

**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

Declaration of the Candidate and Supervisor	i
Acknowledgment	ii
Abstract	iii
Table of Contents	v
List of Figures	vii
List of Tables	viii
List of Abbreviations	ix
1 Introduction	1
2 Literature Review	4
2.1 Crash Models.....	4
2.1.1 Traditional Crash Models	5
2.1.2 Artificial Intelligence (AI) and Machine Learning (ML) based Models	9
2.1.2.1 Decision Trees and Ensemble Learning	10
2.1.2.2 Hyperparameter Tuning.....	12
2.1.3 Explainable AI	14
2.2 Exploratory Data Analysis	14
2.2.1 Correlation in Categorical Data	15
2.2.2 Correlation in Mixed Data	16
2.3 Geospatial Data Visualization.....	16
3 Methodology.....	20
3.1 Developing a Unified Approach for Crash Data Analysis.....	20
3.2 Exploratory Data Analysis	20
3.3 The ML Crash Model.....	21
3.3.1 Feature Selection.....	22
3.3.2 Feature Engineering	22
3.3.3 Drop or Replace Irrelevant Records	23
3.3.4 Merging the Data Tables.....	23
3.3.5 Model Development.....	23
3.4 Data Visualizations	24
3.4.1 Static Choropleth Maps.....	24

3.4.2	Interactive Maps.....	24
4	Case Study	26
5	Results and Discussion	30
5.1	The Feasibility of a Unified Approach.....	30
5.2	Proposal for a Road Safety Management System	30
5.3	Sample Results that Support the Effectiveness of the Techniques	32
5.3.1	Findings from the Exploratory Data Analysis	32
5.3.2	Machine Learning Model Accuracy	38
5.3.3	Feature Importance	39
5.3.4	Hyperparameter Tuning to Improve Model Performance	40
5.3.5	Use of Explainable AI.....	42
5.3.6	Impact of Data Visualization Methods	44
6	Conclusions and Recommendations.....	48
7	References	50
8	Appendices	56
8.1	Review of crash factor data collection	56

LIST OF FIGURES

Figure 2-1 Types of techniques used for crash models.....	4
Figure 2-2 Parts of a road traffic system (based on Asalor, 1984)	5
Figure 2-3 Types of discrete choice models (based on Savolainen et al, 2011).....	8
Figure 2-4 Traditional vs machine learning approaches. Adopted from (Malmasi & Zampieri, 2017)	10
Figure 2-5 Bootstrap aggregating in decision trees	11
Figure 2-6 Gradient boosting in decision trees	12
Figure 2-7 Search space for two hyperparameters of a sample objective function	13
Figure 2-8: Comparison of grid search and random search for hyperparameter turning (Feurer & Hutter, 2019).....	13
Figure 2-9 Examples of covariance for three different data sets (Komorowski et al., 2016)..	15
Figure 2-10 Geospatial distribution of crashes across six times of the day in a US city (Roland et al., 2021)	17
Figure 2-11 Spatial and temporal patterns of Waze accident reports, April - September 2017 (Flynn et al., 2018).....	18
Figure 3-1 The ML model development workflow (Source: Google ML Documentation)	21
Figure 3-2 Data preparation steps for ML model	22
Figure 4-1 The entity relationship schema of the data set	26
Figure 5-1 A high level proposal for a digital platform.....	31
Figure 5-2 Correlation matrix for the crash data set.....	33
Figure 5-3 Attributes with the highest correlation with fatalities	34
Figure 5-4 Association between fatality and link number	35
Figure 5-5 Association between fatality and node number	36
Figure 5-6 Percentage of fatalities by hour of the day.....	37
Figure 5-7 Heatmap of fatal crashes with element type vs hour of the day.....	38
Figure 5-8 Feature importance of LightGBM model.....	39
Figure 5-9 Feature importance of Catboost model	39
Figure 5-10: Percentage of severity by Station No.....	40
Figure 5-11 Possible scenarios in a binary classification model	41
Figure 5-12 SHAP Summary Plot for one of the crashes	43
Figure 5-13 Crash density by district.....	44
Figure 5-14 Crash density in Colombo district.....	45
Figure 5-15 Crash hot spots and clusters at different zoom levels	46
Figure 5-16 Bird's eye view of crash locations recorded in the study data set	47
Figure 8-1 The Haddon Matrix (Williams, 1999).....	56

LIST OF TABLES

Table 2-1 Suggested EDA techniques depending on the type of data (Komorowski et al., 2016)	14
Table 4-1 Number of records by type of entity.....	26
Table 4-2 Breakdown of crashes and casualties by severity level.....	26
Table 4-3 Attributes in the Case Study data set.....	27
Table 5-1 ML model performance scores.....	38
Table 5-2 Hyperparameters before and after tuning.....	41
Table 5-3 Model performance before hyperparameter tuning.....	42
Table 5-4 Model performance after hyperparameter tuning.....	42
Table 5-5 Accuracy of crash GPS locations by casualty type.....	46
Table 8-1 Percentage of unknown crash factors in the data set.....	56

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