MACHINE LEARNING-BASED ANALYSIS OF CRASH FREQUENCY AT RAILROAD GRADE CROSSINGS

Abstract

Railroad-highway grade crossings present significant safety challenges which contributes to thousands of crashes annually in the United States. This study involves advanced machine learning techniques specifically are Random Forest and XG-Boost models which analyzes crash frequency at these grade crossings also comparing their performance with the existing traditional Negative Binomial Regression model. The Key variables which are included in the analysis is for capturing complex, non-linear relationships are Annual Average Daily Traffic, percentage of truck traffic, crossing surface types, train speed, highway speed, number of lanes, terrain type, trains per day, and illumination conditions. The results obtained indicates that Random Forest outperformed XG-Boost model by showing higher accuracy and lower error metrics, including Mean Squared Error (MSE) 40.80, 157.18 and Mean Absolute Error (MAE) 3.50, 3.53 respectively in predictive accuracy and variable importance ranking. However when compared to traditional Negative Binomial Regression method, Random Forest produced a good results by handling non-linear interactions with more consistent with real-world scenarios, percentage of truck traffic emerged as the most significant predictor in Random Forest, emphasizing the impact of heavy vehicle traffic on crash frequency at crossings. Moreover AADT, Train speed, highway speed and type of crossing surface also were ranked highly in Random Forest which aligns with engineering judgements. Variables like train per day and illumination were moderately ranked in machine learning models, demonstrating their ability to capture complex, non-linear relationships overlooked by traditional statistical methods. These findings underscore the importance and superiority of machine learning models especially the Random Forest model which is not only improve prediction accuracy but also provide extensive and comprehensive identification of critical risk factors purposefully for providing actionable insights to improve safety at railroad-highway grade crossings, However, the application of machine learning on extensive and high-quality datasets provide that there is a need for expansion of data collection efforts for future research since these model demand high quality of data for prediction.