BUILDING EXPLANATORY MODELS FOR ROAD CRASH ANALYSIS USING DATA SCIENCE AND MACHINE LEARNING TECHNOLOGIES

H. W. I. U. De Silva

(199601A)

M.Sc. in Transportation

Department of Civil Engineering

University of Moratuwa
Sri Lanka

July 2022

BUILDING EXPLANATORY MODELS FOR ROAD CRASH ANALYSIS USING DATA SCIENCE AND MACHINE LEARNING TECHNOLOGIES

H. W. I. U. De Silva

(199601A)

Thesis/Dissertation submitted in partial fulfillment of the requirements for the M.Sc. in Transportation

Department of Civil Engineering

University of Moratuwa
Sri Lanka

July 2022

DECLARATION OF THE CANDIDATE AND SUPERVISOR

I declare that this is my own work, and this thesis/dissertation does not incorporate without acknowledgment any material previously submitted for a Degree or Diploma in any other University or institute of higher learning, and to the best of my knowledge and belief, it does not contain any material previously published or written by another person except where the acknowledgment is made in the text.

Also, I hereby grant to the University of Moratuwa the non-exclusive right to reproduce and distribute my thesis/dissertation, in whole or in part in print, electronic or other medium. I retain the right to use this content in whole or part in future works (such as articles or books).

Signature: Date:

The above candidate has carried out research for the Masters Dissertation under my supervision.

Name of the Supervisor: **Dr. Loshaka Perera**

Signature of the Supervisor: Date:

ACKNOWLEDGMENT

I wish to extend my sincere gratitude to the supervisor Dr. Loshaka Perera for his guidance, advice, and continued encouragement to complete a successful thesis. His assistance in obtaining the necessary data sets for the analysis was crucial for the effectiveness of the research.

I also thank Prof. J. M. S. J. Bandara and Dr. Dimantha De Silva for all the feedback and guidance provided during the progress evaluation of the research. My sincere appreciation is also extended to all other permanent and visiting faculty staff of the University of Moratuwa for the knowledge and skills imparted throughout the past two years.

The fellow batchmates of the MSc/MEng program also deserve a note of thanks for the support and feedback extended during the research study and group assignments.

Last, but certainly not least, I extend a warm sense of gratitude to my loving wife Patali and the caring daughter Hesara, for accommodating my absence from family time and supporting my studies, without which this research work wouldn't have been a reality!

ABSTRACT

Over three thousand people die annually on the roads of Sri Lanka due to traffic crashes. This is a massive socio and economic problem faced by the country. Road crashes globally cause more than 1.3 million fatalities every year and are the eighth leading cause of death worldwide.

Traditionally, road traffic crash analysis and accident modeling resorted to regression models and discrete choice models based on past data. Many countermeasures have been identified and implemented addressing the issues highlighted through such models.

Since road traffic crashes occur across space and time, the conventional numerical approaches have failed to provide alerts and insights in relation to geospatial regions. Also, having to handcraft these models limits the explainability that can be leveraged with the help of advanced tools and techniques available in modern data science and machine learning disciplines.

Further, the disjointed efforts in building analytical models or geospatial models on available crash data (e.g., crash hotspot identification) limit road agencies' abilities in prioritizing funds allocation for more impactful improvements. Due to the difficulty in identifying patterns in causal factors of accident risks using conventional or isolated methods, the authorities also find it difficult to prioritize their staff strength in high-risk areas.

The combination of exploratory data analysis (EDA), machine learning models, and modern geospatial visualization tools offer a unique opportunity to fill these gaps cost-effectively. This study presents an application of the latest data science and machine learning technologies to build explanatory models that help analyze road crashes. Popular packages written in Python and Javascript programming languages were used. **Pandas** and **SweetViz** libraries provided simple, yet powerful EDA. **GeoPandas** library provided the ability to process GPS locations (latitude and longitude) while **Matplotlib** was used to generate static maps. **Folium** library and the underlying **Leaflet.js** library were applied to generate interactive maps to help visualize crash hot spots. Two leading gradient boosting techniques, namely **LightGBM** and **Catboost** were applied to build models that highlight causal factors via feature importance estimation methods.

The study developed algorithms, methods, and charts to generate attribute correlation and gradient boosted decision tree models to relate accident severity with recorded data sets and interactions of certain aggregate features (e.g., weather, and light condition). The visualization efforts produced road crash density maps by administrative region size and population.

Interactive maps that allow authorities to drill down (or zoom in) to hot spots were also developed.

The programmatic approach developed in this study enables the repeatable application of the explanatory analysis and visualizations to new and old datasets with minimal effort. The findings from the study lay the foundation for a digital system that can be easily converted to an online platform for road and enforcement agencies to obtain reports and alerts on road crash risks and hot spots. The application was tested using crash data in Sri Lanka and the outcomes are presented in this study.

Future work on the fusion of multiple data sources such as real-time weather data and traffic congestion levels onto the same platform can enhance these outcomes to even near real-time crash prediction to further assist proactive accident prevention measures.

Keywords: road safety, road crashes, exploratory data analysis, machine learning crash models, explanatory models, geospatial crash visualization, multi-faceted analysis

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

Abbreviation	Description
AI	Artificial Intelligence
EDA	Exploratory Data Analysis
GDP	Gross Domestic Product
GPS	Global Positioning System
GRSF	Global Road Safety Facility
HAMS	Highway Asset Management System
LightGBM	Light Gradient Boosting Method
ML	Machine Learning
SHAP	Shapley Additive exPlanations
XAI	Explainable AI

CHAPTER 1

1 INTRODUCTION

Deaths caused by road traffic crashes are a global problem and account for more than 1.3 million fatalities annually worldwide. They are also the eighth leading cause of death and the leading cause of death for young adults aged between 5 and 29 years ("Global Status Report on Road Safety 2018," 2018).

Fatalities and severe injuries caused by road crashes in Sri Lanka are also major national problems. According to the statistics published by the National Road Safety Council of Sri Lanka, there have been over 3,000 deaths on average per annum between 2016 and 2019 (*National Council for Road Safety, Sri Lanka*, 2020).

In addition to the loss of lives and the grievances faced by the families of the deceased individuals and the hardships caused by severe injuries, the cost of road crashes is a massive burden on a country's economy. The Global Road Safety Facility (GRSF) of the World Bank estimates that road crashes drain out as much as 4.9% of Sri Lanka's GDP each year (*Sri Lanka's Road Safety Country Profile*, 2021). In a statistical analysis of road crash data, economic indices, and per capita health expenditure in Sri Lanka between 1977 and 2016, Thangamani (2019) found that every 1% increase in the fatality index[1] reduces the economic growth rate of the country by 0.79%. The same study also concluded that the health expenditure per capita increased by 0.87% for every 1% increase in the total casualties (Thangamani, 2019).

Given the enormous social and economic impact of road crashes, it is essential to identify and implement effective measures to control and minimize the extent of this problem. The use of explanatory models with statistical and mathematical rigor to understand the causal factors of crashes is an important prerequisite for such control measures. Road agencies and research scholars worldwide have developed various crash models to analyze road crashes.

A crash model typically tries to establish the relationship between crash frequency or severity and the causal factors such as vehicle factors (type of vehicle, safety features), human factors (influence of alcohol, use of safety features such as helmets), and road and environment factors (road condition, weather, light condition). Approaches such as mathematical models (Asalor, 1984), various regression models, discrete choice models, and artificial intelligence (AI) based

[1] Fatality Index = (Total Fatalities / Total Casualties X 100) (Thangamani, 2019)

techniques have been employed in developing crash models (Abdulhafedh, 2017). In an extensive review of crash modeling approaches, Savolainen et al reported that researchers have used binary and multivariate discrete choice models, AI techniques such as neural networks, and data mining techniques in previous studies (Savolainen et al., 2011).

The statistical and mathematical models provide a solid foundation for the scientific understanding of road crashes. The numerical nature of such models could be complemented by the powerful geospatial analysis tools and data visualization technologies that have advanced rapidly with the big data revolution. These new developments offer visualization that improves the efficacy of human experts (Shneiderman, 2014).

Further, two recent phenomena, namely the exponential growth of available computation power and the abundance of data to train models, caused this rapid advancement in the AI and machine-learning fields (Haenlein & Kaplan, 2019; Singh, 2017).

The combination of powerful machine-learning models and advanced data visualization technologies provides a good platform to build crash models to expand the available methodologies to study the road crash problem.

The objective of this research is to combine the theocratical knowledge from the field of transportation studies and the power of modern data science and machine learning technologies to build programmatic explanatory models. The focus is not necessarily about building a particular model but is more about experimenting and proposing a new approach to crash data analysis. The proposition is to look beyond a single approach such as a statistical model, a machine learning model, or a geospatial analysis model. Instead, the author developed a wholistic approach that leverages all these approaches with the power of modern computing and programming. The key benefit of this proposed integrated approach is the ability to develop multi-dimensional analysis of crash data. The combination of crash data with other sources such as road geometry and weather data also give the ability to extend the models to support predictions of high-risk locations enabling authorities to take proactive action. The computer programmatic approach enables these to be repeatedly applied to new data sets with minimal changes in the model. The computer science applications are based on leading free and open-source technologies in the industry thus making the approach cost-effective and affordable.

The multi-faceted study using exploratory data analysis (EDA), machine-learning models and geospatial analysis forms the core outcome as a foundation for a comprehensive road safety management system.

Chapter 2 provides a detailed literature survey on past and present crash models including the traditional numerical models and more recent AI models. A discussion on exploratory data analysis and some previous applications of geospatial modeling on crash analysis is also included. The methodology used and the new approach developed in the research including the open-source technologies and tools is discussed in Chapter 3.

Chapter 4 describes the data set used.

The feasibility of the unified approach as established from the study is presented in Chapter 5. A high-level proposal for an integrated system based on the unified approach and sample results covering the visualization charts, EDA findings, and feature importance obtained from the machine learning models are also presented. Some techniques that were tried but proved less effective in terms of explainability are also discussed under this section.

The conclusions and future recommendations of the research are presented in Chapter 6.

CHAPTER 2

2 LITERATURE REVIEW

The research focused on building explanatory models for road crash analysis using data science and machine learning technologies. The data science aspect expanded on exploratory and statistical analysis as well as geospatial data visualization. Therefore, the literature survey involved explorations on crash models, data science techniques related to data analysis, and geospatial data visualization. The following sub sections present each of these aspects.

2.1 Crash Models

Crash models attempt to understand the causal factors of traffic crashes using various statistical techniques. They can be categorized based on the type of technique used to build the model or based on the outcome expected from the model.

The type of model can be broadly categorized into a) traditional approaches and b) artificial intelligence (AI) and machine learning (ML) driven approaches. The traditional approaches have been typically either a general mathematical model, a regression model, or a discrete choice model (Abdulhafedh, 2017; Perera & Dissanayake, 2012; Savolainen et al., 2011). Figure 2-1 illustrates this categorization.

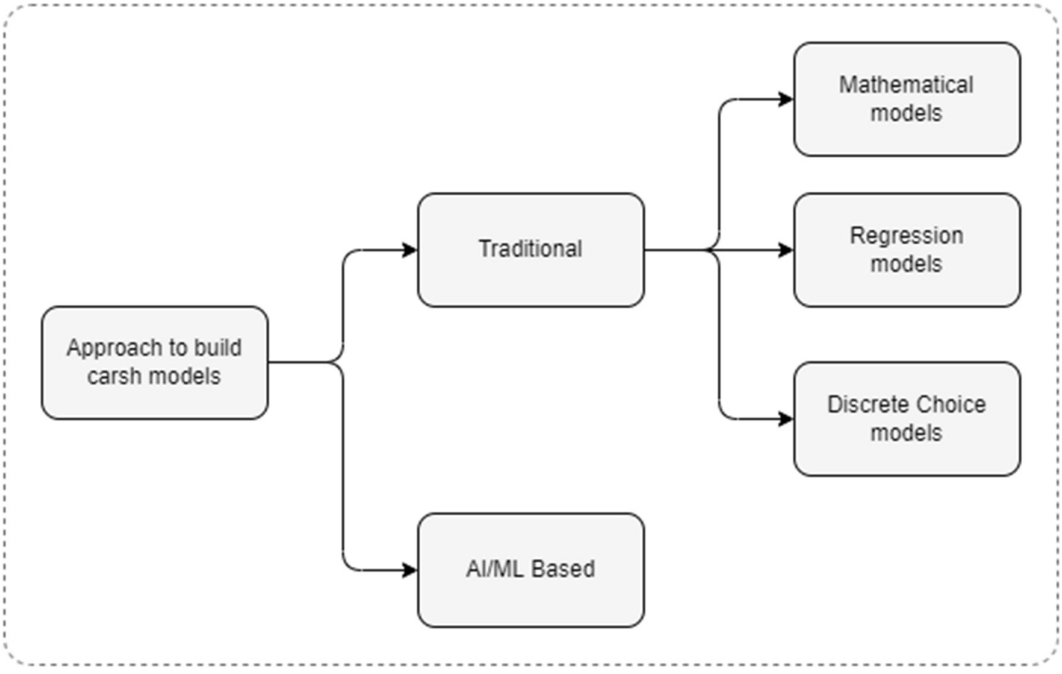


Figure 2-1Types of techniques used for crash models

Mehdizadeh et al. categorized the outcome expected from a crash model into either "a) predictive or explanatory models that attempt to understand and quantify crash risk based on different driving conditions, and (b) optimization techniques that focus on minimizing crash risk through route/path-selection and rest-break scheduling" (Mehdizadeh et al., 2020).

2.1.1 Traditional Crash Models

Asalor developed a generic mathematical model for road crashes. The model considered a road traffic system to be comprised of two primary subsystems, an engineering subsystem (the environment and the vehicle excluding the driver) and a human subsystem (e.g., drivers, pedestrians) (Figure 2-2). The model considered a road crash as a failure of one of these subsystems (Asalor, 1984).

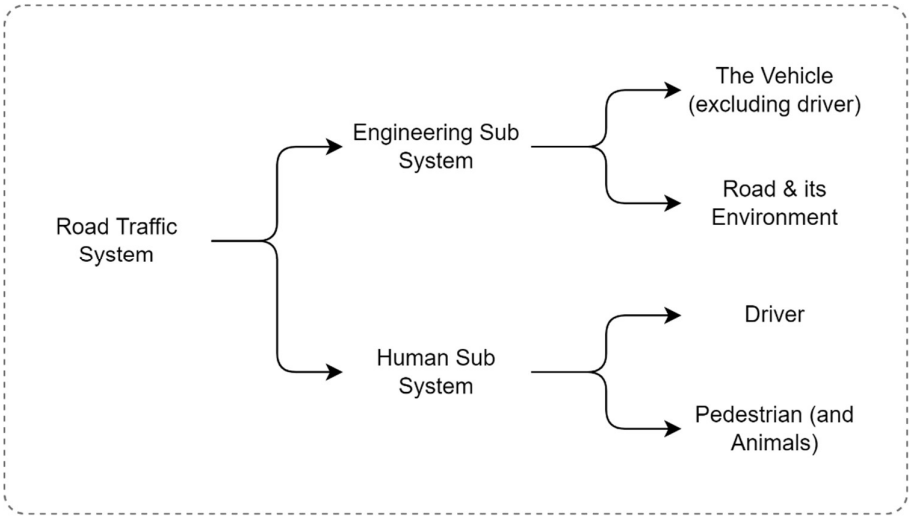


Figure 2-2 Parts of a road traffic system (based on Asalor, 1984)

Later, the researchers developed more elaborate regression-based crash models.

Negative binomial regression (M. A. Abdel-Aty & Radwan, 2000), linear and Poisson regression (Joshua & Garber, 1990), mix of Poisson regression and negative binomial regression (Dissanayake & Amarasingha, 2012) are some examples of regression-based models.

Multiple linear regression models took the general form as shown in equation (2-1), while the Poisson regression models were of the general form as in equation (2-2).

$$y = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_k x_k \quad (2\text{-}1)$$

$$\ln(y) = b_1 x_1 + b_2 x_2 + \cdots + b_k x_k \quad (2\text{-}2)$$

Aty & Radwan stated that researchers attempted three techniques in regression modeling, namely, the use of multiple linear regression, Poisson regression, and negative binomial regression**.** They built a negative binomial regression model to analyze the accidents on a principal arterial in Central Florida (M. A. Abdel-Aty & Radwan, 2000).

Joshua and Garber (Joshua & Garber, 1990) used both linear and Poisson regression models to analyze the relationship between truck crashes and road geometry. Dissanayake and Amarasingha (Dissanayake & Amarasingha, 2012) used Poisson regression and negative binomial regression models for a similar study on truck crashes.

Further variations on regression models such as Poisson-lognormal, zero-inflated Poisson, and zero-inflated negative binomial to overcome the limitations in earlier models have also been developed (Abdulhafedh, 2017).

Discrete choice models such as *logit* and *probit* models and their variants were also commonly applied in accident analysis. Discrete choice models estimate the probabilities of a *binary dependent variable*, for example, *y*, which would take values 0 and 1. This estimation is done using a non-linear function of the independent variables. In the case of logit models, the non-linear function is the *cumulative density function* of the normal distribution (equation 2-3). The probit model uses the *logistic function* as the non-linear function (equation 2-4) (Katchova, 2020).

$$P(y = 1) = \phi(x\beta) = \int_{-\infty}^{x\beta} \phi(z)\, dz \quad (2\text{-}3)$$

$$P(y = 1) = G(x\beta) = \frac{e^{x\beta}}{1 + e^{x\beta}} \quad (2\text{-}4)$$

Tay & Rifaat (2007) used an *ordinal probit model* to study the factors leading to severity of intersection crashes. Abdel-Aty (2003) developed an *ordered probit model* incorporating driver, roadway, and traffic characteristics to model crashes in roadway sections, signalized intersections, and toll plazas.

In a review of 92 journal papers, Savolainen et al (2011) found that the discrete choice models used in crash modeling broadly belonged to three categories, namely, binary outcome models, ordered discrete outcome models and unordered, multinomial discrete outcome models.

Each of these classes of discrete choice models is further sub divided into numerous sub classes depending on factors such as bivariate vs multivariate models, logit vs probit models and so on. Figure 2-3 is an illustration of the extent to which each type of model is used in the papers reviewed by Savolainen et al. The height of a bar is proportional to the frequency of use of that model type in the research reviewed.

These regression and discrete choice models have used multiple sources of data and methods in the modeling process. For example, Dissanayake and Perera (2011) used a survey methodology to identify reasons that are not typically exposed from crash data. Devasurendra et al (Devasurendra et al., 2016) used vehicle registration and population data in addition to crash data. Crash data, highway geometry data, weather data, and traffic data are the most frequently used types of data in crash models. (M. A. Abdel-Aty & Radwan, 2000)(Jiménez-Mejías et al., 2016)(Roland et al., 2021)

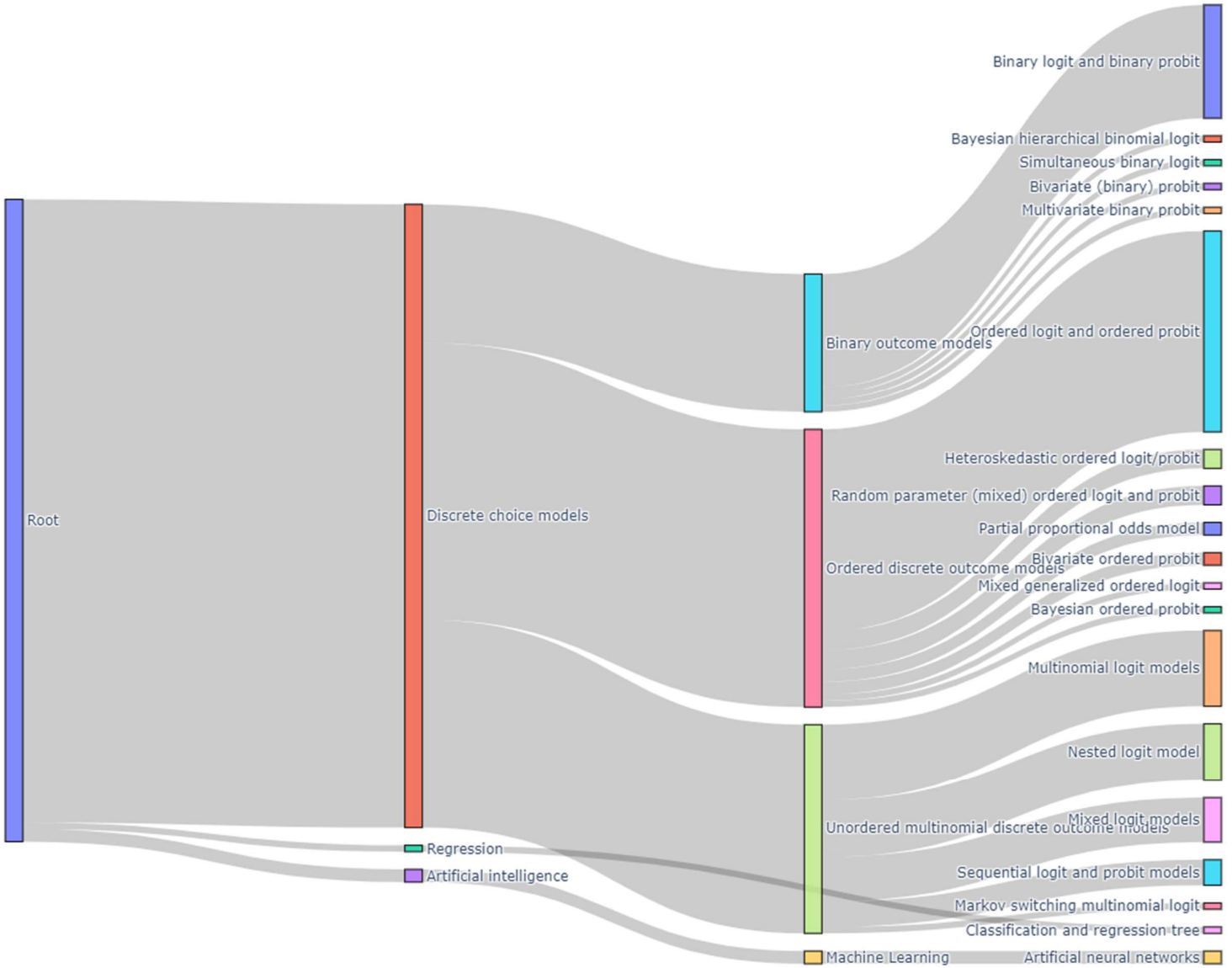


Figure 2-3 Types of discrete choice models (based on Savolainen et al, 2011)

2.1.2 Artificial Intelligence (AI) and Machine Learning (ML) based Models

The field of machine learning has seen exponential growth in complexity in the past two decades. The advances in deep learning, a specialized form of machine learning, helped a computer program (AlphaGo) beat the world champion in Go in 2016 (Silver et al., 2016). Go has a state space of 10^{172} combinations compared to 10^{50} in Chess. Also, the game tree size in Go is 10^{360} while that of Chess is 10^{123} (Yee & Alvarado, 2012). Therefore, Go is considered the most complex board game developed by humans, and AlphaGo's victory thus is evidence of the remarkable progress of machine learning techniques.

Two recent phenomena, namely the exponential growth of available computation power and the abundance of data to train models, caused this rapid advancement in the AI and machine-learning fields (Singh, 2017)(Haenlein & Kaplan, 2019).

The ML based techniques take a different modeling approach.

In traditional modeling, the function (or the model) that relates the independent variables to the dependent variables is crafted by the modeler. However, in the ML approach, that function is figured out by the machine. Equation 2-5 depicts this conceptually (Rokach, 2019).

$$y = f(?) \tag{2-5}$$

Figure 2-4 further elaborates on this difference between the traditional and ML modeling approach. One of the advantages of the ML approach also stems from the auto discovery of the translation function.

Paredes et al (Paredes et al., 2017) found that machine learning models outperform discrete choice models in prediction problems when proper feature engineering is applied. Kleinberg et al (Kleinberg et al., 2015) proved that machine learnings are better suited for prediction problems compared to discrete choice (or econometric) models.

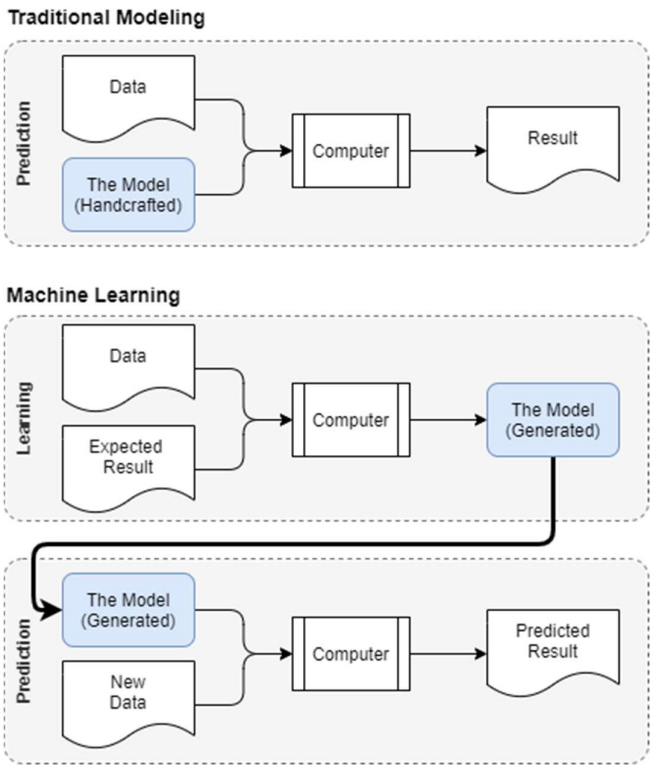


Figure 2-4 Traditional vs machine learning approaches. Adopted from (Malmasi & Zampieri, 2017)

Recent research has adopted machine learning based approaches in crash modeling.

Using random forest and extreme gradient boosting (XGBoost) for crash injury severity prediction at signalized intersections (Kidando et al., 2021), multilayer perceptron neural networks for predicting vehicle crash hotspots (Roland et al., 2021), and feed-forward neural networks (FNN), support vector machine (SVM), fuzzy C-means clustering based feed-forward neural network (FNN-FCM), and fuzzy c-means based support vector machine (SVM-FCM) for predicting crash injury severity (Assi et al., 2020) are some examples.

2.1.2.1 Decision Trees and Ensemble Learning

Decision trees are a popular class of machine learning algorithms due to their wide-ranging practical applications. A decision tree, as the name suggests, recursively partitions the data set into subspaces. Each node in the tree represents a decision typically based on the values of a single feature (or attribute). The leaf nodes represent a prediction (Rokach, 2019).

Chang & Wang applied (Chang & Wang, 2006) Classification and Regression Trees (CART), a widely applied tree-based machine learning method to analyze injury severity using driver, vehicle, and highway data.

A single decision tree may not offer higher performance when the feature or the data space is larger. Therefore, ensemble learning algorithms are developed to combine multiple weak learners (or models) that, collectively, produce a better output. The rationale behind this approach is the *wisdom of the crowds*. This is like how humans seek multiple opinions when faced with a complex decision (Rokach, 2019).

Random Forests and Gradient Boosting are two popular ensemble technologies. They differ in how they produce the individual trees and how they are combined.

Random forests use a method of splitting the data into multiple sub trees and then aggregating them to produce a strong learner. This method is referred to as bootstrap aggregating or bagging (Figure 2-5).

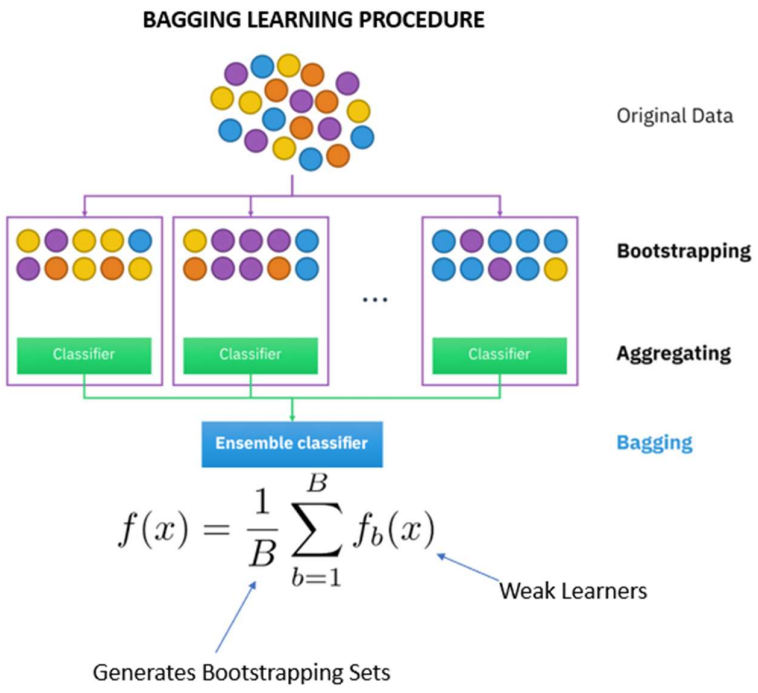


Figure 2-5 Bootstrap aggregating in decision trees

Gradient boosting on the other hand uses a sequential improvement method of connecting a series of sub trees, each boosting (or improving) the accuracy of the previous sub tree (Figure 2-6).

Out of the two, gradient boosting has remained popular due to its powerful techniques for solving complex learning problems (Prokhorenkova et al., 2018).

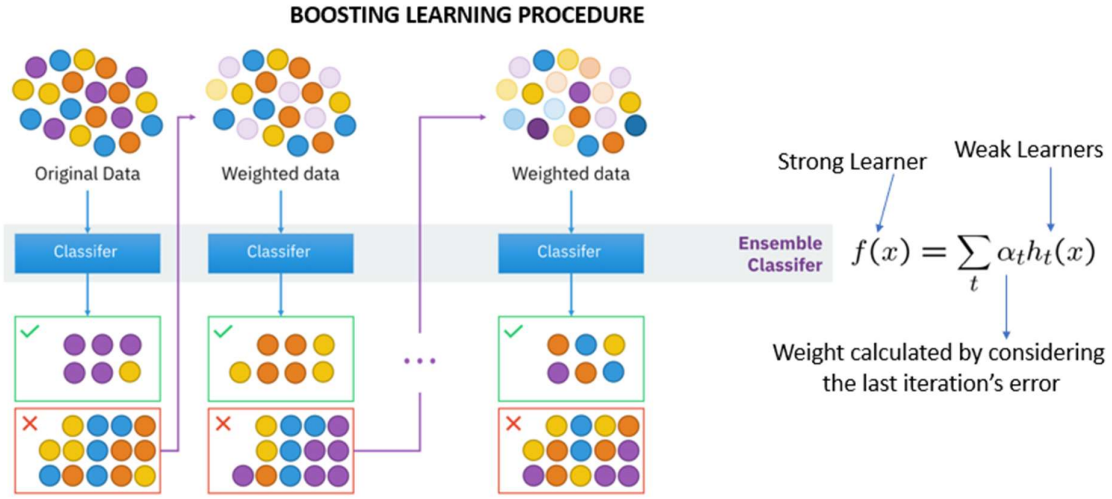


Figure 2-6 Gradient boosting in decision trees

Gradient boosting also has the following benefits in the context of this research (Prokhorenkova et al., 2018):

- Best for categorical data
- Easy to use
- Works well with small data sets (unlike Artificial Neural Networks)

This study employed two leading gradient boosting algorithms, namely LightGBM (Ke et al., 2017) and Catboost (Prokhorenkova et al., 2018) to develop an ML model for crash severity analysis.

2.1.2.2 Hyperparameter Tuning

In machine learning, hyperparameters contain the data that govern the training process itself. In other words, they are the parameters that are not directly learned within the model (e.g., in $y = ax + b$, a and b are parameters that are learned).

The objective function of a machine learning model is a hyper-plane with many global and local optimums. The default training parameters might end up with a local optimum and

therefore, hyperparameter tuning is required to achieve the global best outcome for the model. This is conceptually illustrated in Figure 2-7.

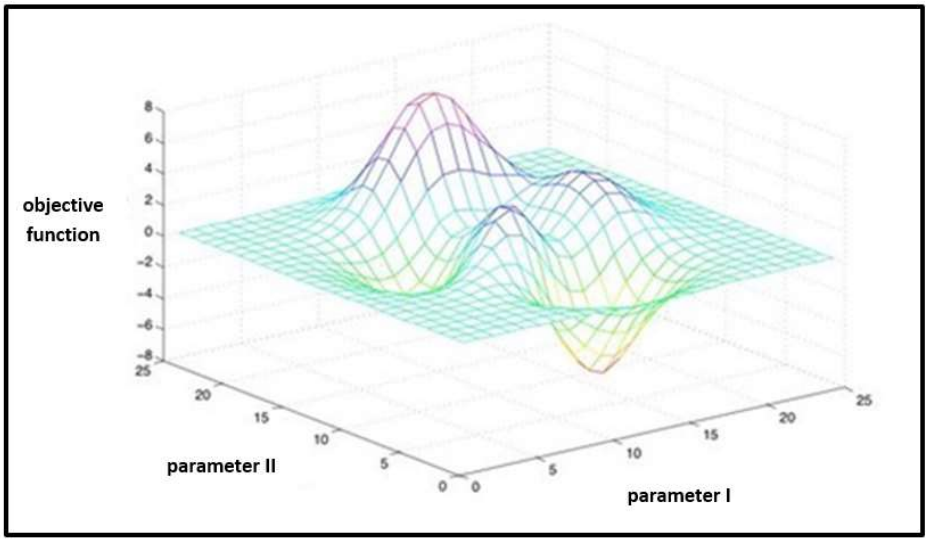


Figure 2-7 Search space for two hyperparameters of a sample objective function

The ML modelers use several techniques to optimize the model by adjusting hyperparameters. The most common techniques are a) manual tuning, b) grid search, and c) random search (Feurer & Hutter, 2019).

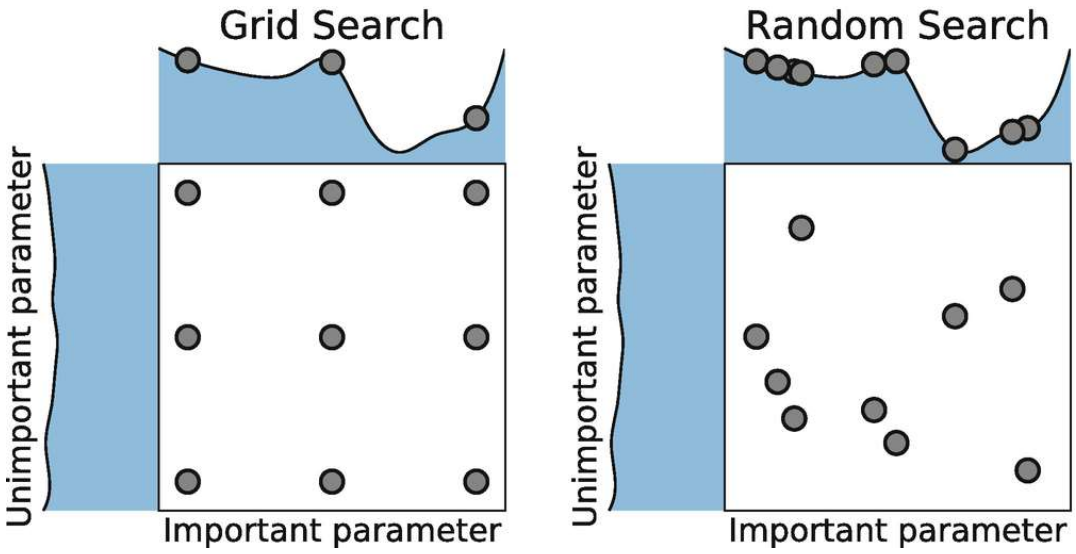


Figure 2-8: Comparison of grid search and random search for hyperparameter turning (Feurer & Hutter, 2019)

2.1.3 Explainable AI

As AI/ML models are making their way into real world decision making, the ability to explain why ML models produce the predictions they do has become a critical requirement (Tjoa & Guan, 2019). For example, why a model thinks a patient is cancer positive or why a banking model rejects a credit card application are important questions that medical teams or customer service teams in a bank should be able to explain.

The branch of explainable AI (XAI) deals with the problem of understanding why a machine learning model behaves the way it is.

SHAP (short for Shapley Additive exPlanations), an outcome of a recent Microsoft research, is one of the recent, but promising approach for XAI. SHAP uses the concept of Shapley values from the game theory to produce advance visualizations that explain ML models (Lundberg & Lee, 2017).

2.2 Exploratory Data Analysis

Exploratory data analysis, or EDA for short, is a term coined by John W. Tukey to describe *the act of looking at data to see what it seems to say*. In EDA, the analysis does not assume any underlying population or a distribution pattern as in standard statistical analysis. Instead, the objective is to look at the data from as many viewpoints as possible using multiple procedures to analyze the data (Morgenthaler, 2009). The procedures may involve graphical or non-graphical tools or methods.

EDA is a fundamental early step in understanding a data set after the data collection step. EDA methods are categorized either as graphical or non-graphical (quantitative) methods or univariate or multi-variate methods (Komorowski et al., 2016). Typical recommended EDA techniques for different types of data are shown in Table 2-1.

Table 2-1 Suggested EDA techniques depending on the type of data (Komorowski et al., 2016)

Type of data	**Suggested EDA techniques**
Categorical	Descriptive statistics
Univariate continuous	Line plot, Histograms
Bivariate continuous	2D scatter plots

Type of data	Suggested EDA techniques
2D arrays	Heatmap
Multivariate: trivariate	3D scatter plot or 2D scatter plot with a 3rd variable represented in different color, shape, or size
Multiple groups	Side-by-side boxplot

An EDA technique of particular significance for road crash analysis is the exploration of covariance and correlation between different spatial and temporal attributes in crash statistics databases.

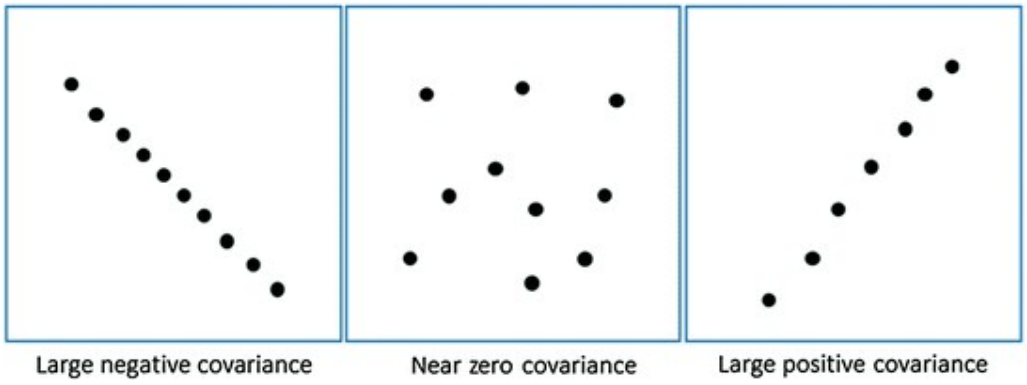


Figure 2-9 Examples of covariance for three different data sets (Komorowski et al., 2016)

The covariance and correlation analysis are useful in identifying causal factors that show the highest probability of increasing crash frequency or crash severity.

2.2.1 Correlation in Categorical Data

Crash data attributes are most of the time categorical in nature (rather than numerical). Day of the week (Monday to Sunday), hour of the day (0..23), light condition (daylight, dusk, dawn, …) and area type (urban, rural) are some examples of categorical data.

The Pearson's Correlation Coefficient (or Pearson's R) is defined for numerical data. This is inappropriate for categorical (or nominal) variables as those variables do not follow the normal distribution nor do they have equal distances between intervals (Kearney, 2017). Hence, a different correlation measure is necessary when analyzing road crash data.

Cramer's V statistic offers a good measure of association between nominal variables. It is a modified version of Pearson's Chi Squared Test and produces a value bounded between [0,1]. Equation 2-6 depicts the calculation of Cramer's V.

$$\phi_C = \sqrt{\frac{\chi^2}{N\,(L-1)}} \tag{2-6}$$

N = Sample size

L = The smaller value of either the number of columns or the number of rows in the contingency table

2.2.2 Correlation in Mixed Data

Some associations in crash data sets are between numerical and categorical data. (e.g., age of the vehicle vs crash severity). Pearson's R (for numerical vs numerical) or Cramer's V (for categorical vs categorical) does not help in this situation.

The Correlation Ratio is used to assess the association between numerical and categorical data. This ratio answers the question of "given a continuous number, how well can you know to which category it belongs to?" (Zychlinski, 2018).

2.3 Geospatial Data Visualization

Crash models in numerical or algorithmic form are useful and their effect can be augmented using geospatial visualizations that go hand in hand.

Road crashes are a phenomenon involving both temporal (e.g., time of day, day of the week) and spatial factors (e.g., road geometry, terrain, weather), and as such, incorporating both groups of factors for explanatory and predictive crash models is important. Roland et al performed a project that tried to aggregate temporal and spatial data and developed a geospatial visualization model using ArcGIS (Esri, 2020). The study combined a machine learning model that had, on average, over 80% accuracy over different train/test split ratios into the data together with a crash hotspot visualization as depicted in Figure 2-10 (Roland et al., 2021).

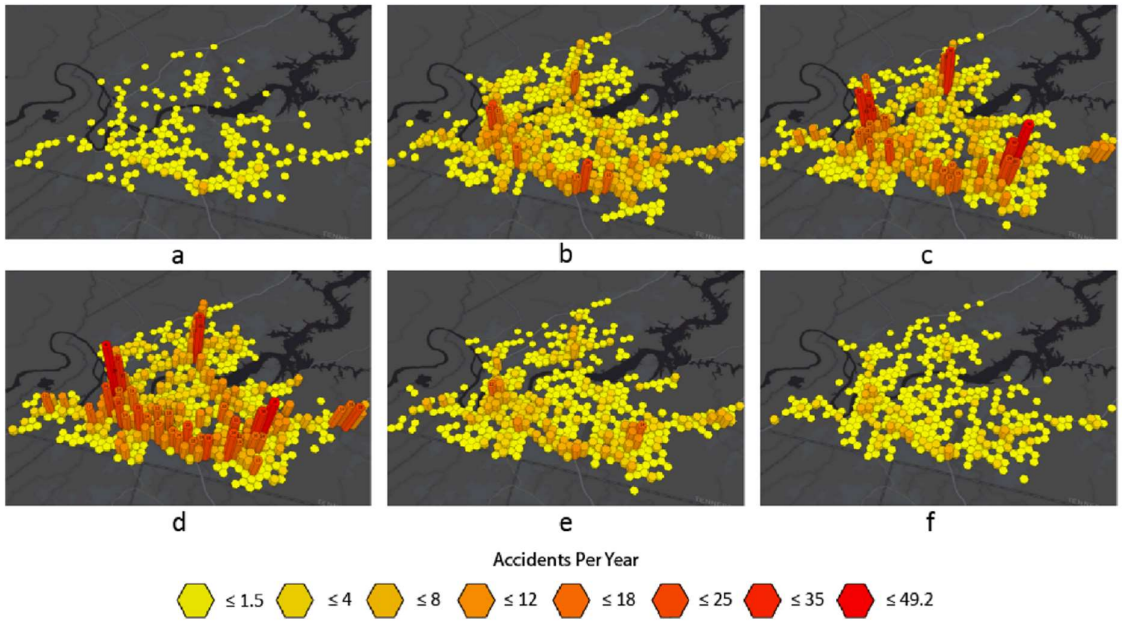


Figure 2-10 Geospatial distribution of crashes across six times of the day in a US city (Roland et al., 2021)

In another study funded by the US Dept of Transportation (DOT), crowdsource data from Waze, a traffic related service (www.waze.com), were aggregated with other sources of data in an attempt to build a near real-time alerting system to enhance the analysis and visualization that informs safe policy decisions (Flynn et al., 2018). Figure 2-11 is an example visualization from this study.

Yuan et al developed a model using a Convolutional Long Short-Term Memory neural network that worked as an image prediction problem using heterogenous spatio temporal data in the state of Iowa in the USA. This research also used geospatial visualization and map technologies to incorporate the multi-dimensional aspects of road crash analysis without limiting to numerical models (Yuan et al., 2018).

There are several technologies that support geospatial visualizations. ArcGIS Pro (Esri, 2020) and Mapbox (Mapbox, 2021) are leading commercial products. Kepler.gl from Uber (Uber, 2021) is an open-source platform that uses powerful WebGL capabilities in modern web browsers.

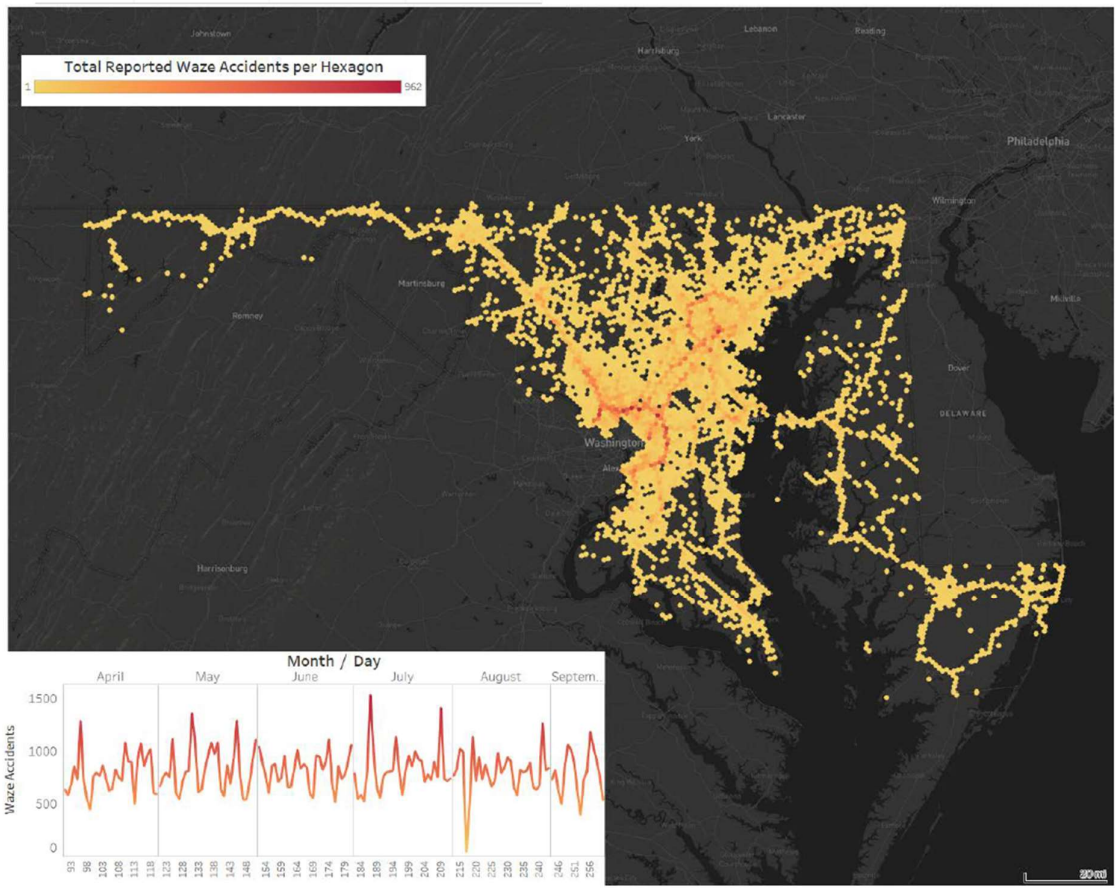


Figure 2-11 Spatial and temporal patterns of Waze accident reports, April - September 2017 (Flynn et al., 2018)

This study employed a group of Python language-based libraries to process the GPS based data and generate static and interactive maps. The **`GeoPandas`** library (Jordahl et al., 2020), an extension to one of Python's core data science library `Pandas`, provides powerful features to manipulate geo data frames.

`GeoJSON` (Butler et al., 2016) is a modern data specification format for geographical data defined by the Internet Engineering Task Force. It is an alternative to the popular Shape (.shp) file format.

`Matplotlib` (Hunter, 2007), the Python ecosystem's core plotting library provides a vast collection of charts and graphs options. The choropleth map plots based on geo data frames provide the basis to generate crash density maps.

`Folium` is a Python library that wraps the Leaflet Javascript framework. Folium exposes interactive maps using Web Mercator map tiles based on Open Street Map data (Agafonkin, 2020).

`Folium` exposes the capabilities of Leaflet to generate GPS point clusters and heat maps at varying zoom levels. Underneath, Leaflet uses a hierarchical greedy clustering algorithm to generate the clusters as and when a map is zoomed in or out (Agafonkin, 2016).

CHAPTER 3

3 METHODOLOGY

3.1 Developing a Unified Approach for Crash Data Analysis

The study used complementary approaches to build explanatory models to analyze and visualize road crash data. The first is the use of exploratory data analysis (EDA) to understand general patterns. Second, a machine learning model using gradient boosting algorithms was built to provide added explainability from the auto detected feature importance through the ML models. The third and final element was static and interactive maps for geospatial analysis.

The unification of various analysis and visualization techniques is the fundamental premise of the methodology developed in the study. The analysis used Sri Lanka Police crash data set for 2019. This data set has been subject to numerous statistical analyses in previous academic and institutional work so some of the results are not entirely new. However, the focus was to develop a new approach for building explanatory crash models and to lay the groundwork for an integrated system that helps authorities to perform regular analysis easily and cost effectively. For an example, an integrated system would look at crashes that are happened not in isolation, but in connection with weather data, road geometric data, light condition data, vehicular flow data, etc. In other words, this development opens the gate to integrate crash data with many other related databases mentioned above so that more rigorous, meaningful (with reasoning) crash analysis as well as predictions can be done.

The sub sections in this chapter will provide details of each of the separate techniques. The next chapter will provide a discussion on how the individual aspects can be combined to a holistic system.

3.2 Exploratory Data Analysis

As described in section 2.2, EDA is the process of looking at the data using multiple views or techniques to see what the data says. Several open-source data science technologies were used to achieve this objective.

Python programming language was used to develop functions and algorithms that can programmatically analyze the data set. **Pandas** ((Reback et al., 2022), **Matplotlib** (Hunter,

2007) and **Seaborn** (Waskom, 2021) libraries were used to produce the basic tabular and graphical analyses.

SweetViz (Bertrand, 2021) library provided a technique to perform a comprehensive correlation analysis over the entire data set including categorical data (see 2.2.1) and categorical vs numerical (see 2.2.2) associations.

Pandas is a widely used, powerful open-source library in the data science domain. It's application in the case study provided the ability to benefit from the extensive community and research driven efforts for data analysis. SweetViz is a new open-source library but provides a concise method to produce a detailed correlation analysis on a data set that would otherwise have taken a lot more effort to produce the same results. These tools were chosen in the EDA step because of these reasons.

Jupyter Notebooks and Google Colab were used to write and execute the Python code.

3.3 The ML Crash Model

The study developed a machine learning model to **explain crash severity** using the potential causal factors such as the collision type, area, type of vehicle etc.

The ML model development followed the typical workflow used for such work. This process is illustrated in

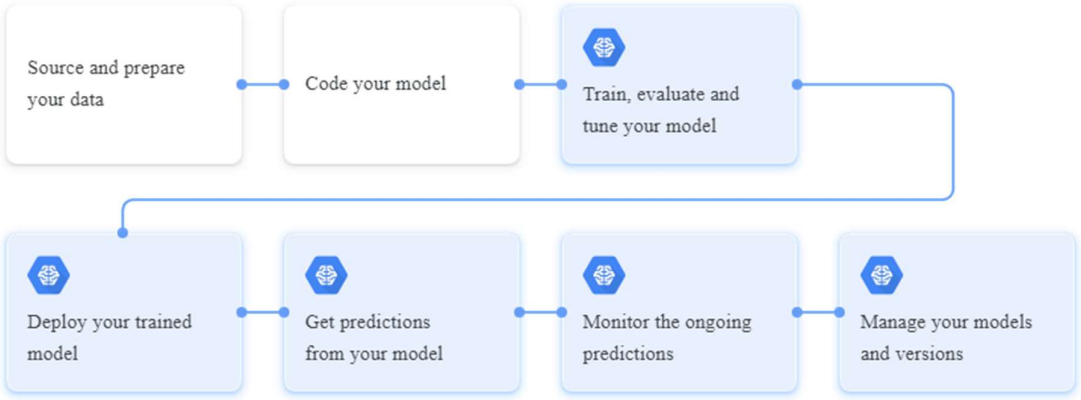


Figure 3-1 The ML model development workflow (Source: Google ML Documentation)

The first step of preparing the data involves the sub steps illustrated in Figure 3-2 and are explained in the sub sections that follow.

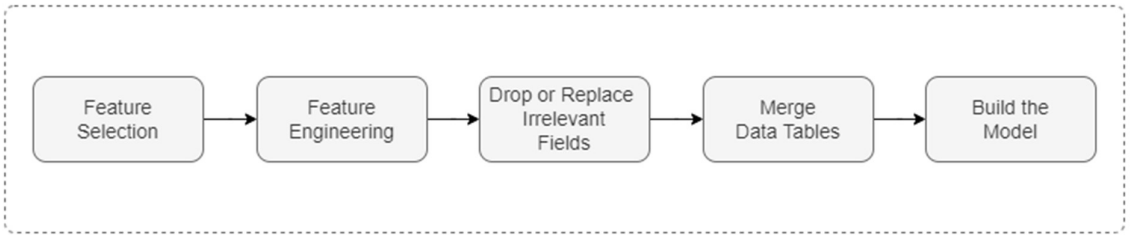


Figure 3-2 Data preparation steps for ML model

3.3.1 Feature Selection

Feature selection is the step of selecting what features (or fields) are useful for the model.

The data set had a total of 71 fields. Certain fields were excluded due to the following reasons.

- **Class imbalance** – fields that had heavily skewed data with 90% or more records consisting of one or two values from a large group (e.g., all the crash factor fields)
- **Non-predictive** – some fields are a result of the crash severity, the target variable, so including those fields causes inaccurate results (e.g., police action, B report number)
- **Key fields and sequence numbers** – certain fields are keys or sequence numbers that are different for each record (e.g., vehicle registration number, accident key)
- **Redundant fields** – some fields represent the same type of information (e.g., road name and road number)

3.3.2 Feature Engineering

Feature engineering is the process of deriving new features from the available fields. Such derived features can represent **hidden attributes** (e.g., the hour of day extracted from the time of the accident) or **combination factors** that can impact the target variable (e.g., the impact of rainy weather in poor light conditions)

The following features were the result of feature engineering.

- Hour – from the Time of crash
- Month – from the Date of crash
- Weather_and_Light – combinations of weather conditions and light conditions

3.3.3 Drop or Replace Irrelevant Records

The target variable is crash severity (or the injury severity) and therefore, the 'Damage Only' crashes were dropped from the final data set.

A minor number of records had null values and they were replaced using a meaningful value based on the type of the field.

3.3.4 Merging the Data Tables

The data set consisted of three distinct tables (see section 4 for details) containing the crash details, the traffic element(s) involved in the crash and the casualty details. A unified data set provides better analysis in the ML model and therefore the tables were merged using the 'Accident Key' as the reference field. The merged data set had 65,005 records with 69 columns (or fields) in each record.

3.3.5 Model Development

Google Colab, an online Jupyter Notebook hosted by Google supporting Python based research projects was used to develop the ML model. The model development used the following Python libraries:

- Pandas
- NumPy
- Matplotlib and Seaborn
- Scikit-Learn
- LightGBM
- Catboost

Pandas and **NumPy** were used for data pre-processing, exploratory analysis, and feature engineering. The rest were used for the ML model development and charting the ML model performance and inferences.

The **Scikit-Learn** estimator interface (`LGBMClassifier, CatBoostClassifier`) was used to develop the model. Use of the Scikit-Learn interface enabled to have a consistent code across the two models.

The case study data set was split in to two buckets, one for the training purposes and the other smaller bucket for testing the trained model. This is the typical practice in ML model development and a **train/test split ratio** of 0.75/0.25 was used.

3.4 Data Visualizations

The crash database records the GPS locations using the EPSG:5234 (Kandawala / Sri Lanka) coordinate reference system (CRS). This data set was in a tabular (CSV) form and therefore, QGIS (QGIS, 2021) was used to convert it to a Shape file format (.shp) suitable to be processed by **GeoPandas**.

The Sri Lanka geo boundary and population data for the entire country and the four administrative regions (Province, District, DSD, and GND) were obtained from the LK Geo Data set (Senaratna, 2021).

QGIS is a strong, open-source alternative for commercial GPS tools such as ArcGIS. The use of GeoJSON and GeoPandas are also because of the powerful support from the open-source community.

3.4.1 Static Choropleth Maps

The crash database GPS data and the geo boundary in the **GeoJSON** format were combined using **GeoPandas**. Methods were developed to generate static choropleth maps for crash density (i.e., crashes per 100,000 persons) for different area types depending on the regions of interest.

Choropleth maps are used due their effectiveness in quick visual identification of the focus areas. These maps provide legend with a color gradient highlighting regions with higher densities using darker colors. Sample maps from the study are shown in section 0.

3.4.2 Interactive Maps

The static maps do not provide the ability to zoom in or zoom out to interactively inspect crash locations and identify hot spots.

Therefore, interactive maps were developed using the **Folium** library and **OpenStreetMaps** data.

Folium `MarkerCluster` objects were used to add the dynamic clustering behavior based on map zoom level. The heat maps to identify crash hotspots were added using the `HeatMap` objects in **Folium**. **GeoPandas** library's `GeoSeries` object functions were used to filter crash data points that belong to administrative regions of interest.

Folium `FeatureGroup` objects were used to on/off different regions or filters as layers in the map.

The combination of marker clusters, heat maps, and feature groups to add/remove different layers of information were used due to the powerful drill down capabilities they provide in the geospatial models.

CHAPTER 4

4 CASE STUDY

The crash database for the year 2019 from the Sri Lanka Police was used for the study. The data set consisted of three tables (or entities) as depicted in Figure 4-1.

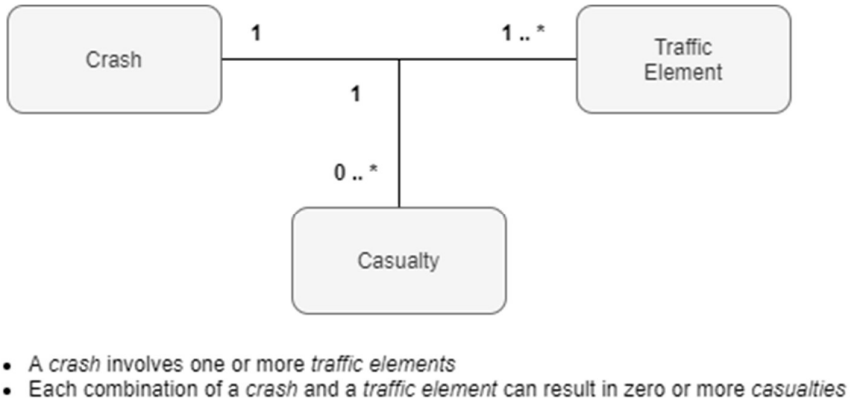


Figure 4-1 The entity relationship schema of the data set

Table 4-1 shows the total number of records for each entity and Table 4-2 has the breakdown by severity level.

Table 4-1 Number of records by type of entity

Type	Number of Records
Crashes (or accidents)	30,346
Traffic elements involved in crashes	58,282
Total Casualties	31,530

Table 4-2 Breakdown of crashes and casualties by severity level

Severity Level	Number of Casualties
Fatal	2,899
Grievous	9,578
Non-grievous	19,053
Damage Only	9,137

Table 4-3 shows the full set of fields in the data set. The Status column shows whether the attribute was included or excluded in the machine learning model, and if excluded, the remarks column shows the reason for exclusion.

The main reasons for excluding attributes were:

- Key and Sequence fields – these are used for database purposes and don't add value to a model
- Redundant attributes – those that represent the same aspect (e.g., Road Name excluded because Road Number is sufficient)
- Non-predictive – Certain fields are a result of the crash severity and are not causal factors (e.g., hospitalization, police action)
- Poor data quality – fields with majority of unknown values (see appendix 8.1)

Table 4-3 Attributes in the Case Study data set

Field	Table	Type	Domain Type	Status	Remarks
Accident Key	Accidents	int64	Sequence	Include	
Number of Vehicles	Accidents	int64	Numerical	Include	
Number of Casualties	Accidents	int64	Numerical	Exclude	Drives Severity
DSDivision	Accidents	int64	Nominal	Include	
Station No	Accidents	int64	Nominal	Include	
Date	Accidents	datetime64	Date	Include	
Time	Accidents	datetime64	Time	Include	
Serial No	Accidents	float64	Sequence	Exclude	Sequence
Highest Severity	Accidents	int64	Ordinal	Include	
UrbanRural	Accidents	int64	Nominal	Include	
WorkDay/Holiday	Accidents	int64	Nominal	Include	
Day of Week	Accidents	int64	Nominal	Include	
RoadNumber	Accidents	object	Nominal	Include	
RoadStreetName	Accidents	object	Nominal	Exclude	Redundant
NearestLowerKmPost	Accidents	int64	Ordinal	Exclude	Redundant
DistanceLowerKmPost	Accidents	int64	Ordinal	Exclude	Redundant
Node Number	Accidents	object	Nominal	Include	

Field	Table	Type	Domain Type	Status	Remarks
Link Number	Accidents	object	Nominal	Include	
DistanceFromNode	Accidents	int64	Numerical	Include	
East coordinate	Accidents	int64	Nominal	Exclude	
North coordinate	Accidents	int64	Nominal	Exclude	
Collision type	Accidents	int64	Nominal	Include	
Second Collision	Accidents	int64	Nominal	Include	
Road Surface	Accidents	int64	Nominal	Include	
Weather	Accidents	int64	Nominal	Include	
LightCondition	Accidents	int64	Nominal	Include	
Location Type	Accidents	int64	Nominal	Include	
Pedestrian Location	Accidents	int64	Nominal	Include	
Traffic Control	Accidents	int64	Nominal	Include	
SpeedLimitPosted	Accidents	int64	Bool	Include	
SpeedLimit_LightVeh	Accidents	int64	Numerical	Include	
SpeedLimit_HeavyVeh	Accidents	int64	Numerical	Include	
PoliceAction	Accidents	int64	Nominal	Exclude	Not predictive
CaseNumber	Accidents	object	Sequence	Exclude	Not predictive
BReport	Accidents	object	Sequence	Exclude	Not predictive
Description of Crash	Accidents	object	String	Exclude	Not predictive
Research Purpose	Accidents	object	Nominal	Exclude	Not predictive
Export Status	Accidents	float64	Nominal	Exclude	Not predictive
Accident Key	Vehicles	int64	Sequence	Exclude	Key
Vehicle Reference Number	Vehicles	int64	Sequence	Exclude	Key
Element Type	Vehicles	int64	Nominal	Include	
Vehicle Registration Number	Vehicles	object	Sequence	Exclude	Sequence
Vehicle Year of Manufacture	Vehicles	object	Numerical	Exclude	Age is better
Age of Vehicle	Vehicles	int64	Numerical	Include	
VehicleOwnership	Vehicles	int64	Nominal	Include	

Field	Table	Type	Domain Type	Status	Remarks
Direction of Moving	Vehicles	object	Nominal	Exclude	Not predictive
Driver/Pedestrian Gender	Vehicles	int64	Nominal	Include	
Driver/Pedestrian Age	Vehicles	int64	Numerical	Include	
Driver License Number	Vehicles	object	Sequence	Exclude	Sequence
ValidityofLicence	Vehicles	int64	Nominal	Include	
Licence Year of Issue	Vehicles	float64	Numerical	Exclude	Age of license is better
Number of Years Since Issue	Vehicles	int64	Numerical	Include	
HumanPreCrashFactor1	Vehicles	int64	Nominal	Exclude	High class imbalance
HumanPreCrashFactor2	Vehicles	int64	Nominal	Exclude	High class imbalance
PedPreCrashFactor	Vehicles	int64	Nominal	Exclude	High class imbalance
RoadPreCrashFactor	Vehicles	int64	Nominal	Exclude	High class imbalance
VehiclePreCrashFactor	Vehicles	int64	Nominal	Exclude	High class imbalance
CrashFactorforSeverity	Vehicles	int64	Nominal	Exclude	High class imbalance
OtherCrashFactor	Vehicles	int64	Nominal	Exclude	High class imbalance
AlcoholTest	Vehicles	float64	Nominal	Include	
DriverRideratFault	Vehicles	float64	Nominal	Exclude	Not predictive
Research Purpose	Vehicles	object	Nominal	Exclude	Not predictive
Accident Key	Casualties	float64	Sequence	Exclude	Key
Casualty Reference Number	Casualties	float64	Sequence	Exclude	Key
Traffic Element Number	Casualties	float64	Sequence	Exclude	Key
Severity	Casualties	float64	Ordinal	Include	
Category	Casualties	float64	Nominal	Include	
CasualtyGender	Casualties	float64	Nominal	Include	
Age	Casualties	float64	Numerical	Include	
Protection	Casualties	float64	Nominal	Include	
Hospitalised	Casualties	float64	Nominal	Exclude	Not predictive

CHAPTER 5

5 RESULTS AND DISCUSSION

5.1 The Feasibility of a Unified Approach

As previously mentioned in sections 1 and 3, building a multi-faceted approach that unifies several modern data science and machine learning technologies was the key objective in this study. The idea was to use the analytical principles form the previous studies on crash modeling that used techniques such as regression and discrete choice models and combine that knowledge with the new developments in the computer science domain.

The machine learning approach provides the ability to use the power of modern computing to figure out the rules that explain how individual causal factors contribute to crash frequency and intensity. This is a powerful and effective way rather than human analysts trying to hand-craft mathematical or analytical models (see Figure 2-4). Extracting the feature importance (see 5.3.3) from the ML models and using the latest developments in explainable AI (XAI) (see 5.3.5) significantly simplifies the human analysts job in understanding what causal factors contribute to crashes in what extent.

The EDA techniques such as SweetViz provide the ability to analyze the correlation in numerical and categorical fields all at once with a few lines of programming code (see 5.3). The powerful visualization techniques such as heatmaps make it possible to emphasize, for example, how the vulnerable road users are at higher risks during the evening peak hours, in an intuitive manner (see Figure 5-7).

Geospatial visualization techniques (see 5.3.6) provided ability to generate both static and interactive maps at varying degrees of detail and scale. Here again, the programmatic approach makes the development of such maps and interactive charts a repeatable process for new and old data sets with minimal effort.

5.2 Proposal for a Road Safety Management System

This study proved the viability of using a unified approach for crash data analysis using the latest computer technologies. The effectiveness and quality of open-source data science and visualization technologies have increased several folds in the last couple of decades. Community foundations such as the NumFOCUS (https://numfocus.org) group are now overseeing the development and maintenance of these technologies.

Similarly, leading companies in the IT industry today lead or support the development of AI and ML capabilities. LightGBM and SHAP (Microsoft), Catboost (Yandex, or widely regarded as the Google of Russia) Google Colab (Google) are some examples.

It was also observed that a single dimensionality of data in the case study (only Police crash database) is inadequate to build a holistic understanding of road crashes. Incorporation of other dimensions such as weather data, road geometry data, and traffic volume etc. can produce a comprehensive analysis.

Further, it can be stated that this study highlighted the following important understandings.

a) The cross-discipline application of computer science knowledge with transportation engineering knowledge can bring in more insights in the road safety analysis problem
b) The combination of exploratory data analysis, crash models using machine learning and geospatial visualization techniques offers great advantages compared to using them in isolation
c) The latest open-source technologies are extremely powerful to build effective and affordable systems to empower road and enforcement agencies

Based on the study findings, the author proposes the development of an integrated digital platform as a Road Safety Management System. The outcome of the research proves the feasibility of such a system and provide the fundamental know-how to build one.

A very high-level design for such a digital platform is presented in Figure 5-1.

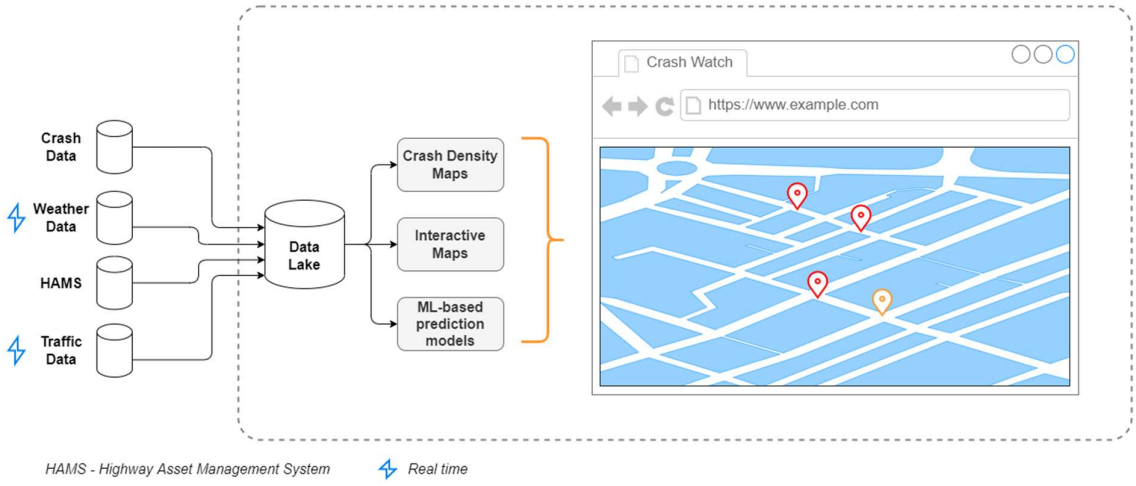


Figure 5-1 A high level proposal for a digital platform

5.3 Sample Results that Support the Effectiveness of the Techniques

The feasibility of a combined approach for crash data analysis and the proposal for a basic architecture for an integrated system that uses this combined approach was presented in sections 5.1 and 5.2 respectively.

The sub sections below provide some sample results obtained using each of the techniques developed in the study. The results are not a comprehensive presentation of the results developed or what's possible, but just a set of samples that illustrates the efficiency and effectiveness of the techniques.

5.3.1 Findings from the Exploratory Data Analysis

Running the SweetViz library on the pre-processed case study data set produced the correlation matrix shown in Figure 5-2.

In this matrix, squares are categorical associations (uncertainty coefficient & correlation ratio) from 0 to 1 and circles are the symmetrical numerical correlations (Pearson's R) from -1 to 1. The row labels indicate how much they provide information to each label at the top.

The first column in Figure 5-2 has the field "Is Fatal". This is a derived field from the Severity field and therefore has a perfect correlation to severity.

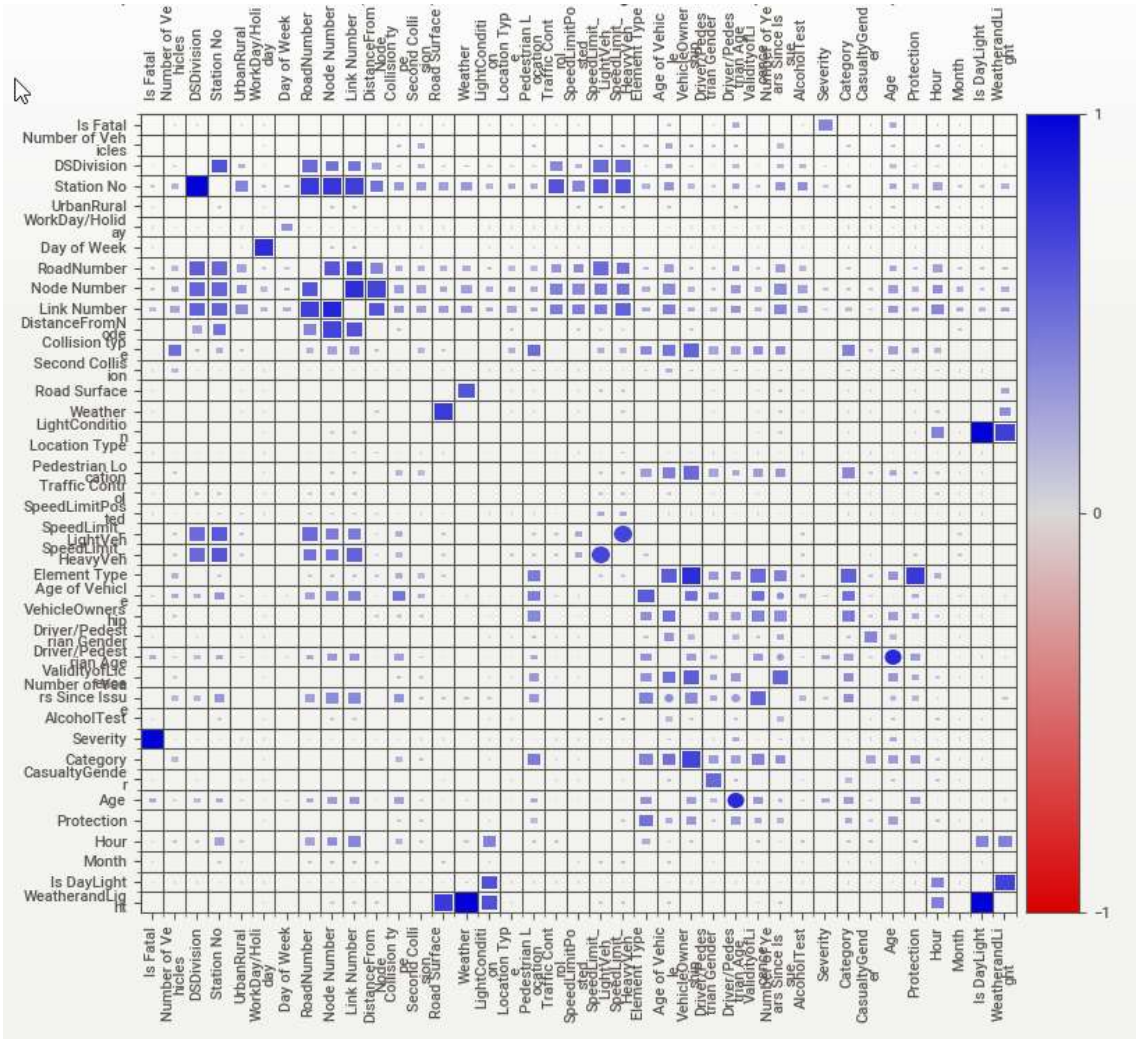
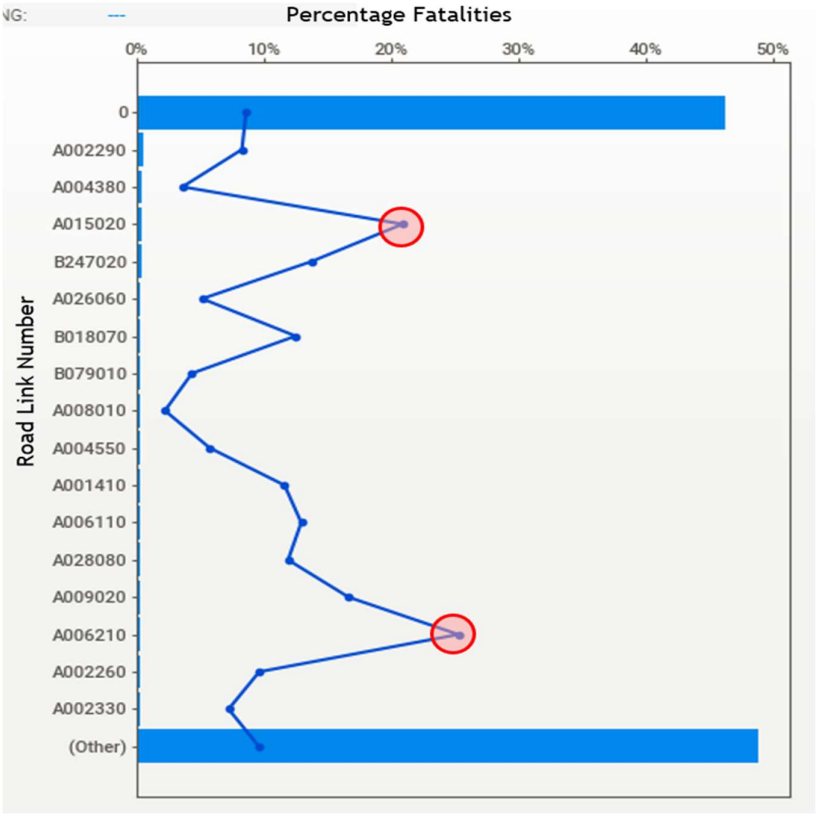


Figure 5-2 Correlation matrix for the crash data set

Ignoring the Severity attribute because of the perfect correlation, the “Is Fatal” field shows the highest association with the attributes “Link Number” (road section) and “Node Number” (intersection) respectively. These two and the other fields in the descending order of correlation are shown in Figure 5-3.

Figure 5-3 Attributes with the highest correlation with fatalities

Digging deep into the Link Number and Node Number attributes reveals that certain links and nodes have a significantly high percentage of fatal crashes out of total crashes reported at those links or nodes. Since these are percentages, the jumps, or peaks in part (a) of Figure 5-4 and Figure 5-5 reveal a finding that warrants further attention.

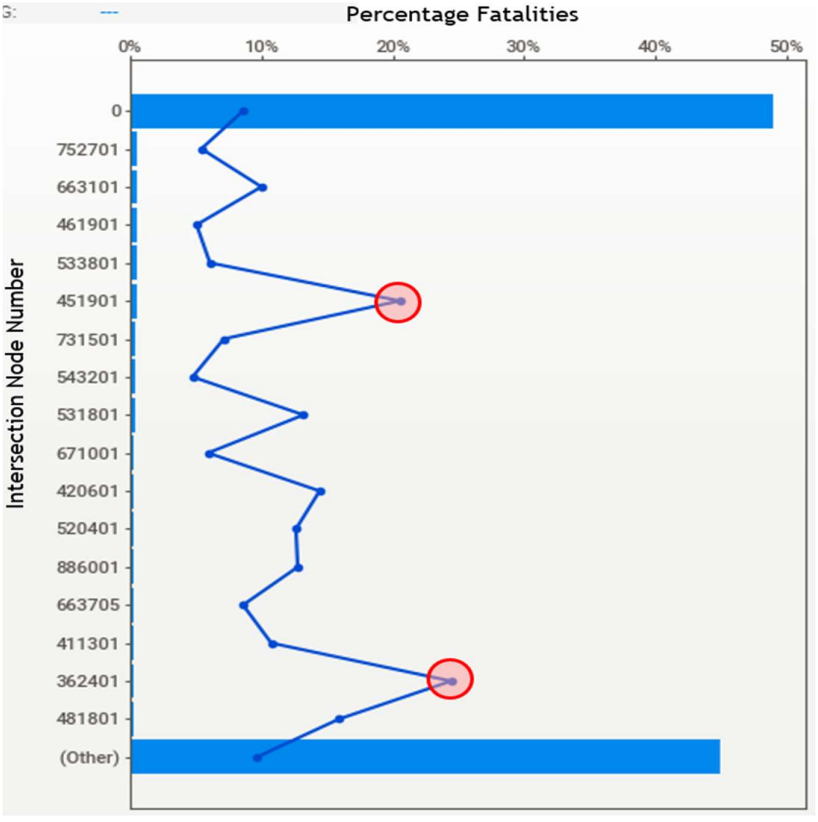


(a)

TOP CATEGORIES			Is Fatal	
0	14,585	46%	1,255	9%
A002290	145	<1%	12	8%
A004380	111	<1%	4	4%
A015020	110	<1%	23	21%
B247020	109	<1%	15	14%
A026060	97	<1%	5	5%
B018070	96	<1%	12	12%
B079010	92	<1%	4	4%
A008010	90	<1%	2	2%
A004550	87	<1%	5	6%
A001410	86	<1%	10	12%
A006110	85	<1%	11	13%
A028080	84	<1%	10	12%
A009020	84	<1%	14	17%
A006210	83	<1%	21	25%
A002260	83	<1%	8	10%
A002330	83	<1%	6	7%
(Other)	15,420	49%	1,482	10%
ALL	31,530	100%	2,899	9%

(b)

Figure 5-4 Association between fatality and link number



(a)

TOP CATEGORIES			Is Fatal	
0	15,446	49%	1,331	9%
752701	184	<1%	10	5%
663101	170	<1%	17	10%
461901	158	<1%	8	5%
533801	148	<1%	9	6%
451901	146	<1%	30	21%
731501	140	<1%	10	7%
543201	126	<1%	6	5%
531801	106	<1%	14	13%
671001	101	<1%	6	6%
420601	97	<1%	14	14%
520401	95	<1%	12	13%
886001	94	<1%	12	13%
663705	93	<1%	8	9%
411301	93	<1%	10	11%
362401	90	<1%	22	24%
481801	88	<1%	14	16%
(Other)	14,155	45%	1,366	10%
ALL	31,530	100%	2,899	9%

(b)

Figure 5-5 Association between fatality and node number

The correlation of fatality with the hour of the day also reveals an interesting pattern. Crashes occurring early morning (between hours 01:00 and 05:00) and late night (around hours 22:00 and 23:00) have a high percentage of fatal crashes.

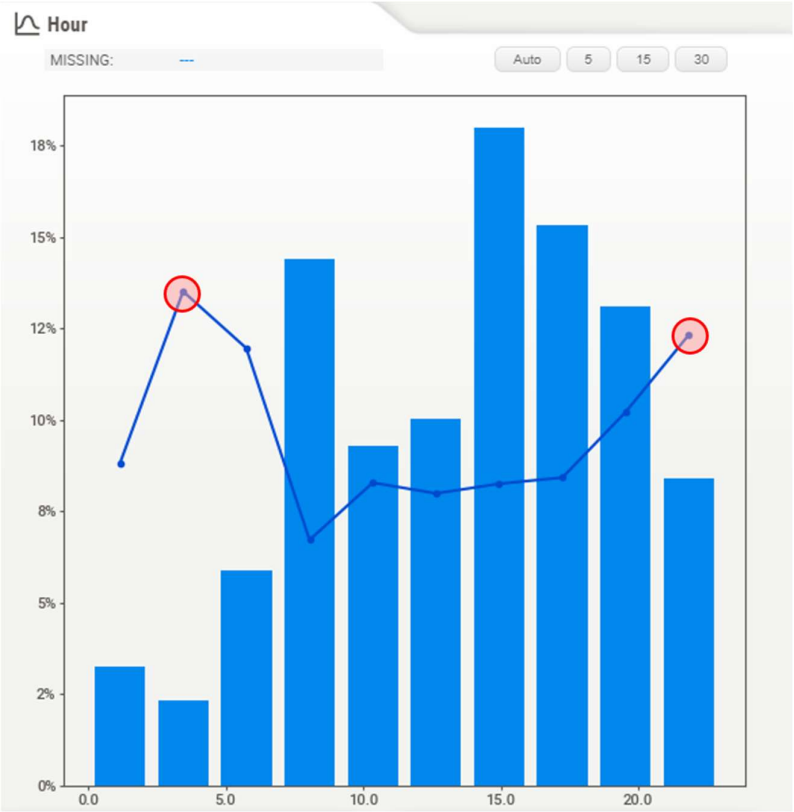


Figure 5-6 Percentage of fatalities by hour of the day

Another analysis done with Pandas and Seaborn produced a heatmap of the number of fatal crashes against the traffic element involved in the fatality and the hour of the day.

As per the heatmap in Figure 5-7, the vulnerable road user group (Motorcyclists and Pedestrians) has a higher intensity of fatalities compared to other groups. This is already known from basic statistics published by the Sri Lanka Police and from other studies such as Shajith et al, about the vulnerable road user groups in Sri Lanka (Shajith et al., 2019). However, the highlighting of the higher fatal crashes for this group in the evening peak (hours 16:00 to 19:00) is a noteworthy finding from this analysis.

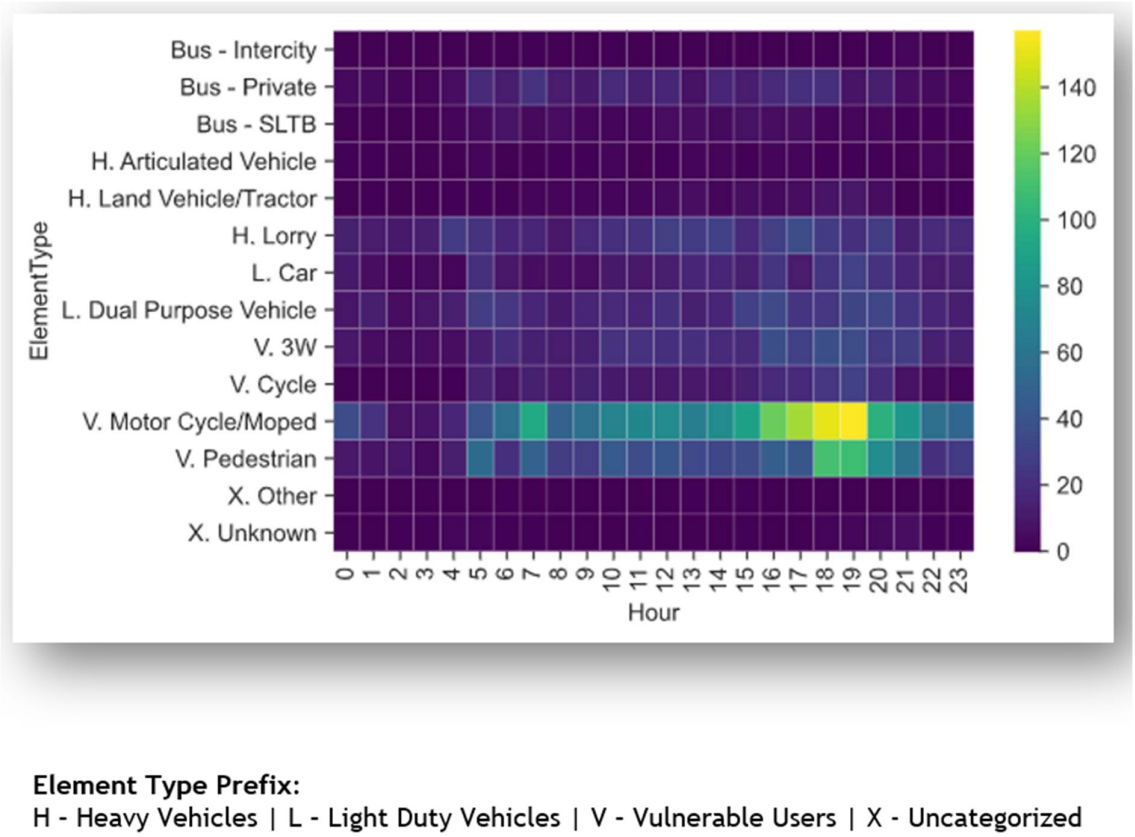


Figure 5-7 Heatmap of fatal crashes with element type vs hour of the day

5.3.2 Machine Learning Model Accuracy

The LightGBM and Catboost classifier estimators were run on the prepared data set. To have an equivalent comparison of the performance, the default hyperparameters were used in both models. The model performance in terms of accuracy is listed in Table 5-1.

As evident, the LightGBM model performed better and produced a 79% accuracy. The interpretation of this number is that the model could accurately predict the crash severity (i.e., whether the crash could cause fatal, grievous, or non-grievous casualties) approximately 80% of the time. The accuracy is significant given this is the basic model without any hyperparameter tuning.

Table 5-1 ML model performance scores

Model	Accuracy
LightGBM	0.7899 (79%)
Catboost	0.6878 (69%)

5.3.3 Feature Importance

One drawback with the ML model is that the translation function is a black box unknown to the modeler. The AI frameworks offer a few techniques to close this gap.

Feature importance, or how strong or weak the contribution of each feature to the predicted result is one way to understand this. Figure 5-8 and Figure 5-9 illustrate the feature importance of the LightGBM and Catboost models respectively.

Each of these figures shows the top 20 contributing features from the two models. It is evident that the two models have used the features in slightly different ways. However, it is important to note that, *Station No* comes as the most important feature in both models.

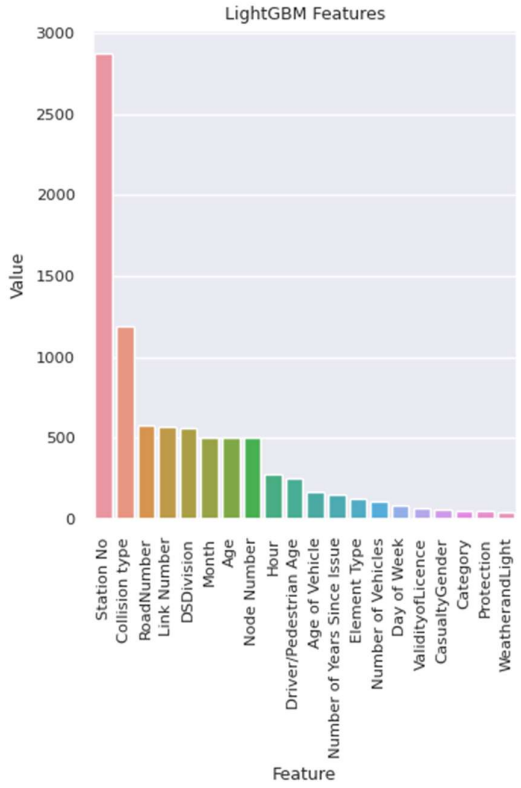


Figure 5-8 Feature importance of LightGBM model

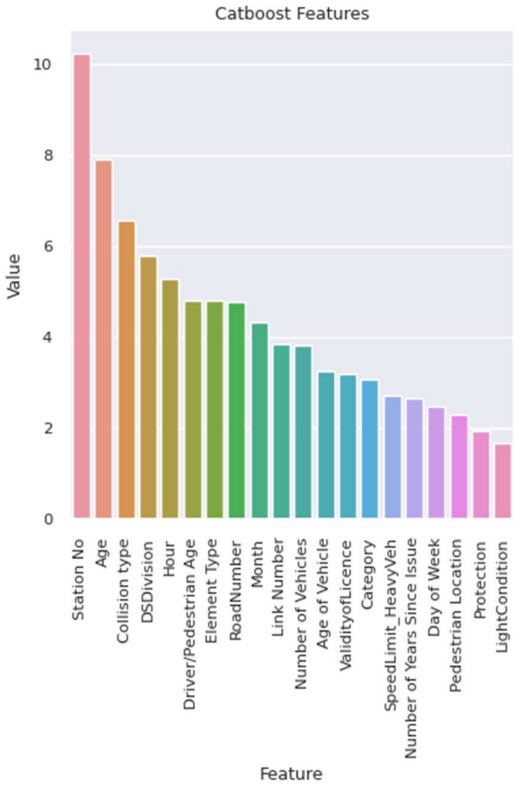


Figure 5-9 Feature importance of Catboost model

Also, 7 of the top 10 features in both models are common and include the following.

- Station No
- Collision Type
- RoadNumber
- DSDivision
- Age

- Hour
- Driver/Pedestrian Age

The interpretation is that these 7 attributes have a high significance in determining the crash severity. This insight generated from a machine algorithm purely from data should be further investigated and validated with human input.

Since *Station No* came as the top contributing attribute in both models, an analysis of the severities by the percentages of each type of severity against each station was done (Figure 5-10). It was revealed that certain stations have $^1/_3$ or more of the crashes resulting in fatalities. The author suggests further analysis of this finding.

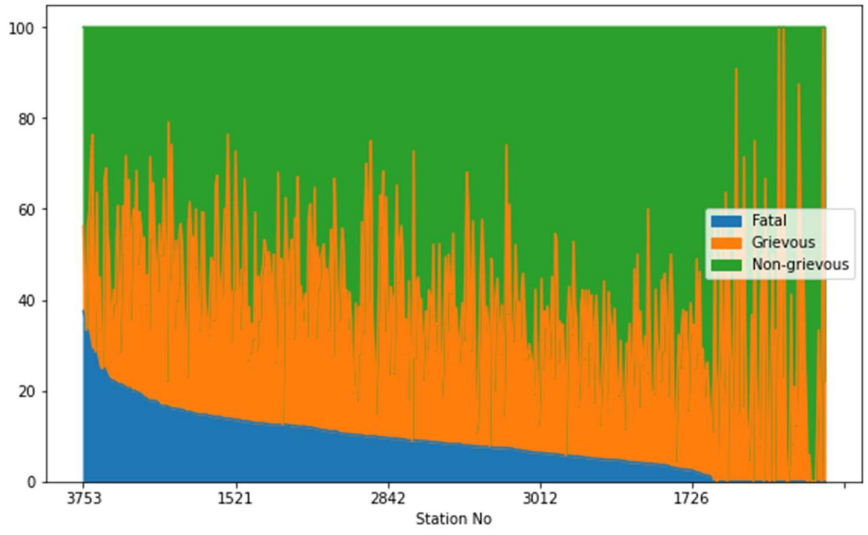


Figure 5-10: Percentage of severity by Station No

5.3.4 Hyperparameter Tuning to Improve Model Performance

Hyperparameter tuning (see section 2.1.2.2) was used to improve the LightGBM model, the one that had the better performance with the default settings (see Table 5-1). Auto tuning option with **Scikit-Learn** library's random search method (**`RandomizedSearchCV`**) was used for the parameter optimization.

Table 5-2 Hyperparameters before and after tuning

Parameter	Purpose	Default Value	Optimized Value
`reg_alpha`	Controls the regularization (smoothing) of the model	0.0	0.01
`num_leaves`	Control the complexity of the tree model by limiting number of tree nodes	31	20
`min_child_samples`	Minimum number of data points needed in a child node	20	15
`max_depth`	Control the depth of the tree	-1	5
`learning_rate`	Control the number of corrections in boosting step, and thereby, the number of trees in the model	0.1	0.05

The result of the auto parameter tuning is listed in Table 5-2.

The effect of hyperparameter optimization can be understood in terms of the key performance measures of a machine learning model.

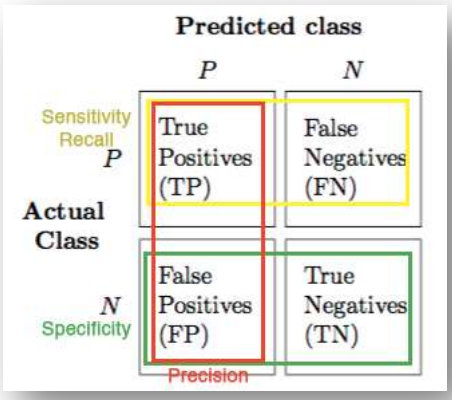


Figure 5-11 Possible scenarios in a binary classification model

Since the ML model's target was a binary classification (whether a crash is fatal or not), there are four outcomes with each prediction in the model. (Figure 5-11). Given this, the model performance measures are:

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$F1\ Score = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

The model performance measures before the hyperparameter tuning are shown in Table 5-3.

Table 5-3 Model performance before hyperparameter tuning

	Precision	Recall	F1-Score
0 (Non-fatal)	0.91	1.00	0.95
1 (Fatal)	0.31	0.02	0.03

The model performance measures after the hyperparameter tuning are shown in Table 5-4.

Table 5-4 Model performance after hyperparameter tuning

	Precision	Recall	F1-Score
0 (Non-fatal)	0.92	1.00	0.95
1 (Fatal)	0.48	0.04	0.08

The auto tuning improved the precision of detecting fatal crashes from 0.31 to 0.48, a 17% gain from the default case. However, the recall showed a negligible gain of 2% to improve from 0.02 to 0.04. The performance of detecting non-fatal crashes is very good in both the default and tuned cases, however, this could be due to the large imbalance of the data set that had 91% of the cases as non-fatal.

The as-is model is not a good indicator of fatal crashes. The single dimension of the data set (i.e., only the Police crash database) and the absence of other data sources such as road geometry, traffic level and weather data etc. can be the reason for the lower recall level of the model.

5.3.5 Use of Explainable AI

The study attempted the use of SHAP (Lundberg & Lee, 2017) as an explainable AI (XAI) technique. The intent was to use the power of latest XAI research to improve the explainability of crash causal factors by using SHAP on the ML model developer under 3.3.

Using SHAP it is possible to generate the contribution of each independent variable (e.g., Age, Hour of Day etc.) to the dependent or predictor variable (e.g., severity of the crash). Figure 5-12 is an example summary plot from SHAP. This plot shows the main independent variables

that impact the severity of a crash either positively or negatively. For example, for higher and higher values of Age, the SHAP value is increased (larger positive value) thus explaining that older age results with higher severity. The fact that SHAP is able to generate this without any human level understanding of what each attribute mean is quite remarkable.

SHAP, however, works well for numerical data and, at present, does not produce meaningful explanations of the categorical (or ordinal) data such as Element Type, Day of Week. This is evident from the grayed-out plots in Figure 5-12.

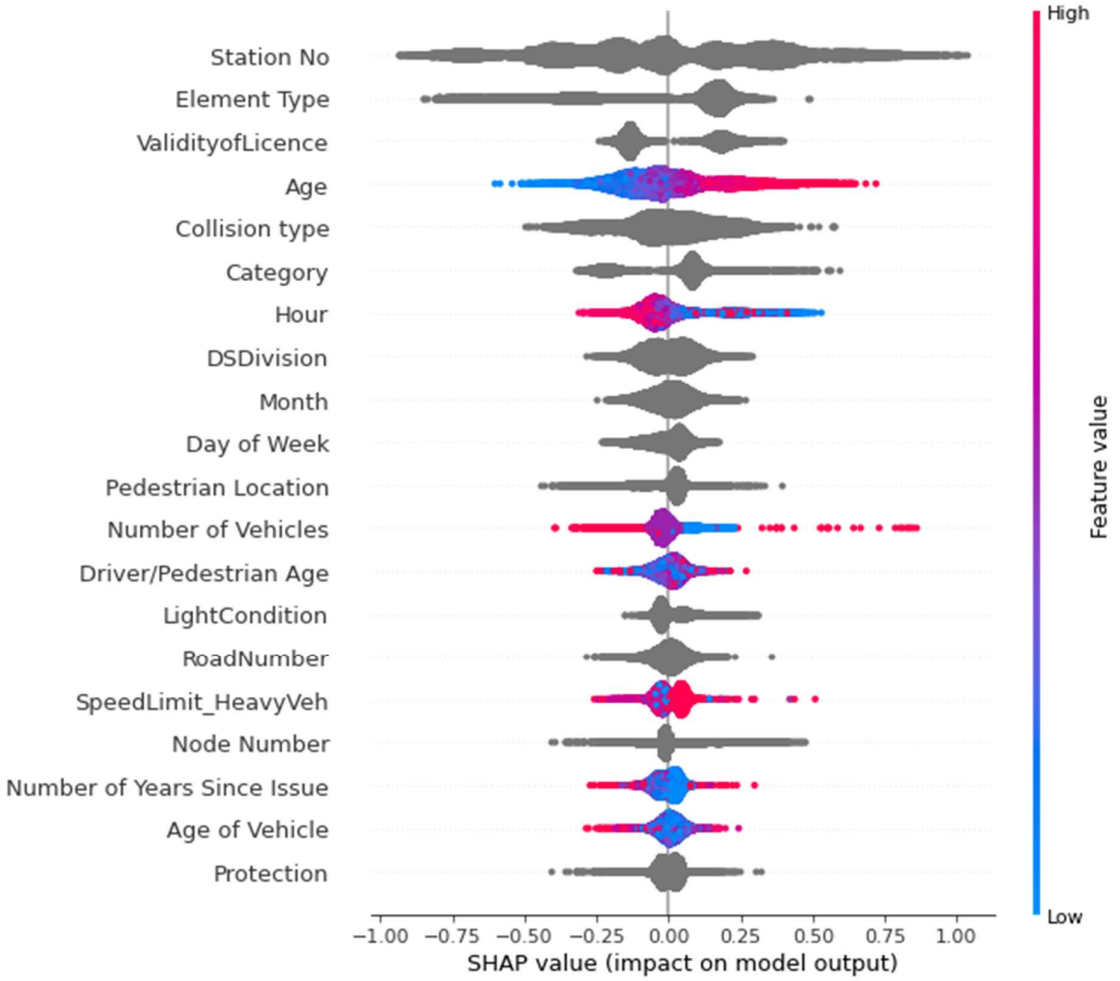


Figure 5-12 SHAP Summary Plot for one of the crashes

As crash data are predominantly categorical the further development of SHAP was not pursued in this study. The author however suggests further exploration of SHAP and other explainable AI techniques to improve the crash analytics developed in this study.

5.3.6 Impact of Data Visualization Methods

Application of the open-source geospatial tools yielded a valuable set of techniques to help process GPS related crash data.

A couple of examples are shown in Figure 5-13 (Crash density by district) and Figure 5-14 (Crash density for Colombo district).

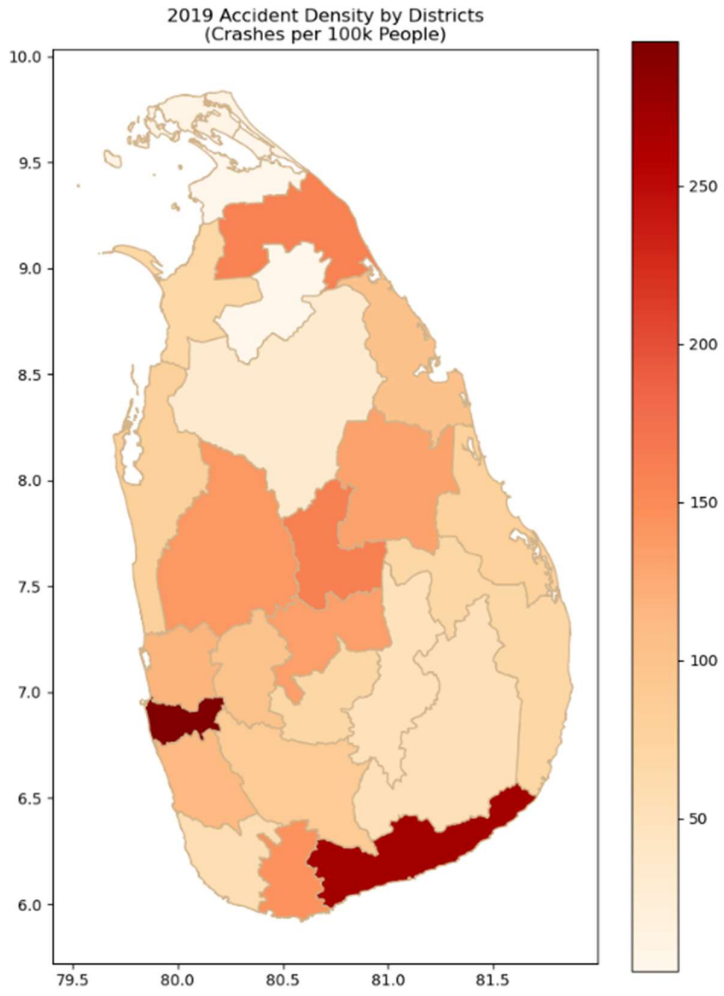


Figure 5-13 Crash density by district

The heat map and crash point clusters at different zoom levels in the interactive maps running inside a browser window are shown in Figure 5-15. Part (a) shows a country level view, in and around Kalutara district, zoomed in one level down is shown in part (b). The ability to zoom in to single crash point level is illustrated in part (c).

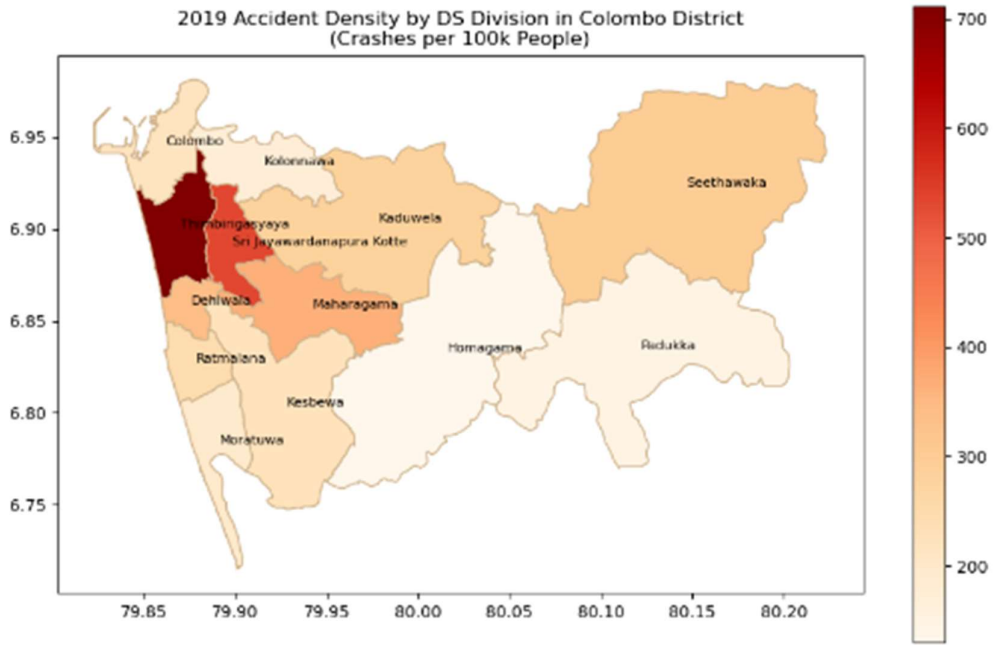
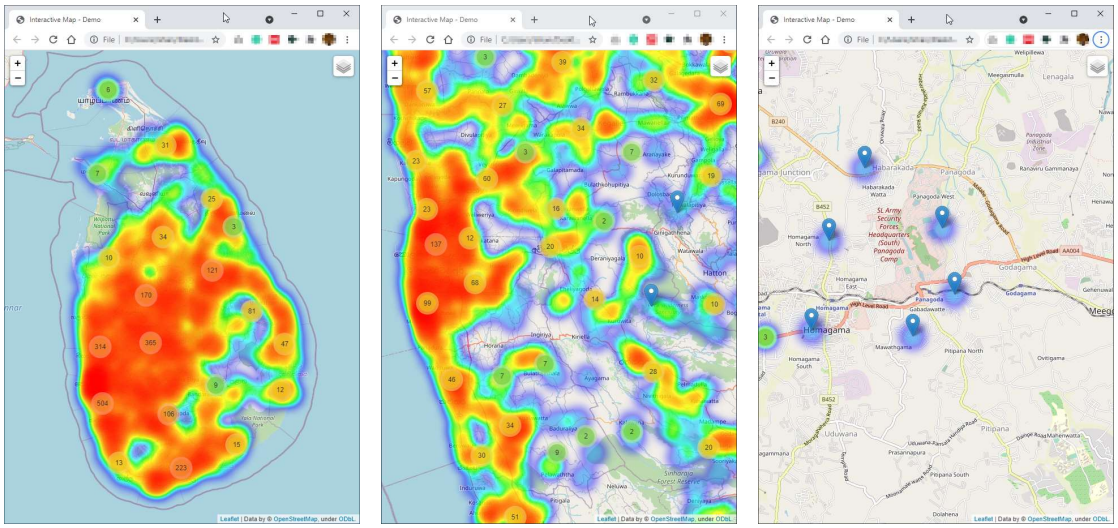


Figure 5-14 Crash density in Colombo district



(a) (b) (c)

Figure 5-15 Crash hot spots and clusters at different zoom levels

Another observation from the geospatial analysis was the level of accuracy of the GPS location recording. Over 1/5th of the Fatal and Grievous crash locations was outside the boundaries of Sri Lanka, while overall, 17.5% of the GPS coordinates were inaccurate. A breakdown of GPS inaccuracy by casualty type is shown in Table 5-5.

Table 5-5 Accuracy of crash GPS locations by casualty type

Severity	Total	Valid	Invalid	Invalid %
Fatal	2,704	2,096	608	22.5%
Grievous	7,775	6,098	1,677	21.6%
Non-Grievous	10,730	8,613	2,117	19.7%
Damage Only	9,137	8,233	904	9.9%
All	30,346	25,040	5,306	17.5%

As illustrated in Figure 5-16, a significant portion of crash locations recorded in the case study data set had inaccurate GPS locations that lie outside the boundaries of Sri Lanka.

This finding is a side effect of the geospatial analysis performed on the case study data set. It's recorded here as an input to authorities for consideration in data quality improvement actions.

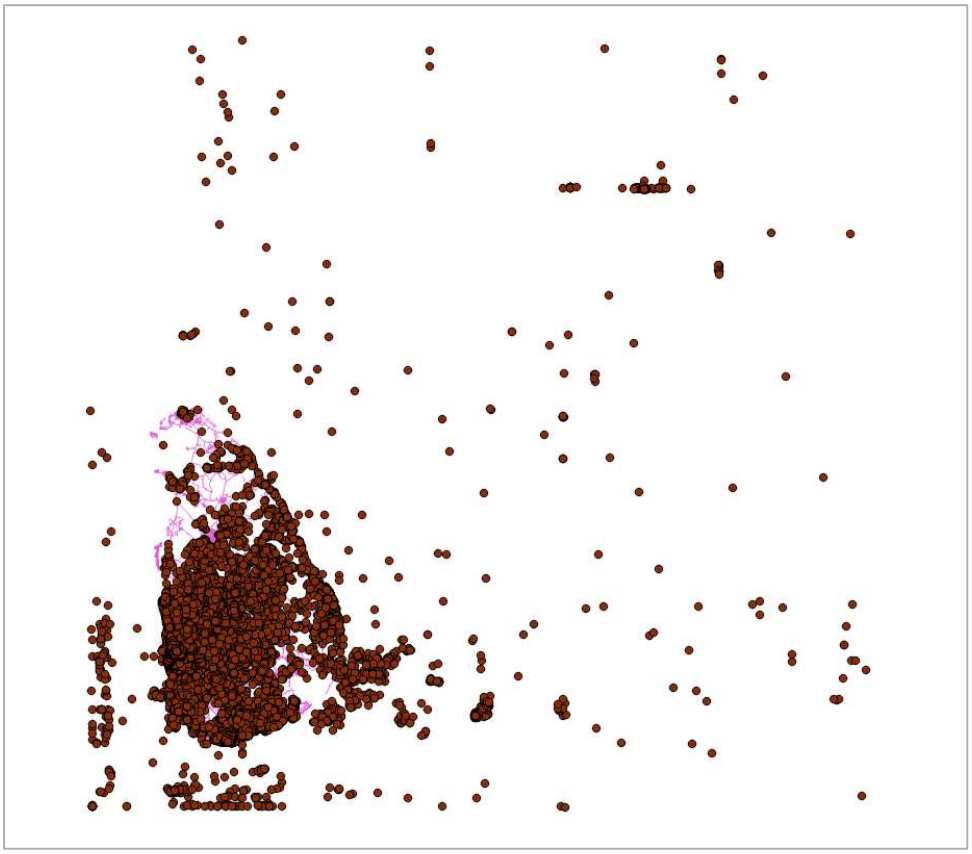


Figure 5-16 Bird's eye view of crash locations recorded in the study data set

CHAPTER 6

6 CONCLUSIONS AND RECOMMENDATIONS

The road crash problem is a severe issue that has impacted societies all around the world in varying degrees. There have been several approaches to understanding the causal factors, based on one or more techniques such as crash modeling, exploratory analysis, and geospatial analysis.

The combination of machine learning based crash models, exploratory data analysis and geospatial visualization technologies using open-source technologies explored in this study lay the foundations for an efficient and affordable digital platform to help authorities in their mission to control road crashes in developing countries such as Sri Lanka.

The power of modern data science and machine learning technologies for efficient and effective analysis of road crash data was evident in this study. The outliers of certain road segments and intersections with a high percentage of fatalities (5.3) were identified with a simple EDA procedure. Visualization heatmap of the fatalities against the time of day highlighted the fatalities happening during the evening peak hours. The fact that certain Police stations have a higher percentage of fatal accidents was identified because the "Station No." attributed appeared at the top of the *feature importance* list in the machine learning models (5.3.3).

The blending of transportation engineering knowledge with the power of computing knowledge and open-source technologies provides a great opportunity for the authorities to build efficient explanatory models that provide fast and new insights that are not possible with traditional analysis approaches.

It can be concluded that the key research contributions from this study are a) the cross-discipline application of computer science knowledge with transportation engineering knowledge, b) the unified analysis provided by EDA, ML crash models, and geospatial visualization techniques, and c) proof of viability of latest open-source technologies in building effective and affordable digital systems to help road and enforcement agencies.

Further, a few recommendations based on some data quality and consistency issues found during the study are presented below.

The data issues are:

- Incomplete or missing data – this was mainly evident in crash factors (8.1)

- Lack of consistency in recording data (e.g., multiple formats of recording road and link numbers and vehicle numbers making data cleansing and analysis difficult)
- Improper GPS location tagging – overall, about 17% of the crash records had completely incorrect GPS locations. The precision of the GPS tagging was also not satisfactory in a considerable portion of the data set

Further, attributes such as "Collision Type" could be split into more than one field without trying to capture all permutations in a single field, which makes recording the attribute difficult.

A review of the data capturing tools and method and the data schema is recommended based on these observations.

Finally, a high-level design of a unified system (Figure 5-1) to improve road crash analysis is presented as a proposal based on the aggregate outcome learnt from the study. Integrating multiple other data sources (traffic data, weather data and road geometry data) with crash data will provide an enhanced crash analysis and visualization platform offering near real-time alerts and insights to control the road crash problem.

7 REFERENCES

Abdel-Aty, M. (2003). Analysis of driver injury severity levels at multiple locations using ordered probit models. *Journal of Safety Research*, *34*(5), 597–603. https://doi.org/10.1016/j.jsr.2003.05.009

Abdel-Aty, M. A., & Radwan, A. E. (2000). Modeling traffic accident occurrence and involvement. *Accident Analysis & Prevention*, *32*(5), 633–642. https://doi.org/10.1016/S0001-4575(99)00094-9

Abdulhafedh, A. (2017). Road Crash Prediction Models: Different Statistical Modeling Approaches. *Journal of Transportation Technologies*, *7*(2), 190–205. https://doi.org/10.4236/JTTS.2017.72014

Agafonkin, V. (2016). *Clustering millions of points on a map with Supercluster | by Mapbox | maps for developers*. Mapbox Blog. https://blog.mapbox.com/clustering-millions-of-points-on-a-map-with-supercluster-272046ec5c97

Agafonkin, V. (2020). *A Web Map from Scratch / Vladimir Agafonkin / Observable*. https://observablehq.com/@mourner/simple-web-map

Asalor, J. O. (1984). A general model of road traffic accidents. *Applied Mathematical Modelling*, *8*(2), 133–138. https://doi.org/10.1016/0307-904X(84)90066-0

Assi, K., Rahman, S. M., Mansoor, U., & Ratrout, N. (2020). Predicting Crash Injury Severity with Machine Learning Algorithm Synergized with Clustering Technique: A Promising Protocol. *International Journal of Environmental Research and Public Health 2020, Vol. 17, Page 5497*, *17*(15), 5497. https://doi.org/10.3390/IJERPH17155497

Bertrand, F. (2021). *SweetViz* (2.1.2). https://github.com/fbdesignpro/sweetviz

Butler, H., Daly, M., Doyle, A., Gillies, S., Hagen, S., & Schaub, T. (2016). *The GeoJSON Format, RFC 7946*. https://datatracker.ietf.org/doc/html/rfc7946

Chang, L. Y., & Wang, H. W. (2006). Analysis of traffic injury severity: An application of non-parametric classification tree techniques. *Accident Analysis & Prevention*, *38*(5), 1019–1027. https://doi.org/10.1016/J.AAP.2006.04.009

Devasurendra, K., Perera, L., & Bandara, S. (2016). An insight to motorized two and three wheel crashes in developing countries: A case study in Sri Lanka. *Journal of Transportation Safety & Security*, *Vol 9*, 204–215. https://doi.org/10.1080/19439962.2016.1236052

Dissanayake, S., & Amarasingha, N. (2012). Effects of Geometric Design Features on Truck Crashes on Limited-Access Highways. *Final Reports & Technical Briefs from Mid-America Transportation Center*.

Dissanayake, S., & Perera, L. (2011). A Survey Based Study of Factors Related to Older driver Highway Safety. *Journal of Transportation Safety & Security*, *3*(2), 77–94. https://doi.org/10.1080/19439962.2010.537437

Esri. (2020). *ArcGIS Pro* (2.6). Esri Inc. https://www.esri.com/en-us/arcgis/products/arcgis-pro/overview

Feurer, M., & Hutter, F. (2019). Hyperparameter Optimization. In F. Hutter, L. Kotthoff, & J. Vanschoren (Eds.), *Automated Machine Learning: Methods, Systems, Challenges* (pp. 3–33). Springer International Publishing. https://doi.org/10.1007/978-3-030-05318-5_1

Flynn, D., Gilmore, M., & Sudderth, E. (2018). *Estimating Traffic Crash Counts Using Crowdsourced Data: Pilot analysis of 2017 Waze data and Police Accident Reports in Maryland*. https://rosap.ntl.bts.gov/view/dot/37256

Global Status Report on Road Safety 2018. (2018). In *World Health Organization* (Issue 1). https://www.who.int/publications/i/item/9789241565684

Haenlein, M., & Kaplan, A. (2019). A Brief History of Artificial Intelligence: On the Past, Present, and Future of Artificial Intelligence. *California Management Review*, *61*(4), 5–14. https://doi.org/10.1177/0008125619864925

Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. *Computing in Science and Engineering*, *9*(3), 90–95. https://doi.org/10.1109/MCSE.2007.55

Jiménez-Mejías, E., Martínez-Ruiz, V., Amezcua-Prieto, C., Olmedo-Requena, R., Luna-Del-Castillo, J. D. D., & Lardelli-Claret, P. (2016). Pedestrian- and driver-related factors associated with the risk of causing collisions involving pedestrians in Spain. *Accident Analysis and Prevention*, *92*. https://doi.org/10.1016/j.aap.2016.03.021

Jordahl, K., Bossche, J. van den, Fleischmann, M., Wasserman, J., McBride, J., Gerard, J., Tratner, J., Perry, M., Badaracco, A. G., Farmer, C., Hjelle, G. A., Snow, A. D., Cochran, M., Gillies, S., Culbertson, L., Bartos, M., Eubank, N., maxalbert, Bilogur, A., … Leblanc, F. (2020). *geopandas/geopandas: v0.8.1*. https://doi.org/10.5281/ZENODO.3946761

Joshua, S. C., & Garber, N. J. (1990). Estimating truck accident rate and involvements using linear and poisson regression models. *Transportation Planning and Technology*, *15*(1), 41–58. https://doi.org/10.1080/03081069008717439

Katchova, A. (2020). *Econometrics Academy - Probit and Logit Models*. Econometrics Academy. https://sites.google.com/site/econometricsacademy/masters-econometrics/probit-and-logit-models

Ke, G., Meng, Q., Finely, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T.-Y. (2017, December). LightGBM: A Highly Efficient Gradient Boosting Decision Tree. *Advances in Neural Information Processing Systems 30 (NIP 2017)*. https://www.microsoft.com/en-us/research/publication/lightgbm-a-highly-efficient-gradient-boosting-decision-tree/

Kearney, M. (2017). *Cramér's V*. https://doi.org/10.4135/9781483381411.n107

Kidando, E., Kitali, A. E., Kutela, B., Ghorbanzadeh, M., Karaer, A., Koloushani, M., Moses, R., Ozguven, E. E., & Sando, T. (2021). Prediction of vehicle occupants injury at signalized intersections using real-time traffic and signal data. *Accident Analysis & Prevention*, *149*, 105869. https://doi.org/10.1016/J.AAP.2020.105869

Kleinberg, J., Ludwig, J., Mullainathan, S., & Obermeyer, Z. (2015). Prediction Policy Problems. *American Economic Review*, *105*(5), 491–495. https://doi.org/10.1257/AER.P20151023

Komorowski, M., Marshall, D. C., Salciccioli, J. D., & Crutain, Y. (2016). Exploratory data analysis. In *Secondary Analysis of Electronic Health Records* (pp. 185–203). Springer International Publishing. https://doi.org/10.1007/978-3-319-43742-2_15/FIGURES/18

Lundberg, S. M., & Lee, S. I. (2017). A Unified Approach to Interpreting Model Predictions. *Advances in Neural Information Processing Systems*, *2017-December*, 4766–4775. https://doi.org/10.48550/arxiv.1705.07874

Malmasi, S., & Zampieri, M. (2017). Detecting hate speech in social media. *International Conference Recent Advances in Natural Language Processing, RANLP*, *2017-September*, 467–472. https://doi.org/10.26615/978-954-452-049-6-062

Mapbox. (2021). *Mapbox*. Mapbox Inc. https://www.mapbox.com/

Mehdizadeh, A., Cai, M., Hu, Q., Yazdi, M. A. A., Mohabbati-Kalejahi, N., Vinel, A., Rigdon, S. E., Davis, K. C., & Megahed, F. M. (2020). A review of data analytic applications in road traffic safety. Part 1: Descriptive and predictive modeling. *Sensors (Switzerland)*, *20*(4). https://doi.org/10.3390/S20041107

Morgenthaler, S. (2009). Exploratory data analysis. *Wiley Interdisciplinary Reviews: Computational Statistics*, *1*(1), 33–44. https://doi.org/10.1002/WICS.2

National Council for Road Safety, Sri Lanka. (2020). https://www.transport.gov.lk/web/index.php?option=com_content&view=article&id=29&Itemid=149&lang=en

Paredes, M., Hemberg, E., O'Reilly, U. M., & Zegras, C. (2017). Machine learning or discrete choice models for car ownership demand estimation and prediction? *5th IEEE International Conference on Models and Technologies for Intelligent Transportation Systems, MT-ITS 2017 - Proceedings*, 780–785. https://doi.org/10.1109/MTITS.2017.8005618

Perera, L., & Dissanayake, S. (2012). Contributing Factors to Older-Driver Injury Severity in Rural and Urban Areas. *Journal of the Transportation Research Forum*, *49*(1), 5–22. https://doi.org/10.5399/OSU/JTRF.49.1.2506

Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., & Gulin, A. (2018). Catboost: Unbiased boosting with categorical features. *Advances in Neural Information Processing Systems*, *2018-December*, 6638–6648. https://github.com/catboost/catboost

QGIS. (2021). *QGIS* (3.22.0). Open Source Geospatial Foundation Project. https://qgis.org/en/site/index.html

Reback, J., jbrockmendel, McKinney, W., den Bossche, J. van, Augspurger, T., Cloud, P., Hawkins, S., Roeschke, M., gfyoung, Sinhrks, Klein, A., Hoefler, P., Petersen, T., Tratner, J., She, C., Ayd, W., Naveh, S., Darbyshire, J., Garcia, M., … Seabold, S.

(2022). *pandas-dev/pandas: Pandas 1.4.1*. Zenodo. https://doi.org/10.5281/zenodo.6053272

Rokach, L. (2019). *Ensemble Learning : Pattern Classification using Ensemble Methods* (Second). World Scientific.

Roland, J., Way, P. D., Firat, C., Doan, T. N., & Sartipi, M. (2021). Modeling and predicting vehicle accident occurrence in Chattanooga, Tennessee. *Accident Analysis & Prevention*, *149*, 105860. https://doi.org/10.1016/J.AAP.2020.105860

Savolainen, P. T., Mannering, F. L., Lord, D., & Quddus, M. A. (2011). The statistical analysis of highway crash-injury severities: A review and assessment of methodological alternatives. *Accident Analysis and Prevention*, *43*(5), 1666–1676. https://doi.org/10.1016/j.aap.2011.03.025

Senaratna, N. (2021). *LK Geo Data*. https://github.com/nuuuwan/geo-data

Shajith, S. L. A., Pasindu, H. R., & Ranawaka, R. K. T. K. (2019). Evaluating the Risk Factors in Fatal Accidents involving Motorcycle – Case Study on Motorcycle Accidents in Sri Lanka. *Engineer: Journal of the Institution of Engineers, Sri Lanka*, *52*(3), 33–42. https://doi.org/10.4038/engineer.v52i3.7363

Shneiderman, B. (2014). The big picture for big data: Visualization. *Science*, *343*(6172), 730. https://doi.org/10.1126/SCIENCE.343.6172.730-A/ASSET/B23773B5-0932-4174-8E8C-071CCF844B85/ASSETS/SCIENCE.343.6172.730-A.FP.PNG

Silver, D., Huang, A., Maddison, C., Guez, A., Sifre, L., Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., Dieleman, S., Grewe, D., Nham, J., Kalchbrenner, N., Sutskever, I., Lillicrap, T., Leach, M., Kavukcuoglu, K., Graepel, T., & Hassabis, D. (2016). Mastering the game of Go with deep neural networks and tree search. *Nature*, *529*, 484–489. https://doi.org/10.1038/nature16961

Singh, A. (2017). Deep Learning Will Radically Change the Ways We Interact with Technology. *Harvard Business Review*. https://hbr.org/2017/01/deep-learning-will-radically-change-the-ways-we-interact-with-technology

Sri Lanka's Road Safety Country Profile. (2021). https://www.roadsafetyfacility.org/country/sri-lanka

Tay, R., & Rifaat, S. M. (2007). Factors contributing to the severity of intersection crashes. *Journal of Advanced Transportation*, *41*(3), 245–265. https://doi.org/10.1002/atr.5670410303

Thangamani, B. (2019). The Economic Impact of Road Accidents: The Case of Sri Lanka. *South Asia Economic Journal*, *20*, 124–137. https://journals.sagepub.com/doi/abs/10.1177/1391561418822210

Tjoa, E., & Guan, C. (2019). A Survey on Explainable Artificial Intelligence (XAI): Towards Medical XAI. *IEEE Transactions on Neural Networks and Learning Systems*, *14*(8), 1–21. https://doi.org/10.1109/tnnls.2020.3027314

Uber. (2021). *Kepler.gl* (2.5.5). Uber Inc. https://kepler.gl/

Waskom, M. L. (2021). seaborn: statistical data visualization. *Journal of Open Source Software*, *6*(60), 3021. https://doi.org/10.21105/joss.03021

Williams, A. F. (1999). *The Haddon matrix: its contribution to injury prevention and control*.

Yee, A., & Alvarado, M. (2012). Pattern recognition and monte-carlo tree search for go gaming better automation. *Lecture Notes in Computer Science*, *7637 LNAI*, 11–20. https://doi.org/10.1007/978-3-642-34654-5_2

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 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, *18*. https://doi.org/10.1145/3219819

Zychlinski, S. (2018, February 24). *The Search for Categorical Correlation*. Towards Data Science. https://towardsdatascience.com/the-search-for-categorical-correlation-a1cf7f1888c9

8 APPENDICES

8.1 Review of crash factor data collection

The Haddon Matrix is widely regarded as the first and foremost approach to the scientific analysis of road crash injuries (Williams, 1999). The crash factors are categorized into three or four groups, namely, human factors, vehicle factors, road or physical environment, and socioeconomic factors. They are also divided along a time dimension to capture the phase of the crash, pre-event, event, and post-event. This two-dimensional arrangement is shown in Figure 8-1.

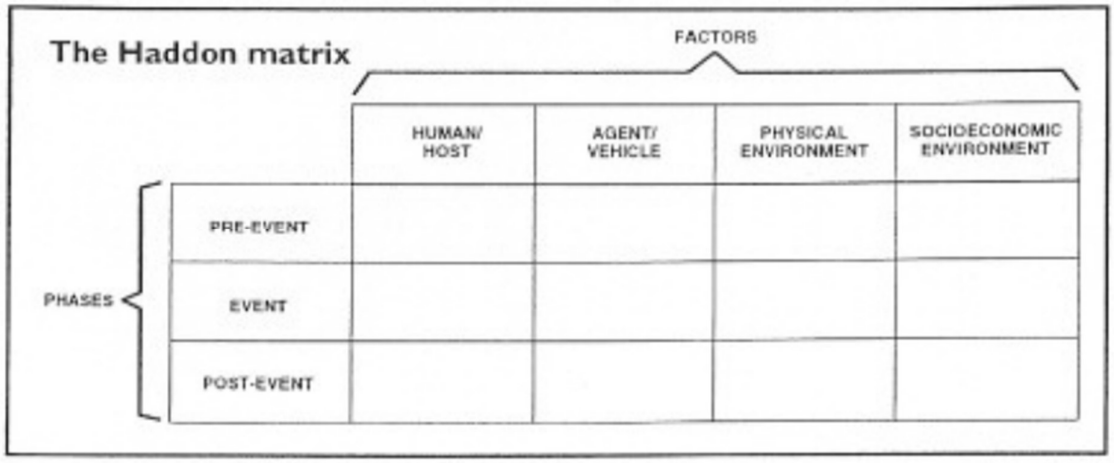


Figure 8-1 The Haddon Matrix (Williams, 1999)

During the exploratory data analysis phase of the study, it was found that crash factor collection in the case study data set was insignificant. There are six attributes related to crash factors but most of the time the factor is not recorded (i.e., unknown) as shown in Table 8-1.

Table 8-1 Percentage of unknown crash factors in the data set

Crash Factor	Percentage of Unknown Cases
Human Pre-Crash Factor 1	42.8 %
Human Pre-Crash Factor 2	72.1 %
Pedestrian Pre-Crash Factor	96.7 %
Road Pre-Crash Factor	98.5 %
Vehicle Pre-Crash Factor	99 %
Severity Causing Factor	86.7 %