher crash likelihoods.

The number of highway lanes positively impacts on crash risk, meaning that crossings intersecting with highways with more lanes will be more likely to have crashes. Crossings intersecting with an interstate national highway (HWYSYS=1) have the smallest crash rate (zero likelihood) among all HWYSYSs, followed by non-federal aid highway (HWYSYS=4), federal aid highway (HWYSYS=3) and crossings intersecting with a non-interstate national highway (HWYSYS=2) have the highest crash likelihoods.

Some interesting findings are also observed. When day through-train traffic is low (1 to 2 trains per day), presence of FLASHMAS and FLASHPAI will have higher crash likelihoods. Crossings with pavement markings within 200 feet have the least crash likelihood among all the other distance-groups of pavement markings.

Effect of COMPOWER on crash rate depends on train traffic. When DAYTHRU is greater than 7, crashes are more likely to happen at crossings with access to commercial power. When DAYTHRU is low (between 1 and 2), the presence of commercial power results in lower crash likelihood.

2.5.2. Gradient Boosting Model Results Analysis

This section presents findings based on the GB model. First, the selection of an optimal model, based on model performance, is presented. Then, based on the optimal model, influential variables are ranked by their importance to the target variable. After that, impacts of top influential variables on crash prediction are analyzed.

2.5.2.1. Model Setup

Table 6 shows how the model performs with various learning rate and tree complexity levels. Class represents target variable class, where class=0 indicates a non-event level or, in other words, no crash happened, while class=1 indicates an event level where a crash occurred. Columns of “training data error pct.” and “testing data error pct.” show the percentage of prediction error for training data and testing data respectively. The last column, “Number of Trees,” shows the number of trees needed to train for an optimal model under corresponding learning rate and tree complexity. It is clear that for a lower learning rate or a lower tree complexity, more trees are needed to achieve the optimal model. However, there is no clear relationship between learning rate and prediction error. Prediction error is associated with both learning rate and tree complexity. Note that when learning rate decreases, prediction error increases when tree complexity is 2. However, when tree complexity is 8, prediction error fluctuates as learning rate decreases. On the other hand, under a constant learning rate, as tree complexity increases, prediction error for training data decreases.

The optimal model should predict well for both training and testing data, in addition, accurate prediction of event level is also critically important. Moreover, the number of trees required to achieve the optimal performance model indicates computing time and should be considered when selecting regularization parameters. By balancing model performance in terms of training error, testing error, event forecasting error, non-event forecasting error and number of trees need to obtain optimization model, this research selected the model with a learning rate of 0.01, tree complexity of 8 and an ensemble of 1,092 trees as the optimal model. Prediction accuracy for the optimal model is 85.7% for non-event level (ACCIDENT=0), and 83.9% for event level (ACCIDENT=1). Variable importance and their impacts on crash at HRGCs will be generated based on the optimal model using 1,092 simple decision trees.

Table 6. Misclassification Rate vs Learn Rate and Complexity of Trees

Learning rate	Tree complexity	Class	Training data error Pct.	Testing data error Pct.	number of trees
0.05	2	0	0.1737	0.1847	1439
		1	0.1691	0.1829	
	4	0	0.1486	0.1611	468
		1	0.1544	0.2927	
	6	0	0.1404	0.1574	394
		1	0.1397	0.2927	
	8	0	0.1216	0.141	181
		1	0.1287	0.3171	
	15	0	0.112	0.1292	181
		1	0.114	0.3171	
0.01	2	0	0.18	0.1938	2853
		1	0.1949	0.2073	
	4	0	0.1491	0.1629	2461
		1	0.1581	0.2683	
	6	0	0.143	0.1565	1424
		1	0.1507	0.2683	
	8	0	0.1399	0.1547	1092
		1	0.1324	0.2561	
	15	0	0.1153	0.1338	711
		1	0.1176	0.3049	
0.005	2	0	0.1808	0.192	5317
		1	0.2132	0.2073	
	4	0	0.1514	0.1656	4528
		1	0.1618	0.2561	
	6	0	0.1446	0.1611	2745
		1	0.1544	0.2561	
	8	0	0.1277	0.1383	1986
		1	0.1287	0.2927	
	15	0	0.1164	0.1328	1387
		1	0.1103	0.3049	

2.5.2.2. Variable Importance

The importance of a variable in a simple single tree is measured by the number of times the variable is used as splitter and the squared improvement attributed to the tree due to the splits by the variable. After summing the number of times used as splitter and the squared

improvement over the ensemble of trees, the average value of the summation is regarded as the variable importance in the model. A high value of variable importance indicates a high level of contribution that a variable makes to the prediction (Friedman & Meulman 2003). The GB model uses the same algorithm to measure variable importance.

Table 7 presents the relative variable importance of each influential factor based on the selected GB model. The “Relative Importance” column shows the importance value of the corresponding variable. It is computed by assigning an importance value of 100 to the most important variable and then giving all other variables a relative importance value (SAS Institute Inc., 2017). “Influence Pct (%)” is an absolute importance factor which indicates how much each variable contributes to the prediction, or its influence power in percentage. The last column is cumulative influence in percentage. Twenty eight factors out of thirty are identified as having impacts on crashes at HRGCs and the top ten factors contribute about 60% of the total crash influence power. It is clear that, except for average annual daily highway traffic, no single variable makes large individual contributions to the prediction. Single factor influence power ranges from 1% to 11%, and the majority of them are less than 5%. In other words, crashes at HRGCs are complicated, and cannot be explained by only a few factors.

Among all 30 influential factors, average annual daily highway traffic, daily through-train traffic, train detection type, nightly through-train traffic, average train speed, and the number of traffic lanes are the top six contributors to crash prediction, with individual influence percentage greater than 5%. Among these six variables, four of them are traffic characteristic variables describing highway and railway traffic exposure situations and contribute about 30% to prediction. Most of the predictors are crossing characteristics (17 out of 30), such as SPSEL,

ADVWARN, and PAVEMRK, which provide information about warning systems and train detecting systems, and they cumulatively contribute to about 50% of the impacts.

Table 7. Variable Importance Based on GB Model

Variable	Relative Importance	Influence Pct (%)	Cumulative Influence Pct (%)
AADT_N	100.00	11.00	11.00
DAYTHRU	73.33	8.07	19.07
SPSEL	67.73	7.45	26.52
NGHTTHRU	58.28	6.41	32.93
AVERAGE_TRAIN_SPEED	54.93	6.04	38.98
TRAFICLN	45.62	5.02	44.00
HWYSYS	37.98	4.18	48.18
ADVWARN	37.19	4.09	52.27
TOTAL_NUMBER_TRACK	35.75	3.93	56.20
PAVEMRK	34.15	3.76	59.96
XANGLE	32.85	3.61	63.57
FLASHPAI	31.24	3.44	67.01
COMPOWER	27.52	3.03	70.04
HIGHWAY_STOP	27.36	3.01	73.05
NEAR_CITY	26.31	2.89	75.94
XBUCK	25.73	2.83	78.77
HIGHWAY_PAVED	21.93	2.41	81.19
FLASHMA	21.42	2.36	83.54
SCHLBUS	19.78	2.18	85.72
SGNLEQP	19.45	2.14	87.86
GATES	18.00	1.98	89.84
TRUCKLN	16.70	1.84	91.68
STOPSTD	14.24	1.57	93.24
DOWNST	13.58	1.49	94.74
WHISTBAN	13.46	1.48	96.22
DAYSWT	13.19	1.45	97.67
NGHTSWT	11.80	1.30	98.97
FLASHOV	9.37	1.03	100.00
FLASHNOV	0.00	0.00	100.00
WIGWAGS	0.00	0.00	100.00

2.5.2.3. Marginal Effect of Contributing Variables

One of the objections frequently found in literature for newer predictive modeling approaches, such as gradient boosting machines is the difficulty of interpretation relative to linear regression models. For that reason, the marginal effect analysis is conducted in this study by showing the partial dependent plots. Partial dependent plots can be viewed as a graphical representation of contributor coefficients for each individual independent variable. Partial dependent plots are not directly drawn from the raw data, but from predictions based on the model. Essentially, they are model-based simulations (Salford Systems, 2017). The values appearing in the y-axis are the modeled values of the response variable. A positive y value indicates that the contributing variable at the corresponding value has a positive influence for the classification in the model, or makes prediction of a “yes crash” more likely and a negative value indicates the opposite. Figure 1 illustrates the use of partial dependence plots to characterize the marginal effects of three types of contributor: traffic, highway, and crossing characteristics.

Figure 10 (a), (b), (c), and (d) present the effects of AADT, DAYTHRU, NGHTTHRU, and AVG_TRAIN_SPEED on crash at HRGCs, respectively. They all exhibit complex, nonlinear patterns with several peaks and valleys. However, it is clear that a roughly increasing pattern exists in all three plots. In addition, “no crash” is more likely to be predicted with an extremely low highway or railway volume. However, in Figure 10 (a), crash rate suddenly reaches a peak value when AADT is about 500, which indicates that a “crash” is very likely to be predicted when AADT is 500 and then drops back to the general trend. In Figure 10 (b), crash likelihood stays roughly constant when DAYTHRU is between 7 and 20. As to Figure 10 (c), crash rate fluctuates at a low rate before NGHTTHRU reaches 11, beyond which, a sudden dramatic increase is observed which indicates crash likelihood increase dramatically if nighttime

through train traffic increases from 11 to 13 and remains high when nightly through train volume is greater than 13. Figure 10 (d) suggests that a crash is less likely when train speed is less than 30 mph and increases dramatically when train speeds increase to 35 mph and then drop fast when speeds increase from 35 to 40 mph. As shown, crossings with trains travelling at speeds between 3 and 13 are more likely to have “no crash” prediction results.

Figure 10 (e) shows the effect of one of the highway characteristic variables: HWYSYS. It is found that crashes tend to happen at crossings intersecting federal-aid highways (coded as 3). In contrast, crossings intersecting with non-Interstate highways (coded as 2) or non-federal-aid highways (coded as 4) are very likely to have no crash.

Effect of crossing characteristic variables is critical for HRGC design. Figure 10 (f) and (g) show the effects of two crossing characteristic variables, SPSEL and TRAFICLN respectively. It indicates that a direct current audio frequency overlay (SPSEL=2) installed at a HRGC helps to reduce the likelihood of crashes. It also suggests that crashes tend to happen at HRGCs with constant warning time systems. As shown in Figure 10 (f), highways with no more than 2 lanes have a negative impact on crash. Moreover, it is noticeable that a highway with 4 lanes has the highest positive impact on crash prediction.

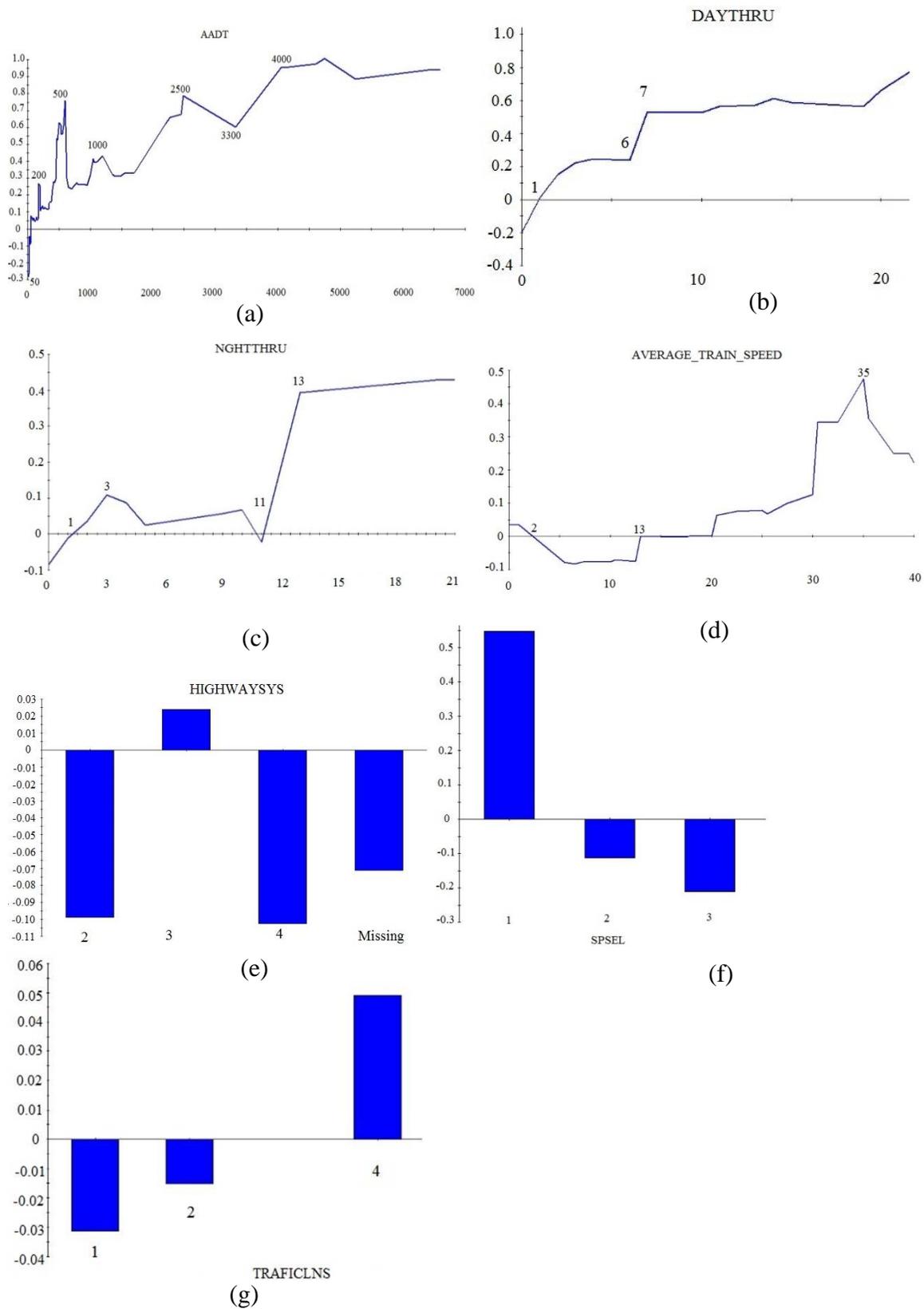


Figure 10. Partial Dependent Plots (GB Model)

2.5.3. Neural Network Model Results Analysis

2.5.3.1. Final Optimal NN Model

With the 30 independent variables as input variables shown in Table 2, the optimal network structure is selected based on the dominant selection criteria, which is sum of squared errors with following procedure steps: 1) Determine the number of candidate dimensions for the hidden layers. To detect non-linear pattern, at least one layer of hidden layer is highly recommended. As the number of hidden layers increases, it forms a more complicated NN, which requires much longer time to process, and the model performance does not necessarily become better (Nielsen, 2015). In this case study, the number of hidden layers is tested from 1 up to 5. It is found that when 3 and more hidden layers are set up, the NN model fails to converge to an optimal result. Regarding to choosing the number of neurons in hidden layer, the rule of thumb is to be less than the number of neurons in the input layer (Nielsen, 2015). Thus, the number of neurons in hidden layers was tested from 1 to 21. The optimal model structure is selected where the sum of squared errors from training and testing data are both the lowest. Based on such selection criteria and procedures, the optimal network structure proved to be 22-10-1 with the optimal training data percentage of 70%.

2.5.3.2. Variable Importance

NN models are often criticized by researchers as a black box and do not generate interpretable parameter for each explanatory variable, in this research further variable importance analysis and marginal effect analysis are conducted to demonstrate relative and quantitative effects of each explanatory variable.

As stated earlier, it is important to identify contributors to crash likelihood, and to predict crash accurately. With these two objectives, two criteria are selected to measure variable importance based on the final optimal NN model: connection weight and mean square error.

Connection-weight-based algorithms, such as connection weights (Olden & Jackson 2002), and Garson's algorithm (Garson 1991) identify important variables as those with the greater value of connection weights. The connection weight of each input variable represents strength of connectivity that each input variable transfer its information to the next layer and eventually to the final outcome variable. The variable transferred most of its information through layers to the final outcome variable will assign the highest importance score with this criterion. These algorithms focus more on causal importance because connection weights are directly related to causal importance of the inputs. Garson's algorithm evaluates the importance of each variable by summing the absolute value of connection weights associated with each neuron throughout each layer. Considering the different variables' scales, before measuring variable importance based on connection-weight-based algorithm, all input variables have to be standardized to avoid connection weight biasing to variables with lower value scales. For example, in the studied data, AADT variable varies from 0 to more than 10,000 compared with variable ADVWARN (Table 2) with range of 0 to 1. Without standardization, even if they would equally impact the target variable in essence, AADT would be given an extremely small connection weight to fit the activation function, and AADT would be identified as a less important variable. Standardization methods include the 0-1 scaling method, range-based method, Z-score scaling method, and standard deviation-based method (Everitt 1993).

On the other hand, mean-square-error-based algorithms, such as forward stepwise addition, backward stepwise elimination, and improved stepwise selection (Gevrey et al. 2003),

measure the importance of an input by eliminating it from the model and then observing how much the error increases. These algorithms focus more on predictive importance because the change in the error function is a direct measure of predictive importance. The backward stepwise elimination algorithm proposed by Gevrey, Dimopoulos and Lek (2003) measures the change in mean square error by sequentially removing input variables from the neural network model. A more significant change in mean square error indicates a greater impact of an input variable on the target variable, thus this variable will assign the highest importance score for this criterion.

In this study, NN variable importance is measured based on both connection-weights-based and mean-square-error-based algorithms. For different application purposes, both methods will determine the relative importance of the variable.

Table 8 summarizes variable importance by the NN model based on Garson's algorithm and backward stepwise elimination algorithm. Among 30 of input variables, 20 of them identified as significant impact variables regarding to both reduce forecasting errors (backward stepwise elimination) and explain the forecasting relationship (Garson's algorithm). The most important variable is assigned with importance value of 1 and then all other variables are assigned with a relative importance (SAS Institute Inc. 2015). Garson's algorithm and the backward stepwise elimination algorithm generate two different variable importance rankings.

Table 8. Variable Importance (NN Model)

Backward Stepwise Elimination		Garson's Algorithm	
Variable	Importance	Variable	Importance
AADT	1	FLASHNOV	1
TRAFICLN	1	FLASHMAS	0.9
Number of tracks	0.6	STOPSTD	0.87
FLASHMAS	0.55	COMPOWER	0.76
DAYTHRU	0.54	NGHTTHRU	0.71
DAYSWT	0.53	XBUCK	0.6
NGHTTHRU	0.53	Train speed	0.4
FLASHNOV	0.53	WIGWAG	0.27
XBUCK	0.52	SCHLBUS	0.21
ADVWARN	0.51	Number of track	0.15
SCHLBUS	0.5	HIGHWAY_PAVED	0.11
GATES	0.49	Number of Lanes	0.1
STOPSTD	0.44	ADVWARN	0.1
Train speed	0.43	AADT	0.1
FLASHPAI	0.4	GATES	0.1
NGHTSWT	0.3	DAYTHRU	0.1
HIGHWAY_PAVED	0.28	NGHTSWT	0.1
COMPOWER	0.27	FLASHPAI	0.1
DOWNST	0.26	DAYSWT	0.1
WIGWAG	0.24	DOWNST	0.1

Table 8 shows the difference of variable importance by the backward stepwise elimination algorithm and Garson's algorithm. As indicated earlier, variable importance defined by Garson's algorithm focuses on explaining the forecast relationship while importance defined by backward stepwise elimination is based on assessment of the change in the mean square error of the network. Variable importance based on Garson's algorithm shows that crossing characteristics play a more important role in crash likelihood prediction at HRGCs than traffic characteristics. As shown in Table 8, four of the top five influential factors are crossing characteristics. In contrast, the variable importance based on the backward stepwise elimination algorithm shows that four of five the most important variables are traffic characteristics: AADT, number of highway lanes, number of rail tracks, and day through train traffic volume.

The results based on backward stepwise elimination algorithm are more consistent with previous researchers' findings, especially the top 5 influencing factors (Austin and Carson 2002; Hu, Li and Lee 2012; Zheng, Lu and Tolliver 2016), however, previous researchers did not identify the importance criteria. Results indicate that AADT and number of highway lanes are equally the most important variables for reduction of the forecasting error. The next 10 important variables have similar importance scores. They describe railway traffic characteristics and the presence of warning devices such as flashing lights and cross bucks. These findings are intuitive and agree with previous studies (Millegan, et al. 2009; Chadwick, Zhou and Saat 2016; McCollister and Pflaum 2007), which indicates the importance criteria researched in the previous studies might be the focus of improving forecasting accuracy.

The variables importance based on two selection criteria are different, and reveal two different research focuses. If the focus of research is to identify influential contributors to crashes, then crossing characteristics are most important variables identified. But if the focus of research is to accurately forecasting crashes, then traffic characteristics are the most important contributors since they introduce the least forecasting errors.

2.5.3.3. Marginal Effects of Influential Variables

To further facilitate the use of neural networks and dispel the myth of black boxes, simply examining the relative importance of variables is not enough. It is also necessary to evaluate the magnitude of the relationship and determine the direction of the relationship between contributors and target variables.

A general method to evaluate the effect of explanatory variables on the response variable is to describe the relationship between the response variable and the studied variable across its range while holding the other variables consistent (Fish & Blodgett, 2003). However, Gevrey,

Dimopoulos, and Lek (2003) pointed out that in the traditional method, a fair value could hardly be determined to keep the other variables consistent with, and the meaning of the value is arguable. For example, in the study of crash likelihood, when binary variables represent the presence of warning devices, to keep them at their mean values is absurd. They argued that in the situation, the explored relationship between response variable and an explanatory is meaningless. Furthermore, the effect of an explanatory variable on the response variable is expected to change when the other variables are kept at different values. Thus in this study, the effects of any single explanatory variable on crash rates is conducted with holding all the other variables either at their highest levels or at their lowest levels. It should be noted that the NN models do not assume that explanatory variables may not be completely independent with each other. In fact, they may depend on each other. Thus, in this study we keep other explanatory variables kept at various levels.

In the following section, a detailed marginal effect analysis is introduced. Because of space limitations, six explanatory variables, representing traffic characteristics and crossing characteristics, are selected for use in the analysis. The selected variables' marginal effects are shown in Figure 11.

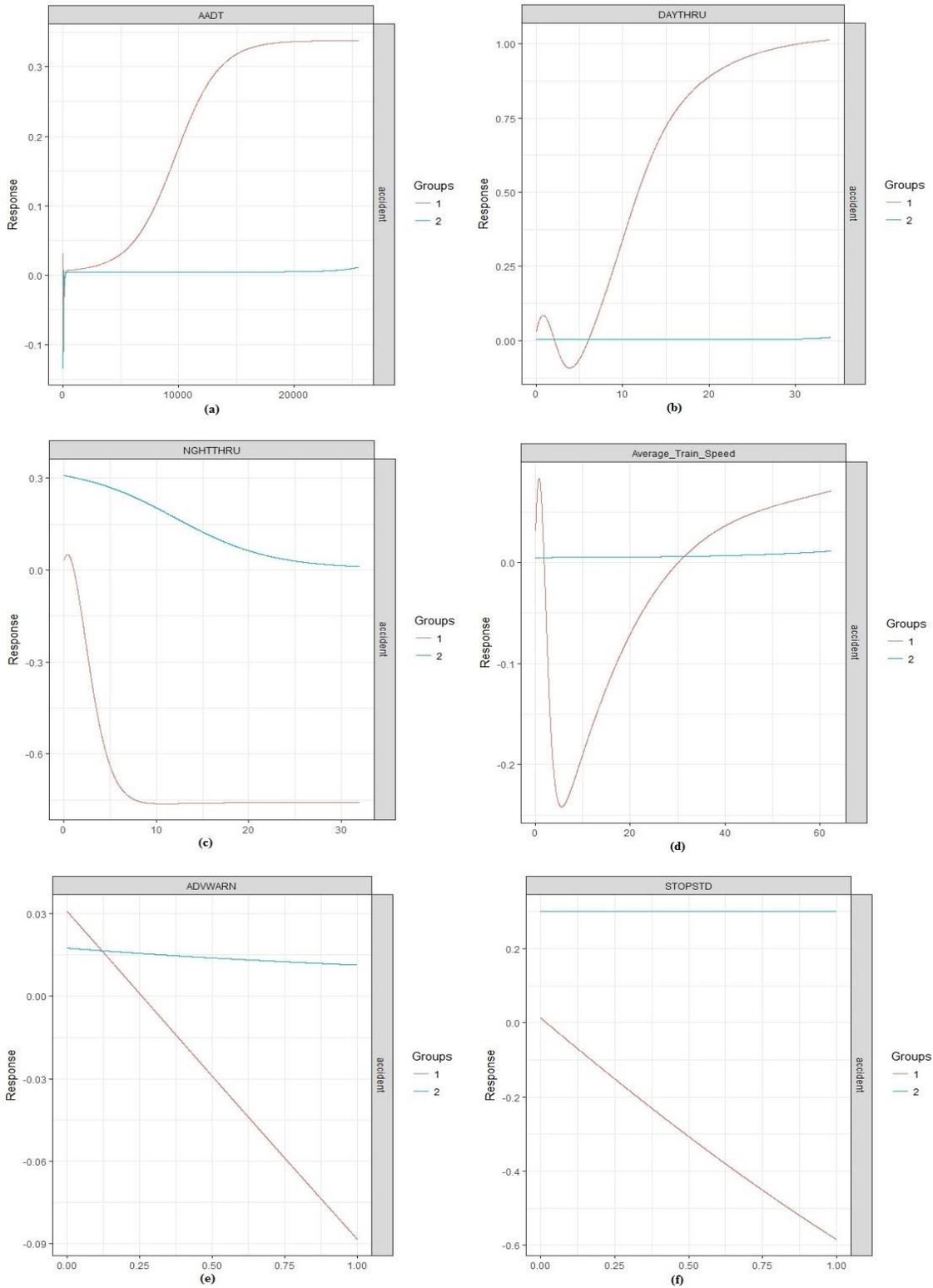


Figure 11. Crash Likelihood vs Explanatory Variables

Figure 11 analyzed contributor variables' impact on crash likelihood within the range of target contributor's reasonable value interval while holding all the other contributors in two group levels. As indicated in the methodology section, Group level 1 means all the other contributors were held at their lowest observed values. Group 2 means the opposite situation. The value of crash likelihood is represented by the vertical axis in Figure 11. A positive vertical value indicates that the corresponding target analysis variable has a positive impact on crash likelihood at HRGCs at the corresponding value, while a negative vertical value indicates that the target variable has a negative impact on crash likelihood at the corresponding value. The line trend indicates the change of direction in crash risk as increase or decrease the likelihood of crashes while increasing the analysis variables. From Figure 11, one can clearly tell that dynamic nonlinear relationships exist between numerical variables and target variable. Moreover, one marginal effect on target variable showing different relationship when all the other variables held at different levels.

As shown in Group 1 (all other unstudied variables such as warning devices are set at minimal level) AADT and day through train volume (Figure 11 (a) and (b)) have positive impacts on crash likelihood within their value range and crash likelihood increases exponentially as AADT and day through train volume goes up. However, there is a dip observed at lower values of both of the variables. For example, as AADT changes from 10,000 to 20,000, crash likelihood at HRGCs increases from 0.15 to 0.34. At the dip, crash likelihood is below zero, indicating that low volume of AADT and day through train traffic will have negative impact on crash likelihood at HRGCs. In other words, it is less likely that a crash will happen. Group 2 in Figure 11 (a) and (b) show that AADT and day through train volume will still have positive

impacts on crash likelihood. However, the impact is not significant (slightly greater than zero) when all other contributors are at their highest value or, in other words, at their highest influence.

Curve trending in Figure 11 (c) shows that increasing night through train volume will decrease crash likelihood with both control levels. A higher train volume at night means less train volume in the daytime, which reduces confliction between train and highway traffic. Thus, allocating more train movement at night could help to reduce crashes at HRGCs. However, note that the crash likelihoods are all positive for control group 2 (all other contributors are held at their highest values). This indicates that when all other factors are at their highest levels, for example, AADT is very high and all available warning devices are present at crossing, night through train is a positive contributor to crash likelihood. In other words, train traffic exposure at night will still contribute to crash likelihood positively. However, as night through train volume increases, crash likelihood becomes less sensitive to night through train volume. Note, the crash likelihoods are almost all negative for control group 1, which indicates that when AADT at its lowest level and there are no warning devices, even night time train traffic exposure will negatively impact on crash risk. In other words, even when there is night time train traffic volume, the crossing is less likely to have a crash.

For average train speed, group1 in Figure 11 (d) indicates that when other unstudied variables were set at their minimal level, an extremely low average train speed (less than 3) has a positive impact on crash likelihood, and the impact increases slightly as average train speed increases from 0 to 2, but declines to 0 when average train speed changes from 2 to 3. When average train speed is between 3 and 30 mph, it has a negative impact on crash likelihood. However, within this range, there is a “U” shaped relationship between average train speed and crash likelihood, reaching the lowest value at an average train speed of 7 mph. When average

train speed is over 30 mph, its impact on crash likelihood is positive and increases as average train speed increases. Group 2 in Figure 11 (d) shows that average train speed will still have a minor positive impact on crash likelihood and increases slightly when train speed is increasing, when all other contributors are at their highest values or highest impact.

Figure 11 (e) and (f) present effectiveness of static advance warning signs and stop signs. Group 1 in Figure 11 (e) demonstrates that a static advance warning sign is significantly effective in reducing crash likelihood at HRGCs when all other variables are held at their lowest value. When a static advance warning sign is not present ($ADVWARN=0$), the impact on crash likelihood is positive, which means a crash is more likely to happen at such condition. When a static advance warning sign is present ($ADVWARN=1$), crash likelihood is negative, indicating that presence of warning sign is significantly effective in preventing crashes at HRGCs. Similar conclusions can be drawn from Group 1 in Figure 11 (e) with regard to the presence of a stop sign which decreases crash likelihood. Note that from Group 2 in Figure 11 (e) and (f), that when all the other contributors are held at their highest observed values, or in other words, at their highest impact values, the presence of a stop sign has almost no impact on the likelihood of a crash and the presence of a static advance warning sign decreases the likelihood of crash slightly.

2.5.4. Model Prediction Accuracy

As indicated above, most of previous literatures do not present model prediction accuracy (Zheng, Lu, & Tolliver, 2015). In addition, the traditional prediction accuracy measurements fail to evaluate model prediction accuracy comprehensively. Prediction result classification table is presented in Table 9. Based on the model prediction accuracy measurements proposed in section 2.4.7, prediction results of the DT, GB, and NN models are summarized in Table 10. As stated by Hastie (2014), GB models usually surpass DT models in training accuracy because in a GB

model, new trees learn from errors of previous trees with the training data. In addition, the NN model is well-known for its strong learning and predicting capability. With the proposed four prediction accuracy measurements, all three models' prediction accuracy will be evaluated comprehensively.

Table 9. DT, GB, NN Model Prediction Classification Table

			Condition	
			Present	Absent
DT	Predict (Training)	Positive	243	975
		Negative	46	3306
	Predict (Testing)	Positive	32	127
		Negative	4	408
NN	Predict (Training)	Positive	245	749
		Negative	44	3532
	Predict (Testing)	Positive	30	111
		Negative	6	424
GB	Predict (Training)	Positive	256	681
		Negative	33	3600
	Predict (Testing)	Positive	31	99
		Negative	5	440

Table 10. Model Prediction Accuracy Summary

			Traditional accuracy	Proposed accuracy	Projected target level percentage	True target level percentage
Training Data	Event Prediction	DT	84.1%	20.0%	26.7%	6.0%
		NN	84.8%	24.6%	21.8%	6.0%
		GB	88.6%	27.3%	20.5%	6.3%
	Nonevent Prediction	DT	77.2%	98.6%	73.3%	94.0%
		NN	82.5%	98.8%	78.2%	94.0%
		GB	84.1%	99.1%	79.5%	93.7%
Training Data	Event Prediction	DT	88.9%	20.1%	27.8%	6.0%
		NN	83.3%	21.3%	24.7%	6.0%
		GB	86.1%	23.8%	22.6%	6.3%
	Nonevent Prediction	DT	76.3%	99.0%	72.2%	94.0%
		NN	79.2%	98.6%	80.2%	94.0%
		GB	81.6%	98.9%	77.4%	93.7%

As shown in Table 10, traditional prediction accuracy indicates that all three models’ prediction accuracy is higher than 77.2% for both of crash and non-crash prediction. Compared with previous study results ((Chang & Chen, 2005), (Chang & Wang, 2006), (Chang & Chien, 2013)) which often failed to forecast rare events despite the model providing relatively high overall forecasting accuracy rates as a result of the extremely imbalance dataset, all three data mining models not only provide accurate predictions for rare events (crashes) but also accurately estimate zero-crashes equally well.

As discussed in section 2.4.7, the traditional prediction accuracy and the proposed prediction accuracy reflect the model performance from two aspects. In the aspect of true alarm rate, which is indicated by traditional prediction accuracy, all three models predict crash-class better than non-crash-class in both of training and testing data. In addition, all three models

perform well in training data. For crash-class prediction, traditional prediction accuracy of the GB model is 88.6% and 86.1% in training and testing data, respectively, which is the highest among all three models. Taking 88.6% as an example, it represents that 88.6% of true event class observations are accurately predicted by the GB model. For testing data event class prediction, the DT model performs the best in terms of traditional prediction accuracy (88.9% comparing with 83.3% and 86.1%). As to non-crash-class prediction, the GB model provides a better true alarm rate than the other two models in both of training and testing data. Thus, in the aspect of traditional prediction accuracy, the GB model is superior to the DT and NN model unless of crash predictions in testing data. It is worth to mention that the true alarm rate for crash class is greater than 84% in both of training and testing data set, which is an outstanding improvement, compared with previous studies (Chang & Wang, 2006; Chang & Chen, 2005; Chang & Chien, 2015).

Proposed prediction accuracy, on the other hand, focuses on how accurate the model's prediction capabilities. As it shown in Table 10, proposed prediction accuracy shows great difference between crash-class prediction and non-crash-class prediction in both of the models. In training data, the DT model predicts crash and non-crash class with accuracy of 20.0% and 98.6%, respectively. The GB model predicts 937 observations to be crash level. However, only 27.3% of them are correct even though the correct 27.3% of events account for 88.6% of actual observed events. Meanwhile, 99.1% of model non-event forecasts are correct. Same pattern can be found in the DT and NN model prediction results: 20% and 98.6% vs 24.6% and 98.8% for crash and non-crash class prediction. It is clearly to see that even though GB model is better than the other two models in terms of forecasting accuracy for both training and testing data and event/non-event forecasts, all three models over-estimate event rates and underestimate non-

event rates by comparing column 6 and 7, and crash level forecasting accuracy is still relatively low, 20%, 24.6%, and 27.3% for and DT, NN, and GB model respectively. Because of the overestimation issue (26.7%, 21.8 and 20.5% event rate verses 6.3% actual rate), even though the model forecasting rate is low (27.3%, 24.6% and 20%) the model can correctly estimate 84.1%, 84.8% and 88.6% of the actual crashes respectively.

It is critical to understand true model forecasting accuracy by evaluate the model performance with all four proposed indicators. As found in this case study, the traditional approach over estimates model prediction accuracy of event level, and underestimates model prediction accuracy of non-event level. It may result in an inappropriate resource allocation. When evaluating a model's performance, if the true alarm rate is more important to decision makers, the traditional prediction accuracy measurement should be valued more than other prediction measurements. If a more accurate model prediction is preferred, the proposed prediction accuracy measurement should be focused on during model selection.

2.5.5. Section Summary

Three non-parametric data mining models, DT, GB, and NN model are tested in HRGC crash prediction, and are demonstrated to be solid predicting models in HRGC crash study. By conducting marginal effect analysis, the outputs of the GB and NN model are more interpretable. In addition, non-linear relationship between crash likelihood and contributor variables is successfully identified by both of GB and NN model. Due to the unstable tree structure, a marginal effect analysis cannot be conducted in a DT model. However, the non-linear relationship has no impact on DT model performance. The DT model, on the other hand, generates a tree structure decision flow that can be easily followed by users. By splitting the raw

data into training and testing dataset, and setting up regularization parameters and stopping criteria, over-fitting problem is successfully prevented in the three models.

All three proposed data mining models identify variable importance based on different criteria. Variable importance identified by the GB model agrees most of with that identified by the DT model. For example, both of the models identify the same top four contributory factors with different rankings: AADT, DAYTHRU, NGHTTHRU, and AVERAGE_TRAIN_SPEED. The GB model identified AADT and daytime train traffic as the top two most influential crash factors. Note that both models found that all four top contributors have positive impacts on crash predictions. The GB model demonstrates that 28 variables are related with HRGC accident prediction, compared with 23 variables proved by the DT model and 17 variables identified by the NN model. The NN model analyzes contributor variables' importance based on two criteria, importance to crashes and importance to model predictive accuracy. When choosing importance to crashes as criteria, warning devices type and presence are demonstrated to be more important than others. When pursuing better predictive model, traffic exposure variables are proved to be the most important. In addition, highway traffic related variables are found to be the most important contributor variables in the NN model, compared with that the DT and GB model roughly equivalent highway traffic and railway traffic factors' contribution.

It is proved that the three data mining models' performance is not affected by the non-linear relationship between the target variable and associated factors. In fact, all three models perform well with improved prediction accuracy, especially on crash prediction. The GB model performs the best in terms of prediction accuracy based on proposed evaluation method with four prediction accuracy measurements considered. Thus, the GB model is selected in the following commercial truck involved crash injury severity level analysis.

CHAPTER 3. APPLYING GRADIENT BOOSTING IN COMMERCIAL TRUCK CRASH SEVERITY LEVEL ANALYSIS

3.1. Introduction

Trucking is a well-known important element for freight movement and economic development. According to a 2012 commodity flow survey, trucks move 73.1% of commodities by value, 71.3% by tons, and 42.0% by ton-miles (*USDOT/BTS, 2012*). Truck crashes not only interrupt traffic flow, but also cause economic loss. Moreover, truck crashes contribute to a large number of injuries and fatalities due to additional risks, such as a larger vehicle size, heavier weight, and possible hazardous material release. In 2014, the total number of fatalities in truck crashes was 3,903 (*Federal Motor Carrier Safety Administration, 2014*). Compared with the total number of fatalities in strictly passenger car crashes, 28,559, truck crashes do not seem as alarming. However, truck crashes are overall more likely to result in more severe outcomes such as a fatality. In 2014, there were 14 fatalities in large truck crashes per 100 million vehicle miles traveled by large trucks, while only 10.5 fatalities in passenger vehicle crashes per 100 million vehicle miles traveled by passenger vehicles. Additionally, there were 29.4 injury crashes involving large trucks per 100 million vehicle miles traveled by large trucks, compared with 58.5 for passenger vehicles (*Federal Motor Carrier Safety Administration, 2014*).

The need to improve commercial trucking company safety performance has been a major social concern in the United States for decades. Transportation agencies and other stakeholders must identify the complete picture of factors that contribute to the severity levels of commercial truck collision and provide directions for commercial truck operation policies that will reduce the severe crash rates of commercial trucks.

Previous studies on modeling truck crash severities provide great insights and findings (Lemp, J. et al., 2011; Zhu and Srinivasan, 2011). However, some factors are overlooked and not considered in those studies. Intuitively thinking, characteristics of management, organization, culture, strategies, and financial situations in a trucking company should be closely associated with the company's safety performance. For example, safety culture shapes the attitude and behavior of their employees. Building a strong safety culture has a great effect on incident reduction (U.S. Department of Labor). Furthermore, a strong safety culture will result in better trained employees who will react better when they encounter a potential crash situation, and thus may result in a less severe crash outcome. Moreover, sufficient capital and profit promote truck maintenance and technology development, so that equipment is well-performing, which will minimize risk of equipment failure. In return, incident likelihood and crash severity level would be reduced. Although several studies have been carried out to investigate contributing variables to truck crash severity outcomes, the literature review revealed that it is still not clear how some commercial trucking company and driver characteristics impact crash severity levels.

This paper seeks to investigate commercial truck crash severity and contributing factors, especially trucking company characteristics, through the application of a data mining model to commercial trucking crash data. The paper is organized with a literature review, data description, methodology, results analysis, and conclusions of the research.

3.2. Literature Review

Vehicle crash studies have been completed by a substantial number of researchers focusing on crash frequencies and injury severity (Dong et al., 2014; Gabauer and Li, 2015; Wu et al., 2016; Chen et al., 2016; Wood, Donnel, Fariss, 2016; Lu and Tolliver, 2016). The majority of them are focused on vehicle crashes in urban road tunnels. Meng and Qu (2012) examined rear-

end vehicle crash frequency in urban road tunnels. Wu et al (2016) conducted a crash severity study examining the factors of weather condition, class of highway and drug use and their impact on single-vehicle crashes. An integrated study of crash frequency and severity was conducted by Chiou and Fu (2013). Freeway geometrics, traffic characteristics, neighborhood, and freeway facilities were found to significantly contribute to vehicle crash frequency and severity.

As a common understanding, vehicle types such as passenger cars or commercial trucks should have a different impact on crash severity outcomes. There are numerous studies focusing on truck crashes only. Most of them examined one specific influential factor of truck crashes, such as wind speed, driver turnover rate, presence of portable message sign, the time of day, truck configurations, and driver body mass (Young and Liesman, 2007; Staplin and Gish, 2005; Bai, Yang, Li, 2015; Curnow, 2002; Braver, et al., 1997; Anderson, et al., 2012). Braver, et al. (1997) applied Logistic regression model to explain multi-trailer tractor involved crashes. They found out that day of week, time of day, urban/rural area, and specific highway as significant explanatory variables among different types of trucks' crash likelihood. Their results indicated that under unfavorable operating conditions, the crash likelihood of double-trailer trucks would increase. Curnow (2002) studied Australia truck crash database, and concluded that 10 percent of articulated truck drivers involved in serious injury crashes were fatigued at the time of crash. Staplin and Gish (2005) analyzed Motor Carrier Management Information System (MCMIS), and developed quantitative relationship between the risk of truck drivers involved in crashes and the drivers' job change frequency. Their findings revealed that risk of crash risk started to rise for the drivers who changed their jobs more than twice with different employers each year for two years or longer. For drivers who changed jobs three or more times per year, the risk of multiple crashes was more than doubled than those with lower job change rate. In the study by Young and Liesman (2007),

Wyoming freight vehicle crash data were analyzed to establish quantitative relationship between freight vehicles overturning crash likelihood and wind speed at weight station. They tested six variables, including road surface condition, crash location curvature, wind direction, wind gust, crash location distance to weather station, and wind speed, and four of them were significant for overturning crash risk estimation based on statewide data except wind direction and distance to weather station. Their results indicated that wind speed, wind gust, and road curvature made positive contribution to likelihood of being overturning crashes, while slippery road surface made negative contribution. Anderson et, al., (2012) proposed that obesity was associated with heavy truck crash risk among newly recruited commercial drivers. Multivariate Poisson regression and Cox proportional hazards models were used to estimate the relationship between crash risk and Body Mass Index (BMI). Their research demonstrated that drivers with obesity problems were of significantly higher risk of crashes than the drivers whose BMI were normal. Bai, Yang, and Li (2015) focused on researching the effective location of a portable changeable message to reduce the risk of truck-related crashes in work zones. They proposed a hypothesis that the difference of speed changes between trucks and passenger cars was considered as one of the major reasons which caused truck-related crashes in work zones. They collected passenger car and truck speed data with portable changeable message sign placed at 750, 550, and 400 ft away from W20-1 sign. By comparing the speed variance of passenger car speed and truck speed, it was demonstrated that the message sign located at 550 ft away from W20-1 sign was the most effective location to significantly reduce the speed variability between trucks and passenger cars. Most of those studies focus on the effects on truck crash frequency but only a limited number of studies contribute to understanding truck crash severities (*Khattak, Schneider, Targa, 2003; Naik et al., 2016; Campbell, 1991; Uddin and Huynh, 2017; Zou, Wang, Zhang, 2017*). Uddin and Huynh

(2017) studied influential factors of crash severity involving hazardous materials trucks. Explanatory variables were defined, including occupant, crash, vehicle, roadway, environmental, and temporal characteristics. The model results evidenced that the occupants being male, truck drivers, crashes occurring in rural locations, under dark-unlighted, under dark-lighted conditions, and on weekdays were contributed to increased probability of major injuries. Conversely, the older occupants (age 60 and over), trucks making a turn, rear-end collision, collision with an object, crashes occurring on non-interstate highway, higher speed limit highway (65 mph), and flat terrain were associated with decreased probability of major injuries. Pahukula, Hernandez, and Unnikrishnan (2015) improved truck safety by studying effect of contributor variables on truck crash injury severity in large populated urban area in five different time periods in a day: early morning (0:00-4:00), morning (5:00-9:00), mid-day (10:00-15:00), afternoon (16:00-20:00), evening (21:00-23:00). The results of the individual models demonstrated considerable differences among the five time periods. Key contributors were identified as traffic flow, light conditions, surface conditions, time of year, and percentage of trucks on the road. Ever-changing traffic flow patterns throughout the day were evident in the mid-day model. Speeding and changing lanes contributed to large truck-involved crashes between 10:00 a.m. and 3:00 p.m. (i.e., typically uncongested time period). The summer indicator variable in the afternoon model also suggested that traffic volume impacted injury severity. Crashes between June and August were found to decrease the likelihood of a severe injury. Naik et al. (2016) investigated the impact of weather conditions on single-vehicle truck crash injury severity based on Nebraska DOT crash data and detailed 15-min weather data from National Climate Data Center from 2009 to 2011. Their results indicated that greater wind speed increased the likelihood of greater severity level. Rain and warm air temperature added to the severity of injuries, while higher level

of humidity was associated with less severe severity. Icy or snowy road surface was found out related with less severe crashes. Campbell (1991) collected survey data about trucks involved in fatal crashes, and analyzed the effect of drivers' age on likelihood fatal crashes. It was demonstrated that young drivers under age 27 had higher likelihood of fatal crashes than elder drivers. Especially, drivers under age 21 were the most vulnerable group of involved in fatal crashes compared with all drivers. Khattak, Schneider, Targa (2003) investigated effect of associated factors with truck involved single-vehicle crash severity levels. A total of 44 variables were defined as contributor variables, describing vehicle, crash, roadway, driver, and environmental features. The major contribution of their study is that they quantified the direct and indirect effects of key factors on injury severity levels sustained by truck occupants. It was found out that crashes with greater severity levels were associated with curves and especially dangerous driving behaviors, including reckless driving, speeding, passing violation and alcohol/drug use. Zou, Wang, and Zhang (2017) link truck crash severity with spatial location and time of day. Their results reveal that individual truck crashes are spatially dependent events for single and multi-vehicle crashes. Single-vehicle crashes in the afternoon and at night tend to be less severe, while multi-vehicle crashes at the same time are more severe.

Among all previous researchers, an understanding of the influence of attributes of the trucking company and driver's license on crash injury severity is still unclear. Several studies discussed that the little research on trucking company characteristics' impact on crash severity is due to the lack of available company data (*Chen, 2008*). This research focuses on risk factors for commercial truck crash severity, in particular how company related characteristics affect crash severity, with a more comprehensive truck crash dataset available through the Federal Motor Carrier

Safety Administration (FMCSA). The detailed information regarding this database is described later in the data description section.

The literature search also reveals that most prior studies are based on logit, probit, and their extension statistical models (*Lemp, J. et al., 2011; Zhu and Srinivasan, 2011; Wu et al., 2016; Charbotel et al., 2003*). However, as it mentioned in previous chapters, these statistical models are all based on certain assumptions. One of the common assumptions is that the effects of contributing factors are assumed identical across different severity levels. These assumptions are inappropriate and do not hold true in most circumstances. Once violated, numerous errors will be generated. In addition, truck crashes are affected by a set of heterogeneous variables (*Kumar and Toshniwai, 2015*). A good crash injury severity model is expected to be able to extract hidden, valuable information from large, complex datasets. Thus, instead of applying statistical models, the non-parametric gradient boosting (GB) model, a data mining technique, is selected in this study to overcome the shortcomings and achieve more convincing conclusions. The GB model does not have any pre-defined data assumptions like other statistical models do. Moreover, the GB model inherits most of the tree-based data mining models' advantages. It is also superior than most of the tree-based data mining models with its missing data handling techniques, robustness with data noise and resistance to over-fitting (*Friedman and Meulman, 2003; Salford Systems*). The GB model proves its success in crash prediction analysis (*Chung, 2013; Saha, Alluri, Gan, 2015*), however, it has never been used in a truck crash injury severity explanatory study. Therefore, the authors decided to adopt a GB model to comprehensively analyze influential factors on truck crash injury severity.

3.3. Data Description

In this study, truck crash data was obtained from the Federal Motor Carrier Safety Administration (FMCSA). Crash data file, census file, and inspection files from the Motor Carrier Management Information System (MCMIS) are selected for the research. The MCMIS datasets contain 1) records from state police crash reports including information on drivers, crash conditions, environment factors when the crash happened, and crash involved truck conditions; 2) motor carrier corporation variables and operational factors; and 3) motor carrier safety inspection records. This study examines truck crash related data for crashes that occurred in the states of North Dakota and Colorado in the past six years (from 2010 to 2016). The selection of the two states is due to the availability of data, research interest, and data size limitation, however, the research can be extended to national level or include additional states if it is of interest.

The authors exclude irrelevant, privacy variables and four redundancy variables from the raw data before performing mathematical analysis. Summarized in Table 11, 38 variables are removed from analysis.

Table 11. Summary of Unanalyzed Variables

Variable	Rationale for removal
Carrier related variables	
Address	Irrelevant variable
Zip code	Irrelevant variable
Country	Irrelevant variable
Phone	Irrelevant variable
Identification number	Irrelevant variable
Last updated date	Irrelevant variable
May have undeliverable physical address	Irrelevant variable
May have undeliverable mailing address.	Irrelevant variable
Carrier name	Irrelevant variable
USDOT number	Irrelevant variable
City	Irrelevant variable
Crash related variables	
Crash ID	Irrelevant variable
Crash year	Redundant variable with variable "Year"
Crash quarter	Irrelevant variable
Federal recordable	Irrelevant variable
Officer badge	Irrelevant variable
Crash Number	Irrelevant variable
Crash Date	Irrelevant variable
Crash Time	Redundant variable with variable "Time of Day"
Officer Badge	Irrelevant variable
Record Status	Irrelevant variable
Matched Status	Irrelevant variable
SAFETYNET Input Date	Irrelevant variable
MCMIS Upload Date	Irrelevant variable
Number Days to SAFETYNET	Irrelevant variable
Number Days to MCMIS	Irrelevant variable
Counter	Irrelevant variable
Vehicle Configuration Desc	Redundant variable with variable "Vehicle configuration"
GVW Rating Desc	Redundant variable with variable "GVW"
City	Irrelevant variable
City code	Irrelevant variable
County	Irrelevant variable
County code	Irrelevant variable

Table 11. Summary of Unanalyzed Variables (continued)

Variable	Rationale for removal
Number assigned to motor carriers engaging in interstate or foreign operations	Irrelevant variable
Registered as a common carrier: A- Active registration, I- Inactive registration, N- no registration	Irrelevant variable
Driver related variables	
First name	Irrelevant variable
Last name	Irrelevant variable
Mid name	Irrelevant variable

The detailed information of the data analyzed in this research is shown in Table 12. In general the data variables can be grouped into the following five categories:

- 1) Trucking company characteristics (e.g., total number of trucks, inspection value, registered date, and location);
- 2) Crash characteristics (e.g., first injury or damaging-producing event, day of week, time of day, and number of injuries);
- 3) Environment characteristics (e.g., road type, light condition, road surface condition, and weather condition);
- 4) Driver characteristics (e.g., age, driver license class, and driver license state); and
- 5) Truck characteristics (e.g., cargo type, configuration, and gross vehicle weight).

There are 24 variables selected to be investigated and tested. Twenty one of them are categorical variables (labeled with ‘\$’ in Table 2) and two of them are numeric variables. In this study, the target variable (injury severity) is classified as: 0=property damage only; 1=injury only (no fatalities); 2=only one fatality; 3=two or more fatalities.

The total number of recorded truck-involved crashes that happened in ND and CO from 2010 to 2016 is 16,389. Of all the crashes, 72.13% (11,822) resulted in property damage only

(severity=0); 24.22% (3,969) were injury only (severity=1); 1.97% (323) caused one fatality (severity=2); and 1.68% (275) caused two or more fatalities (severity=3).

Company size is divided into five categories: 1=single truck companies; 2=small truck companies; 3=medium size truck companies; 4=large truck companies; and 5=very large truck companies.

Table 12. Variable Description

Variable	Total	Number of missing	Missing percentage	Description
Trucking Company Characteristics				
Carrier State\$	16,389	0	0	State/District/Province of the principal place of business of the carrier registered
Company Size\$	13,221	3,168	19.33	1, 2, 3, 4, 5
Indicator\$	13,269	3,120	19.04	'S' = Safety; 'I' = Insufficient Data; 'N' = Intrastate Safety; 'R' = Random
Inspection Value	13,269	3,120	19.04	Inspection value. Ranging from 0 to 100 with 100 indicate the worst performance
Interstate Carrier\$	15,376	1,013	6.18	Is carrier an interstate carrier? Yes/No
New Entrant\$	16,389	0	0	Is carrier a new registered carrier? Yes/No
Crash Characteristics				
Day of Week\$	16,389	0	0	Sun.; Mon.; Tue.; Wed.; Thu.; Fri.; Sat.
First Harmful Event\$	16,114	275	1.68	The first injury or damage-producing event, including: Involving Animal; Involving Fixed Object; Involving Motor Vehicle In Transport; Involving Other Movable Object; Involving Parked Motor Vehicle; Involving Pedalcycle; Involving Pedestrian; Involving Train; Involving Unknown Movable Object; Work Zone Maintenance Equipment; Eqp Failure; Cargo Loss Or Shift; Cross Median/Centerline; Downhill Runaway; Explosion Or Fire; Jackknife; Other; Overturn (Rollover); Ran Off Road; Separation Of Unit; Unknown
Time of Day\$	16,250	139	0.85	12:00 AM - 2:59 AM; 3:00 AM - 5:59 AM; 6:00 AM - 8:59 AM; 9:00 AM - 11:59 AM; 12:00 PM - 2:59 PM; 3:00 PM - 5:59 PM; 6:00 PM - 8:59 PM; 9:00 PM - 11:59 PM
Tow Away\$	16,389	0	0	Is accident vehicle towed away? Yes/No
Number of Vehicles	16,388	1	0.01	The total number of vehicles or vehicle combinations involved in the crash. Numeric variable.
Environment Characteristics				
Light Condition\$	16,371	18	0.11	Dark – Lighted; Dark - Not Lighted; Dark - Unknown Roadway Lighting, Dawn; Daylight; Dusk; Other; Unknown
Road Surface Condition\$	16,382	7	0.04	Dry; Ice; Other; Sand, Mud, Dirt, Oil Or Gravel; Slush; Snow; Unknown; Water(Standing, Moving); Wet
Traffic Way Type\$	16,388	1	0.01	Not Reported; One-Way Trafficway, Not Divided; Two-Way Trafficway, Divided, Positive Barrier; Two-Way Trafficway, Divided, Unprotected Median; Two-Way Trafficway, Not Divided
Weather Condition\$	16,378	11	0.07	Blowing Sand, Soil, Dirt, Or Snow; Fog; No Adverse Conditions; Other; Rain; Severe Crosswinds; Sleet, Hail; Snow; Unknown
Driver Characteristics				
Driver's Age\$	16,389	0	0	<26; 26 – 35; 36 – 45; 46 – 55; 56 – 65; 66 – 75; >75

Table 12. Variable Description (continued)

Variable	Total	Number of missing	Missing percentage	Description
Driver's License Class\$	15,816	573	3.5	A, B, C, D
Driver's License State\$	16,020	369	2.25	The license state/district/province of the driver.
Valid Driver's License\$	16,148	241	1.47	If driver's license is valid or not. Yes/No
Truck Characteristics				
GVWR\$	16,382	7	0.04	Gross Vehicle Weight Rating in pounds: < 10,000; 10,001-26,000;>26,000
Cargo Body Type\$	16,333	56	0.34	Auto Transporter; Bus Seats For 9-15 People, Including Driver; Bus Seats For > 15 People, Including Driver; Cargo Tank; Concrete Mixer; Dump; Flatbed; Garbage/Refuse; Grain, Chips, Gravel; Intermodal; Logging; Not Applicable/No Cargo Body; Other; Pole; Van/Enclosed Box; Vehicle Towing Another Vehicle
Vehicle Configuration\$	16,370	19	0.12	Light Truck(Only If Vehicle Displays Hm Placa; Single-Unit Truck (2-Axle, 6 Tire); Single-Unit Truck (3 Or More Axles); Tractor/Double; Tractor/Semi-Trailer; Tractor/Triples; Truck Tractor (Bobtail); Truck/Trailer; Unknown
Vehicle License State\$	16,356	33	0.2	The license state/district/province of the truck.
Target Variable				
Severity	16,389	0	0	0=no injuries and no fatalities; 1=injuries and no fatalities; 2=one fatality and no injuries; 3=one fatality and injuries, or two or more fatalities

3.4. Result Analysis

The raw crash data from the states of North Dakota and Colorado are fit into the GB model and twenty-five contributor variables are tested as predictors of injury severity. Out of these variables, twenty-one are found to be associated with injury severities. However, it can be troublesome to state that 21 variables contribute to injury severity, thus the relative variable importance analysis for causal importance of inputs is also conducted. Table 13 presents variable importance under various injury severities. 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 improvement on mean squared error attributed to the tree due to the splits by the variable. After summing the importance score computed in a simple single tree over the ensemble of trees, the average value

of the summation is scaled, so that the most important variable scores 100. Then, the scaled average value is regarded as the variable's importance in the model. A high value of variable importance indicates a high contribution that a variable makes to the prediction (*Friedman and Meulman, 2003*). As noted in Table 13, top 11 variables account for more than 80% of injury forecasting. For property damage only, the most important variable is 'Carrier State', which indicates that the variable of 'Carrier State' makes the most contributions as compared to the other variables in explaining property damage only crashes. And "Tow Away" is the second important contributor, and it accounts 72% of the importance that "Carrier State" contributes to. The column "cum %" in Table 13 indicates the absolute cumulative contribution of the variables.

Table 13. Variable Importance under Each Level of Severity

Damage only (Severity=0)			Injury (Severity=1)		
Variable	Score	Cum %	Variable	Score	Cum %
Carrier State\$	100	16%	Carrier State\$	100	16%
Tow away\$	71	28%	First Harmful Event\$	85	30%
First Harmful Event\$	53	36%	Tow away\$	50	39%
Cargo Body Type\$	50	44%	Vehicle Configuration\$	44	46%
Time of Day\$	46	52%	Traffic Way Type\$	44	53%
Day of Week\$	40	58%	Cargo Body Type\$	43	60%
Driver Age\$	33	64%	Road Surface Condition\$	40	67%
Weather Condition\$	31	69%	Light Condition\$	32	72%
Number of Vehicles	26	73%	Number of Vehicles	30	77%
Vehicle Configuration\$	26	77%	Weather Condition\$	25	81%
Company Size\$	22	81%	Driver Age\$	22	85%
Light Condition\$	19	84%	Time of Day\$	19	88%
Road Surface Condition\$	17	87%	Day of Week\$	16	91%
GVWR\$	15	89%	Company Size\$	15	93%
Indicator\$	14	92%	GVWR\$	9	95%
Driver's License Class\$	13	94%	Indicator\$	8	96%
Inspection Value	13	96%	Interstate Carrier\$	7	98%
Traffic Way Type\$	10	98%	Inspection Value	6	99%
Interstate Carrier\$	9	99%	Driver's License Class\$	6	100%
New Entrant\$	2	100%	Valid Driver's License\$	1	100%
Valid Driver's License\$	1	100%	New Entrant\$	0	100%
One Fatality (Severity=2)			Two or More Fatalities (Severity=3)		
Variable	Score	Cum %	Variable	Score	Cum %
Carrier State\$	100	17%	Carrier State\$	100	16%
First Harmful Event\$	69	28%	Number of Vehicles	66	27%
Cargo Body Type\$	55	38%	First Harmful Event\$	50	35%
Time of Day\$	45	45%	Cargo Body Type\$	45	43%
Day of Week\$	39	52%	Time of Day\$	43	50%
Vehicle Configuration\$	37	58%	Weather Condition\$	41	57%
Driver Age\$	35	64%	Day of Week\$	33	62%
Road Surface Condition\$	33	69%	Light Condition\$	30	67%
Weather Condition\$	24	73%	Tow away\$	25	72%
Company Size\$	22	77%	Vehicle Configuration\$	25	76%
Driver's License Class\$	21	81%	Inspection Value	25	80%

Table 13. Variable Importance under Each Level of Severity (continued)

Variable	Score	Cum %	Variable	Score	Cum %
Inspection Value	17	84%	Driver Age\$	24	84%
Number of Vehicles	16	86%	Traffic Way Type\$	22	88%
Indicator\$	15	89%	Company Size\$	20	91%
Light Condition\$	15	92%	Road Surface Condition\$	17	94%
Tow away\$	14	94%	Indicator\$	10	96%
Traffic Way Type\$	14	96%	New Entrant\$	8	97%
Interstate Carrier\$	7	98%	GVWR\$	5	98%
GVWR\$	5	99%	Interstate Carrier\$	5	99%
Valid Driver’s License\$	5	100%	Driver’s License Class\$	4	100%
New Entrant\$	2	100%	Valid Driver’s License\$	1	100%

As illustrated in Table 13, variables contribute differently when explaining different crash severities. A variable showing significant importance for a certain severity level may be less crucial for another. For instance, ‘Cargo Body Type’ is the second most important factor for predicting fatality crashes, but is much less important for predicting property damage only crashes (severity=0). However, it is clear that ‘Carrier State’ is the most influential factor for all severity levels. ‘First Harmful Event’ also plays an important role in predicting all severity levels. Some other interesting findings are observed in the analysis: 1) time of day and day of week play more important roles in explaining damage only crash and fatality crash but less important in explaining injury; 2) Driver age plays a more important role for damage only and one fatality but less important for injury and more fatalities; 3) Vehicle configuration plays a more important role in injury and one fatality than in damage only and more fatalities; and 4) trucking company size, road surface condition, safety inspection value, valid driver’s license, and driver’s license class all significantly impact crash severities at different levels. GB successfully identified the contribution variables to crash severities and prioritized their importance roles. Marginal effect

analysis of each influential variable is also analyzed to provide a further detailed understanding of how they contribute to various crash severities.

Marginal effects of practically important variables are summarized in Table 14. For categorical variables with various levels, due to the space limitation, Table 14 only shows selected levels for a significant contributor categorical variable with the most significant impacts. Moreover, levels with more outstandingly significant impact are bolded. For example, from Table 13, one can tell that weather condition is the 8th significant contributor variable for damage only crash. In Table 14 only the positive effect of the weather condition of snow is listed, which has much more significant impact on damage only crashes than any other weather conditions. And the level of snow is bolded. Conversely, all levels of the 'First Harmful Event' which have a negative effect on severities are un-bolded. This is because none of them make more outstanding contributions than others. The first column 'variable' lists influential variables whose impact on severity prediction is valuable. Positive effect (P) means that the corresponding categories for the influential variable will increase the probability of certain severity level (column severity=0, 1, 2, 3), while negative effect (N) means it will decrease that likelihood.

Examining 'Carrier State' as one example, if a carrier is registered in the state of MA, MS, OH, or WI this has a significantly positive effect on damage only and if a carrier is registered in the ND or TX, this has a significantly negative effect on damage. All other unlisted carrier states have no significant different contribution for damage only crashes. And there is no bolded state among all listed positive or negative impact carrier states indicating their positive/negative impact effects are not significantly different within their corresponding category.

Table 14. Marginal Effect of Influential Variables

Variable	Effect	Damage only (Severity=0)	Injury (Severity=1)	One Fatality (Severity=2)	Two or More Fatality (Severity=3)
Trucking Company Characteristics					
Carrier State\$	P	MA, MS, OH, WI	AL, OR, WI, MS	MO, KS	MI, MB, NC, ND, PA
	N	ND, TX	KS, MO	GA, NY, PA	AL, MA, MS, FL, OH, OR
Inspection Value	P	<30	<25	<45	30-70, >90
	N	>30	>25	>50	80-90
Company Size\$	P	1, 5 (very small or very large)	1, 4 (very small or large size)	1, 3 (very small, medium)	4, 5 (large, very large)
	N	2, 4 (small or large)	2, 3, 5 (small, medium, or very large)	2, 4, 5 (small, large, very large)	1, 2, 3 (very small, small, medium)
Interstate Carrier\$	P	N	N	Y	Y
	N	Y	Y	N	N
New Entrant\$	P	N/A	N	Y	Y
	N	N/A	Y	N	N
Indicator\$	P	N, R	I, N, R	S	N, S
	N	I, S	S	I, N, R	I, R

Table 14. Marginal Effect of Influential Variables (continued)

Crash Characteristics					
First Harmful Event\$	P	Involving Animal; Involving Fixed Object; Involving Other Movable Object; Involving Train; Involving Unknown Movable Object; Work Zone Maintenance Equipment; Eqp Failure; Cargo Loss Or Shift; Cross Median/Centerline; Downhill Runaway; Explosion Or Fire; Jackknife; Other; Overturn (Rollover); Separation Of Unit; Unknown	Involving Fixed Object; Involving Pedalcycle; Involving Unknown Movable Object; Work Zone Maintenance Equipment; Eqp Failure; Cargo Loss Or Shift; Downhill Runaway; Explosion Or Fire; Jackknife; Other; Overturn (Rollover); Ran Off Road; Separation Of Unit; Unknown	Involving Animal; Involving Fixed Object; Involving Other Movable Object; Involving Pedalcycle; Involving Pedestrian; Involving Train; Involving Unknown Movable Object; Work Zone Maintenance Equipment; Downhill Runaway; Ran Off Road; Other	Involving Motor Vehicle In Transport; Involving Parked Motor Vehicle; Involving Pedestrian; Involving Train; Work Zone Maintenance Equipment; Cross Median/Centerline;
	N	Involving Motor Vehicle In Transport; Involving Parked Motor Vehicle; Involving Pedalcycle; Involving Pedestrian; Involving Train; Ran Off Road;	Involving Animal; Involving Motor Vehicle In Transport; Involving Other Movable Object; Involving Parked Motor Vehicle; Involving Pedestrian; Involving Train; Cross Median/Centerline;	Involving Motor Vehicle In Transport; Involving Parked Motor Vehicle; Eqp Failure; Cargo Loss Or Shift; Cross Median/Centerline; Explosion Or Fire; Jackknife; Overturn (Rollover); Separation Of Unit;	Involving Animal; Involving Fixed Object; Involving Other Movable Object; Involving Pedalcycle; Involving Unknown Movable Object; Eqp Failure; Cargo Loss Or Shift; Downhill Runaway; Explosion Or Fire; Jackknife; Other; Overturn (Rollover); Ran Off Road; Separation Of Unit;

Table 14. Marginal Effect of Influential Variables (continued)

Number of Vehicle in Crash	P	<2	<2	>2	>4
	N	>2	>2	<2	<4
Time of Day\$	P	9-12AM; 12PM-3PM; 3-6 PM	12 AM - 3 AM; 6AM - 9 AM; 12 PM - 3 PM; 3PM - 6 PM;	3AM - 6 AM; 9:00 PM - 12PM	0 AM - 3 AM; 3 AM - 6 AM; 12PM - 3 PM; 6 PM - 9 PM;
	N	The rest	3 AM - 6 AM; 9 AM - 12 AM; 6 PM - 9 PM; 9 PM - 12 PM	0AM - 3 AM; 6 AM - 9 AM; 9 AM - 12 AM; 12 PM - 3 PM; 3:00 PM - 6 PM; 6 PM - 9 PM;	6 AM - 9 AM; 9AM - 12 AM; 3 PM - 6 PM; 9 PM - 12 PM
Day of Week\$	P	Mon. Wed. Thu. Fri.	Mon. Wed. Fri.	Tue. Sat. Sun.	Mon. Wed. Thu. Sat. Sun.
	N	Tue. Sat. Sun.	Tue. Thu. Sat. Sun.	Mon. Wed. Thu. Fri.	Tue. Fri.
Tow Away\$	P	Y	N	N	Y
	N	N	Y	Y	N
Environment Characteristics					
Weather Condition\$	P	Rain; Severe Crosswinds; Sleet, Hail; Snow; Unknown	Blowing Sand, Soil, Dirt, Or Snow; Fog; Rain; Severe Crosswinds; Sleet, Hail; Snow; Unknown	No Adverse Conditions;	Fog; Other; Severe Crosswinds; Sleet, Hail; Unknown
	N	Blowing Sand, Soil, Dirt, Or Snow; Fog; No Adverse Conditions; Other;	No Adverse Conditions; Other;	Blowing Sand, Soil, Dirt, Or Snow; Fog; Other; Rain; Severe Crosswinds; Sleet, Hail; Snow; Unknown	Blowing Sand, Soil, Dirt, Or Snow; No Adverse Conditions; Rain; Snow

Table 14. Marginal Effect of Influential Variables (continued)

Road Surface Condition\$	P	Ice; Slush	Ice; Sand, Mud, Dirt, Oil Or Gravel; Slush; Snow; Unknown; Water(Standing, Moving);	Dry; Ice; Other; Sand, Mud, Dirt, Oil Or Gravel; Unknown; Water(Standing, Moving); Wet	Dry; Wet
	N	Dry; Other; Sand, Mud, Dirt, Oil Or Gravel; Snow; Unknown; Water(Standing, Moving); Wet	Dry; Other; Wet	Slush; Snow;	Ice; Other; Sand, Mud, Dirt, Oil Or Gravel; Slush; Snow; Unknown; Water(Standing, Moving);
Light Condition\$	P	Dark – Lighted; Dark - Unknown Roadway Lighting, Dawn; Daylight; Other; Unknown	Dark – Lighted; Dark - Unknown Roadway Lighting, Dawn; Daylight; Other; Unknown	Dark – Lighted; Dark - Not Lighted; Dusk	Dark - Not Lighted;
	N	Dark - Not Lighted; Dusk	Dark - Not Lighted; Dusk;	Dark - Unknown Roadway Lighting, Dawn; Daylight; Other; Unknown	Dark – Lighted; Dark - Unknown Roadway Lighting, Dawn; Daylight; Dusk; Other; Unknown
Trafficway Type\$	P	Not Reported; One-Way Trafficway, Not Divided; Two-Way Trafficway, Divided, Positive Barrier;	One-Way Trafficway, Not Divided; Two-Way Trafficway, Divided, Unprotected Median;	Two-Way Trafficway, Divided, Unprotected Median;	One-Way Trafficway, Not Divided; Two-Way Trafficway, Divided, Unprotected Median; Two-Way Trafficway, Not Divided
	N	Two-Way Trafficway, Divided, Unprotected Median; Two-Way Trafficway, Not Divided	Not Reported; Two-Way Trafficway, Divided, Positive Barrier; Two-Way Trafficway, Not Divided	Not Reported; One-Way Trafficway, Not Divided; Two-Way Trafficway, Divided, Positive Barrier; Two-Way Trafficway, Not Divided	Not Reported; Two-Way Trafficway, Divided, Positive Barrier;

Table 14. Marginal Effect of Influential Variables (continued)

Driver Characteristics					
Driver Age\$	P	26-45	26-45, 66+	<25, 45-65, 75+	75+, 25-
	N	the rest	the rest	the rest	26-45
Driver's License Class\$	P	B, C	B, C	A, D	B, C, D
	N	A, D	A, D	B, C	A
Valid Driver's License\$	P	N	Y	N	N
	N	Y	N	Y	Y
Truck Characteristics					
Cargo Body Type\$	P	Auto Transporter; Bus Seats For 9-15 People, Including Driver; Bus Seats For > 15 People, Including Driver Dump; Intermodal; Logging; Pole; Van/Enclosed Box;	Auto Transporter; Bus Seats For 9-15 People, Including Driver; Bus Seats For > 15 People, Including Driver; Concrete Mixer;	Auto Transporter; Cargo Tank; Concrete Mixer; Dump; Flatbed; Garbage/Refuse; Grain, Chips, Gravel; Intermodal; Logging; Other; Pole; Vehicle Towing Another Vehicle	Cargo Tank; Dump; Flatbed; Garbage/Refuse; Grain, Chips, Gravel; Logging; Not Applicable/No Cargo Body; Other; Van/Enclosed Box; Vehicle Towing Another Vehicle

Table 14. Marginal Effect of Influential Variables (continued)

	N	; Cargo Tank; Concrete Mixer; Flatbed; Garbage/Refuse; Grain, Chips, Gravel; Not Applicable/No Cargo Body; Other; Vehicle Towing Another Vehicle	Cargo Tank; Dump; Flatbed; Garbage/Refuse; Grain, Chips, Gravel; Intermodal; Logging; Not Applicable/No Cargo Body; Other; Pole; Van/Enclosed Box; Vehicle Towing Another Vehicle	Bus Seats For 9-15 People, Including Driver; Bus Seats For > 15 People, Including Driver; Not Applicable/No Cargo Body; Van/Enclosed Box	Auto Transporter; Bus Seats For 9-15 People, Including Driver; Bus Seats For > 15 People, Including Driver; Concrete Mixer; Intermodal; Pole
Vehicle Configuration\$	P	Light Truck(Only If Vehicle Displays Hm Placa; Single-Unit Truck (2-Axle, 6 Tire); Tractor/Triples; Truck/Trailer; Unknown	Light Truck(Only If Vehicle Displays Hm Placa; Single-Unit Truck (3 Or More Axles); Tractor/Triples; Truck/Trailer; Unknown	Tractor/Double; Tractor/Semi-Trailer; Truck Tractor (Bobtail); Truck/Trailer; Unknown	Single-Unit Truck (2-Axle, 6 Tire); Tractor/Double; Tractor/Semi-Trailer
	N	Single-Unit Truck (3 Or More Axles); Tractor/Double; Tractor/Semi-Trailer; Truck Tractor (Bobtail);	Single-Unit Truck (2-Axle, 6 Tire); Tractor/Double; Tractor/Semi-Trailer; Truck Tractor (Bobtail);	Light Truck(Only If Vehicle Displays Hm Placa; Single-Unit Truck (2-Axle, 6 Tire); Single-Unit Truck (3 Or More Axles); Tractor/Triples;	Light Truck(Only If Vehicle Displays Hm Placa; Single-Unit Truck (3 Or More Axles); Tractor/Triples; Truck Tractor (Bobtail); Truck/Trailer; Unknown
GVWR\$	P	10,001-26,000;	10,001-26,000; >26,000	>26,000	>26,000
	N	>26,000	>26,000	10,001-26,000;	10,001-26,000;

With regards to trucking company characteristics, it is found that trucks registered in different states perform differently in terms of crash severity. For single fatality analysis, crashes with trucks for carriers registered in MO and KS are significantly more prone to have a single fatality crash, while trucks from carriers registered in GA, NY, and PA are prone to not have a single fatality crash. Other states not listed have no significant difference with the likelihood of a single fatality crash. Potential rationale to the observation that trucks registered in MO and KS are more prone to single fatality crashes could be that MO and KS truck companies have higher portion of larger and heavier trucks than other states. Larger and heavier trucks need longer braking time in an emergency. On the other hand, this observation could be also due to that drivers hired by MO and KS truck companies are very unfamiliar driving environment while driving in other states, or react less appropriate in other state driving situation. Further investigation and validation is recommended from two aspects: percentage of truck configurations and driver driving behaviors from those two states. If further studies validate the hypothesis, additional training to truck drivers from those state driving heavier and larger truck may be an effective tool to reduce crashes. The training can focus on driving environment and driving habits in the destination state or en-route states or specific driving training focus on extreme large and heavy trucks.

Referring to row 7 and 8 of Table 14, different trucking company sizes in terms of the number of trucks they own have different influence on different crash severity levels. Examining single-truck companies, one can see that they have a high risk to be involved in damage only, injury or single fatality crashes, but have a low risk to be involved in multiple fatality crashes. This observation could result from that single-truck company drivers may become more tired due to more frequent shifts, because of fewer drivers in single-truck companies, so that they are more prone to crashes even though they are not likely to have very severe crashes such as multiple

fatality crash. To fully understand this finding, further analysis is needed. It is notable that small truck companies, those that own two to five trucks, are found to be the best safety-performance companies (crash severity wise) given the fact that they are estimated to have a negative impact on all levels of crash severity. In other words, small sized companies are significantly different than any other size company in their involvement in any level of crash severity. The underlying reasons for this observation are unclear. A potential rationale could be due to the management culture in a small sized trucking company (2 to 5 trucks) and operation efficient and effectiveness of such sized companies. Trucks from medium sized companies are more likely in involvement in single fatality crashes and less likely involving in injury or more severe crashes. Trucks from large size companies are more likely involved in either injury only or multiple fatality crashes. Very large companies, those that own more than 100 trucks, tend to have a high risk to be involved in either damage only or multiple fatality crashes. It is inferred that in a fatal crash, trucks from large size companies and very large companies are more likely to cause multiple deaths. A potential rationale could be larger truck companies own more heavier and larger trucks, and those trucks are hard to maneuver and need longer time to perform brake operation in an emergency situation.

As expected, the company inspection value has a significant impact on their safety conditions. FMCSA has defined three general categories based on the inspection value: higher risk carriers have inspection values of 75 or greater, medium risk carriers have inspection values between 50 and 74, and low risk carriers have inspection values less than 50. The results indicate that low inspection values, less than 30 or 25 respectively, are positively associated with the likelihood of a less severe crash result, damage only or injury. In other words, low inspection values generally indicate better performing companies; however, the bench mark value is around 25 to 30 rather than 50 in terms of crash severity. For fatality analysis, the bench mark value is 45,

which is close to the 50 used in the FMCSA inspection categories. The most interesting finding is for multiple-fatality crashes. A truck with a company inspection value between 30 and 70 or greater than 90, is significantly more prone to multiple-fatality crashes. For a company with an inspection value greater than 90, it is not a surprise. The inspection value is based on the prior safety record of the company. Those companies with higher values have had more crashes and more violations in past inspections, and thus are more likely to continue to have safety issue. Another rationale for this observation could be that those companies have more extreme larger trucks, but to verify the hypothesis, further study is suggested. However, for a company with a value between 30 and 70, it is very surprising. Further single factor investigation is needed to understand how the inspection value is associated with crash severity.

It is found that trucks owned by interstate companies are more prone to have fatality crashes, while those owned by an intrastate company are more likely to be involved in damage or injury only crashes. Interstate company truck drivers usually have longer driving distance, which could cause drowsy driving run-off-the-road fatal crashes. Thus, further data collection to support significant test on drowsy driving contribution to crash severity levels between interstate and intrastate company truck drivers is recommended. If the hypothesis that severe crashes associated with interstate companies are significantly attributed by drowsy driving is tested to be true, several countermeasures can be suggested to apply to interstate companies, such as law enforcement or regulation to have at least two drivers per shift to reduce risk of drowsy driving. Or requirement for adopting new advanced technologies to those interstate trucking companies could be another possible countermeasure to prevent drowsy driving. For example, with machine learning advances, drowsy warning system based on a facial movement recognition machine learning algorithm could be helpful to alert drowsy drivers.

In addition, interstate companies may hire more drivers registered in the same state as the companies. These drivers from other states may be not familiar with local driving behavior when driving in other states. For example, comparing with ND and CO, most of ND highways are rural highway and ND drivers are less aggressive than CO drivers. Thus, when a ND truck driver drives in CO, he or she may not familiar with the driving behavior, and may not react in time. Because of limitation of data availability, it is suggested that police officers to record driver license registration state for the accident involved drivers. If the hypothesis is verified, truck drivers are recommended to be trained to drive under various driving-behavior environments.

Newly registered companies are found to be non-significant for damage-only crashes. However, trucks owned by newly registered companies have a higher risk of fatality crashes. This is intuitive because newly registered companies are usually less experienced in fleet management and safety practices.

Regarding crash characteristics, the first harmful event is one of the most significant explanatory variables in crash outcome prediction. A conclusion can be drawn that the huge difference of speed and weight between vehicles involved in crashes is one of the major contributors to fatality crashes. Under such cases, the more vulnerable road users expose themselves to a high risk of fatal crashes. For example, when a truck hits a passenger car, the fatal outcome could be due to the huge impact at the moment of collision. As expected, the more vehicles involved in a truck crash increases the probability of a more severe outcome. Crash severity level also changes over the time in a day. It is notable that early morning (3AM-6AM) is considered as the most dangerous time, given the fact that both single fatality and multiple-fatality crashes are more likely to happen during this period. This may be the result of difficulty in making appropriate responses when it is dark or lack of sleep (*Pahukula, Hernandez,*

Unnikrishnan, 2015). During weekends, crashes are more likely to be fatal crashes, while on Fridays crashes, are prone to be non-fatal.

Regarding to environment characteristics, not surprisingly, the weather condition is one of the significant factors affecting crash severity levels. Interestingly, fatal crashes are less likely to happen on a snowy or rainy day when drivers are more cautious than usual. Nevertheless, single fatality crashes are more likely to happen with no adverse weather condition. The reasons for such finding can be that more truck traffic and/or higher travel speed under good weather and drivers tend to really focus on driving, slow down and keep their eyes on the road under bad weather. Thus, traffic exposure data can be really helpful for better understand the relationship between weather and severity. Validation of such hypothesis can result to warning sign on a good weather to remind drivers to driver under speed limits or apply speed enforcement under good weather. Fog and severe crosswinds negatively affect drivers' visualization and make large trucks hard to control. Thus, under these conditions, the probability of multiple-fatality crashes is predicted to increase. An icy road surface is a definite factor of crashes. Drivers usually drive with more attention than usual; however, an icy road surface can make crashes inevitable. Thus, an icy road surface increases the risk of damage-only, injury and single-fatality crashes, but decreases the likelihood of multiple-fatality crashes, most likely because drivers drive with a lower speed under such a condition. A slushy road condition raises the risk of damage-only and injury crashes. On the other hand, the likelihood of fatal crashes increases under a dry or wet road surface condition. Night time is considered as a "dangerous" time, because night time is a positively significant contributor for all fatality level crashes. It is noteworthy that the risk of fatal crashes increases at night with no lighting condition, under which condition visualization is

negatively impacted, which is supported by previous studies (*Lemp, Kockelman, Unnikrishnan, 2011; Kockelman, Murray, Ma, 2007*).

The marginal effect of traffic way type indicates that a median barrier effectively prevents fatal crashes, because fatal crashes are more prone to happen when two-way traffic is not separated.

Regarding driver's characteristics, a driver's age is a significant factor for predicting crash severity levels. Young drivers (< 25 years old) and old drivers (>75 years old) are found to be the most vulnerable groups for multiple-fatality crashes. The underlying reason could be that young people have less driving experience, and may be more prone to dangerous actions. On the other hand, older people do not react as quickly as younger persons, and their overall health condition could also impact their risk of fatalities (*Chen et al., 2015; Campbell, 1991*). The driver's license class is another practically significant variable. Class A, B, and C are significant in improving truck safety performance with regards to crash severity level. For example, drivers with class D licenses are predicted to be more likely to be involved in fatal crashes. An invalid driver license is predicted to increase risk of damage-only, single-fatality, and multiple-fatality crashes. Not surprisingly, people driving with an invalid driver's license could be less responsible, more aggressive, and possibly have a bad driving record, considering it is illegal to drive with an invalid driver license.

Regarding truck characteristics, the cargo body type is a factor impacting injury severity. Cargo tank, flatbed, and grain trucks, or trucks towing another vehicle increase the probability of high injury severities (severity=2, 3). The heavy weight of trucks increases operation difficulty when an emergency happens. Severity level is predicted to be positively related with gross vehicle weight.

CHAPTER 4. CONCLUSIONS

4.1. Summary and Conclusions

As computer technologies develop, more quality data become available and provide foundation for data driven decision making in transportation safety. GLMs are the most popular models favored by researchers and decision makers, as they establish quantitative relationship between target variable and explanatory factors, and the results are easy to be interpreted and can be directly used. However, current GLMs have limitations on depending on pre-defined assumptions, and require the nature of data fit a certain distribution, which can hardly be met all the time, especially when handle safety big data with complex pattern and structure. When the assumptions are violated, research results based on GLMs can be questionable. On the contrary, data mining models have strong capability in handling complex database, and require no assumptions as GLMs do. In addition, data mining models can define non-linear patterns, following the data structure in nature, and require minimal data preprocessing, such as missing value replacement. They have also been proved to be powerful in various industry fields. However, there are only a limited number of safety researches based on data mining models, as data mining models are always criticized to be a weak explanatory tool in spite of its strong predictive capability. In this study, three data mining models, DT, GB, and NN, are tested in highway rail grade crossing crash likelihood analysis. Furthermore, the GB model was selected in commercial truck crash injury severity analysis. By conducting a few more analysis, including contributor variables' marginal effect, variable importance evaluation, and prediction accuracy analysis, the data mining models are demonstrated to be valid, and can serve as alternative tool in transportation safety study.

In the HRGC crash likelihood study, three data mining models are tested and evaluated. The results proved that all three data mining models are robust tool to predict and explain HRGC crashes. All three models provided good predictive accuracy on both crash and non-crash prediction, and the GB model performs the best in terms of predictive accuracy.

The DT model generated a tree-structured output, which can easily be followed by users and understood how the DT model processed the data. Identified by the DT model, railroad and highway traffic volume and train speed are the most important influential factors. Crashes are more likely to happen at crossings with high traffic volume and train speed. Also, it is shown that the presence of advance train warning systems and train detecting devices are helpful in reducing crash likelihood. It indicates that crash likelihood is the highest at crossings intersecting with a non-interstate national highway.

Forming an ensemble of simple decision trees, the GB model improves the DT model in respect of forecasting accuracy. More importantly, it can also provide easy-to-interpret marginal effect analysis of contributor variables. The marginal effect analysis successfully revealed and demonstrated the non-linear and complex relationship between crash likelihood and related variables, which evidenced that it is fallacious to express the non-linear relationship by linear models. The top five important influential variables identified by the optimal GB model are AADT, day through train volume, warning device type, night through train volume, and average train speed.

With different application purposes, variable importance analysis is further conducted based on two criteria: connection weight and mean square error. If research objective is to identify contributors to crash likelihood, the variable importance based on connection weight is recommended. If application purpose is to predict crash accurately, variable importance based on

mean square error is suggested. Aiming to further explore relationship between crash likelihood and explanatory variables and to reveal the black-box of the NN model, a quantitative marginal effect analysis is conducted. In addition, relationship between crash likelihood and associated factors is tested when other variables held at different levels. The research results demonstrate that the relationship between crash likelihood and associated factors is non-linear, and reinforce the statement that independent variables are not completely independent with each other.

In the study of commercial truck crash injury severity prediction, the GB model was applied to detail a comprehensive analysis of the impacts of a set of heterogeneous factors (trucking company, crash, environment, truck driver, and truck characteristics) on injury severity caused by truck crashes by analyzing six recent years of Federal Motor Carrier Safety Administration data. The target variable (crash severity) is classified into four categories: property damage only, injury only, one fatality, and two or more fatalities. Based on a GB model, twenty-two variables are proved to be significantly related with severity. For the first time, trucking company and driver characteristics are proved to have significant impact on truck crash injury severity. Some of the results in this study reinforce previous studies' conclusions. For example, wet road surface, bad visualization (dark or low light conditions, or fog/poor weather conditions), strong crosswind, heavy gross vehicle weight (over 26,000lbs), and collisions with opposite traffic are estimated to increase the likelihood of more severe outcomes. Young drivers (under 25 years old) and old drivers (over 75 years old) are predicted to be the most likely groups to be involved in crashes resulting in fatalities. Also, truck crash severity level gets higher when more vehicles are involved in truck crashes. One interesting finding is that fatal crashes are likely to happen when the weather is good or the road surface has no adverse conditions, perhaps because adverse conditions make people vigilant to potential risk. Another unique contribution of this study is to

demonstrate the significant effect of the trucking company and driver characteristics on injury severities. Based on crash data from ND and CO, it is estimated that carriers registered in MI, MB, NC, ND, and PA increase the likelihood of the most severe outcomes. Companies owning 2 to 5 trucks are predicted to have the lowest probability of crash risk. Carriers with inspection values of 30-70, or greater than 90, increased the possibility of high injury severities. Newly-registered carriers and interstate carriers are estimated to be associated with a higher possibility of fatal crashes. Drivers with a regular driver license (Class D) only are at greater risk of being involved in fatal crashes.

4.2. Limitation and Future Study

Care should be taken when interpreting the object of this study. This research aims to promote application of data mining models in transportation safety study, instead of addressing that data mining models are superior to GLMs in all aspects. However, studies applying a combined data mining models and GLMs are recommended as data mining models and GLMs have different pros and cons in terms of predictive and explanatory capability. This study did not examine all data mining models in safety study. In addition, as different data mining models have different features and probably are feasible in different types of researches. Thus, more data mining models, such as association rules and clustering analysis are recommended to be tested in safety research. In practice, it is possible that combinations of a few data mining models are used. For example, classification and clustering are similar techniques. By using clustering models to identify objects with more similar attributes, a further classification can be refined by a decision tree model. How to model rare event in crash studies is always a big concern and a challenge that all researchers have to face. Even though this article improved accuracy when forecasting rare events, rare event modeling still has a lot to be improved.

Isolating the analysis of effect of each variable is a challenge. The marginal effect analysis conducted in this study considered only a limited number of situations, including keeping all other unstudied variables at their mean, maximum and minimum values. There are various situations are suggested for future studies. Using AADT for instance, when studying AADT's marginal effect on HRGC crash likelihood, it is recommended to set other variables' level based on a real case situation. In addition, variable importance analysis in the NN model is analyzed by the model performance difference when taking only one variable out each time to the optimal model performance. However, due to correlation among variables, it is possible that a combination of variables may have greater impact on model performance than measuring these variables individually. Thus, future study is also recommended to explore variable importance by removing multiple variables each time.

In the HRGC study, the selected contributor variables were traffic, crossing design, and highway characteristics. According to FRA (2017), some of the crashes at HRGCs are caused by aggressive driving behaviors. Aggressive drivers are more likely to ignore warning signs, and have no patience to wait as to attempt to race or beat the train. However, data about driver and driver behavior, such as driver gender, driving record, and behaviors that causing the crash, was unavailable, and it is necessary to study driver related variables' effect on crash likelihood. Auto insurance companies have equipment installed on insured vehicles to track the drivers' driving behavior, such as if the drivers make full stops at stop signs and average driving speed. These data can be used to measure and define aggressive drivers. A further study can be made to see if the aggressive drivers defined by auto insurance company data are more risky of crashes at HRGCs.

In the commercial truck crash injury severity study, several valuable findings are discovered. However, the underlying reasons leading to the observations are not analyzed in this study due to data limitations. Several potential rationales are addressed. Further studies in the following aspects are recommended to demonstrate the hypothetical rationales and extend this study:

- (1) To collect truck configuration information at corporation level is recommended. Such information can be extremely helpful for better understand relationship between truck company size and crash severities, such as small truck companies owning 2-5 truck perform best in terms of crash severity.
- (2) To collect traffic exposure at corporation level is also recommended to further understand relationship between truck company size, inspection values and crash severities. For example, companies have inspection value greater than 90 tends to continuously perform badly, could it be due to those companies own the most heavy trucks and have the most traffic exposures? To answer those questions, traffic exposure is critical and unfortunately is lack of in this study.
- (3) Travel speed changes under various weather conditions and for various truck configurations is also need collected to further research on weather impact on crash severity
- (4) Driver behavior and driving-environment awareness in corporation level is also needed for further research to clarify the reasons for relationship between trucking company registration state and severity.
- (5) The data analyzed in this study is only from states of CO and ND. It is recommended to obtain data from other states to expand the research to national level;

(6) Alternative algorithms are also suggested to be used to validate the findings in this research.

REFERENCES

- Abdulhafedh, A. "Crash Frequency Analysis." *Journal of Transportation Technologies*. Vol. 6, 2016, pp 169-180.
- Weigend, A. "Big Data, Social Data and Marketing." World Marketing Form, 2013.
http://weigend.com/files/speaking/Weigend_WorldMarketingForum_MEX_2013.06.27.pdf. Accessed on Aug. 20, 2017.
- Association of American Railroads, 2017. Highway-rail Grade Crossing Safety.
<https://www.aar.org/BackgroundPapers/Highway-Rail%20Grade%20Crossing%20Safety.pdf>. Accessed on Dec. 23, 2017
- Assumptions of Linear Regression. Statistics Solution 2017.
<http://www.statisticssolutions.com/assumptions-of-linear-regression/>. Accessed Jul. 1, 2017.
- Arthur, L. "What Is Big Data?" Forbes. 2013.
<https://www.forbes.com/sites/lisaarthur/2013/08/15/what-is-big-data/#7eef96c15c85>. Accessed Nov. 1, 2017.
- Austin, R., Carson, J. "An Alternative Accident Prediction Model for Highway-rail Interfaces." *Accident Analysis and Prevention*, Vol. 34, 2002, pp. 31–42.
- Chadwick, S., Zhou, N., and Saat, M. "Highway-rail grade crossing safety challenges for shared operations of high-speed passenger and heavy freight rail in the U.S." *Safety Science*. Vol. 28, 2016, pp. 128-137
- Chang, L. "Analysis of freeway accident frequencies: Negative binomial regression versus artificial neural network." *Safety Science*, Vol. 43, 2005, pp. 541-557.

Chang, L., and Wang, H. "Analysis of Traffic Injury Severity: an Application of Non-parametric Classification Tree Techniques." *Accident Analysis and Prevention*, Vol 38, 2006, pp. 1019-1027.

Chang, L., Y., and Chen, W., C. "Data Mining of Tree-based Models to Analyze Freeway Accident Frequency." *Journal of Safety Research*, Vol. 36, 2005, pp. 365-375.

Chiou, Y. "An Artificial Neural Network-based Expert System for the Appraisal of Two-car Crash Accidents." *Accident Analysis and Prevention*, Vol. 38, 2006, pp. 777-785.

Taylor, C. "Structured vs Unstructured Data." Datamation, 2017.

<https://www.datamation.com/big-data/structured-vs-unstructured-data.html>. Accessed on Sep. 21, 2017.

Codur, M, Y. and Tortum, A. "An Artificial Neural Network Model for Highway Accident Prediction: a Case Study of Erzurum, Turkey." *Traffic and Transportation*, Vol. 27, 2015, No. 3, 217-225.

Data Mining Techniques, 2018. TIBC Statistica. <http://www.statsoft.com/Textbook/Data-Mining-Techniques#mining>. Accessed Jan. 21, 2018.

Ellingwood, J. "An Introduction to Big Data Concept and Terminology." Digital Ocean. Sep. 28, 2016. <https://www.digitalocean.com/community/tutorials/an-introduction-to-big-data-concepts-and-terminology>. Accessed on Sep 12, 2017.

Federal Railroad Administration Highway-rail Grade Crossing Safety Fact Sheets, 2013.

USDOT. <file:///C:/Users/zz1802/Downloads/FRA%20Highway-Rail%20Grade%20Crossing%20Safety%20Fact%20Sheet.pdf>. Accessed on Nov. 11, 2017.

- Fish, K.E., and Blodgett, J.G. "A visual method for determining variable importance in an artificial neural network model: An empirical benchmark study." *Journal of Targeting, Measurement & Analysis for Marketing*, Vol. 11(3), 2003, pp. 244-254.
- Garson, G.D. "Interpreting neural network connection weights." *Artif. Intell. Expert*. Vol. 6, 1991, pp. 47-51.
- Forsyth, J. "U.S. commuters spend about 42 hours a year stuck in traffic jams." Reuters, Aug. 25, 2015. <https://www.reuters.com/article/us-usa-traffic-study/u-s-commuters-spend-about-42-hours-a-year-stuck-in-traffic-jams-idUSKCN0QV0A820150826>. Accessed on Sep.13, 2017.
- Gevrey, M., I. Dimopoulos and S. Lek. "Review and comparison of methods to study the contribution of variables in artificial neural network models." *Ecol. Model*. Vol. 160, 2003, pp. 249-264.
- Hu, S., Li, C., and Lee, C. "Model Crash Frequency at Highway-rail Grade Crossings Using Negative Binomial Regression." *Journal of the Chinese Institute of Engineers*, Vol. 35, 2012, pp. 841-852.
- Karlaftis, M., G., and Vlahogianni, E., I. "Statistical Methods versus Neural Networks in Transportation Research: Differences, Similarities and Some Insights." *Transportation Research Part C*. Vol. 19, 2011, 387-399.
- Kashani, A., T., Rabieyan, R., and Besharati, M., M. "A Data Mining Approach to Investigate the Factors Influencing the Crash Severity of Motorcycle Pillion Passengers." *Journal of Safety Research*, Vol. 51, 2014, pp. 93-98.

- Koslowski, T. "Your Connected Vehicle is Arriving." MIT Technology Review. 2012.
<https://www.technologyreview.com/s/426523/your-connected-vehicle-is-arriving/>.
Accessed on Sep. 29, 2017.
- Laney, Douglas. "3D Data Management: Controlling Data Volume, Velocity and Variety".
Gartner. Retrieved 6 February 2001.
- Li, X., Zhang, Y., Xie, Y. "Predicting Motor Vehicle Crashes Using Support Vector Machines
Models." *Accident Analysis and Prevention*, Vol. 40, 2008, pp. 1611-1618.
- Lu, P., and Tolliver, D. "Accident Prediction Model for Public Highway-rail Grade Crossings."
Accident Analysis and Prevention, Vol. 90, 2016, pp. 73-81.
- McCollister, G., and Pflaum, C. "A Model to Predict the Probability of Highway Rail Crossing
Accidents." *Institution of Mechanical Engineers*. Vol. 221, 2007, Part F.
- McCulloch, W., and Pitts, W. "A Logical Calculus of the Ideas Immanent in Nervous Activity."
Bulletin of Mathematical Biophysics. Vol. 5, 1943, pp. 115-133.
- Millegan, H., et al. "Evaluation of Effectiveness Stop Sign Treatments at Highway Rail Grade
Crossings". *Transportation Research Record: Journal of the Transportation Research
Board*, No. 2122, 2009, pp. 78-85.
- National Highway Traffic Safety Administration, 2016. "Traffic Safety Facts." U.S. Department
of Transportation.
- Oh, A., Washington, S., P., and Nam, D. "Accident Prediction Model for Railway-highway
Interfaces." *Accident Analysis and Prevention*, Vol. 38, 2006, pp. 346-356.
- Olden, J.D. and D.A. Jackson. "Illuminating the "black box": a randomization approach for
understanding variable contributions in artificial neural networks." *Ecological Modeling*.
Vol. 154, 2002, pp. 135-150.

Paying Our Way: A New Framework for Transportation Finance. National Surface Transportation Infrastructure Financing Commission, February 2009. http://financecommission.dot.gov/Documents/NSTIF_Commission_Final_Report_Advance%20Copy_Feb09.pdf. Accessed January 2, 2013.

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

SAS Institute Inc. Decision Tree Node. SAS Enterprise Miner 14.1 Reference Help <https://support.sas.com/documentation/cdl/en/emgsj/67981/PDF/default/emgsj.pdf>. Accessed Jun. 10, 2017

SAS Institute Inc. The HPSPLIT Procedure. SAS User’s Guild: High-Performance Procedures http://support.sas.com/documentation/cdl/en/stathpug/66410/HTML/default/viewer.htm#stathpug_hpsplit_details32.htm. Accessed Jun. 10, 2017.

Shrank, D., B. Eisele, and T. Lomax. TTI’s 2012 Mobility Report. Texas A&M Transportation Institute, Texas A&M University System, December 2012

Special Report 307: Policy Options for Reducing Energy and Greenhouse Gas Emissions from U.S. Transportation. Transportation Research Board of the National Academies, Washington, D.C., 2011

Texas A&M Transportation Institute. 2015 Urban Mobility Scorecard. <https://static.tti.tamu.edu/tti.tamu.edu/documents/mobility-scorecard-2015.pdf>. Accessed on Sep.14, 2017.

The Digital Universe of Opportunities: Rich Data and the Increasing Value of the Internet of Things. <http://www.emc.com/leadership/digital-universe/2014iview/index.htm>. Accessed on Sep. 19, 2017.

The Organization for Economic Cooperation and Development (OECD), 2015. "Big Data and Transport: Understanding and Assessing Options." International Transport Forum. https://www.itf-oecd.org/sites/default/files/docs/15cpb_bigdata_0.pdf. Accessed on Aug. 18, 2017.

Transportation—Are We There Yet? The Bottom Line Report—2009. American Association of State Highway and Transportation Officials, Washington, D.C., 2009.

Transportation for Tomorrow. National Surface Transportation Policy and Revenue Study Commission, December 2007. http://transportationfortomorrow.com/final_report/. Accessed January 2, 2013.

Transportation Research Board 2013 Executive Committee. "Critical Issues in Transportation." Transportation Research Board, 2013. <http://onlinepubs.trb.org/Onlinepubs/general/criticalissues13.pdf>. Accessed on May. 29, 2017.

U.S. DOT Launches New Railroad Crossing Safety Ad. Federal Railroad Administration. Jan. 13, 2017. <https://www.fra.dot.gov/eLib/details/L18522>. Accessed Jul. 1, 2017

Everitt, B.S. Cluster Analysis. 1993. Third Edition. (New York and Toronto: Halsted Press, of John Wiley & Sons Inc.).

Williams, R. Marginal Effects for Continuous Variables, <https://www3.nd.edu/~rwilliam/stats3/Margins02.pdf>, Accessed on Jul. 2, 2017.

World Health Organization. Burden of disease project. Global burden of disease estimates for 2001. <http://www3.who.int/whosis/menu.cfm?path=burden>

- Xie, Y., Lord, D., and Zhang, Y. "Predicting Motor Vehicle Collision Using Bayesian Neural Network Models: an Empirical Analysis." *Accident Analysis and Prevention*, Vol. 39, 2007, pp. 922-933.
- Yan, X., Richards, S., and Su, X. "Using Hierarchical Tree-based Regression Model to Predict Train-vehicle Crashes at Passive Highway-rail Grade Crossings." *Accident Analysis and Prevention*, Vol. 42, 2010, pp. 64-74.
- Zeng, Q., and Huang, H. "A Stable and Optimized Neural Network Model for Crash Injury Severity Prediction." *Accident Analysis and Prevention*, Vol. 73, 2014, pp. 351-358.
- Zhang, L., and Meng, X. "An Approach to Predict Road Accident Frequencies: Application of Fuzzy Neural Network".
<http://onlinepubs.trb.org/onlinepubs/conferences/2011/RSS/2/Zheng,L.pdf>. Accessed on Jul. 2, 2017.
- Zhang, Y., Xie, Y., and Li, L. "Crash Frequency Analysis of Different Types of Urban Roadway Segments Using Generalized Additive Model." *Journal of Safety Research*, Vol. 43, 2012, pp. 107-114.
- Zheng, Z., Lu, P., and Tolliver, D. "Accident Prediction for Highway-Rail Grade Crossings using Decision Tree Approach: An Empirical Analysis." *Journal of the Transportation Research Board*. Vol. 2545, 2016. pp. 115-122.

Web reference

Statista, Road Accidents in the U.S. – Statistics & Facts.

<https://www.statista.com/topics/3708/road-accidents-in-the-us/> Accessed in June, 2017.

PBS, Traffic Accidents in the U.S. Cost \$871 Billion a Year, Federal Study Finds. Lowy, J., Associated Press. <https://www.pbs.org/newshour/nation/motor-vehicle-crashes-u-s-cost-871-billion-year-federal-study-finds> Accessed in Jan, 2018

Congressional Budget Office, 2015. Public Spending on Transportation and Water Infrastructure, 1956 to 2014. <https://www.cbo.gov/publication/49910> Access in Jan, 2018

Barnat, R, 2014. Strategic Management: Formulation and Implementation. <http://www.introduction-to-management.24xls.com/en125> Accessed in Dec. 2017

Ghodke, D., 2015. Getting the Real Deal: Why Data Mining is Important. Firstpost. <https://www.firstpost.com/business/getting-the-real-deal-why-data-mining-is-important-2471526.html> Accessed in Nov. 2017

PennState Eberly College of Science, 2017. Introduction to Generalized Linear Models. <https://newonlinecourses.science.psu.edu/stat504/node/216/> Accessed in Nov, 2017

SAS Institute Inc. Data Mining What It is and Why It Matters. https://www.sas.com/en_us/insights/analytics/data-mining.html Accessed in Nov, 2017

Techopedia, Association Rule Mining. <https://www.techopedia.com/definition/30306/association-rule-mining> Accessed in Nov, 2017

Stefanowski, J., 2008. Data Mining – Clustering. Institute of Computing Sciences Poznan University of Technology. <https://www.techopedia.com/definition/30306/association-rule-mining> Accessed in Nov, 2017

Fu, Y., N.A. Data Mining: Tasks, Techniques, and Applications. University of Missouri – Rolla. <http://academic.csuohio.edu/fuy/Pub/pot97.pdf> Accessed in Nov, 2017

Li, R., 2015. Top 10 Data Mining Algorithms, Explained. KDnuggets.

<https://www.kdnuggets.com/2015/05/top-10-data-mining-algorithms-explained.html/3>

Accessed in Oct, 2017

Priyadharshini, G., S., 2017. Data Mining vs Statistics – How are They Different? Simplilearn

Solutions. <https://www.simplilearn.com/data-mining-vs-statistics-article> Accessed in Oct,

2017.

Brusilovkey, P. N.A. Data Mining vs Statistics. Business Intelligence Solutions.

<http://www.bisolutions.us/Data-Mining-vs-Statistics.php> Accessed in Oct, 2016

Shmueli, G., 2010. To Explain or to Predict? Statistical Science. Vol, 25 289-310

<https://www.stat.berkeley.edu/users/aldous/157/Papers/shmueli.pdf> Accessed in Oct,

2016

Cortez, P., and Embrechts, M., J., 2011. Opening Block Box Data Mining Models Using

Sensitivity Analysis. Computational Intelligence and Data Mining.

<https://pdfs.semanticscholar.org/de03/d9438be30501a4d219f50333ac72d3ad4c6d.pdf>

Accessed in Oct, 2016

Federal Railroad Administration, 2017. U.S. DOT Launches New Railroad Crossing Safety Ad.

U.S. Department of Transportation Office of Public Affairs. Jan.13, 2017.

<https://www.fra.dot.gov/eLib/details/L18522> Accessed in Oct. 2016.

SAS Institute Inc. The HPSPLIT Procedure. SAS User's Guild: High-Performance Procedures

Accessed in Oct. 2016