A MACHINE LEARNING MODEL FOR VEHICLE CRASH TYPE PREDICTION

BY

XIYUE LI

THESIS

Submitted in partial fulfillment of the requirements
for the degree of Master of Science in Civil Engineering
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2020

Urbana, Illinois

Adviser:

Professor Hadi Meidani

ABSTRACT

Travel safety research works include the studies of risk factor investigation, crash detection, and crash frequency prediction. However, the existing studies are focused on the macro level, paying little attention to the specific crash type.

In this study, eXtreme Gradient Boosting (XGBoost) method is applied to predict the occurrence of different types of crashes. A two-layer model is proposed. The first layer is used to distinguish potential crashes from crash-free observations and the second layer is used for crash type recognition. The results show that the proposed model can detect the potential accident and identify the crash type successfully, with accuracy levels of over 99% and 62%, respectively. Besides the crash type prediction model, this study provides a detailed analysis of the impacts of different risk factors on different types of crashes.

From the traffic management perspective, the results of this study can prepare traffic managers for the potential threatens well in advance. From travelers' perspective, the results of this study can be used to warn travelers of potential dangers before the trip so that a better trip planning can be made as well as alert them of the potential dangers during the trip. All of these actions are important for the travel safety management and can help protect people's life and property.

Keywords: Crash type, Crash prediction, Machine learning

ACKNOWLEDGMENTS

First and foremost, I would like to express my greatest gratitude to Dr. Hadi Meidani. Thank you so much for all the guidance and support. Thank you for believing in me and training me to think like a researcher. Thank you for giving me the opportunity to be a member of your research group. I enjoyed studying in this group and have learned a lot here.

I would like to give special thanks to Mohammad Amin Nabian. Thank you for your valuable suggestions for this study and for all the other aspects. Your suggestions help me a lot. Thanks also go to Qiming Zhu. Thank you for helping me improve my coding skills and have a deeper understanding of the algorithms.

I would also like to give special thanks to Rini Gladstone and Negin Alemazkoor. Thank you for your assistance during the Transportation Research Board annual meeting. I would also like to take this opportunity to thank all my friends and colleagues. Thank you for your care and love.

Finally, I would like to express my sincere gratitude to my husband, Zhonghua Zheng and my parents Mr. Wenqiang Li and Mrs. Yongqing Liang. Thank you for your encouragement, support and love.

TABLE OF CONTENTS

CHAPTER 1: INTRODUCTION

Traffic accident is one of the leading causes of injuries and deaths. The number of police reported crashes in 2018 is estimated to be 6,734,000 in the United States, leading to 36,560 fatalities and 2,710,000 injuries, according to the Traffic Safety Facts Research Note published by the National Highway Traffic Safety Administration. Enormous loss such as productivity reduction, medical bill, legal and court expenditure, emergency cost, insurance fee, property damage, congestion cost and workplace loss that are caused by traffic crashes bring huge burdens to individuals and the whole society (Kahn & Gotschall, 2015).

The transportation agencies have taken many measures and have made significant investment to improve travel safety. For example, the Federal Highway Administration has released the Highway Safety Improvement Program, from which more than 2 billion dollars are delivered every year to improve the travel safety level. The program stresses the importance of using a data-driven and strategic approach to reduce the number of fatalities and severe injuries through infrastructure-related improvements and provides 4 directions for substantive roadway safety study, namely, crash frequency, crash rate, crash type and crash severity. Researchers also spend a significant amount of effort in helping reduce traffic accident rate. Many methods have been proposed for crash prediction. However, these studies mainly focus on the macro level, identifying whether there would be a crash or predicting the crash frequency in the study area. The study of different types of crashes has not drawn enough attention. While it’s important to forecast potential crashes or find the crash-prone area, it’s also important to clearly understand the distinctions between different types of crashes as well as predict them accurately so that corresponding countermeasures for the particular type of crash can be provided in time to save people’s lives and property.

Highway Safety Improvement Program Manual provides a guidance of three steps for identifying potential traffic crashes. The three steps are (1) Identifying key crash types, which helps select the crash types that have a high frequency of occurrence or would lead to severe consequences (2) Determining the characteristics of facilities where the key crashes occur (3) Setting thresholds for defining facilities of high risks and implementing the countermeasures. Research about different types of crashes is carried out in this study following the aforementioned guidance with some modifications. The rest of this paper is organized as follows: Chapter 2 introduces some previous studies which are closely related to vehicle crash prediction. Chapter 3 identifies the key crash types and investigates the impacts of risk factors on different types of crashes using the data resources obtained from the Virginia SmarterRoads platform. Chapter 4 introduces the proposed machine learning model for predicting different types of crashes. Chapter 5 discusses the results of the proposed model and compares it with some other models. Chapter 6 makes a conclusion of this study and points out 5 questions to be investigated in the future study.

CHAPTER 2: LITERATURE REVIEW

Researchers have investigated various methods and plentiful data resources for the improvement of travel safety. Some researchers explored the relationship between ambient factors and traffic crashes and identified the most closely related factors to help transportation engineers design safer roads and present safer driving conditions. Others developed models to forecast the potential crashes so that the travelers can have a better trip planning and transportation engineers can prepare in advance. As indicated by (Hossain, Abdel-Aty, Quddus, Muromachi, & Sadeek, 2019), a proactive traffic management system would be needed, and the basic requirement of it is having a reliable crash prediction model. In this chapter, a summarization of widely adopted methods regarding crash prediction is made from the perspective of statistical method, machine learning method, and deep learning method, which is an evolution of machine learning method. In most studies, the crash prediction problem was explored mainly from two aspects, crash detection and crash frequency prediction. Both of them are included in this chapter.

2.1 Statistical Method

Due to that fact that the crash frequencies are random, discrete and non-negative numbers, many researchers chose to use the Poisson model to predict crashes. The Poisson model can be represented as $\mathrm{p}(\mathrm{n}) = \frac{\lambda^{\mathrm{n}} \exp{(-\lambda)}}{\mathrm{n}!}$ where $\mathrm{p}(\mathrm{n})$ is the probability of having n crashes over the study period, λ is the expected crash frequency, X is the vector of explanatory variables and β is the coefficient vector. However, the Poisson model requires that the mean and variance to be the same ($\mathrm{mean} = \lambda, \mathrm{variance} = \lambda$), which is not realistic for the crash data. Negative Binomial model releases this constraint by adding an error term to calibrate the expected crash frequency such that $\lambda = \exp{(\beta \mathrm{X} + \varepsilon)}$ where $\exp(\varepsilon)$ follows Gamma distribution. It has been proved in many studies that this method is more appropriate than the Poisson method to model the vehicle crash data of which variance exceeds the mean (Hadi, Aruldhas, Chow, & Wattleworth, 1995; Poch & Mannering, 1996).

Highway Safety Manual adapted the Negative Binomial model and named it as Safety Performance Function (SPF) approach, to provide guidance for transportation engineers to estimate the crash frequency. To be more specific, for the task of estimating crashes on highway segments, the function is expressed as $\exp\,(\mathrm{a} + \beta\ln(\mathrm{AADT}) + \ln(\mathrm{Segment\ Lenght}))$. For intersections, the function is expressed as $\exp\,(\mathrm{a} + \beta_1\ln(\mathrm{AADT_{major}}) + \beta_2\ln(\mathrm{AADT_{minor}}))$ where $\mathrm{AADT_{major}}$ and $\mathrm{AADT_{minor}}$ are the annual average daily traffic of major intersections and minor intersections, respectively. However, the safety performance function used for one specific site cannot be directly applied for another site since the facility conditions may be different and the traffic pattern in different regions are different. Crash Modification Factors (CMFs) and Calibration factors (C) must be applied to adjust the predicted crash frequency according to site-specific and local conditions. The adjusted predicted crash frequency can be calculated by $\mathrm{C} \times \mathrm{CMFs} \times \mathrm{Predicted\ crash\ frequency\ under\ base\ condition}$. For predicting long-term expected crash frequency, Empirical Bayes approach is adopted. The long-term expected crash frequency is calculated as $\omega \times \mathrm{Adjusted\ predicted\ crash\ frequency} + (1-\omega) \times \mathrm{Observed\ crash\ frequency}$, where the weight factor ω reflects the reliability of the predicting model.

Although being widely adopted, the Poisson model and the Negative Binomial model have some deficiencies. For example, as indicated by (Chang, 2005), if the rule of the Gamma distribution is not satisfied, the Negative Binomial model would not be valid for crash prediction. Some researchers suggested using the Poisson-lognormal model, which has a more relaxed constraint. The error term, $\exp(\varepsilon)$, of the Poisson-lognormal model follows the Lognormal distribution instead of the Gamma distribution. To overcome the problems of excessive non-crash observations, some researchers proposed using Zero-inflated Poisson and Zero-inflated Negative Binomial model (Lee, Stevenson, Wang, & Yau, 2002; Miaou, 1994; V. Shankar, Milton, & Mannering, 1997). To incorporate the spatial and temporal correlations, additional random effect variables were introduced in the Poisson model and the Negative Binomial model (Johansson, 1996; V. N. Shankar, Albin, Milton, & Mannering, 1998). Observations were divided into several groups according to the when or where the crash happened. The expected crash frequency of the road section i belonging to the group j was calculated as $\lambda_{\mathrm{ij}} = \exp\,(\beta \mathrm{X_{ij}})\exp\,(\varepsilon_{\mathrm{j}})$. The group-specific effect of the observation group j was reflected by ε_{j} and $\exp\,(\varepsilon_{\mathrm{j}})$ followed Gamma distribution. Some researchers also suggested using the Negative Multinomial Regression model to address the correlation problem. (Caliendo, Guida, & Parisi, 2007) compared the capability of Poisson, Negative Binomial and Negative Multinomial regression model regarding accident detection on multilane rural roads. The result of the study showed that the Negative Multinomial regression model considering over-dispersion impacts had the best prediction performance.

There are many other innovative statistical models proposed in the past two to three decades. Each of them has its distinguished contribution to the travel safety study. However, for various reasons such as model complexity or transferability, the applications of them are yet very limited. The inherent limitations of the crash data as well as the strengths and deficiencies of the proposed statistical models were comprehensively analyzed in the study of (Lord & Mannering, 2010) and (Mannering & Bhat, 2014).

2.2 Machine Learning Method

Machine Learning methods have some advantages over the statistical methods. First, while the statistical methods predefine the underlying relationship between the explanatory variables and the dependent variables, which would lead to the failure of the model if the assumptions are violated, machine learning methods do not require inherent assumptions. Second, machine learning methods can handle the problem of associations between the explanatory variables (Chang, 2005) and are able to capture complicated relationships while it might be difficult to be achieved in statistical models. Third, as mentioned in the study of (Lord & Mannering, 2010; Mannering & Bhat, 2014), the progress of transportation related research would be greatly promoted by new data resources provided by the emerging technologies. New data resources such as video surveillance data, satellite images, social media data, mobile sensing data and GPS trajectory data are available for travel safety study in some districts. Statistical models might not be able to handle these data resources very well. Last but not least, Machine Learning methods are able to predict vehicle crashes over a large area and a long time period, which might be difficult for statistical models.

A number of studies were conducted to unveil the potential of using machine learning methods to predict crashes. The most widely adopted models are Logistic Regression, K Nearest Neighbor, Support Vector Machine and tree-based models. (Abdel-Aty, Uddin, Pande, Abdalla, & Hsia, 2004) proposed a matched case-control logistic regression method to detect freeway crashes using real-time traffic flow data. Several non-crash observations observed by the same loop detector of the crash site were collected and matched with the observations when crashes happened. A Logistic Regression model was applied to predict the crash occurrence. (Lv, Tang,

& Zhao, 2009) applied K-Nearest Neighbor (KNN) method to predict highway vehicle crashes using real-time traffic flow data. The average Euclidean distance of different classes was calculated to help select the accident precursors before training the KNN classifier. Support Vector Machine (SVM) method, of which the decision is made by finding the hyperplane that has the max distance to the closet element of each class, was also examined by many researchers. (X. Li, Lord, Zhang, & Xie, 2008) applied the SVM model and concluded that the SVM mothod outperforms the Negative Binomial method in terms of predicting vehicle crash frequency on rural frontage roads. (S. Chen, Wang, & van Zuylen, 2009) ensembled several SVM models to help detect freeway traffic accidents. Different combination schemes based on bagging, boosting and cross-validation committee were tested in the study and the results indicated that the ensemble technique can improve the accuracy of a single SVM in most cases. (Dong, Huang, & Zheng, 2015) investigated the possibility of using the SVM method for predicting crash risk (frequency) of different traffic analysis zones. The spatial association of analysis zones was revealed by 4 different spatial dependence matrices including the matrix of adjacency, the matrix of shared boundary length, the matrix of geometry centroid distance, and the matrix of crash-weighted centroid distance. The study showed that SVM with radial based kernel outperformed the SVM with linear kernel and the Bayesian spatial model. However, the SVM model has some limitations (Yu & Abdel-Aty, 2013). First, it’s necessary to have the feature selection process before building the model. If all the available variables are fed into the model without the knowledge of which variables are informative, the model may not be able to perform the task well. The second limitation is that the model may have unsatisfactory performance when handling datasets with a large number of samples.

While the methods mentioned above use all of the input features collectively to decide the final result, the tree-based machine learning method has multiple decision steps that are organized in a hierarchical structure. One feature is used to make a binary decision in each step, and the final decision is made according to the leaf node. Thanks to this nature, it can graphically represent the analyzing process and thus provide straightforward guidance for the engineers to understand the interrelation. Many researchers examined the tree-based machine learning method for crash prediction. For instance, (Chang & Chen, 2005) proposed using classification tree method to predict accident rate on freeway segments. Highway geometry characteristics, traffic characteristics, and environment conditions are investigated. The graphical representation of the tree provided by the authors indicated that high traffic volume, high precipitation, high grade and large curvature would lead to a relative high crash accident rate.

While a single tree is built on the entire dataset and may not be robust, ensemble methods combine many weak learners to be a single strong learner. There are two ensemble methods, bagging and boosting. The bagging method uses the bootstrap sampling strategy to sample several subsets from the entire dataset with replacement and train a decision tree on each subset. The final result is the majority of or the average of the decisions made by trees. Boosting method create new learner in sequence, each one is formed to help improve the learner built in the previous step. The most representative model that applied the bagging method is the Random Forest model. Not only the observations but also the features are sampled randomly in the training process of the Random Forest model. Though being more robust than a single decision tree, it still has some deficiencies. Since the final prediction of random forest is based on the majority vote or the average of all outcomes, it may not be able to capture the precise value for the regression problem or may vote the wrong result if the parameters are not tuned well for the classification problem. Many studies have found that XGB, which is an implementation of the boosting method being proposed by (T. Chen & Guestrin, 2016), outperforms the Random Forest model. For example, (Schlögl, Stütz, Laaha, & Melcher, 2019) compared the classification capability of regression methods, SVM, bagging (Random Forest and Extremely Randomized Trees) methods, boosting (XGBoost) method and Bayesian Neural Network for incident detection. While both bagging and boosting methods have remarkably better performance than the other methods, the XGBoost method was better than the Random Forest method in general. Researchers have also adopted the XGBoost method for many other travel safety related tasks such as incident detection (Parsa, Movahedi, Taghipour, Derrible, & Mohammadian, 2020), crash severity forecast (Mokoatle, Marivate, & Esiefarienrhe, 2019) and accident duration prediction (Shan, Yang, Zhang, Shi, & Kuang, 2019). In the study of (Shan et al., 2019), the accident duration prediction problem is solved by using an ensembled XGBoost model. Several XGBoost binary classifiers were built and ensembled by a neural network. The study provided a new idea for exploring the complex traffic accident data.

2.3 Neural Network Method

Several studies have examined the neural network method, from shallow neural networks to deep neural networks. (Chang, 2005) compared a 3-layer Artificial Neural Network (ANN) model with the Negative Binomial model for crash frequency prediction. The overall performance of the ANN model was better than the Negative Binomial model for highway sections with one or more accidents, while the Negative Binomial model had a better performance for sections with zero accidents. (Xie, Lord, & Zhang, 2007) compared Bayesian Neural Network (BNN) and Back Propagation Neural Network (BPNN) with the Negative Binomial model for traffic accident frequency prediction on rural frontage roads. The authors found that both of the neural network models had better performance than the Negative Binomial model while the BNN model outperformed the BPNN by incorporating the Bayesian inference. However, the study of (X. Li et al., 2008) compared this model with the Support Vector Machine (SVM) model and found that the two models had similar performance while less time was needed for SVM.

Comparing with the shallow neural networks, Convolutional Neural Network (CNN) can capture the spatial information and have been widely adopted in image-enabled problems. (Wenqi, Dongyu, & Menghua, 2017) mapped the explanatory variables into state matrices and fed them into a CNN model with two hidden layers in order to detect crashes on highway segments. The model had better performance than the BPNN model. (Huang, Wang, & Sharma, 2020) organized real-time traffic as images and fed them into a CNN model to detect crashes on highway segments given a specified traffic condition. The temporal information was also added into the model through the last fully connected layer. (Q. Chen, Song, Yamada, & Shibasaki, 2016) trained a deep model of Stack denoise Autoencoder (SdAE) by using traffic accident data and GPS data to help understand the relationship between human mobility and traffic crashes. The model can generate real-time citywide accident risk map when given real-time GPS data. CNN method also enables traffic crash detection using video data (Formosa, Quddus, Ison, Abdel-Aty, & Yuan, 2020). However, the vision-based method requires the backup of a large amount of computing resources and information storage space.

While the CNN model can capture the spatial information, Long Short Term Memory (LSTM) has very good performance on capturing the periodic information and has been widely adopted in many sequence learning problems. With respect to the crash prediction problem, (Ren, Song, Liu, Hu, & Lei, 2017) proposed an updated LSTM model to predict average crash frequency per hour for each predefined grid cell (1KM*1KM) in Beijing for the same time period of recent 3 days. The model stacked 4 LSTM layers and 3 fully connected layer together. Average accident frequency in previous time periods was fed into the first LSTM layer and the location information of the grid cell was fed into the first fully connected layer. (J. Yuan, Abdel-Aty, Gong, & Cai, 2019) used a 2-layer LSTM model to predict whether there would be a crash for signalized intersections in Florida for next 5 to 10 minutes. The inputs of the model included individual vehicle speed data, signal timing data and vehicle-counts data aggregated in 5 minutes as well as weather records in the nearest time period. The model had better performance than the conditional logistic model.

Researchers have tried to capture spatial patterns and temporal dependency at the same time by using the CNN and LSTM together. (Z. Yuan, Zhou, & Yang, 2018) represented the entire Iowa State by a map with 128-by-64 grids and proposed a Hetero-ConvLSTM model to help predict daily crash frequency in each grid cell. Input features included road network, road conditions, satellite image, rainfall, weather conditions, traffic volume and time information, each being represented by a 128-by-64-by-1 tensor. The authors addressed the spatial heterogeneity issue by training several ConvLSTM models and combing their results together. (P. Li, Abdel-Aty, & Yuan, 2020) concatenated a two-layer LSTM model and a two-layer CNN model in parallel. Real-time signal timing, queuing and waiting time, traffic volume, average vehicle speed and weather-related variables were feed into the parallel LSTM-CNN model to help predict crash risk on arterials in real time. The authors also compared the t-Distributed Stochastic Neighbor Embedding (t-SNE) of raw data with the extracted features from the last layer of the LSTM-CNN model. The crash and non-crash events were almost separable when being presented by the extracted features, while were tangled together when being represented by the raw dataset. (Bao, Liu, & Ukkusuri, 2019) joined CNN, LSTM and ConvLSTM as a synthesis model to help predict the sum of the severity level of all potential crashes in each predefined grid cell in New York. Variables that are spatially varied but temporally static were fed into the CNN model, variables that were temporally varied but spatially static were fed into the LSTM model, variables that were both spatially and temporally varied were fed into the

ConvLSTM model. The output of the three sub-models were combined together as one dense vector and then transformed into the final output through several fully connected layers.

2.4 Summary

While the great contribution of the surveyed models to accident prevention cannot be ignored, it is worth noting that predicting different type of crashes so that specific countermeasure can be made is also important. However, limited research has been conducted for this purpose.

Among all the existing studies related to the crash type prediction, most of them aimed at predicting a specific type of crash under a particular condition. For example, (Pande & Abdel-Aty, 2006) identified the traffic flow regime of which rear end crash prone to happen by using methods including Kohonen Clustering Algorithm, Classification Tree, Multilayer Perceptron, and Normalized Radial Basis Function neural networks. (Abdel-Aty & Haleem, 2011) proposed using a Multivariate Adaptive Regression Splines model to predict vehicle angel crashes for unsignalized intersections. Random Forest model was applied to help screen the covariates. There are also works exploring the method of predicting multiple crash types though the number is extremely limited. For instance, in the study of (Christoforou, Cohen, & Karlaftis, 2011) freeway crashes were classified by the number of vehicles involved and the crash type. A binomial probit was built for each dependent variable separately and a multivariate probit model was applied jointly to estimate the probability of having each type of crash. Although the study had the merit of using multivariate probit model, of which dependent variables can be correlated, it had an important disadvantage, having many unknown coefficients to be estimated. Although the structure of the synthetic model was simple, estimating the coefficients could be burdensome. For different regions, different coefficients would be needed since the traffic pattern varies from region to region. Thus, it might be inconvenient for some agencies to put this method into use.

Also, it is worth noticing that most of the models mentioned in this chapter used traffic flow data as the primary input. While it must be admitted that real time traffic flow can reflect the true traffic condition, there are several deficiencies when applying it for crash prediction. The first and most important one is that real time traffic surveillance system is not as prevalent as expected. Realistically, there are many roads where traffic flow data are not available. The second one is that real time traffic flow data are not always accurately reported and sometimes not be reported, which would lead to an inaccurate result. The problem of sensor malfunction has been a trouble for a long time and has not been completely solved. In this sense, real time traffic flow data may not be a reliable resource for crash prediction. The third one is that traffic managers may not be able to capture the risk instantly and may not have enough time to prepare countermeasures if the real time traffic flow data are the main inputs of the crash prediction model.

Based on the above analysis, it can be concluded that a crash type prediction model that can be conveniently implemented and be widely applied in any target region and during any time interval of interest is worthy of exploration.

CHAPTER 3: DATA ANALYSIS

The datasets used for this study are from the Virginia SmarterRoads platform. Three shapefiles are included, (1) Crashes shapefile, which records motor vehicle crashes that involve a fatality, injury, or property damage of over $1500. Figure 1 shows the percentage of crashes in each district of Virginia from 2010 to 2016. It is obvious that North Virginia, Richmond and Hampton Roads districts have much more crashes than the other districts. (2) Average Daily Traffic shapefile, which lists average daily traffic volume of roadway segments. The ADT information is visualized in Figure 2. (3) Speed Limits shapefile, which lists the truck speed limits and the car speed limit of each road segment. The speed limit information is visualized in Figure 3.

Note: a% represents the ratio of the quantity of traffic accidents in this district to the total amount of traffic accidents in Virginia

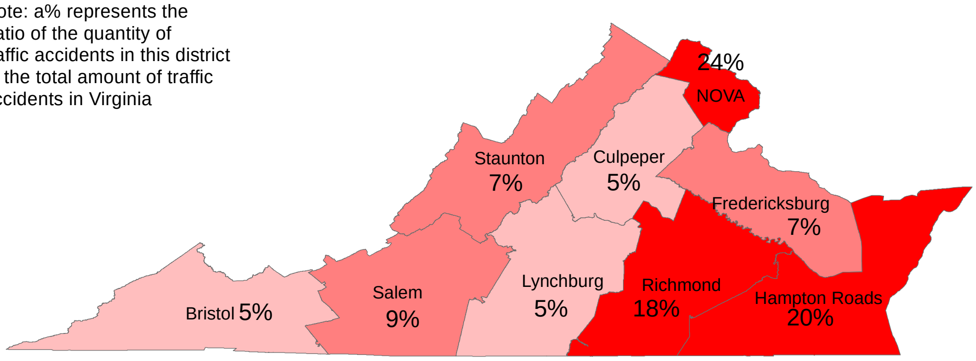


Figure 1: Crash rates

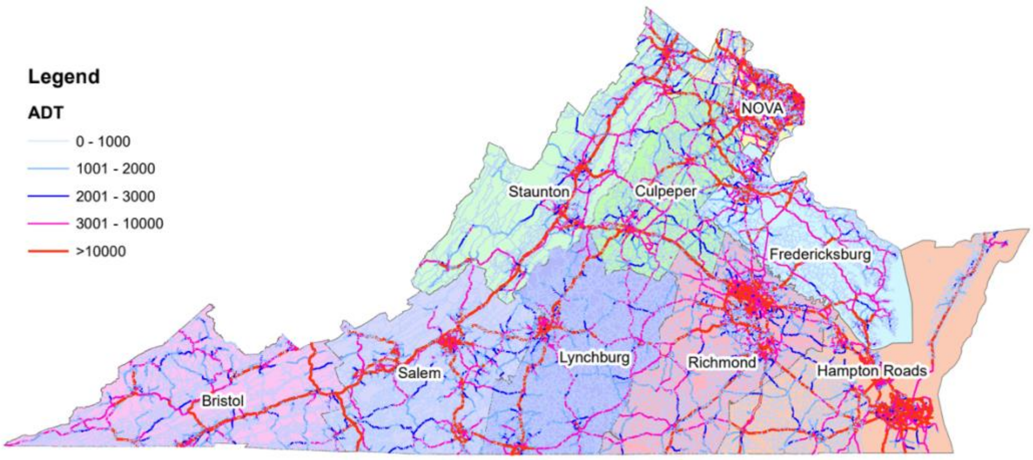


Figure 2: ADT

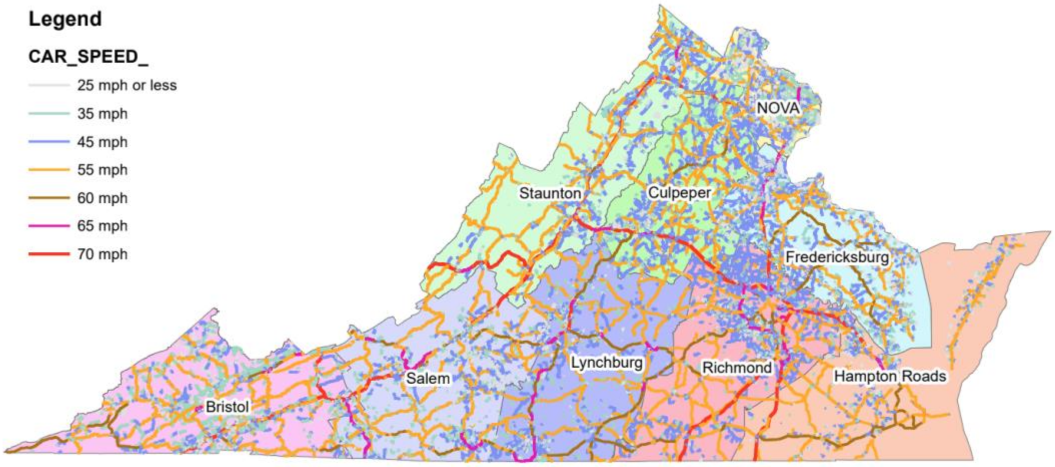


Figure 3: Speed limit

3.1 Crash Type Statistics

Virginia Department of Transportation (VDOT) keeps records of 16 types of crashes. Rear end crash is one of the most frequent traffic accidents and it accounts for more than 1/3 of all the crashes. Angle crash and Fixed object (off road) crash follow closely, each accounting for about 20% of all the crashes. The other types of crashes account for a relatively small percentage, less than 10%, of all the crashes. Severity levels are categorized into 5 classes, namely fatal crash, injury crash, pedestrian fatal crash, pedestrian injury crash and property damage crash. Crashes that involve pedestrians account for a small percentage of all the crashes and are not in the scope of this study. Figure 4 shows the percentage of different types of crashes that caused fatal crash, injury crash and property damage crash.

Among all causes that lead to fatal crashes, fixed object crash shows the highest percentage, accounting for half of the fatal crashes. Angle crashes account for the second largest proportion. As for injury crashes, rear end crashes, angle crashes and fixed object crashes account for a similar proportion while rear end crashes account for a slightly higher proportion. For property damage crashes, the proportion of rear end crashes is the highest, followed by angle crashes and fixed object crashes.

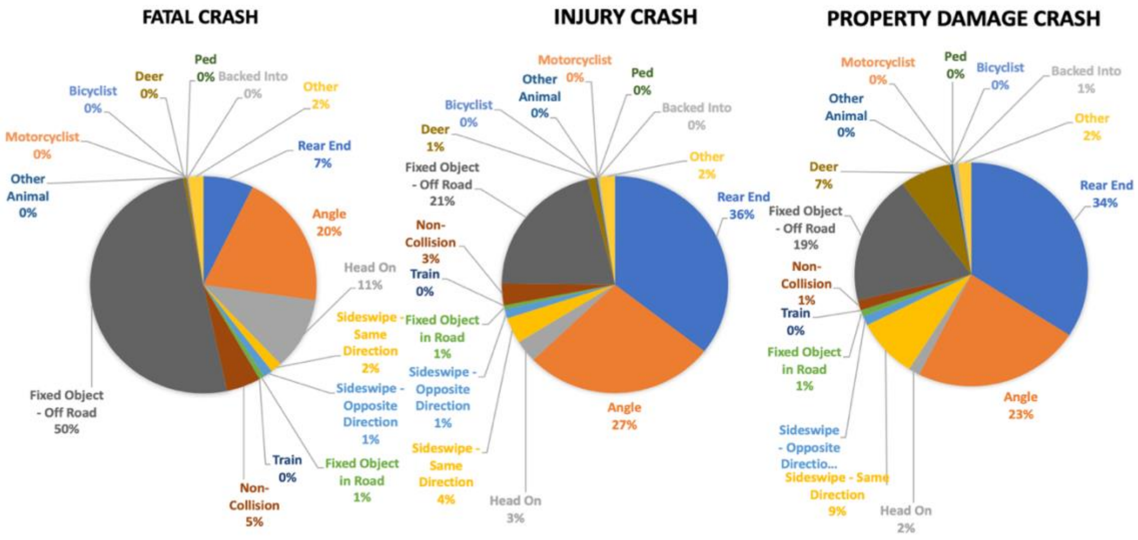


Figure 4: Crash statistics by severity level

It can be concluded that rear end crash, angle crash and fixed object crash are the three major types of crashes. So, this study mainly focuses on these three types of crashes. Figure 5 shows the proportion of different types of crashes in each district. It is worth noting that different districts have different crash patterns. For example, in Northern Virginia, Richmond and Hampton Roads districts, the most frequently happened accidents are rear end crashes. Angle crash ranks the second. In Staunton, Culpeper and Salem, fixed object crash is the most common type of crashes. In Fredericksburg, the most frequently happened accidents are rear end crashes, followed by fixed object crashes and angle crashes. In Lynchburg and Bristol, the most frequently happened crashes are fixed object crashes, while rear end crashes and angle crashes share the same proportion.

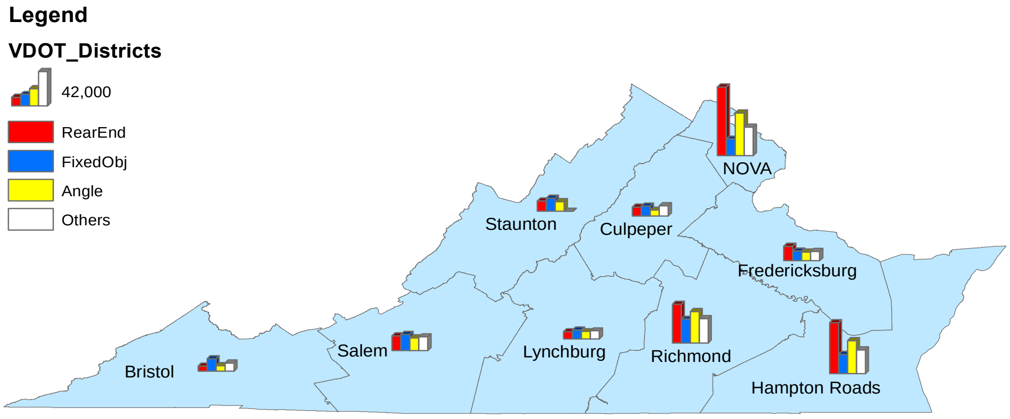


Figure 5: Statistics of major crashes by district

3.2 Risk Factor Analysis

Analyzing crash records is the first step as well as the major method of understanding the characteristics of different types of crashes and evaluating travel safety for the future. While the real time traffic flow, video records and other real time traffic surveillance data may not be available for all the roads, information including time and date, light condition, location, posted speed limit, and road condition are always available in the traffic crash reports. In this section, a risk factor analysis for these 6 factors is performed to provide an insight of interpreting various types of crashes.

3.2.1 Hour Index

Figure 6 illustrates crash counts in different time periods. It is not surprising that the crash counts follow the same pattern as the traffic flow in a typical workday, having two peaks. More vehicles on roads would increase the possibility of having crashes. It can be observed that from 00:00 to 06:00, fixed object crash takes the highest proportion among all crashes. As it gets closer and closer to regular working time, rear end crash becomes the major type of crashes. As time pass by, there are more and more angle crashes. From 21:00 to 24:00, the major type of crashes becomes fixed object crash again.

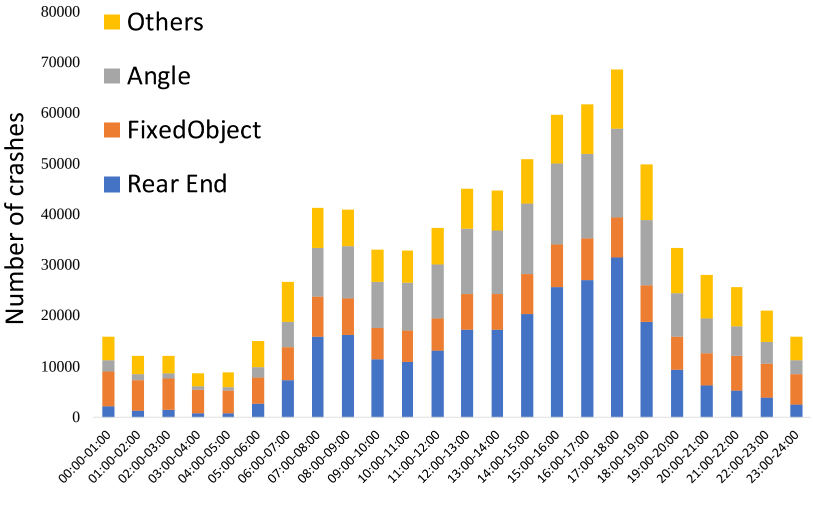


Figure 6: Crash statistics by hour

3.2.2 Light Condition

Light condition would affect drivers' visibility and driving behaviors. For example, when driving at night without a sufficient amount of light, drivers could not see the surrounding objects clearly, and thus it would take more time to react to the roadside objects. Meanwhile, driving in darkness can be tiring and challenging due to the problem of poor visibility. However, when driving during daylight time, some drivers would be overconfident and thus have a low level of alertness, which may lead to a high risk of getting involved in accidents. Figure 7 presents the statistics of light conditions when different types of traffic crash happened.

During dawn time, while the proportion of rear end crashes is the highest among all types of crashes, fixed object crashes also account for a big proportion. A possible reason might be that drivers are in a hurry to work and drive fast. They focus on the traffic ahead without paying attention to the nearby obstacles. During the daylight period, the most common type of accidents is rear end crash. During the time periods when the light condition is good, such as dusk and road lighted time, while rear end is still the most common type of accidents, angle crashes also account for a large proportion. In darkness without road lights, the fixed object crash is the most common type of accidents. This may be caused by the limited range of vision in darkness and drivers' inattention. Drivers may feel tired or even sleepy when driving at night, paying less attention to roadside objects. The lack of stimulations from light and color and the low visibility in darkness make the situation worse. To reduce such accidents, keeping a good light condition, having strict sanctions against fatigue and inattentive driving, and having a vehicle-based night vision system are necessary.

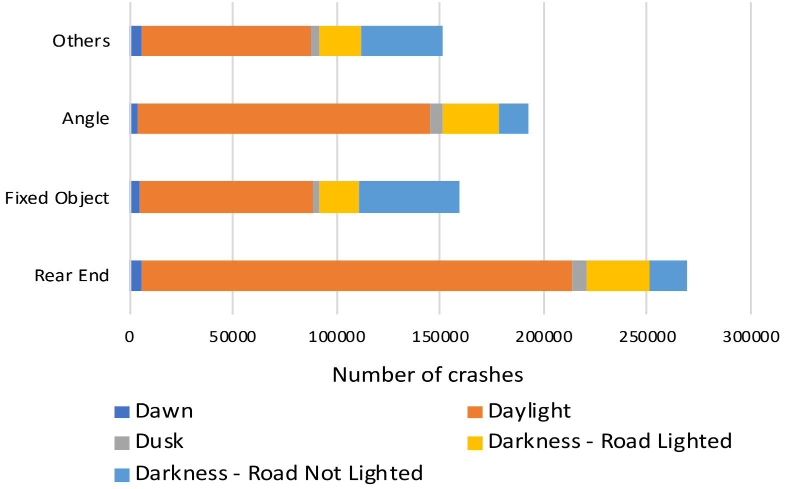


Figure 7: Crash statistics by light condition

3.2.3 Roadway Type

Different roadways are designed based on different design specifications under different budgets to serve for different transportation purposes, so their service levels are different. Figure 8 shows the statistics of different types of accidents that happened on different roadways.

It can be observed that on frontage roads and urban roads, angle crashes are the most frequently occurred accidents. The possible reason for this might be that comparing with the other types of roads, there are more turns, intersections and signal lights on these roads. Meanwhile, these roads are more crowded, having more vehicles within limited spaces. So, the probability of having angle crashes on these roads is higher than on the other roads. To reduce such traffic accidents, having a large buffer area around corners so that drivers have a good view and enough time to respond to other vehicles is important. On interstate highways, state routes and U.S. routes, rear end crashes are the most frequently occurred accidents. The possible reason for this could be that drivers are more likely to drive fast on these roads. Meanwhile, drivers have to drive longer time to reach the exist if they are driving on these roads. So, there is a great possibility of feeling tired and being distracted. To reduce such traffic accidents, having strict sanctions against overspeed, distractions and drunk driving as well as having more rest stations are important. On secondary routes, chances of having rear end crashes, fixed object crashes, angle crashes and other types of crashes are close.

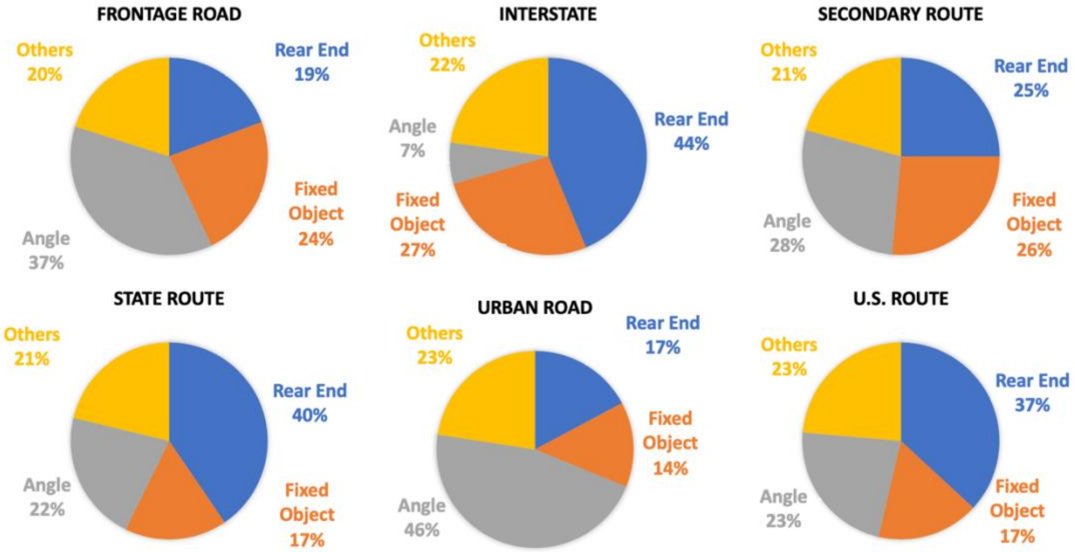


Figure 8: Crash statistics by roadway type

3.2.4 Car Speed Limit

It has been proved by many researchers that driving speed is closely related with traffic accidents. Driving with high speed may lead to hydroplaning and vehicle being out of control. It also takes more time for drivers to react to the nearby vehicles and objects. However, driving with a low speed on roads where drivers are expected to drive with a relative high speed may also cause troubles to the traffic flow. Ideally, the actual driving speed of the accident vehicles and the average speed of the other vehicles around the accident zones should also be analyzed. However, only car speed limits and truck speed limits are provided in the dataset. Car speed limits are used for crash analysis here. Figure 9 shows the statistics of crashes by speed limits. It can be observed that when driving on roads of a relative low speed limit, drivers should take extra care of angle collisions. When driving on roads of a moderate speed limit, drivers should take extra care of rear end collisions. When driving on roads of a high speed limit, drivers should be more careful about fixed object collisions.

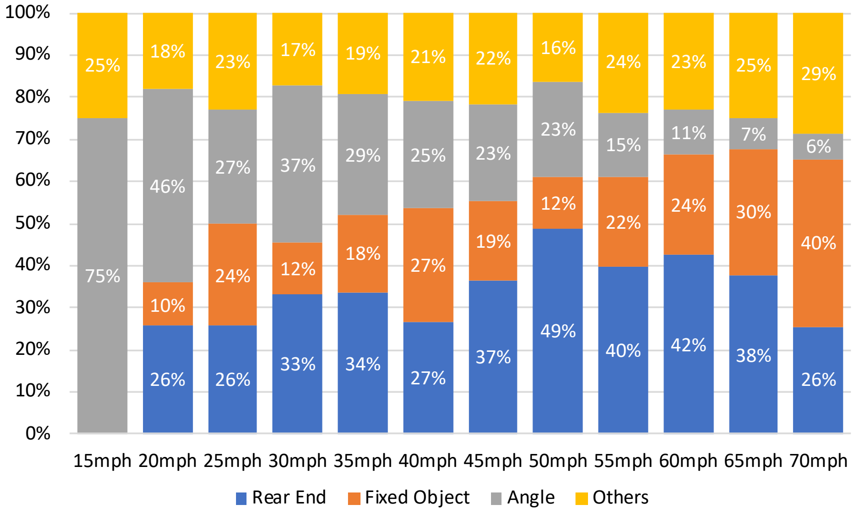


Figure 9: Crash statistics by speed limit

3.2.5 Weather Condition

There have been many studies about the impacts of weather conditions on traffic crashes. However, the investigation of the association between weather conditions and crash types is insufficient. Figure 10 shows the statistics of different types of crashes by weather condition.

It can be observed that in clear or cloudy days, the most frequent crashes are rear end crashes. Avoiding driving overspeed is still the top priority task. In days with fog or blowing particles, drivers are struggling to watch the traffic ahead, paying less attention to the obstacles nearby, so fixed object crashes happen most frequently. Transportation managers may need to restrict the use of roads during the time of such conditions. In mist and rainy days, both rear end and fixed object crashes are prone to happen. In snowy and sleet days, fixed object crashes happen most frequently. A road de-icing system would be crucial to reduce this type of accident. It is also worth noting that most crashes happen when the weather is clear or cloudy. It is surprising but reasonable since people would be less cautious when the weather is good. So, restricting the speed, eliminating distractions, and reminding the drivers to be concentrated would be important.

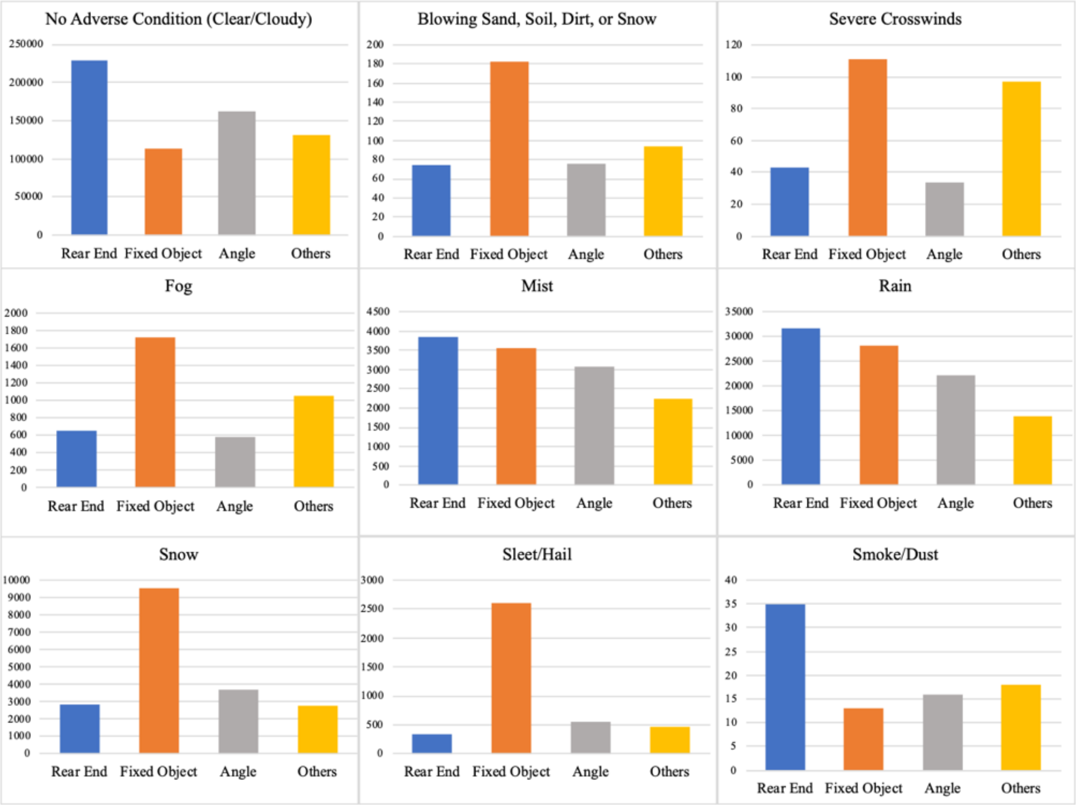


Figure 10: Crash statistics by weather condition

3.2.6 Surface Condition

While weather condition affects driving safety through having impacts on drivers' sight, road surface condition would have a direct impact on the friction between road and tire. When the road surface is dry, more attention should be given to preventing rear end crashes. On wet surfaces, preventions of all types of crashes are all important. When the road surface is icy, de-icing materials should be placed in time. Mud, oil, water and some other liquid materials would also make it difficult to control the direction and stop a car. Pavement cleaning would be essential to prevent fixed object crashes caused by these factors. Statistics of accident type by road condition are summarized in Figure 11.

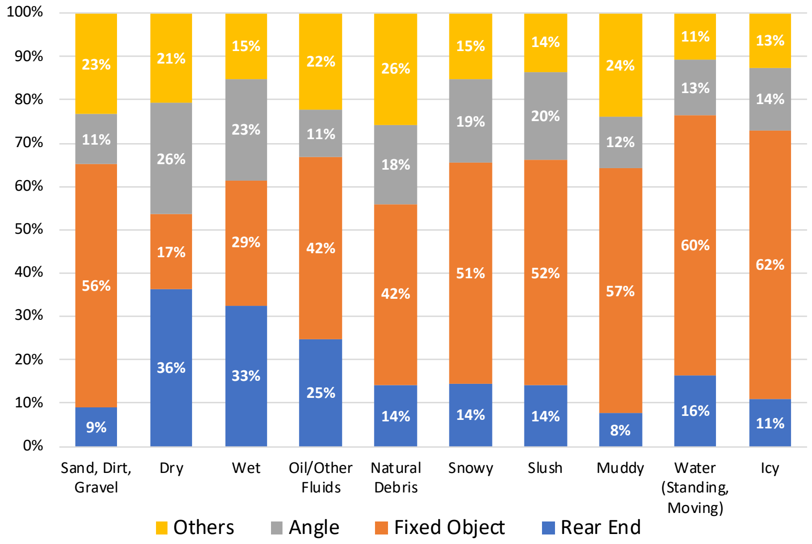


Figure 11: Crash statistics by road condition

CHAPTER 4: CRASH PREDICTION MODEL

4.1 Model Design

A two-layer model is proposed for predicting different types of crashes. The first layer is an XGBoost binary classifier, which is used for distinguishing potential crashes from crash-free observations. The second layer is an XGBoost multiclass classifier and is used for crash type recognition. If the feedback of first layer alarms that there would be a crash, the second layer would be triggered and tell what type of crash is most likely to happen. Sub-models used for the two layers are trained separately and stacked together to be an integrated model. The architecture of the model is shown in Figure 12.

4.2 Data Processing

In addition to the six risk factors that are analyzed in the previous chapter, the dataset of Average Daily Traffic (ADT) is also used to provide the traffic condition information for crash type prediction. The crash records are obtained from the Virginia SmarterRoads platform while the non-crash data records are generated artificially following two rules. The first rule is that the artificial data should follow the distribution as close to the real distribution as possible. The second rule is that the model should be simple and general. The non-crash records are created by the following steps.

First, the distribution of crash records in each hour is calculated. It is an array of 1-by-24, being obtained by dividing the number of crash records in a certain hour by the total number of crashes. Then, the relative non-crash records distribution is obtained by making “unit 1” minus the crash frequency in each time period. Again, an array of 1-by-24 dimension is attained, under the assumption that for each time period, the total probability of having a crash record and having a non-crash record is “unit 1”. Then, the hourly distribution of non-crash records is acquired by dividing each cell of the relative non-crash records distribution array by the sum of all the cells. Finally, an hour index is assigned to each record according to the hourly distribution of non-crash records. To assign the light condition, a cross table is made from crash data in order to find out the relationship between light condition and time (hour index). Then the light condition is assigned to each record according to its hour index and the corresponding light condition distribution. A route type ID is assigned to each record according to the distribution of non-crash records by each type of routes, using the same method for assigning the hour index.

Speed limit attribute was assigned according to the route type. A weather condition index is assigned to each record randomly since all kinds of weather condition would be encountered when applying the model for real world. The surface condition is assigned to each record according to the corresponding weather condition.

The crash records and non-records are concatenated together and shuffled randomly before being split as the training dataset and testing dataset. As a decision tree based algorithm, XGBoost can handle the correlation issue of input variables. It is safe to include all the available variables into the model and let the model figure out which variable are more important. Input variables are listed in Table 1. All the variables are label encoded before being fed into the model.

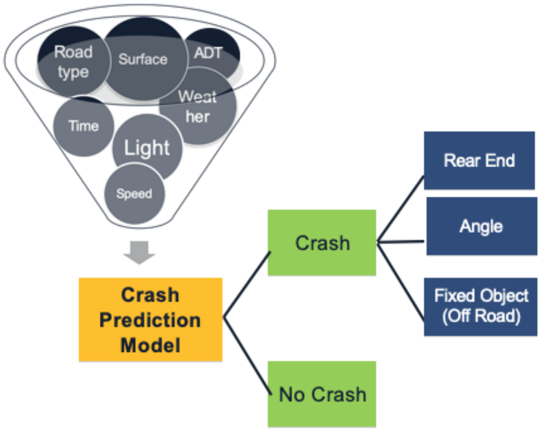


Figure 12: Model structure

Table 1: Input Variables

Variables	Type	Details
Hour	Numerical	24 hours in a day
Light Condition	Categorical (Norminal)	(1) Daylight (2) Dusk (3) Dawn (4) Darkness - Road Not L ighted (5) Darkness - Road Lighted (6) Darkness - Unknown Road Lighting
Route Type	Categorical (Norminal)	(1) Secondary Route (2) U.S. Route (3) State Route (4) Interstate (5) Frontage Route (6) Urban Route
Speed Limit	Numerical	Speed limit of the road segment
ADT	Numerical	ADT of the road segment
Weather	Categorical (Norminal)	(1) Rain (2) Fog (3) Snow (4) Sleet/Hail (5) Mist (6) Smoke/Dust (7) Severe Crosswinds (8) Blowing Sand Soil Dirt, or Snow (9) No Adverse Condition (Clear/Cloudy)
Road Surface Condition	Categorical (Norminal)	(1) Dry (2) Wet (3) Snowy (4) Icy (5) Muddy (6) Oil/Other Fluids (7) Water (Standing, Moving) (8) Slush (9) Sand, Dirt, Gravel

4.3 Extreme gradient boosting - XGBoost

Given a dataset of n samples with m features, the goal of the XGBoost model is to predict the dependent variable y by summing up the values obtained from K decision trees. The final predicted value of the sample x_i can be formulated as

$$\hat{y}_i = \phi(x_i) = \sum_{k=1}^{K} f_k(x_i)\, , f_k \epsilon\ F \tag{1}$$

Where f_k represents a decision tree and $F = \{f(x) = w_{q(x)}\}(q: \mathbb{R}^m \rightarrow T, w \in \mathbb{R}^T)$ represents the space made up of decision trees. T is the total number of leaves in one tree. The structure of each tree is represented by q, and the weights of leaves are represented by w. The final prediction is obtained by summing up the values obtained from the leaf where the sample locates in each tree.

The overall objective is to minimize Equation (2).

$$\mathcal{L}(\phi) = \sum_i l(\hat{y}_i{}^t, y_i) + \sum_k \Omega(f_k) \tag{2}$$

Where $l(\hat{y}_i{}^t, y_i)$ measures the difference between the predicted value and the ground truth, and Ω constrains the model complexity so that the overfitting problem can be avoided. The constraint term consists of two parts, the number of leaves and the weight of each leaf:

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda \|w\|^2 \tag{3}$$

In XGBoost, trees are built in sequence and each tree is built to help overcome the defects existed in the previous trees. The value of the i^{th} sample obtained at the t^{th} iteration is calculated as $\hat{y}_i{}^t = \hat{y}_i{}^{t-1} + f_t(x_i)$. To optimize the objective function, the second-order approximation is used. The objective function represented in Equation (2) is reformulated as

$$\mathcal{L}^{(t)} = \sum_{i=1}^{n} [l(y_i, \hat{y}_i{}^{t-1}) + g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i)] + \Omega(f_t) \tag{4}$$

Where $g_i = \partial_{\hat{y}^{(t-1)}} l(y_i, \hat{y}^{(t-1)})$ and $h_i = \partial^2_{\hat{y}^{(t-1)}} l(y_i, \hat{y}^{(t-1)})$. The constant term $l(y_i, \hat{y}_i{}^{t-1})$ can be removed in the optimization process. By solving the above equations, the optimal weight w^* of leaf j for a fixed tree structure $q(x)$ is obtained:

$$w_j^* = -\frac{\sum_{i \in I_j} g_i}{\sum_{i \in I_j} h_i + \lambda} \tag{5}$$

The optimal value of Equation (4) is then achieved with the constant omitted:

$$\tilde{\mathcal{L}}^{(t)}(q) = -\frac{1}{2}\sum_{i=1}^{T}\frac{(\sum_{i\in I_j} g_i)^2}{\sum_{i\in I_j} h_i + \lambda} + \gamma T \quad (6)$$

Equation (6) is used for evaluating a specified tree structure q. Since enumerating all the possible tree structure is not feasible, a greedy algorithm is used for the branch splitting process. Starting from a single leaf, branches are added in the tree iteratively according to the loss reduction given in Equation (7). A large loss reduction indicates that splitting the node is beneficial.

$$\mathcal{L}_{split} = \frac{1}{2}[\frac{(\sum_{i\in I_L} g_i)^2}{\sum_{i\in I_L} h_i + \lambda} + \frac{(\sum_{i\in I_R} g_i)^2}{\sum_{i\in I_R} h_i + \lambda} - \frac{(\sum_{i\in I} g_i)^2}{\sum_{i\in I} h_i + \lambda}] \quad (7)$$

Where I_L and I_R are the instances of the left node and the right node if choose to branch, while I represents the node before splitting and $I = I_L U I_R$.

The above equations describe the key process of the XGBoost model. There are some other algorithms and techniques that have been adopted in XGBoost to help prevent over-fitting, reduce the computation cost as well as improve the predicting capability (T. Chen & Guestrin, 2016). The model can be used for both regression and classification tasks and has been widely used to solve real-world problems.

4.4 Experiment Design

4.4.1 Crash Detection

For the binary classification problem, the following evaluation metrics are used:

(a) Accuracy, which reflects the percentage of correctly identified samples, either truly positive or truly negative.

$$Average\ accuracy = \frac{tp + tn}{tp + tn + fp + fn} \quad (8)$$

tp is the number of samples that are classified as positive and are real positive.

tn is the number of samples that are classified as negative and are real negative.

fp is the number of samples that are classified as positive but are real negative.

fn is the number of samples that are classified as negative and are real positive.

(b) Recall, which is also called True Positive Rate or Sensitivity, reflects the percentage of samples that are classified as positive correctly out of all true positive samples.

$$Recall = \frac{tp}{tp + fn} \tag{9}$$

(c) Precision, which is the percentage of samples that are correctly identified as positive out of all samples that are identified as positive.

$$Precision = \frac{tp}{tp + fp} \tag{10}$$

(d) F1-score, which is defined as the harmonic mean of the precision and recall

$$F - 1 = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{11}$$

Grid-search technique is applied to find the best hyperparameters combination. For each combination, 3-fold cross validation is implemented and the average score of each evaluation criterion are documented. The hyperparameters tested and the combination that has the best overall performance are listed in Table 2. Evaluation scores of different combinations when using the learning rate 0.05 are visualized in Figure 13.

Table 2: Crash detection model hyperparameter tuning

Hyperparameter	**Range**	**Best Choice**
Learning rate	0.001,0.005,0.01,0.05,0.1	0.05
Max depth	3,4,5,6,7,8	8
Number of estimators	5,10,15,20,50,100	100

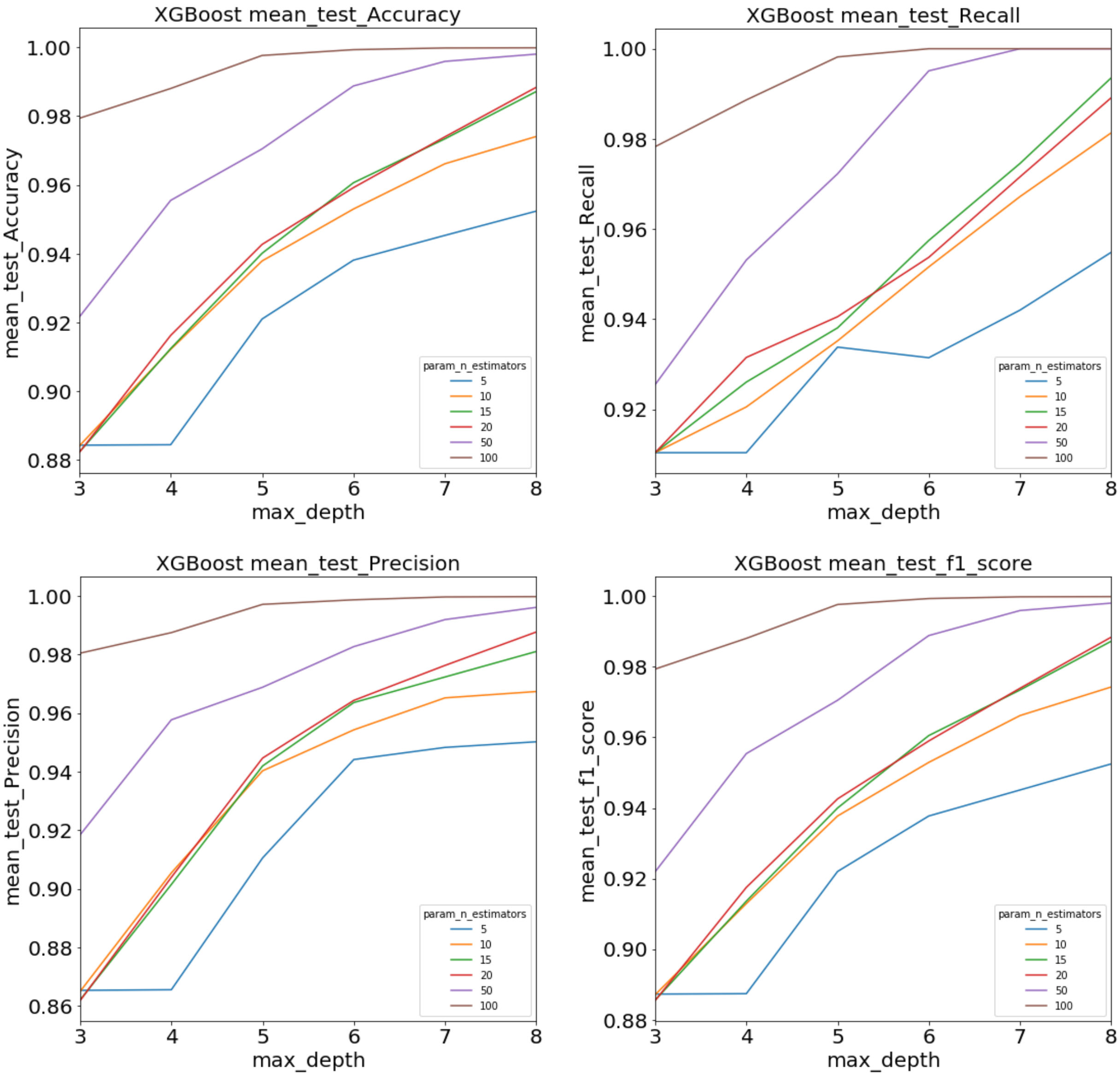


Figure 13: Crash detection model evaluation

4.4.2 Crash Type Identification

For the multi-class classification problem, the following evaluation metrics are used:

(a) Average accuracy of classifying each class

$$Accuracy_M = \frac{\sum_{i=1}^{3} \frac{tp_i + tn_i}{tp_i + tn_i + fp_i + fn_i}}{3} \tag{12}$$

(b) Macro recall, being used to define the capability of the classifier to identify each class

$$Recall_M = \frac{\sum_{i=1}^{3} \frac{tp_i}{tp_i + fn_i}}{3} \tag{13}$$

(c) Macro precision, being used to characterize the average effectiveness of each class

$$Precision_M = \frac{\sum_{i=1}^{3} \frac{tp_i}{tp_i + fp_i}}{3} \tag{14}$$

(d) Macro F1-score, the harmonic mean of the macro precision and the macro recall

$$F - 1_M = \frac{2 \times Precision_M \times Recall_M}{Precision_M + Recall_M} \tag{15}$$

Grid-search and cross-validation techniques are applied again to find the best hyperparameters combination. Since the multi-class classification task is more difficult than the binary classification task, a more complex model would be needed. The hyperparameters tested and the best combination found are listed in Table 3. The model would achieve the best performance using 150 estimators with the max depth as 10, having the learning rate as 0.1. Evaluation scores of different combinations when learning rate is 0.1 are visualized in Figure 14.

Table 3: Crash type identification model hyperparameter tuning

Hyperparameter	**Range**	**Best Choice**
Learning rate	0.01,0.05,0.1	0.1
Max depth	3,5,8,10,15,20,25	10
Number of estimators	10,15,20,50,100,150	150

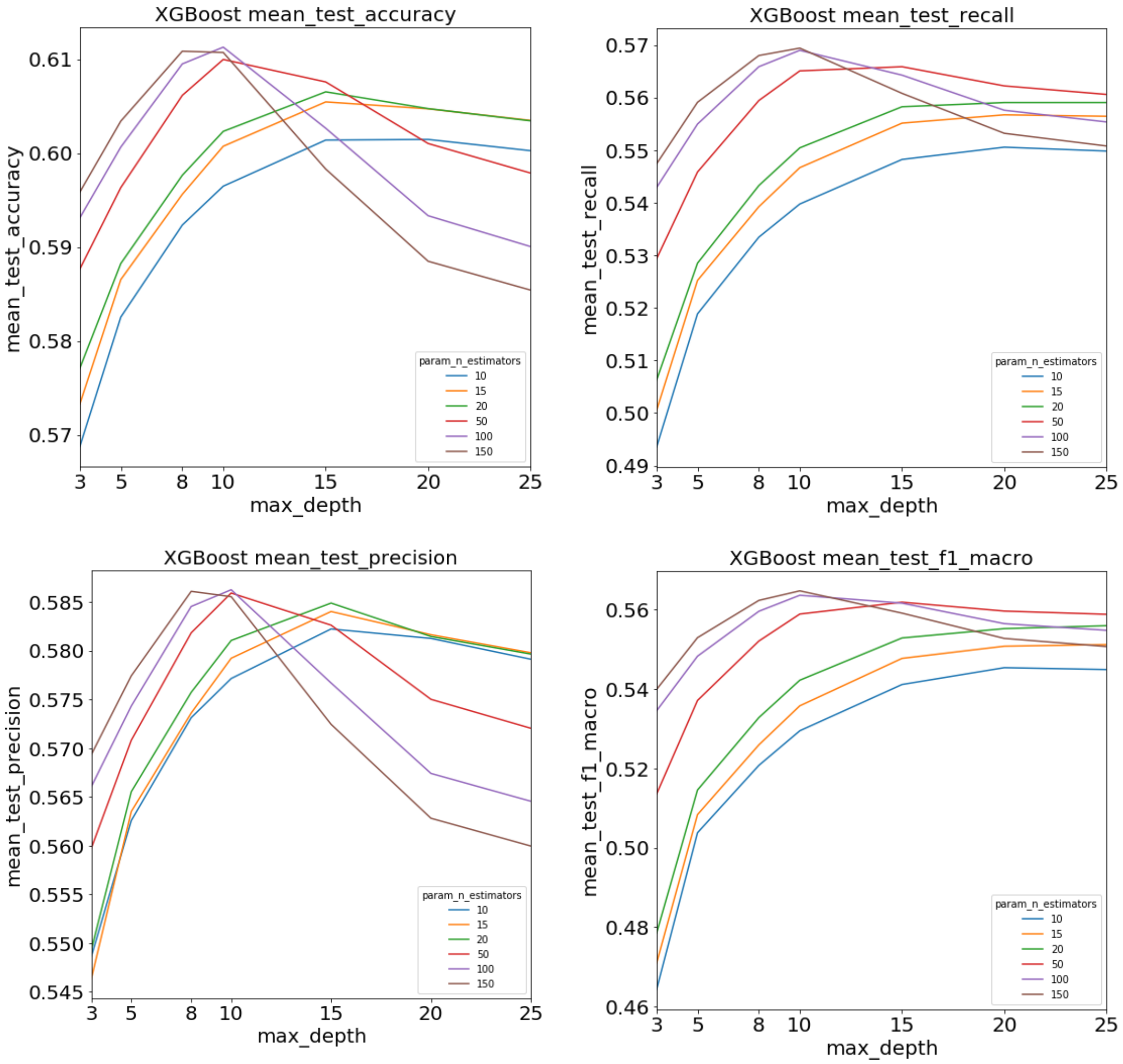


Figure 14: Crash type identification model evaluation

CHAPTER 5: DISCUSSION

5.1 Feature Importance

To better understand the associations between risk factors and traffic crashes, the relative feature importance of the model according to the Gini metric is calculated and shown in Figure 15 and Figure 16.

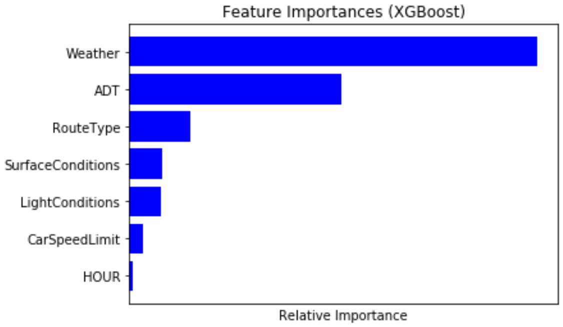


Figure 15: Relative feature importance of the crash detection model

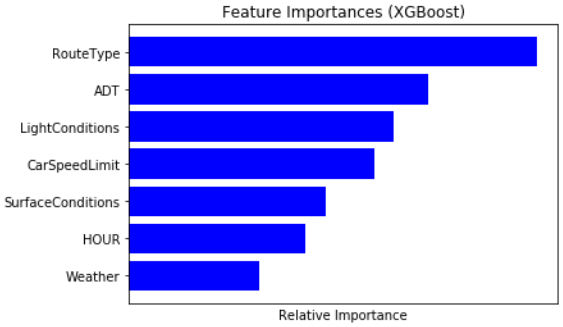


Figure 16: Relative feature importance of the crash type identification model

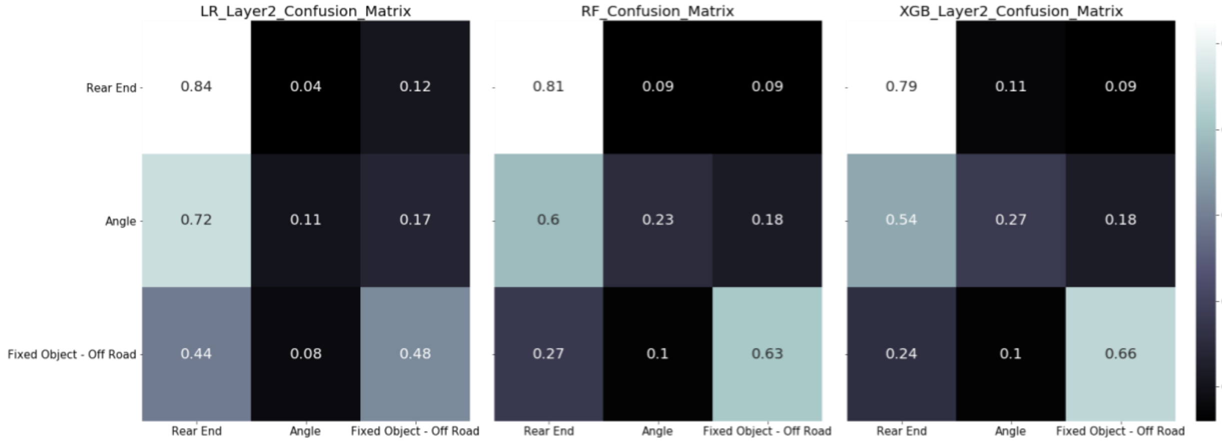
It can be observed that for the crash detection task, weather has the most, followed by the ADT. The other factors make a relatively small contribution. For the crash type identification task, route type plays an important role while the other features also provide much useful information. While it’s important to find the relative contribution of each feature, it’s also important to remember that the accidents are the consequences of interactions between various factors. For example, the study of (Golob & Recker, 2003) showed that rear end crashes are more likely to happen on dry roads during daylight time while fixed object crashes are more likely to happen on wet roads. Thus, when preparing countermeasures for different types of crashes, the combination of factors should be considered.

5.2 Model Comparison

Logistic Regression model and Random Forest model are compared with the XGBoost model. For the crash detection layer, all the three methods achieve good performance, correctly classifying almost all the crash and crash-free samples. For the crash type identification layer, the performance varies from model to model. The performance measurements of the three models are listed in Table 4. XGBoost model achieves the best performance, followed by the Random Forest model. Logistic Regression is inferior to both of the tree-based models, which indicates that the simple model cannot capture the complex associations between non-behavior risk factors and automobile crashes. Although the average accuracy of the XGBoost model and Random Forest model are very close to each other, and both models have room for improvement, it is worth noticing that the XGBoost model can sense the difference between different classes more accurately than the Random Forest model. The confusion matrices of the three models are shown in Figure 17.

Table 4: Model performance comparison

	Measurement Metrics			
	$\mathrm{Accuracy}_{\mathrm{M}}$	$\mathrm{Recall}_{\mathrm{M}}$	$\mathrm{Precision}_{\mathrm{M}}$	$\mathrm{F}-1_{\mathrm{M}}$
Logistic Regression	0.54	0.48	0.51	0.45
Random Forest	0.61	0.56	0.58	0.55
XGBoost	0.62	0.57	0.59	0.57



(a) Logistic Regression (b) Random Forest (c) XGBoost

Figure 17: Confusion matrix of the crash type identification layer

Besides the machine learning models, deep learning method has also been tested for this study, including a simple ANN model as well as the Convolutional-LSTM model proposed by (Shi et al., 2015) and applied by (Z. Yuan et al., 2018) for crash frequency prediction. For the ANN model, the inputs are the same as the inputs of the machine learning model listed in Table 1. For the Convolutional-LSTM model, a spatial grid is created for the study area, each one representing a space of one square mile. The input variables are aggregated into each grid cell. The model has the same structure as the model proposed by (Z. Yuan et al., 2018). However, both of the models do not have satisfactory performance. The reasons for the poor performance are analyzed as follows.

For the ANN model, it generates similar performance as the XGBoost model, with the cost of spending much longer time finding the best model structure and the corresponding parameters. Therefore, the shallow neural network may not be a good choice for the task performed in this study. For the Convolutional-LSTM model, the model performance is much worse than the XGBoost model. The major limitation comes from insufficient data resources. Without sufficient spatial information, the convolutional part of the model couldn’t capture the change in space even though it has the fantastic capability and has been proved to be powerful in many areas. Instead, a lot of information would be lost during the aggregation process. For example, using the average ADT in the grid cell instead of the ADT of each road segment as the input would make the prediction more difficult for the model. Meanwhile, the lack of temporal information makes the LSTM part of the model ineffectual. Detailed information such as visibility and precipitation per hour, real-time traffic flow data as well as pavement design, which have proven to be very useful in many studies related with travel safety, would be necessary to improve the result.

CHAPTER 6: CONCLUSION

The result of this study can be applied to help improve travel safety level by predicting crashes and identifying the most likely type of crash, with an accuracy level of over 99% and 62% respectively. From traffic management perspective, the prediction results can prepare traffic managers for the potential threatens well in advance. Directed and effective countermeasures can be made according to different type of crashes. From travelers' perspective, the prediction results can be used to warn travelers of potential dangers before the trip so that they can make a better trip planning as well as alert them during the trip through the method of message signs and in-vehicle broadcast. All of these actions are important for the proactive traffic management and can help protect people's life and property. The other merit of this study is that the proposed model can be widely applied in any region of interest since the it is built basing on the crash data that are commonly recorded.

However, the model still has room for improvement, such as improving the prediction accuracy and incorporating more conditions of crash types. To improve the accuracy level, more data resources such as land use, demographic, road geometry can be incorporated into the model. These data resources are usually available for each Metropolitan Planning Organization (MPO). As for real time traffic surveillance data, how to incorporate it into the model should be carefully considered. On the one hand, it has to be admitted that real time traffic flow can reflect the traffic conditions of the prediction period more accurate than the ADT data. On the other hand, using real time surveillance records for some roads might be unfair for the other roads since the fact that not all the roads are equipped with real time traffic surveillance devices. In this sense, establishing a comprehensive traffic surveillance system and having well-established crash reporting system would promote the development of proactive traffic management system.

In the future study, the following 5 questions are going to be investigated. The first question is how to get real crash-free data. Comparing with generating the fake crash-free data, obtaining the crash-free data from the real-world records would be more appropriate. For example, the weather information can be obtained from meteorological stations (Theofilatos, Chen, & Antoniou, 2019). The second question is what else risk factors should be included and how to incorporate them into the model. Various kinds of risk factors have been examined in past studies. A summary of the results of these studies is desired. Meanwhile, it is important to stress that the risk factors should be selectively included in the prediction model so that the model can be conveniently implemented and be widely applied. The third question is how to clearly visualize or report the impact of the interactions between different risk factors on the crashes. As mentioned before, there are various kinds of factors that can impact road safety, and the accidents are the consequences of interactions between these factors. Visualizing or reporting the intricate interactions clearly and intuitively is an arduous task but would be a great help for transportation engineers to prepare the countermeasures. The fourth question is that what are the appropriate countermeasures for each specified condition. The last question is that if there is any other effective method that could predict the crashes and identify the crash type more accurately. Summing up, a proactive crash forecast and management system would be essential for travel safety improvement.

REFERENCES

Abdel-Aty, M., & Haleem, K. (2011). Analyzing angle crashes at unsignalized intersections using machine learning techniques. *Accident Analysis and Prevention*, *43*(1), 461–470. https://doi.org/10.1016/j.aap.2010.10.002

Abdel-Aty, M., Uddin, N., Pande, A., Abdalla, M. F., & Hsia, L. (2004). Predicting freeway crashes from loop detector data by matched case-control logistic regression. *Transportation Research Record*, (1897), 88–95. https://doi.org/10.3141/1897-12

Bao, J., Liu, P., & Ukkusuri, S. V. (2019). A spatiotemporal deep learning approach for citywide short-term crash risk prediction with multi-source data. *Accident Analysis and Prevention*, *122*(July 2018), 239–254. https://doi.org/10.1016/j.aap.2018.10.015

Caliendo, C., Guida, M., & Parisi, A. (2007). A crash-prediction model for multilane roads. *Accident Analysis and Prevention*, *39*(4), 657–670. https://doi.org/10.1016/j.aap.2006.10.012

Chang, L. Y. (2005). Analysis of freeway accident frequencies: Negative binomial regression versus artificial neural network. *Safety Science*, *43*(8), 541–557. https://doi.org/10.1016/j.ssci.2005.04.004

Chang, L. Y., & Chen, W. C. (2005). Data mining of tree-based models to analyze freeway accident frequency. *Journal of Safety Research, 36*(4), 365–375. https://doi.org/10.1016/j.jsr.2005.06.013

Chen, Q., Song, X., Yamada, H., & Shibasaki, R. (2016). Learning deep representation from big and heterogeneous data for traffic accident inference. *30th AAAI Conference on Artificial Intelligence, AAAI 2016*, 338–344.

Chen, S., Wang, W., & van Zuylen, H. (2009). Construct support vector machine ensemble to detect traffic incident. *Expert Systems with Applications*, *36*(8), 10976–10986. https://doi.org/10.1016/j.eswa.2009.02.039

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, *13-17-Augu*, 785–794. https://doi.org/10.1145/2939672.2939785

Christoforou, Z., Cohen, S., & Karlaftis, M. G. (2011). Identifying crash type propensity using real-time traffic data on freeways. *Journal of Safety Research*, *42*(1), 43–50. https://doi.org/10.1016/j.jsr.2011.01.001

Dong, N., Huang, H., & Zheng, L. (2015). Support vector machine in crash prediction at the level of traffic analysis zones: Assessing the spatial proximity effects. *Accident Analysis and Prevention, 82*, 192–198. https://doi.org/10.1016/j.aap.2015.05.018

Formosa, N., Quddus, M., Ison, S., Abdel-Aty, M., & Yuan, J. (2020). Predicting real-time traffic conflicts using deep learning. *Accident Analysis and Prevention*, *136*(January). https://doi.org/10.1016/j.aap.2019.105429

Golob, T. F., & Recker, W. W. (2003). Relationships among urban freeway accidents, traffic flow, weather, and lighting conditions. *Journal of Transportation Engineering*, *129*. https://doi.org/110.1061/(ASCE)0733-947X(2003)129:4(342)

Hadi, M. A., Aruldhas, J., Chow, L. F., & Wattleworth, J. A. (1995). Estimating safety effects of cross-section design for various highway types using negative binomial regression. *Transportation Research Record*, (1500), 169–177.

Hossain, M., Abdel-Aty, M., Quddus, M. A., Muromachi, Y., & Sadeek, S. N. (2019). Real-time crash prediction models: State-of-the-art, design pathways and ubiquitous requirements.

Accident Analysis and Prevention, *124*(July 2018), 66–84. https://doi.org/10.1016/j.aap.2018.12.022

Huang, T., Wang, S., & Sharma, A. (2020). Highway crash detection and risk estimation using deep learning. *Accident Analysis and Prevention*, *135*(April 2019), 105392. https://doi.org/10.1016/j.aap.2019.105392

Johansson, P. (1996). Speed limitation and motorway casualties: A time series count data regression approach. *Accident Analysis and Prevention*, *28*(1), 73–87. https://doi.org/10.1016/0001-4575(95)00043-7

Kahn, C. A., & Gotschall, C. S. (2015). The economic and societal impact of motor vehicle crashes, 2010 (Revised). *Annals of Emergency Medicine*, *66*(2), 194–196. https://doi.org/10.1016/j.annemergmed.2015.06.011

Lee, A. H., Stevenson, M. R., Wang, K., & Yau, K. K. W. (2002). Modeling young driver motor vehicle crashes: Data with extra zeros. *Accident Analysis and Prevention*, *34*(4), 515–521. https://doi.org/10.1016/S0001-4575(01)00049-5

Li, P., Abdel-Aty, M., & Yuan, J. (2020). Real-time crash risk prediction on arterials based on LSTM-CNN. *Accident Analysis and Prevention*, *135*(July 2019), 105371. https://doi.org/10.1016/j.aap.2019.105371

Li, X., Lord, D., Zhang, Y., & Xie, Y. (2008). Predicting motor vehicle crashes using Support Vector Machine models. *Accident Analysis and Prevention*, *40*(4), 1611–1618. https://doi.org/10.1016/j.aap.2008.04.010

Lord, D., & Mannering, F. (2010). The statistical analysis of crash-frequency data: A review and assessment of methodological alternatives. *Transportation Research Part A: Policy and Practice*, *44*(5), 291–305. https://doi.org/10.1016/j.tra.2010.02.001

Lv, Y., Tang, S., & Zhao, H. (2009). Real-time highway traffic accident prediction based on the k-nearest neighbor method. *2009 International Conference on Measuring Technology and Mechatronics Automation, ICMTMA 2009*, *3*, 547–550. https://doi.org/10.1109/ICMTMA.2009.657

Mannering, F. L., & Bhat, C. R. (2014). Analytic methods in accident research: Methodological frontier and future directions. *Analytic Methods in Accident Research*, *1*, 1–22. https://doi.org/10.1016/j.amar.2013.09.001

Miaou, S. P. (1994). The relationship between truck accidents and geometric design of road sections: Poisson versus negative binomial regressions. *Accident Analysis and Prevention*, *26*(4), 471–482. https://doi.org/10.1016/0001-4575(94)90038-8

Mokoatle, M., Marivate, V., & Esiefarienrhe, M. (2019). Predicting road traffic accident severity using accident report data in South Africa. *ACM International Conference Proceeding Series*, 11–17. https://doi.org/10.1145/3325112.3325211

Pande, A., & Abdel-Aty, M. (2006). Comprehensive analysis of the relationship between real-time traffic surveillance data and rear-end crashes on freeways. *Transportation Research Record*, (1953), 31–40. https://doi.org/10.3141/1953-04

Parsa, A. B., Movahedi, A., Taghipour, H., Derrible, S., & Mohammadian, A. (Kouros). (2020). Toward safer highways, application of XGBoost and SHAP for real-time accident detection and feature analysis. *Accident Analysis and Prevention*, *136*(December 2019), 105405. https://doi.org/10.1016/j.aap.2019.105405

Poch, M., & Mannering, F. (1996). Negative binomial analysis of intersection-accident frequencies. *Journal of Transportation Engineering*, Vol. 122, pp. 105–113. https://doi.org/10.1061/(ASCE)0733-947X(1996)122:2(105)

Ren, H., Song, Y., Liu, J., Hu, Y., & Lei, J. (2017). A Deep Learning Approach to the Citywide Traffic Accident Risk Prediction. *21st International Conference on Intelligent Transportation Systems (ITSC)*, 3346–3351. Retrieved from http://arxiv.org/abs/1710.09543

Schlögl, M., Stütz, R., Laaha, G., & Melcher, M. (2019). A comparison of statistical learning methods for deriving determining factors of accident occurrence from an imbalanced high resolution dataset. *Accident Analysis and Prevention*, *127*(March), 134–149. https://doi.org/10.1016/j.aap.2019.02.008

Shan, L., Yang, Z., Zhang, H., Shi, R., & Kuang, L. (2019). Predicting duration of traffic accidents based on ensemble learning. *Lecture Notes of the Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering, LNICST*, *268*, 252–266. https://doi.org/10.1007/978-3-030-12981-1_18

Shankar, V., Milton, J., & Mannering, F. (1997). Modeling accident frequencies as zero-altered probability processes: An empirical inquiry. *Accident Analysis and Prevention*, *29*(6), 829–837. https://doi.org/10.1016/S0001-4575(97)00052-3

Shankar, V. N., Albin, R. B., Milton, J. C., & Mannering, F. L. (1998). Evaluating median crossover likelihoods with clustered accident counts an empirical inquiry using the random effects negative binomial model. *Transportation Research Record*, (1635), 44–48. https://doi.org/10.3141/1635-06

Shi, X., Chen, Z., Wang, H., Yeung, D. Y., Wong, W. K., & Woo, W. C. (2015). Convolutional LSTM network: A machine learning approach for precipitation nowcasting. *Advances in Neural Information Processing Systems*, *2015-Janua*, 802–810.

Theofilatos, A., Chen, C., & Antoniou, C. (2019). Comparing Machine Learning and Deep Learning Methods for Real-Time Crash Prediction. *Transportation Research Record*, *2673*(8), 169–178. https://doi.org/10.1177/0361198119841571

Wenqi, L., Dongyu, L., & Menghua, Y. (2017). A model of traffic accident prediction based on convolutional neural network. *2017 2nd IEEE International Conference on Intelligent Transportation Engineering, ICITE 2017*, 198–202. https://doi.org/10.1109/ICITE.2017.8056908

Xie, Y., Lord, D., & Zhang, Y. (2007). Predicting motor vehicle collisions using Bayesian neural network models: An empirical analysis. *Accident Analysis and Prevention*, *39*(5), 922–933. https://doi.org/10.1016/j.aap.2006.12.014

Yu, R., & Abdel-Aty, M. (2013). Utilizing support vector machine in real-time crash risk evaluation. *Accident Analysis and Prevention*, *51*, 252–259. https://doi.org/10.1016/j.aap.2012.11.027

Yuan, J., Abdel-Aty, M., Gong, Y., & Cai, Q. (2019). Real-Time Crash Risk Prediction using Long Short-Term Memory Recurrent Neural Network. *Transportation Research Record*, *2673*(4), 314–326. https://doi.org/10.1177/0361198119840611

Yuan, Z., Zhou, X., & Yang, T. (2018). Hetero-ConvLSTM: A deep learning approach to traffic accident prediction on heterogeneous spatio-temporal data. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 984–992. https://doi.org/10.1145/3219819.3219922