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